SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2017, Vol. 60 012109:1–012109:13
doi: 10.1007/s11432-015-0900-x

Phrase-based hashtag recommendation for microblog posts

Yeyun GONG, Qi ZHANG*, Xiaoying HAN & Xuanjing HUANG

Shanghai Key Laboratory of Intelligent Information Processing,
School of Computer Science, Fudan University, Shanghai 201203*, China*

Received May 15, 2016; accepted July 4, 2016; published online November 17, 2016

Abstract In microblogs, authors use hashtags to mark keywords or topics. These manually labeled tags can be used to benefit various live social media applications (e.g., microblog retrieval, classification). However, because only a small portion of microblogs contain hashtags, recommending hashtags for use in microblogs are a worthwhile exercise. In addition, human inference often relies on the intrinsic grouping of words into phrases. However, existing work uses only unigrams to model corpora. In this work, we propose a novel phrase-based topical translation model to address this problem. We use the bag-of-phrases model to better capture the underlying topics of posted microblogs. We regard the phrases and hashtags in a microblog as two different languages that are talking about the same thing. Thus, the hashtag recommendation task can be viewed as a translation process from phrases to hashtags. To handle the topical information of microblogs, the proposed model regards translation probability as being topic specific. We test the methods on data collected from real-world microblogging services. The results demonstrate that the proposed method outperforms state-of-the-art methods that use the unigram model.

Keywords recommendation, topic model, translation, phrase extraction, hashtag

Citation Gong Y Y, Zhang Q, Han X Y, et al. Phrase-based hashtag recommendation for microblog posts. Sci China Inf Sci, 2017, 60(1): 012109, doi: 10.1007/s11432-015-0900-x

1 Introduction

With the booming Internet development, social networks have experienced rapid growth. A large number of Internet users are also members of at least one social networking site[1)]. Over the past few years, microblogging has become one of the most popular service of the social media services. Hence, microblogs have also been widely used as sources for public opinion analysis [1], prediction [2], reputation management [3], and many other applications [4–6]. In addition to the limited number of characters in the content, microblogs also contain metadata tags (hashtags), which are a string of characters preceded by the symbol (#). Hashtags are used to mark keywords or topics in a microblog, and they can occur anywhere in the content. Hashtags have proven to be useful in many applications, including microblog

* Corresponding author (email: qi_zhang@fudan.edu.cn)

1) As reported on Aug 5, 2013 by the Pew Research Center's Internet & American Life Project.

retrieval [7], query expansion [8], and sentiment analysis [9]. However, only a few microblogs include hashtags labeled by their users. Hence, the automatic recommendation of hashtags has become an important research topic that has received considerable attention in recent years.

Due to space limitations, sometimes hashtags may not appear in the microblog content. To solve problems related to the vocabulary gap, some studies have proposed the use of a translation model and have achieved significant improvements [10, 11]. The underlying assumption is that the content and tags of a resource can be regarded as describing the same topic, but are written in different languages. Thus, to address the vocabulary gap problem, these studies regard tag suggestion as a translation process from document content to tags.

Compared with traditional text settings, microblogs pose challenges due to their open access. Topics tend to be more diverse in microblogs than in formal documents [12], and due to the 140 character limit, microblogs are often short texts. Recently, there has been much progress with regard to modeling topics for short texts [13, 14] and microblog recommendations [15].

However, with respect to the hashtag suggestion task, existing topical translation models are word-based [15, 16]. These methods assume that all the words in a phrase will respectively align with the hashtags, which is not usually the case. For example, the phrase "Earthquake of Japan" should correspond exactly to the hashtag "Earthquake of Japan," while in actual word-based topical translation models, it was aligned with the hashtag "Earthquake" or "Japan," respectively. As such, here we propose phrase-based topical translation models that are based at the phrase level for the purposes of translation, and in which words in the same phrase share the same topic. The underlying assumption of our proposed phrase-based model is that a phrase is a natural unit that conveys a complete topic and contains real meaning that a human being wishes to express.

The main contributions of this work can be summarized as follows:

• We have observed that aligning single words with hashtags frequently results in the loss of the inherent meaning that microblog users wish to express, while regarding phrases as units could enhance model performance.

• We propose a novel phrase-based topical translation model to perform the hashtag recommendation task.

• We gathered a large collection of microblogs from a real microblogging service. This collection can benefit other researchers who are investigating the same or other topics related to microblogs data. We also conducted experiments on this real dataset. The results show that our phrase-based topical translation model outperforms state-of-the-art methods in microblog hashtag suggestion.

2 Related work

Various interesting studies have been carried out in social media. In social media like twitter, posts are typically presented in chronological order, which is inconvenient for users who wish to find posts they care about. Therefore, many recent studies have recommended the use of microblogs [17–23]. Refs. [17, 19] employed collaborative ranking by using the theme of the tweet content, social links, and some latent factors to identify personal interests. Ref. [21] proposed a joint model that integrated a collaborative filtering process with a propagation process to perform a microblog recommendation task. Ref. [23] introduced a graph-based method that employs a co-ranking technique to rank microblog content and authors.

Social media plays an important role in the social interaction of its users. Many research efforts have focused on detecting user interest communities and friends [24–28]. Ref. [25] introduced a weighted minimum-message ratio (WMR) model that leverages the number of messages between people. Ref. [26] introduced a new probabilistic matrix factorization framework that considers the opinions of the user's friends. Ref. [28] introduced a task for making event-based group recommendations. The authors presented a social-geo-aware matrix factorization framework that combines implicit patterns. Chen et al. [24] perform personalized community recommendations by adopting a collaborative filtering-based

Table 1 Main notations used in the proposed model

Variable	Description
D	Microblog data set
V	Vocabulary of the words in the corpus
T	Vocabulary of the hashtags in the corpus
Z	Topics set
v_d	Words in the microblog d
s_d	Phrases in the microblog d
t_d	Hashtags in the microblog d
ψ_w^z	Probability of word w occurring in topic z
θ_z^d	Probability of topic z occurring in microblog d
$\phi_t^{z,s}$	Alignment probability between hashtag t and phrase s under topic z

method that takes into account multiple kinds of social network situations.

