Automatic Hashtag Recommendation for Microblogs using Topic-specific Translation Model

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

Abstract

Microblogging services continue to grow in popularity, users publish massive instant messages every day through them. Many tweets are marked with hashtags, which usually represent groups or topics of tweets. Hashtags may provide valuable information for lots of applications, such as retrieval, opinion mining, classification, and so on. However, since hashtags should be manually annotated, only 14.6% tweets contain them (Wang et al., 2011). In this paper, we adopt topic-specific translation model(TSTM) to suggest hashtags for microblogs. It combines the advantages of both topic model and translation model. Experimental result on dataset crawled from real world microblogging service demonstrates that the proposed method can outperform some state-of-the-art methods.

Title and Abstract in Chinese

基于特定话题下翻译模型的微博标签推荐

微博服务变得越来越流行，用户可以通过微博提交大量的及时信息。很多条微博被用户通过标签标记，这些标签代表了微博的话题类别。标签可以为很多应用提供有价值的信息，比如检索，情感分析，分类等等。微博的标签本应该由用户自行标记，然而，根据统计只有14.6%的微博包含标签。在这篇论文中，我们提出了一种基于特定话题的翻译模型， 来为每条微博自动推荐标签。此模型综合了话题模型和翻译模型的优点。在基于真实微博语料的实验中，我们提出的方法超过了很多经典的方法。

Keywords: Microblogs, Tag recommendation, Topic model.

Keywords in Chinese: 微博，标签推荐，话题模型.

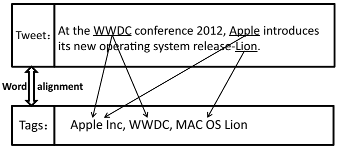
1 Introduction

Hashtags, which are usually prefixed with the symbol # in microblogging services, represent the relevance of a tweet to a particular group, or a particular topic (Kwak et al., 2010). Popularity of hashtags grows concurrently with the rise and popularity of microblogging services. Many microblog posts contain a wide variety of user-defined hashtags. It has 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), and many other applications. However, not all posts are marked with hashtags. How to automatically generate or recommend hashtags has become an important research topic.

The task of hashtag recommendation is to automatically generate hashtags for a given tweet. It is similar to the task of keyphrase extraction, but it has several different aspects. Keyphrases are defined as a short list of phrases to capture the main topics of a given document (Turney, 2000). Keyphrases are usually extracted from the given document. However, hashtags indicate where a tweet is about a particular topic or belong to a particular group. So words and hashtags of a tweet are usually diverse vocabularies, or even hashtags may not occur in the tweet. Take the tweet in Table 1 for instance, the word "Lion" is used in the tweet, while users annotate with the hashtag "Mac OS Lion". That is usually refered to as a *vocabulary gap* problem.

Tweet
At the WWDC conference 2012, Apple introduces its new operating system release-**Lion**.
Annotated tags
Apple Inc, WWDC, **MAC OS Lion**

Table 1: An example of a tweet with annotated hashtags.



At the WWDC conference 2012, Apple introduces its new operating system release-Lion.

Figure 1: The basic idea of word alignment method for suggesting hashtags.

To solve the *vocabulary gap* problem, most researchers applied a statistic machine translation model to learn the word alignment probabilities(Zhou et al., 2011; Bernhard and Gurevych, 2009). Liu et al. (2011) proposed a simple word alignment method to suggest tags for book reviews and online bibliographies. In this work, tags are trigged by the important words of the resource. Figure 1 shows the basic idea of using word alignment method for tag suggestion.

Due to the open access in microblogs, topics tend to be more diverse in microblogs than in formal documents. However, all the existing models did not take into account any contextual information in modeling word translation probabilities. Beyond word-level, contextual-level topical information can help word-alignment choice because sometimes translation model is vague due to their reliance solely on word-pair co-occurrence statistics. For example, the word "apple" should be translated into "Apple Inc" in the topic of *technology*, or "juice" in the topic of *drink*. Thus the idea is using topic information to facilitate word alignment choice.

Based on this perspective, in this paper, we propose a topic-specific translation model(TSTM) to recommend hashtags for microblogs. This method regards hashtags and tweets as *parallel* description of a resource. We first investigate to combine topic model and word alignment model to estimate the topic-specific word alignment probabilities between the words and hashtags. After that, when given an unlabeled dataset, we first identify topics for each tweet and then compute importance scores for candidate tags based on the learned topic-specific word-

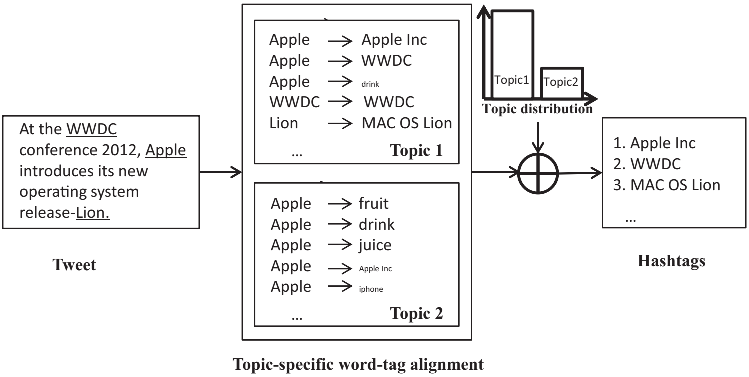


Figure 2: The basic idea of topic-specific word alignment for tag recommendation.

tag alignment probabilities and topic distribution. Figure 2 illustrates the basic idea of our model. In Figure 2, for simplicity, we suppose there are totally two topics, topic 1(information technology) and topic 2(food). We use the font size of tags to indicate the word-tag alignment probability for each specific topic. With the topic distribution and word-tag alignment probabilities for each topic, we can compute the importance score for each candidate tag.

