Joint Estimation of Topics and Hashtag Relevance in Cross-Lingual Tweets

Procheta Sen
CVPR Unit
Indian Statistical Institute
Kolkata, India
senprocheta@gmail.com

Debasis Ganguly
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
dganguly@computing.dcu.ie

Gareth J.F. Jones
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
gjones@computing.dcu.ie

ABSTRACT

Twitter is a widely used platform for sharing news articles. An emerging trend in multi-lingual communities is to share non-English news articles using English tweets in order to spread the news to a wider audience. In general, the choice of relevant hashtags for such tweets depends on the topic of the non-English news article. In this paper, we address the problem of automatically detecting the relevance of the hashtags of such tweets. More specifically, we propose a generative model to jointly model the topics within an English tweet and those within the non-English news article shared from it to predict the relevance of the hashtags of the tweet. For conducting experiments, we compiled a collection of English tweets that share news articles in Bengali (a South Asian language). Our experiments on this dataset demonstrate that this joint estimation based approach using the topics from both the non-English news articles and the tweets proves to be more effective for relevance estimation than that of only using the topics of a tweet itself.

CCS Concepts

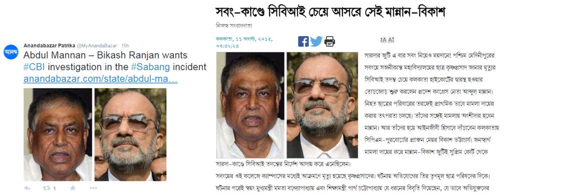
•**Computing methodologies** → **Natural language processing;**

Keywords

Cross-lingual Tweet tagging; bilingual topic modelling; joint estimation of topic and tag relevance

1. INTRODUCTION

In recent years, social media has played a major role in distributing news among people, often crossing language barriers. For example, a significant number of news articles are shared with the help of Twitter, a micro blogging platform. Sometimes, non-native English speakers compose tweets in English in order to share a news article published in their own native language. The most likely motive for this *cross-lingual* news sharing across the social media is to allow a non-English news article to receive wider visibility outside its own local community.



Anandabazar Patrika @MyAnandaBazar · 15h
Abdul Mannan – Bikash Ranjan wants #CBI investigation in the #Sabang incident anandabazar.com/state/abdul-ma...

(a) Tweet

সবং-কাণ্ডে সিবিআই চেয়ে আসরে সেই মান্নান-বিকাশ
নিজস্ব সংবাদদাতা
কলকাতা, ১১ অগস্ট, ২০১৫, ০৩:৫২:২৪

সারদার জুটি এ বার সবং নিয়েও ময়দানে! পশ্চিম মেদিনীপুরের সবংয়ে সজনীকান্ত মহাবিদ্যালয়ের ছাত্র কৃষ্ণপ্রসাদ জানার মৃত্যুর সিবিআই তদন্ত চেয়ে কলকাতা হাইকোর্টের দ্বারস্থ হওয়ার তোড়জোড় শুরু করলেন প্রদেশ কংগ্রেস নেতা আব্দুল মান্নান। নিহত ছাত্রের পরিবারের তরফেই প্রাথমিক ভাবে মামলা দায়ের করার তৎপরতা চলছে। তাঁদের সঙ্গেই মামলায় অংশীদার হবেন মান্নান। আর তাঁদের হয়ে আইনজীবী হিসাবে দাঁড়াবেন কলকাতায় সিপিএম-পুরবোর্ডের প্রাক্তন মেয়র বিকাশ ভট্টাচার্য। জনস্বার্থ মামলা দায়ের করে মান্নান-বিকাশ জুটিই সুপ্রিম কোর্ট থেকে সারদা-কাণ্ডে সিবিআই তদন্তের নির্দেশ আদায় করে এনেছিলেন।

সবংয়ের ওই কলেজে ক্যাম্পাসের মধ্যেই আক্রমণে মৃত্যু হয়েছে কৃষ্ণপ্রসাদের। ঘটনায় অভিযোগের তির তৃণমূল ছাত্র পরিষদের দিকে।

ঘটনার পরেই স্বয়ং মুখ্যমন্ত্রী মমতা বন্দ্যোপাধ্যায় এবং শিক্ষামন্ত্রী পার্থ চট্টোপাধ্যায় যে ধরনের বিবৃতি দিয়েছেন, যে ভাবে অভিযুক্তদের

(b) News article

Figure 1: A sample tweet in English that contains a link to a foreign language (Bengali, in this example) news article.