In addition to the above, there have been a number of studies that have concentrated on making recommendations regarding music [29–31], news [32, 33], affiliation [34], and the like. With respect to music recommendations, Bu et al. [29] proposed a method that leverages various types of social media information as well as musical acoustic-based content. Ref. [31] integrated location-aware weighting of similarities and music content. Shmueli et al. [33] introduced a collaborative filtering approach that incorporates social network and content information for making recommendations to users via a personalized ranked list of news stories. To address problems associated with making affiliation recommendations, Vasuki et al. [34] combined friendship networks with affiliation networks between users and then grouped them to make recommendations.

Methods for tag recommendation tasks have been proposed from various aspects [11, 15, 35–37]. Heymann et al. [35] collected data from a social bookmarking system to investigate the tag recommendation problem. The authors adopted a metric based on entropy that could capture the generation process of a specific tag. Ref. [38] proposed a tag recommendation task method that is based on the mixture model. Krestel et al. [36] abstracted a shared topical structure from a collaborative tagging effort by multiple users to make tag recommendations, using latent Dirichlet allocation. Ref. [39] employed tag and content information in their model. Due to the fact that the same tags are often used to tag similar webpages, Ref. [16] used a topic-aware translation model to deal with the various meanings of words in different contexts. Ref. [40] learned users' perceptions by establishing topic-term relationships for suggesting suitable hashtags. Liu et al. [11] put forward a topical word trigger method to solve the vocabulary problem in the key phrase extraction task. The authors regarded the key phrase extraction problem as a translation process related to latent topics. Inspired by these efforts, our proposal integrates the advantages of the topic model, the translation process, and phrase extraction to implement hashtag suggestion tasks on social media.

3 Notations

We used D to represent the microblog set. $d \in D$ denotes a microblog comprising a sequence of words (v_d) and hashtags (t_d). $v_d = \{v_{dm}\}_{m=1}^{|v_d|}$ represents the set of words in the microblog d. $t_d = \{t_{dn}\}_{n=1}^{|t_d|}$ is the set of hashtags in the microblog. $s_d = \{s_{dm}\}_{m=1}^{|s_d|}$ is the set of phrases in the microblog. All the distinct words in the dataset comprise a vocabulary denoted by $V = \{v_1, v_2, \ldots, v_{|V|}\}$. We use $T = \{t_1, t_2, \ldots, t_{|T|}\}$ to denote the vocabulary of hashtags. $|T|$ is the number of distinct hashtags. These notations are summarized in Table 1.

4 Word-based topical translation model

The second word-based topical translation model (TTM) was proposed in [15]. TSTM assumes that a

Algorithm 1 Topical translation model

Sample π from Beta(.|δ)
Sample background-words distribution $\psi_B \sim$ Dirichlet(.|β)
for each topic $z \in Z$ **do**
  Sample topic-words distribution ψ^z from Dirichlet(.|β^v)
end for
Sample topic distribution θ from Dirichlet(.|α)
for each microblog $d \in D$ **do**
  Sample z_d from Multinomial(.|θ)
  for each word in the microblog d, $v_{dm} \in v_d$ **do**
    Sample y_{dm} from Bernoulli(π)
    if $y_{dm} = 0$ **then**
      Sample a word v_{dm} from Multinomial(.|ψ^B)
    end if
    if $y_{dm} = 1$ **then**
      Sample a word v_{dm} from Multinomial(.|ψ^{z_d})
    end if
  end for
  for each hashtag in the microblog d, $t_{dn} \in t_d$ **do**
    Sample a hashtag t_{dn} from $P(.|v_d, z_d, \phi)$
  end for
end for

microblog contains multiple topics, and that each topic in the microblog corresponds to a distribution of words. For long documents, this is a reasonable assumption. Since microblog posts are limited to 140 characters, a post does not usually have a mixture of topics. In contrast to TSTM, TTM assumes that topics are sampled at the document level. Each microblog is assigned a single topic. On the microblog content side, each topic corresponds to a distribution of words. On the hashtag side, each word and topic corresponds to a distribution of hashtags. In TTM, we assume that the user first chooses a topic for the microblog from the topic distribution and then generates the topic words or background words from the words distribution of the microblog. The algorithm is shown in Algorithm 1.

5 Inference of word-based models

To estimate the parameters in the TSTM and TTM models, we adopt the collapsed Gibbs sampling method [41] to obtain samples of latent variables.

5.1 TSTM

The joint probability distribution of the hashtags t, microblog words v, and topics z can be factorized from the generation process of Algorithm 2:

$$p(v, t, z|\alpha, \beta, \gamma) = p(z|\alpha)p(v|z, \beta)p(t|z, v, \gamma). \tag{1}$$

We use $N_z^{v,t}$ to represent the number of times word v is aligned with hashtag t for topic z. We can expand the distribution $p(t|z, v, \phi)$ and obtain the following equation:

$$p(t|z, v, \phi) = \prod_{d\in D}\prod_{t_n\in t_d} p(t_n|z_n, v_d) = \prod_{z\in Z}\prod_{t\in T}\prod_{v\in V}(\phi_t^{z,v})^{N_z^{v,t}}, \tag{2}$$

where, $\phi_t^{z,v}$ is proportional to the generation probability of hashtag t, given topic z and word v. Integrating over all the values of ϕ, we obtain the posterior distribution $p(t|z, v, \gamma)$ of the hashtag, as follows:

$$p(t|z, v, \gamma) = \int \prod_{z\in Z}\prod_{v\in V}\frac{1}{\Delta\gamma}\prod_{t\in T}(\phi_t^{z,v})^{N_z^{v,t}+\gamma_t-1} d_{\phi^{z,v}} = \prod_{z\in Z}\prod_{v\in V}\frac{\Delta(N_{\phi^{z,v}}+\gamma)}{\Delta(\gamma)}, \tag{3}$$

where $\Delta(\gamma) = \frac{\prod_{t=1}^{|T|}\Gamma(\gamma_t)}{\Gamma(\sum_{t=1}^{|T|}\gamma_t)}$ and $N_{\phi^{z,v}} = \{N_z^{v,t}\}_{t\in T}$.

Algorithm 2 Topic-specific translation model

for each topic $z \in Z$ **do**
  Sample topic-words distribution ψ^z from Dirichlet(.|β)
  for each word $v \in V$ **do**
    Sample word-topic-hashtags distribution $\phi^{z,v}$ from Dirichlet(.|γ)
  end for
end for
for each microblog $d \in D$ **do**
  Sample topic distribution θ_d from Dirichlet(.|α)
  for each word in microblog d, $v_{dm} \in v_d$ **do**
    Sample a topic z_{dm} from Multinomial(.|θ_d)
    Sample a word v_{dm} from Multinomial(.|ψ^z)
  end for
  for each hashtag in microblog d, $t_{dn} \in t_d$ **do**
    Sample a topic $z_{dn} \sim$ Multinomial(.|θ_d)
    Sample a hashtag $t_{dn} \sim p(.|z, v_d, \phi^{z,v})$
  end for
end for

By conducting a similar derivation for $p(v|z,\beta)$ and $p(z|\alpha)$, we can express the joint distribution:

$$p(v,t,z|\alpha,\beta,\gamma) = \prod_{z\in Z}\prod_{v\in V}\frac{\Delta(N_{\phi^{z,v}}+\gamma)}{\Delta(\gamma)}\prod_{z\in Z}\frac{\Delta(N_{\psi^z}+\beta)}{\Delta(\beta)}\prod_{d\in D}\frac{\Delta(N_m+\alpha)}{\Delta(\alpha)}. \tag{4}$$

We compute the conditional probability $p(z_m = k|z_{\neg m}, v, t)$ for the Gibbs sampler, where $z_{\neg m}$ represents all the topics z, $\neg m$ except for the topic of the mth word or hashtag.

Given the state of all the variables except the latent topic z_m of word v_m, we calculate the conditional probability by

$$p(z_m = k|z_{\neg m}, v, t) \propto \frac{N^{v_m}_{k,\neg m}+\beta}{N^{(\cdot)}_{k,\neg m}+\beta|V|}\frac{N_{k,\neg m}+\alpha}{N_{(\cdot),\neg m}+\alpha|Z|}\prod_{t\in t_d}\frac{N^{v_m,t}_{k,\neg m}+\gamma}{N^{v_m,(\cdot)}_{k,\neg m}+\gamma|T|}, \tag{5}$$

where $N^{v_m}_{k,\neg m}$ represents the number of times word v_m is labeled with topic k. $N^{v_m,t}_{k,\neg m}$ represents the number of times the word v_m aligned with the hashtag t when both of their labels have the topic k. $N_{k,\neg m}$ denotes the total number of topic k. $\neg m$ represents the word v_m that is not considered when calculating the counts, and $(\cdot)$ indicates that every condition will be considered when calculating the counts. For example, $N_{(\cdot),\neg m} = \sum_{k\in Z} N_{k,\neg m}$.

Given the state of all the variables except the latent topic z_n of hashtag t_n, we can calculate the probability of the topic by

$$p(z_n = k|z_{\neg n}, v, t) \propto \frac{N_{k,\neg m}+\alpha}{N_{(\cdot),\neg m}+\alpha|Z|}\left(\sum_{v_m\in v_d}\frac{N^{v_m,t_n}_{k,\neg m}+\gamma}{N^{(\cdot),t_n}_{k,\neg m}+\gamma|T|}\right). \tag{6}$$

5.2 TTM

The joint probability distributions of the hashtags t, microblog words v, topics z, and the topic or background words indicate that variables y can be factorized by the generation process of Algorithm 1:

$$p(v,t,z,y|\alpha,\beta,\gamma,\rho) = p(z|\alpha)p(y|\rho)p(v|z,y,\beta)p(t|z,v,y,\gamma). \tag{7}$$

As in the previous section, we can also extend the probability distribution using the following equation:

$$p(v,t,z,y|\alpha,\beta,\gamma,\rho) = \frac{\Delta(N_\eta+\rho)}{\Delta(\rho)}\prod_{z\in Z}\frac{\Delta(N_{\psi^z}+\beta)}{\Delta(\beta)}\frac{\Delta(N_{\psi^B}+\beta)}{\Delta(\beta)}\prod_{z\in Z}\prod_{v\in V}\frac{\Delta(N_{\phi^{z,v}}+\gamma)}{\Delta(\gamma)}\prod_{d\in D}\frac{\Delta(N_d+\alpha)}{\Delta(\alpha)}. \tag{8}$$

Given the state of all the variables except the variable y_m, we can calculate the probability of the variable y_m by

$$p(y_m = q|v, t, z, y_{\neg m}) \propto \frac{N_{\neg m,q} + \rho}{N_{\neg m,(\cdot)} + 2\rho} \frac{N^{v_m}_{\neg m,l} + \beta}{N^{(\cdot)}_{\neg m,l} + \beta|V|}, \tag{9}$$

if $q = 0$, $N^{v_m}_{\neg m,l}$ represents the number of times the word v_m was under the background word label. And if $q = 1$, $N^{v_m}_{\neg m,z}$ indicates the number of times the word v_m occurred as a topic word. $N_{\neg m,1}$ is the total number of topic words and $N_{\neg m,0}$ is the total number of background words. $\neg m$ indicates that the word v_m is not considered when calculating the counts.