The widespread generation of such user generated *cross-lingual* tweets has given rise to the problem of selecting the appropriate hashtags for these tweets (which we call *cross-lingual* tweets), so that they can effectively be retrieved at a later period of time. More often than not Twitter users do not select hashtags that effectively describe the core concepts of their tweet. This observation has motivated research into constructing automated approaches for identifying the relevant tags of a tweet [4], or recommending alternative potentially relevant hashtags to users [2, 7].

In general, it is difficult to model the relevance of hashtags with the tweet text alone, because of their short length. The motivation of our work in this paper is to make use of the text of the related news article to improve the relevance estimation of the tags. Our motivation is that intuitively speaking, it is the *aboutness* of the shared article which should influence the choice of appropriate hashtags for the tweet. Since the language of a news document (say F) shared from a tweet is different from that of the tweet itself (say E)[1], the problem of joint topic modeling is more challenging.

An example of a cross-lingual tweet is shown in Figure 1. The word '#CBI' ('Central Bureau of Investigation') is a relevant tag for the tweet shown in Figure 1. However, it is difficult to predict this word as a relevant tag, since the only word in the tweet topically related to the word 'CBI' is 'investigation'. Contrastingly, the Bengali news article, shown in Figure 1, contains more words that are topically related to the word 'CBI', such as 'mrityur'[2] (of death), 'court', 'nihoto' (dead), 'mamla' (case), 'abhijoger' (of accusation), 'abhijukto' (accused), etc. The key idea of our proposed approach is to use these additional topically relevant words in a news document to improve relevance estimation of the tweet hashtags. Our experimental results show that our method improves the F-score of the relevance of tags by 1.79% and the perplexity of the topic model by 23.9% compared to a mono-lingual baseline [4].

ICTIR '16, September 12-16, 2016, Newark, DE, USA
© 2016 ACM. ISBN 978-1-4503-4497-5/16/09...$15.00
DOI: http://dx.doi.org/10.1145/2970398.2970425

[1] We follow this naming convention for the rest of the paper.

[2] We use Roman transliteration to represent Bengali words.

2. RELATED WORK

A graphical model for hashtag relevance prediction for microblogs (such as Flickr, Hatena) was proposed in [4]. This first estimates the topic distribution of the content words and the tags of a document, similar to LDA [1], and then models the likelihood of relevance of the tags based on the topics. The major limitation of this model is that it only works for monolingual documents, and not on cross-lingual document pairs, which we address in this paper. Our work extends the work in [4] to model the relevance of hashtags using documents in a language different from that of the tweets.

The work in [2] suggests hashtags by using topic models to disambiguate the sense of the content words that are candidate tags. The main disadvantage of a word alignment based model is that its effectiveness largely depends on the availability of a parallel corpus. In contrast, we propose a generative approach which does not depend on any linguistic resources. A convolutional neural network model for predicting hashtags was proposed in [7]. The main difficulty with such a deep learning based approach is that a large training set is required for its effective training. A personalized hashtag recommendation method for tweets, based on the tweet content and user preferences, was proposed in [6].

3. PROPOSED METHOD

Our model is motivated by the mono-lingual model of hashtag relevance prediction [4], which we name mono-lingual tag relevance (MTR) model. Before discussing our proposed method, we briefly introduce the MTR model.

Overview of MTR. The solidly outlined circles in Figure 2 represent the variables of the MTR model [4]. The observed variables (shown shaded) in this model correspond to words and tags of a tweet. The topic assigned to a tweet word, w^E, is sampled from the latent topic variables z^E. The generative process of an observed hashtag t is more involved, because in order to jointly model the topics and relevance of a tag, the model assumes the existence of a latent variable r. A value of $r = 1$ indicates relevance of a tag t to the content of a tweet, in which case, t is sampled from a latent topic distribution c^E, which in turn depends on z^E (topics of the content words of the tweet). Otherwise, $r = 0$ indicates that a tag is not related to the content, in which case, t is drawn from a global distribution not related to the content of the tweet. Note that the dimension of τ is $K + 1$ to account for one additional global topic distribution unrelated to the content of a tweet.

Extending MTR to BTR. We now describe the latent variables that we propose to add to the MTR so as to model the relevance of cross-lingual tweets, in our bilingual tweet relevance (BTR) model. The additional variables are shown with a dotted outline. The key difference between this and the MTR model is that in the scenario of cross-lingual tweets, we have a document pair (a tweet with its shared article) instead of only a single document (the tweet itself).