Given the state of all the variables except the latent topic z_d of the microblog d, we can calculate the probability of the topic by

$$p(z_d = k|v, t, z_{\neg d}, x) \propto \frac{N_{k,\neg d} + \alpha}{N_{(\cdot),\neg d} + \alpha|Z|} \prod_{v_m \in v_d} \frac{N^{v_m}_{k,\neg m} + \beta}{N^{(\cdot)}_{k,\neg m} + \beta|V|} \prod_{t_n \in t_d} \left(\sum_{v_m \in v_d} \frac{N^{v_m,t_n}_{k,\neg d} + \gamma}{N^{v_m,(\cdot)}_{k,\neg d} + \gamma|T|} \right), \tag{10}$$

where $N_{k,\neg d}$ represents the number of times the microblogs were categorized under topic k. $\neg d$ indicates that the microblog d is not considered when calculating the counts.

6 Phrase-based topical translation model

In the previous section, we introduced two word-based topical translation models. In this section, we introduce two phrase-based topical translation models. TSTM and TTM intuitively handle the translation process by regarding words and hashtags as two different languages. However, corresponding each single word with a hashtag is generally not reasonable, since the actual topic that a phrase unit intends to express usually differs from the single words contained in the phrase. In addition, the original LDA was built upon the "bag-of-words" assumption, in which the order of words is completely ignored. As a result, when inferring the topic assignment z_{dm} for word v_{dm}, the topic of a word found elsewhere in the same document has the same impact as a word nearby. To address these problems, we propose a phrase-based topical translation model that regards phrases and hashtags as two different languages.

We adopted the phrase-mining algorithm proposed in [42] to extract phrases. This algorithm obtains the counts of frequent contiguous patterns, then probabilistically considers these patterns while applying contextual constraints to determine meaningful phrases. The method involves two major steps in conducting phrase mining. First, it mines the corpus for frequent candidate phrases and their aggregate counts. Second, it merges the words in each document into quality phrases.

We segment documents into a sequence of phrases generated from the phrase mining step, then represent the documents as a "bag of phrases". Similar to the word-based topical translation model, we also model the microblog from two different aspects in the phrase-based topical translation model. Considering that a microblog contains multiple topics, we propose the phrase topic-specific translation model (PTSTM). In this model, each phrase shares the same topic, and a unique phrase has an alignment probability with each hashtag. In PTSTM, we assume that when a user generates a microblog, he first generates the phrases of the post and then tags them with suitable hashtags. The generation of each phrase in the microblog can be broken down into two steps. First, the user selects a topic for the microblog from its topic distribution. Secondly, he selects each word of the phrase one at a time from the topic words distribution. Finally, he tags the hashtags according to the selected topics and phrases. A graphical representation of this process is shown in Figure 1 and the algorithm is shown in Algorithm 3.

Another assumption is that each microblog is aligned with a single topic. Under this assumption, we propose the phrase topical translation model (PTTM). In this model, phrases in the same microblog share the same topic, and the words in the phrases are either topic words or background words. In PTSTM, we assume that when a user generates a microblog, he first generates the phrases of the post and then tags them with suitable hashtags. In contrast to PTSTM, in PTTM, we assume that the user

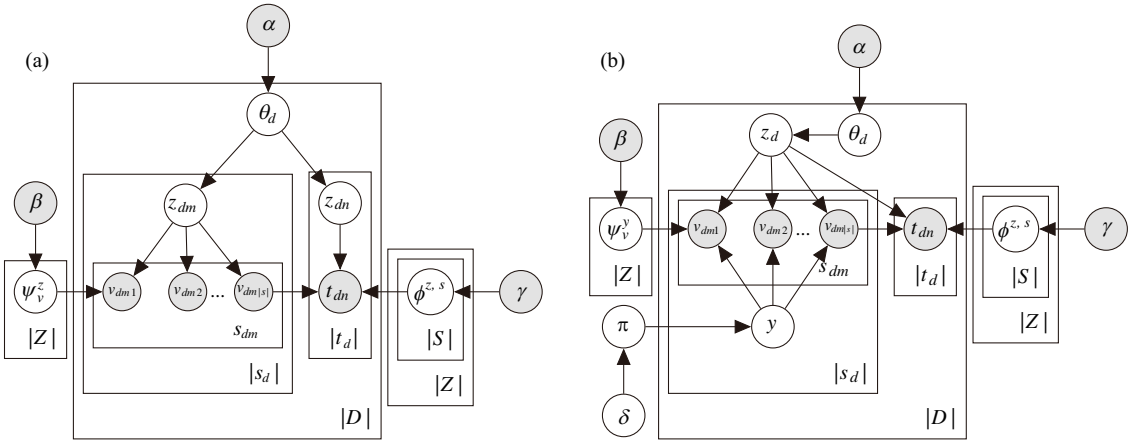

Figure 1 (a) Graphical representation of PTSTM. (b) Graphical representation of PTTM. Shaded circles are observations or constants. Unshaded circles are latent variables.

Algorithm 3 Phrased topic-specific translation model

for each topic $z \in Z$ **do**
  Sample topic-words distribution ψ^z from Dirichlet(.|β)
  for each word $v \in V$ **do**
    Sample topic-word-hashtags distribution $\phi^{z,v}$ from Dirichlet(.|γ)
  end for
end for
for each microblog $d \in D$ **do**
  Sample $\theta_d \sim$ Dirichlet(.|α)
  for each phrase in microblog d,$s_m \in s_d$ **do**
    Sample a topic z_m from Multinomial(.|θ_d)
    for each word in phrase s_m, $v_{dmi} \in s_m$ **do**
      Sample a word v_{dmi} from Multinomial(.|ψ^z)
    end for
  end for
  for each hashtag t in microblog d **do**
    Sample a topic $z_{dn} \sim$ Multinomial(.|θ_d)
    Sample a hashtag $t_{dn} \sim p(.|z, s_d, \phi^{z,s})$
  end for
end for

first chooses a topic for the microblog from the topic distribution. Then he generates each phrase in two steps. First, he decides to generate a topic phrase or a background phrase, then he generates the topic words or background words of the phrase from the words distribution.

A graphical representation of the PTTM process is shown in Figure 1 and the algorithm is shown in Algorithm 4.

In these phrase-based topical translation models, given the observed phrases and hashtags in a collection of microblogs, our task is to estimate the topic distribution θ_d for each microblog d in PTSTM or the topic distribution θ for all the microblogs in PTTM. In both PTSTM and PTTM, we estimate the distribution of phrases ψ^z for each topic z and the distribution of hashtags $\phi^{z,s}$ for each topic and phrase.

7 Inference of phrase-based models

To estimate the PTSTM and PTTM model parameters, we also adopted collapsed Gibbs sampling [41] to obtain samples of latent variables.

7.1 PTSTM

Using the generation process of Algorithm 3 and the inference process in Subsection 5.1, we can easily

Algorithm 4 Phrase topical translation model

Sample π from Beta(.|δ)
Sample background-words distribution ψ_B from Dirichlet(.|β)
for each topic $z \in Z$ **do**
  Sample topic-words distribution ψ^z from Dirichlet(.|β^v)
end for
Sample topic distribution θ from Dirichlet(.|α)
for each microblog $d \in D$ **do**
  Sample a topic z_d from Multinomial(.|θ)
  for each phrase in microblog d, $s_m \in s_d$ **do**
    Sample y_m from Bernoulli(.|π)
    if $y_m = 0$ **then**
      for each word in phrase s_m, $v_{dmi} \in s_m$ **do**
        Sample a word v_{dmi} from Multinomial(.|ψ^B)
      end for
    end if
    if $y_m = 1$ **then**
      for each word in phrase s_m, $v_{dmi} \in s_m$ **do**
        Sample a word v_{dmi} from Multinomial(.|ψ^{z_d})
      end for
    end if
  end for
  for each hashtag $t_{dn} \in t_d$ **do**
    Sample a hashtag t_{dn} from $P(.|s_d, z_d, \phi)$
  end for
end for

obtain the expression of the joint distribution as follows:

$$p(s,t,z|\alpha,\beta,\gamma) = \prod_{z\in Z}\prod_{s\in S}\frac{\Delta(N_{\phi^{z,s}}+\gamma)}{\Delta(\gamma)}\prod_{z\in Z}\frac{\Delta(N_{\psi^z}+\beta)}{\Delta(\beta)}\prod_{d\in D}\frac{\Delta(N_m+\alpha)}{\Delta(\alpha)}. \tag{11}$$

Given the state of all the variables except the latent topic z_m of phrase s_m in the microblog d, we can calculate the probability of the topic by

$$p(z_m = k|z_{\neg m}, s, t) \propto \prod_{v_{dmi}\in s_m}\frac{N_{k,\neg m}^{v_{dmi}}+\beta}{N_{k,\neg m}^{(\cdot)}+\beta|V|}\frac{N_{k,\neg m}+\alpha}{N_{(\cdot),\neg m}+\alpha|Z|}\prod_{t\in t_d}\frac{N_{k,\neg m}^{s_m,t}+\gamma}{N_{k,\neg m}^{s_m,(\cdot)}+\gamma|T|}, \tag{12}$$

where $N_{k,\neg m}^{v_{dmi}}$ is the number of times word v_{dmi} occurred under the topic k. $N_{k,\neg m}^{s_m,t}$ is the number of times phrase s_m aligned with hashtag t under the topic k. $N_{k,\neg m}$ is the total number of times topic k occurred. $\neg m$ indicates that the phrase s_m is not considered when calculating the counts.

Given the state of all the variables except the latent topic z_n of hashtag t_n in the microblog d, we can calculate the probability of the topic by

$$p(z_n = k|z_{\neg n}, v, t) \propto \frac{N_{k,\neg m}+\alpha}{N_{(\cdot),\neg m}+\alpha|Z|}\left(\sum_{s_m\in s_d}\frac{N_{k,\neg m}^{s_m,t_n}+\gamma}{N_{k,\neg m}^{(\cdot),t_n}+\gamma|T|}\right). \tag{13}$$

7.2 PTTM

Using the generation process of Algorithm 4 and the derivation in Subsection 5.2, we can obtain the expression of the joint distribution as follows:

$$p(s,t,z,y|\alpha,\beta,\gamma,\rho) = \frac{\Delta(N_\eta+\rho)}{\Delta(\rho)}\prod_{z\in Z}\frac{\Delta(N_{\psi^z}+\beta)}{\Delta(\beta)}\frac{\Delta(N_{\psi^B}+\beta)}{\Delta(\beta)}\prod_{z\in Z}\prod_{s\in S}\prod_{v\in s}\frac{\Delta(N_{\phi^{z,v}}+\gamma)}{\Delta(\gamma)}\prod_{d\in D}\frac{\Delta(N_d+\alpha)}{\Delta(\alpha)}. \tag{14}$$

Given the state of all the variables except the variable y_m, we can calculate the probability of the variable y_m by

$$p(y_m = q|s,t,z,y_{\neg m}) \propto \frac{N_{\neg m,q}+\rho}{N_{\neg m,(\cdot)}+2\rho}\prod_{v_{dmi}\in s_m}\frac{N_{\neg m,l}^{v_{dmi}}+\beta}{N_{\neg m,l}^{(\cdot)}+\beta|V|}, \tag{15}$$

If $q = 0$, then $N_{\neg m,l}^{v_{dmi}}$ represents $N_{\neg m,B}^{v_{dmi}}$, which is the number of times word v_{dmi} occurred under the background label. If $q = 1$, then $N_{\neg m,l}^{v_{dmi}}$ represents $N_{\neg m,z}^{v_{dmi}}$, which is the number times word v_{dmi} occurred under a topic label. $N_{\neg m,1}$ is the total number of times words occurred under the topic label and $N_{\neg m,0}$ is the total number of times words occurred under the background label. $\neg m$ indicates that the phrase s_m is not considered when calculating the counts.