In our proposed BTR model, a shared document-topic distribution, θ, generates the topics of the news document in language F and the tweet in language E. The additional latent variable representing the topics of the news document is z^F, the words of which (denoted by w^F), are drawn from an additional multinomial distribution from the vocabulary of language F, shown as ϕ^F. Moreover, in order to model the fact that the relevance of a hashtag t for a cross-lingual tweet also depends on the topical content of the shared news document, we introduce a latent variable c^F which depends on z^F in a similar way as c^E depends on z^E in MTR. To see the dependence between c^F and t note that the topic distribution of the non-English words contributes to modifying the sampling probability of the relevance variables, r_is, as shown in Equation 5.

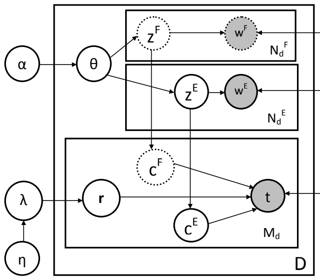


Figure 2: Plate diagram of our proposed generative model for jointly modeling relevance of hashtags of cross-lingual tweets.

BTR Estimation. After describing the key extensions of BTR with respect to MTR, we now provide the estimation details of the BTR model. From Figure 2, it can be seen that the joint distribution of the observed variables, i.e. the words and tags of a tweet in language E (d^E) and those of the document it shares in language F (d^F), depend on the latent variables and the hyper-parameters shown in Equation 1.

$$\begin{aligned} &P(W^F, W^E, T, Z^F, Z^E, C^F, C^E, R, \alpha, \beta, \gamma, \eta) = \\ &P(Z^F|\alpha)P(Z^E|\alpha)P(W^F|Z^F, \beta)P(W^E|Z^E, \beta) \\ &P(T|C^F, C^E, R, \gamma)P(R|\eta)P(C^F|Z^F)P(C^E|Z^E) \end{aligned} \tag{1}$$

The latent topics Z^E and Z^F, given the content words W^E, W^F and the hashtags T, are computed using Gibbs sampling [3]. Following the standard Gibbs sampling exposition for polylingual topic models [5], and using the conditional dependence of C^E on Z^E and that of C^F on Z^F, the sampling probabilities for the latent topic of the j^{th} word of the d^{th} document pair d^E (i.e z^E_j) and d^F (i.e z^F_j) are calculated as shown in Equations 2 and 3, where M^E_{kd} and M^F_{kd} denote the number of tags that are assigned to topic k using d^E and d^F respectively.

$$P(z^E_j = k|Z^E_{\setminus j}) \propto \frac{N^E_{kd\setminus j} + N^F_{kd} + \alpha}{N^E_{d\setminus j} + N^F_d + \alpha K} \frac{N^E_{kw_j\setminus j} + \beta}{N^E_{k\setminus j} + \beta W^E} \Big(\frac{N^E_{kd\setminus j}}{N^E_{d\setminus j}}\Big)^{M^E_{kd}} \tag{2}$$

$$P(z^F_j = k|Z^F_{\setminus j}) \propto \frac{N^E_{kd} + N^F_{kd\setminus j} + \alpha}{N^E_d + N^F_{d\setminus j} + \alpha K} \frac{N^F_{kw_j\setminus j} + \beta}{N^F_{k\setminus j} + \beta W^F} \Big(\frac{N^F_{kd\setminus j}}{N^F_{d\setminus j}}\Big)^{M^F_{kd}} \tag{3}$$

Equations 2 and 3 allow joint modeling of the topics from the content of a document pair.

Equations 4 and 5 show how the latent variable for hashtag-topic relevance r_i (i.e relevance of the i^{th} tag of the d^{th} document d^E) are sampled. It can be seen from Equation 5 that the probability of relevance of a tag increases with the likelihood of the co-occurrence of that tag with a topic, which is estimated by the global tag-topic co-occurrence counts for the corresponding languages E and F, denoted respectively by $M^E_{c,t}$ and $M^F_{c,t}$. $M_{0\setminus i}$ in 4 and 5 denotes the number of non-relevant tags in the tweet collection excluding the i^{th} tag of the d^{th} document d^E.