Given the state of all the variables except the latent topic z_d of microblog d, we can calculate the probability of the topic by

$$p(z_d = k|s, t, z_{\neg d}, y) \propto \frac{N_{k,\neg d} + \alpha}{N_{(\cdot),\neg d} + \alpha|Z|} \prod_{s_m \in s_d} \prod_{v_{dmi} \in s_m} \frac{N_{k,\neg m}^{v_{dmi}} + \beta}{N_{k,\neg m}^{(\cdot)} + \beta|V|} \prod_{t_n \in t_d} \left(\sum_{s_m \in s_d} \frac{N_{k,\neg d}^{s_m,t_n} + \gamma}{N_{k,\neg d}^{s_m,(\cdot)} + \gamma|T|} \right), \quad (16)$$

where $N_{k,\neg d}$ is the number of times microblogs under the topic k occurred. $\neg d$ indicates that microblog d is not considered when calculating the counts.

After a sufficient number of sampling iterations to burn in the Markov chain, we estimate ϕ^z by $\phi_t^{z,s} = \frac{N_z^{s,t} + \gamma}{N_z^{s,(\cdot)} + |T|\gamma}$.

From the model, we can calculate the degree of the alignment probability ϕ by $|T| \times |S| \times |Z|$. $|T|$ is the vocabulary size of the hashtags, and S is the vocabulary size of the phrases. $|Z|$ is the number of times topics occurred. There is a serious data sparsity problem when estimating the probability ϕ. Hence, we use topic-free phrase alignment probability to smooth the translation probability as follows:

$$\Phi_t^{z,s} = \lambda \phi_t^{z,s} + (1 - \lambda) p(t|s), \quad (17)$$

where $p(t|s)$ represents the topic-free phrase alignment probability between the phrase s and the hashtag t. Here, we use the IBM model-1 [43] to obtain $P(t|s)$. We use λ as the trade-off of $\phi_t^{z,s}$ and $p(t|s)$, where $0.0 \leqslant \lambda \leqslant 1.0$. If λ equals 0.0, $\Phi_t^{z,s}$ will be equal to the topic-free phrase translation probability. If λ equals to 1.0, $\Phi_t^{z,s}$ will be equal to the topic phrase translation probability $\phi_t^{z,s}$. $P(h|w, v)$ is the topic-free word alignment probability of hashtag h, given the text word w and visual feature v. Here, we again use the IBM model-1 [43], which is a widely used word alignment model to obtain $P(h|w, v)$. λ, which ranges from 0.0 to 1.0, is a trade-off between these two probabilities. When λ equals 0.0, $P_{\text{smooth}}(h|w, v, z)$ will reduce to topic-free word alignment probability. When λ is set to 1.0, there is no smoothing in $P_{\text{smooth}}(h|w, v, z)$.

8 Hashtag extraction

Given a test dataset, we first use collapsed Gibbs sampling to sample the latent variables of all the phrases in each microblog. After the latent variables of the phrases in each microblog are stable, we can compute the scores for candidate hashtags of microblog d in the unlabeled data by

$$\Pr(t_{dn}|s_d) \propto \sum_{z \in Z} \sum_{s_m \in s_d} \phi_t^{z,s_m} P(s_m|s_d) P(z|s_d). \quad (18)$$

We employ the inverse document frequency (IDF) score of phrase s_m in the microblog to compute the weight of the phrase $P(s_m|s_d)$. $P(z|s_d)$ is equal to θ_z^d. Based on the ranking scores, we can then recommend top-ranked hashtags for each microblog.

9 Experiments

With respect to the experiments, we first introduce the method we used to collect the data and their relevant statistics. Next, we describe the experimental configurations and the baseline methods we used in this work. Finally, we report and analyze our experimental results.

In this section, we introduce the experimental results and data we collected for training and evaluation. First, we describe the method we used to generate the collection and its related statistics. Then, we describe our experimental configurations and baseline methods. Finally, we provide our evaluation results and analysis.

Table 2 Statistical information of the evaluation dataset. Ave_v represents the average number of words per microblog, Ave_t represents the average number of manually labeled hashtags per microblog, $|V|$ represents the vocabulary size of the words, and $|T|$ represents the vocabulary size of the hashtags.

#microblog	Ave_v	Ave_t	$\|V\|$	$\|T\|$
50000	25.34	1.10	128558	3174

Table 3 Evaluation results of different methods

Method	P	R	F_1
NB	0.382	0.347	0.364
IBM1	0.354	0.322	0.337
TSTM	0.385	0.351	0.367
PTSTM	0.404	0.368	0.385
TTM	0.480	0.437	0.457
PTTM	**0.514**	**0.468**	**0.490**

9.1 Data collection

We collected public microblogs using Sina Weibo's API[2)] from randomly selected users. Through this we get a dataset which contained 282.2 million microblogs posted by 1.1 million users. We derived the microblogs that tagged with hashtags. Thus we got a collection of 2.69 million microblogs. We filtered out the microblogs labeled with hashtags which occurred lower than 100 times in our corpus. Finally, we constructed a collection containing 1.03 million microblogs labeled with hashtags. The specific number of hashtags in the corpus was 3204. We randomly selected 50000 microblogs from the collection as the corpus. The hashtags annotated by their users were treated as the gold standard. We randomly selected 80% microblogs as the training set, the rest of the dataset as test set. Table 2 showed the detailed statistics.

9.2 Experiment configurations

To evaluate model performance, we used a precision (P), recall (R), and F1 score (F_1) matrix. The precision is the percentage of the number of correct hashtags of the total number of hashtags recommended by the system. Recall is the percentage of the number of correct hashtags of the total number of hashtags recommended by the system. The F1 score is the harmonic mean of the precision and recall. We conducted 500 iterations of the Gibbs sampling of our model. We tuned the topic number from 10 to 50, then set the topic number $|Z|$ to 20. We also set other hyperparameters as follows: $\alpha = 0.5$, $\beta^v = 0.1$, $\gamma = 0.1$. We set the parameter λ to 0.8. To estimate the translation probability without topical information, we used GIZA++ 1.07 [44].

We then compared our model results with those of four baseline models.