$$P(r_i = 0|R_{\setminus i}) \propto \frac{M_{0\setminus i} + \eta}{M_{\setminus i} + 2\eta} \frac{M_{0t_i\setminus i} + \gamma}{M_{0\setminus i} + \gamma T} \tag{4}$$

$$P(r_i = 1|R_{\setminus i}) \propto \frac{M_{\setminus i} - M_{0\setminus i} + \eta}{M_{\setminus i} + 2\eta} \frac{M^E_{c_it_i\setminus i} + M^F_{c_it_i\setminus i} + \gamma}{M^E_{c_i\setminus i} + M^F_{c_i\setminus i} + \gamma T} \tag{5}$$

The assignment of a topic to a content unrelated hashtag is given by the maximum likelihood estimates from the corresponding documents, i.e. $P(c_i^E = k|r_i = 0, C^E_{\setminus i}, R_{\setminus i}) = N^E_{kd}/N^E_d$ and a similar expression with the corresponding variables for d^F. On the other hand, the assignment of a topic to a content related hashtag is estimated according to Equation 6 and 7.

$$P(c_i^E = k|r_i = 1, C_{\setminus i}, R_{\setminus i}) \propto \frac{M^E_{kt_i\setminus i} + \gamma}{M^E_{k\setminus i} + \gamma T}\frac{N^E_{kd}}{N^E_d} \qquad (6)$$

$$P(c_i^F = k|r_i = 1, C_{\setminus i}, R_{\setminus i}) \propto \frac{M^F_{kt_i\setminus i} + \gamma}{M^F_{k\setminus i} + \gamma T}\frac{N^F_{kd}}{N^F_d} \qquad (7)$$

$M^F_{kt_i\setminus i}$ of Equation 7 denotes the number of times tag t_i is assigned the topic k, using the documents in F (and similarly $M^E_{kt_i\setminus i}$ for the documents in E). Hence, Equations 6 and 7 make a tag-topic association more likely if the topic itself occurs frequently in *both* d^F and d^E (instead of d^E alone as in MTR).

The values of the latent variables of the model, i.e. the topics of the English tweets and the non-English news documents along with the relevance of the tags, i.e. the r_i variables are estimated by executing Gibbs sampling iterations.The r values are eventually used to measure how effectively the relevance of the tags are predicted.

4. EXPERIMENTAL SETUP

DataSet. One of the difficulties in collecting a dataset of cross-lingual tweets is that despite the presence of numerous bilingual Twitter users, due to the limitations of the Twitter streaming API, it is difficult to implement a streaming service that can automatically track such tweets. A much simpler solution is to track a particular Twitter account that is known to post such cross-lingual tweets. Consequently, for the purpose of building the dataset for our investigation, we collected tweets from the Twitter account of a leading Bengali (a South Asian language) news daily Anandabazar Patrika (ABP)[3]. The Twitter account of ABP[4] is a bilingual account that posts tweets sharing Bengali news articles both in Bengali and English. For our research, we collected only the English tweets posted by the ABP twitter account by making use of the language identifier settings of the Twitter API. To collect the data, Twitter data streaming was executed for about 3 months[5]. Out of a total of 13,299 tweets collected from this account, $1,370$ tweets were cross-lingual and hence used for our experiments (see Table 1).

A filtering step ensured that each tweet in our dataset has a relevant news article (in Bengali) linked to it. Tweets with no linked news articles were discarded. The characteristics of the cross-lingual tweet dataset used for our experiments are shown in Table 1. To measure the effectiveness of BTR, we need a reference set of relevant tags for each tweet. Since all of the tweets in our dataset were posted by a news publisher, the tags are mostly carefully selected according to content relevance. However, to make a more realistic dataset where tweets have a mixture of both relevant and non-relevant tags, we randomly assigned a number of tags to each tweet. These tags are considered to be the non-relevant during evaluation, i.e. the objective of the model is to predict that these additional tags are non-relevant. The source of these randomly assigned tags is the entire tag vocabulary of the dataset, which makes the chance of adding a truly relevant tag to a tweet very unlikely.

Baseline. The objective of our experiments is to show that jointly modeling hashtag relevance by additionally using the content of the document shared from a tweet can improve relevance prediction effectiveness. Consequently, as a baseline for our experiments, we use the MTR model, which makes use of the tweet text only to predict hashtag relevance. Bilingual LDA [5] cannot be used as a baseline because it is a model for generating only the content words in two languages with latent topic distributions for each language. It does not however model tag relevance.