- **Naive Bayes (NB).** We employed Naive Bayes in the hashtag recommendation task. In this model, we calculated the posterior probability of each hashtag given in the microblogs.
- **IBM1.** We used IBM model-1 to obtain the translation probability from words to hashtags.
- **TSTM.** In the topic-specific translation model (TSTM) proposed in [16], hashtags are extracted based on topic words. We implemented this model and used it to solve the problem.
- **TTM.** We used the topic translation model proposed in [15] to model each microblog with single topics.

9.3 Experimental results

We evaluated the proposed methods by comparing their results with those of state-of-the-art methods and identified the impacts of the key parameters. Table 3 shows the results for the dataset by all the methods. "PTSTM" denotes the first method proposed in this paper, which is based on the topic-specific translation model. "PTTM" represents the second method proposed in this paper, which is based on the topical translation model. From the results, we can see that the methods proposed in this paper outperformed

2) http://open.weibo.com/.

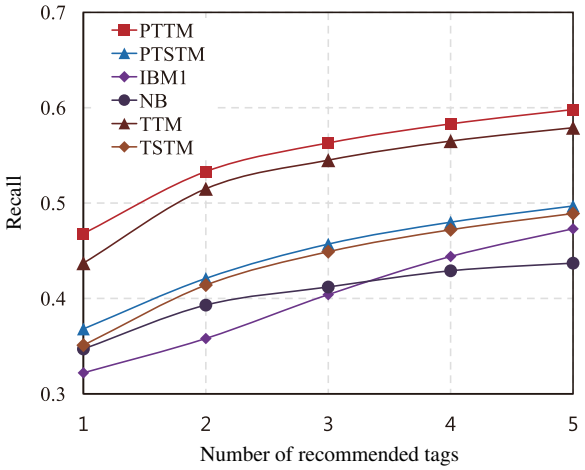

Figure 2 (Color online) Recall with different number of recommendation tags.

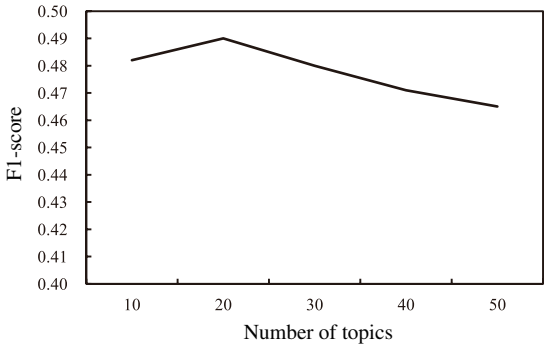

Figure 3 Impact of the topic number $|Z|$ in PTTM.

the other methods. Comparing the results of "PTSTM" and "TSTM", "PTTM" and "TTM" separately, we can conclude that the phrase-based translation model achieved better performance than the word-based translation model. From the results of "PTTM" and "PTSTM", we can see that for microblogs, assigning single topics to each post achieved better performance. We believe that the main reason for this result was that a shorter document like a microblog post typically focuses on one single topic. From the results of "TSTM" and "IBM1," we can conclude that the word topic can improve the results of the translation model. The above results demonstrate that the proposed "PTTM" model compares favorably with other state-of-the-art baselines in performing the hashtag suggestion task.

We then analyzed the recall curve of NB, IBM1, TSTM, TTM, PTSTM and PTTM for different numbers of suggested hashtags, ranging from 1 to 5. Figure 2 shows the results for the evaluation dataset. The axis of the abscissa represents the number of hashtags ranging from 1 to 5. The axis of the ordinate represents the value of the different evaluation matrixes. The highest curve of the graph indicates the best performance. From these results, we can see that the "PTTM" recall curve reaches the highest point of all the recall curves, indicating that the proposed method performed significantly better than the other methods.

From the description in the previous section, we found that several hyperparameters are important to consider in the proposed model. We evaluated the impacts of the two crucial hyperparameters $|Z|$ and λ.

Figure 3 illustrates the influence of the topic number. In this figure, we can see that the best performance was achieved when topic $|Z|$ equals 20. We can also achieve a reasonable performance when the topic number ranged from 10 to 30, which makes it easy to choose a suitable topic number. However, the performance decreases slowly when the number of topics increased from 30. We believe that one of the main reasons for this may be data sparsity. With a larger number of topics, there is a more serious data sparsity problem in estimating topic translation probability.

In Table 4, we describe the impact of the parameter λ in Eq. (17). When λ is set to 0.0, the method results equal those of the topic-free translation model. From the results when λ is equal to 0.0, we can conclude that this task benefits from the inclusion of topical information. When λ is equals to 1.0, the method uses no smoothing, and from the results, we can see that smoothing is an important factor.

10 Conclusion

In this paper, we proposed and examined the performance of a novel topical translation method for recommending hashtags for microblogs, in which we assume that the hashtags and content in a microblog describe the same theme using different languages. As such, we converted the hashtag recommendation task into a machine translation process. While a list of words in the microblog are often used to describe

Table 4 Impact of the parameter λ in PTTM

λ	P	R	F_1
1.0	0.501	0.456	0.477
0.8	**0.514**	**0.468**	**0.490**
0.6	0.503	0.458	0.479
0.4	0.454	0.413	0.433
0.2	0.410	0.373	0.391
0.0	0.354	0.322	0.337

individual topics, using these words to represent the topic is often hard to interpret or ambiguous. Since phrases provide a more accurate description of the topic, we introduced a phrase-aware translation model to address this problem. To handle topical inferences, we proposed using a phrase-based topical translation model to facilitate the translation process. Under this framework, we use specific phrase triggers to bridge the gap between textual content and the hashtags of each microblog. We collected data from real microblog services from which we constructed a large dataset to verify the effectiveness of our model. Our experimental results show that the proposed method achieved better performance than state-of-the-art methods.