Table 1: Characteristics of the cross-lingual tweet dataset.

Attribute	Value
# English Tweets and corresponding Bengali news articles	1370
Vocabulary size of English tweets	4550
Tag vocabulary size of tweets	1325
Vocabulary size of Bengali news articles	55201
Avg. # words in an English tweet (without URLs and stopwords)	3.32
Avg. # words in a Bengali news article	40.29
Avg. # tags per tweet	2.05
Overlap between tweet words and tag words	39.40%

Machine translation (MT) is one way of bridging the vocabulary gap, with which one would be able to apply the mono-lingual tag relevance prediction model. However, there are two main reasons for not using MT in our experiments. Firstly, the availability of parallel corpora for Bengali-English translation is limited. The only parallel corpus that we are aware of comprises 48K parallel sentences in health and tourism[6]. Secondly, the key motivation of our approach is to be able to bridge the vocabulary gap with the help of a completely unsupervised approach even without the presence of any translation resource at all.

Parameters. For collapsed Gibbs sampling of both BTR and MTR, we use 1000 iterations as prescribed in [3]. The LDA hyper-parameters for the Dirichlet priors of the document-topic and the topic-term distributions were set to $50/K$ (K being the number of topics) and 0.01 respectively according to [3]. For BTR, the topic-term distribution priors for both E and F, were set to 0.01. The hyper-parameters η and γ were set to 0.01 for both BTR and MTR.

Evaluation Metrics. To compare BTR against MTR, we use two standard evaluation measures. The first measure, called *perplexity* (shown in Equation 8), uses the posterior document-topic (θ) and the topic-tag distributions (τ) to measure how stable the posterior estimates are. A lower value of perplexity indicates higher posterior likelihood of the observed variables, i.e. the content words and tag words. Our second evaluation measure is targeted towards directly measuring the effectiveness of the hashtag relevance prediction. We calculate the *F-score* by comparing the estimated r_i values for both the relevant and the non-relevant tags with their true values. A higher F-score indicates that there is a better agreement between the true r_i values and the estimated ones, and that the model is more effective in distinguishing relevant tags from non-relevant ones.

$$H = \exp\left(-\frac{\sum_{d=1}^{M} log(P_d(t_1 \ldots t_{M_d}|\theta,\tau))}{\sum_{d=1}^{M} N_d}\right), P_d(t) = \sum_{k=1}^{K}\theta_{dk}\tau_{kt} \qquad (8)$$

5. RESULTS

In our initial experiments, we set the number of topics (K), used for estimating both MTR and BTR, to 10. To evaluate tag relevance, we artificially add one non-relevant hashtag to each tweet. Table 2 shows the results with this settings. Firstly, it can be seen that BTR produces a lower perplexity score in comparison to MTR in Table 2), which indicates that the posterior distributions are more

[3] http://www.anandabazar.com/

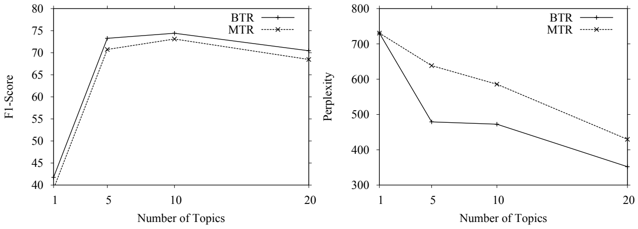
[4] https://twitter.com/MyAnandaBazar

[5] For dataset and code see https://bitbucket.org/procheta/cltagrel

[6] http://sanskrit.jnu.ac.in/projects/ilci.jsp?proj=ilci

Method	Evaluation Metrics	
Name	Perplexity	F-score
MTR	585.76	0.7311
BTR	**472.51**	**0.7442**

Table 2: Relevance prediction effectiveness (F-score) of cross-lingual tweets. #topics was set to 10. One non-relevant hashtag was added for each tweet.



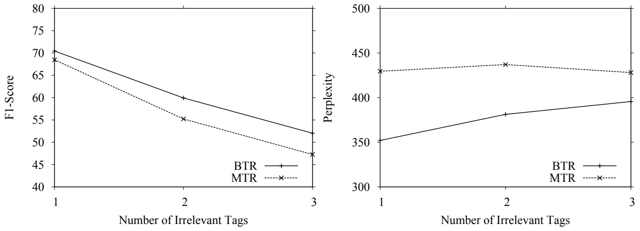
(a) F-measure (b) Perplexity

Figure 3: Sensitivity of F-score and perplexity on #topics.

stable for BTR. Secondly, in terms of predicting the relevance of tags, we see that BTR achieves a higher F-score than MTR. The improvements in F-score are statistically significant as measured by the Wilcoxon test with 95% confidence measure. This verifies our hypothesis that making use of the shared news article helps to provide the additional context for the tag relevance to perform better. The topics estimated jointly over each document pair (i.e. the tweet and the news) are more robust than those estimated over a single tweet, because in BTR, the distribution $P(C^F|Z^F)$ (see Figure 2) improves the prediction of hashtag relevance. In contrast to MTR, this additional factor helps to identify more relevant tags, which is also evident from the higher F-score.

Sensitivity to the number of Topics. For our next set of experiments, we vary the number of topics to see how this affects the perplexity and the F-score values of the hashtag relevance models. The results are shown in Figure 3. It can be seen from Figure 3a that with a degenerate case of only 1 topic, the performance of both MTR and BTR is low. In fact, with a lower number of topics, MTR yields better results than BTR. However, it can be seen from Figure 3a, that the results improve when the number of topics in BTR is increased. It can also be seen that the effectiveness of BTR is optimal with $K = 10$, i.e. when 10 topics are used to estimate the model. A further increase in the number of topics exhibits a steady decrease in F-score for both the models. The effect of the number of topics on perplexity is shown in Figure 3b. We observe that the perplexity of BTR is consistently lower than that of MTR.

Sensitivity to the number of non-relevant tags. For our next set of experiments, we added an increasing number of non-relevant tags to the tweets. The purpose of this set of experiments was to examiner the robustness of the models in the presence of noisy data. The results are shown in Figure 4. As expected, the effectiveness of both the models decreases with an increasing number of non-relevant tags. However, it can be seen from Figure 4a that BTR consistently outperforms MTR even with an increasing number of non-relevant tags, which indicates that BTR is able to recognize relevant tags more effectively than its monolingual counterpart in the presence of non-relevant tags. Figure 4b shows that the perplexity of both the models increases with an increasing number of non-relevant tags. For MTR, this perplexity variation effect with increasing noise is consistent with the observations reported in [4] on different datasets such as Hatena, Delicious and Flickr.



(a) F-measure (b) Perplexity

Figure 4: Sensitivity to the number of non-relevant tags.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we addressed the problem of predicting the relevance of hashtags for a tweet which shares an article in a language different from the language of the tweet itself (we call such tweets cross-lingual tweets). We hypothesize that the relevance of hashtags of cross-lingual tweets depends on the topical content of the articles that they share. We proposed a generative model to jointly model the topics in each document pair comprised of the tweet and the shared article in two different languages. The key idea is that the topics extracted from the content words of the foreign language article can improve on the hashtag relevance prediction performance of the tweets. Our experiments, conducted on a set of cross-lingual tweets, verify this claim. Our proposed model (BTR) consistently outperforms its monolingual counterpart (MTR) over a varying range of number of topics and non-relevant tags. As a part of our future work, we would like to extend our proposed model to a non-parametric version that would not require a preset number of topics. Another idea is to incorporate other user generated signals, such as retweet count, favourites count etc. as a part of the generative model to improve modeling relevance of hashtags.

Acknowledgements: This research is supported by Science Foundation Ireland (SFI) as a part of the ADAPT Centre at DCU (Grant No: 13/RC/2106).

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] Z. Ding, Q. Zhang, and X. Huang. Automatic hashtag recommendation for microblogs using topic-specific translation model. In *Proc. of COLING'12*, pages 265–274, 2012.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101:5228–5235, 2004.

[4] T. Iwata, T. Yamada, and N. Ueda. Modeling social annotation data with content relevance using a topic model. In *Proc. of NIPS '09*, pages 835–843, 2009.

[5] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *Proceedings of the EMNLP '09*, pages 880–889, 2009.

[6] E.-P. L. Su Mon Kywe, Tuan-Anh Hoang and F. Zhu. On recommending hashtags in twitter networks. In *Proceedings of ICSI 2012*, pages 337–350, 2012.

[7] J. Weston, S. Chopra, and K. Adams. #tagspace: Semantic embeddings from hashtags. In *Proc. of EMNLP '14*, pages 1822–1827, 2014.