Acknowledgements This work was partially funded by National Natural Science Foundation of China (Grant Nos. 61473092, 61472088, 61532011), National High Technology Research and Development Program of China (Grant No. 2015AA015408), and Shanghai Science and Technology Development Funds (Grant Nos. 13dz2260200, 13511504300).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Bermingham A, Smeaton A F. Classifying sentiment in microblogs: is brevity an advantage? In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM, 2010. 1833–1836
2 Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci, 2011, 2: 1–8
3 Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inf Retr, 2008, 2: 1–135
4 Becker H, Naaman M, Gravano L. Learning similarity metrics for event identification in social media. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM, 2010. 291–300
5 Guy I, Avraham U, Carmel D, et al. Mining expertise and interests from social media. In: Proceedings of the 22nd International Conference on World Wide Web. New York: ACM, 2013. 515–526
6 Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010. 851–860
7 Efron M. Hashtag retrieval in a microblogging environment. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2010. 787–788
8 Bandyopadhyay A, Mitra M, Majumder P. Query expansion for microblog retrieval. In: Proceedings of the 20th Text Retrieval Conference, TREC, 2011
9 Wang X, Wei F, Liu X, et al. 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. New York: ACM, 2011. 1031–1040
10 Bernhard D, Gurevych I. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg: Association for Computational Linguistics, 2009. 2: 728–736
11 Liu Z Y, Liang C, Sun M S. Topical word trigger model for keyphrase extraction. In: Proceedings of the 24th International Conference on Computational Linguistics, Mumbai, 2012. 1715–1730
12 Zhao W X, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval. Berlin: Springer, 2011. 338–349
13 Diao Q M, Jiang J, Zhu F, et al. Finding bursty topics from microblogs. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012. 536–544
14 Zhao W X, Jiang J, He J, et al. Topical keyphrase extraction from twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, 2011. 379–388
15 Ding Z Y, Qiu X, Zhang Q, et al. Learning topical translation model for microblog hashtag suggestion. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2013. 2078–2084

16 Ding Z, Zhang Q, Huang X. Automatic hashtag recommendation for microblogs using topic-specific translation model. In: Proceedings of the 24th International Conference on Computational Linguistics, Mumbai, 2012. 265
17 Chen K L, Chen T Q, Zheng G Q, et al. Collaborative personalized tweet recommendation. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2012. 661–670
18 Debnath S, Ganguly N, Mitra P. Feature weighting in content based recommendation system using social network analysis. In: Proceedings of the 17th International Conference on World Wide Web. New York: ACM, 2008. 1041–1042
19 Guy I, Zwerdling N, Ronen I, et al. Social media recommendation based on people and tags. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2010. 194–201
20 Konstas I, Stathopoulos V, Jose J M. On social networks and collaborative recommendation. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2009. 195–202
21 Pan Y, Cong F, Chen K, et al. Diffusion-aware personalized social update recommendation. In: Proceedings of the 7th ACM Conference on Recommender Systems. New York: ACM, 2013. 69–76
22 Ronen I, Guy I, Kravi E, et al. Recommending social media content to community owners. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2014. 243–252
23 Yan R, Lapata M, Li X. Tweet recommendation with graph co-ranking. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, Stroudsburg, 2012. 516–525
24 Chen W Y, Zhang D, Chang E Y. Combinational collaborative filtering for personalized community recommendation. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008. 115–123
25 Lo S, Lin C. Wmr–a graph-based algorithm for friend recommendation. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, 2006. 121–128
26 Ma H, King I, Lyu M R. Learning to recommend with social trust ensemble. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2009. 203–210
27 Moricz M, Dosbayev Y, Berlyant M. Pymk: friend recommendation at myspace. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM, 2010. 999–1002
28 Zhang W, Wang J, Feng W. Combining latent factor model with location features for event-based group recommendation. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013. 910–918
29 Bu J J, Tan S L, Chen C, et al. Music recommendation by unified hypergraph: combining social media information and music content. In: Proceedings of the International Conference on Multimedia. New York: ACM, 2010. 391–400
30 Kaminskas M, Ricci F. Contextual music information retrieval and recommendation: state of the art and challenges. Comput Sci Rev, 2012, 6: 89–119
31 Schedl M, Schnitzer D. Hybrid retrieval approaches to geospatial music recommendation. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2013. 793–796
32 Li Q, Wang J, Chen Y P, et al. User comments for news recommendation in forum-based social media. Inform Sci, 2010, 180: 4929–4939
33 Shmueli E, Kagian A, Koren Y, et al. Care to comment? Recommendations for commenting on news stories. In: Proceedings of the 21st International Conference on World Wide Web. New York: ACM, 2012. 429–438
34 Vasuki V, Natarajan N, Lu Z, et al. Affiliation recommendation using auxiliary networks. In: Proceedings of the 4th ACM Conference on Recommender Systems. New York: ACM, 2010. 103–110
35 Heymann P, Ramage D, Garcia-Molina H. Social tag prediction. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2008. 531–538
36 Krestel R, Fankhauser P, Nejdl W. Latent dirichlet allocation for tag recommendation. In: Proceedings of the 3rd ACM Conference on Recommender Systems. New York: ACM, 2009. 61–68
37 Rendle S, Marinho L, Nanopoulos A, et al. Learning optimal ranking with tensor factorization for tag recommendation. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009. 727–736
38 Song Y, Zhuang Z, Li H, et al. Real-time automatic tag recommendation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2008. 515–522
39 Lu Y T, Yu S I, Chang T C, et al. A content-based method to enhance tag recommendation. In: Proceedings of the 21st International Jont Conference on Artifical Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 2009. 2064–2069
40 Tariq A, Karim A, Gomez F, et al. Exploiting topical perceptions over multi-lingual text for hashtag suggestion on twitter. In: Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference, St. Pete Beach, 2013. 474–479
41 Griffiths T L, Steyvers M. Finding scientific topics. Proc Natl Acad Sci, 2004, 101: 5228–5235
42 El-Kishky A, Song Y, Wang C, et al. Scalable topical phrase mining from text corpora. Proc VLDB Endowment, 2014, 8: 305–316
43 Brown P F, Pietra V J D, Pietra S A D, et al. The mathematics of statistical machine translation: parameter estimation. Comput Linguist, 1993, 19: 263–311
44 Och F J, Ney H. A systematic comparison of various statistical alignment models. Comput Linguist, 2003, 29: 19–51