The remainder of this paper is organized as follows: related work and state-of-the-art approaches are reviewed in Section 2. The proposed approach is detailed in Section 3. Experimental results and analysis are described and discussed in Section 4. The last section concludes the paper.

2 Related work

Our approach relates to two research areas: tag suggestion and keyphrase extraction. In this section, we discuss them in detail.

2.1 Tag suggestion

Previous work on tag suggestion can be roughly divided into three directions, including collaborative filtering(CF) (Rendle et al., 2009; Herlocker et al., 2004), discriminative models (Ohkura et al., 2006; Heymann et al., 2008), and generative models(Krestel et al., 2009; Iwata et al., 2009). Our proposal is complementary to these efforts, because microblogs differ from other media in some ways: (1) microblog posts are much shorter than traditional documents. (2) topics tend to be more diverse than in formal documents. So these methods cannot be directly applied to hashtag recommendation in microblogs.

2.2 Keyphrase extraction

Keyphrase extraction from documents is the most similar task to this research. Existing methods can be categorized into supervised and unsupervised approaches. Unsupervised approaches usually selected general sets of candidates and used a ranking step to select the

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
$w_d = \{w_{dn}\}_{n=1}^{N_d}$	words in the dth tweet
$z_d = \{z_{dn}\}_{n=1}^{N_d}$	topic of each word in the dth tweet
$t_d = \{t_{dm}\}_{m=1}^{M_d}$	hashtags in the dth tweet
$c_d = \{c_{dm}\}_{m=1}^{M_d}$	topic of each hashtag in the dth tweet

Table 2: Notations of our model.

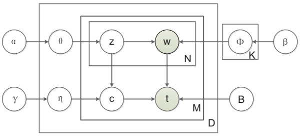


Figure 3: Graphical model representation of our model.

most important candidates (Mihalcea and Tarau, 2004; Wan and Xiao, 2008). Supervised approaches used a corpus of training data to learn a keyphrase extraction model that is able to classify candidates as keyphrases (Turney, 2003; Hulth., 2003).

3 Proposed method

3.1 Preliminaries

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 2. Given an unlabeled data set, the task of hashtag recommendation is to discover a list of hashtags for each tweet.

The proposed topic-specific 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 some topics based on the topic distribution. Then he chooses a bag of words one by one based on the word distribution for each chosen topic. During the generative process for hashtags, a topic is first chosen from topics that have previously generated the content. And hashtags are chosen according to the chosen topic and important words in the content.

Formally, let θ denotes the topic distribution and ϕ_k denotes the word distribution for topic k. Let η_d denote the distribution of topic choice when assigning hashtags for the dth tweet and the choice probability of topic k is sampled randomly from topics of content, as follows, $\eta_{dk} = \frac{N_k^d + \gamma}{N_{(.)}^d + K\gamma}$, where N_k^d is the number of words that are assigned to topic k in the dth tweet. And then each hashtag t_{dm} is annotated according to topic-specific translation possibility $P(t_{dm}|w_d, c_{dm}, \mathbf{B})$, where $P(t_{dm}|w_d, c_{dm}, \mathbf{B}) = \sum_{n=1}^{N_d} P(t_{dm}|c_{dm}, w_{dn}, \mathbf{B})P(w_{dn}|w_d)$ and $\mathbf{B}$ 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, $P(w_{dn}|w_d)$ indicates the importance of the word in the dth tweet, which will be described in detail in section 3.4.2 .

In summary, the generation process of annotated tweets is described as follows:

1. Draw topic probability $\theta \sim$ Dirichlet (α);
2. Draw topic probability $\eta \sim$ Dirichlet (γ);
3. For each topic $k = 1, ..., K$

   Draw word probability $\phi_k \sim$ Dirichlet (β)
4. For each tweet $d = 1, ..., D$

   (a) For each word $n = 1, ..., N_d$

   Draw topic $z_{dn} \sim$ Multinomial (θ_d)

   Draw word $w_{dn} \sim$ Multinomial (Φz_{dn})

   (b) For each hashtag $m = 1, ..., M_d$

   Draw topic $c_{dm} \sim$ Multinomial (η_d)

   Draw hashtag $t_{dm} \sim P(t_{dm}|w_d, c_{dm}, \mathbf{B})$

where α, β and γ are Dirichlet distribution parameters.

Figure 3 shows a graphical model representation of the proposed model.

3.2 Learning and inference

We use collapsed Gibbs sampling(Griffiths and Steyvers, 2004) to find latent variables. The sampling probability of a latent topic for each word and hashtag in the tweet is sampled respectively. Due to the space limit, we leave out the derivation details and the sampling formulas.

After the topics of each word and hashtag become stable, we can estimate topic-specific word alignment table B by: $B_{t,w,c} = \frac{N^t_{c,w}}{N^{(.)}_{c,w}}$. where $N^t_{c,w}$ is a count of the hashtag t that co-occurs with the word w for topic c in tweet-hashtag pairs.

The possibility table $B_{t,w,c}$ have a potential size of WTK, assuming the vocabulary sizes for words, hashtags and topics are W, T and K. The data sparsity poses a more serious problem in estimating $B_{t,w,c}$ 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 sparseness: $B^*_{t,w,c} = \lambda B_{t,w,c} + (1 - \lambda)P(t|w)$, where $P(t|w)$ is topic-free word alignment probability from the word w and the hashtag t, λ is trade-off of two probabilities. Here we explore IBM model-1 (Brown et al., 1993), which is a widely used word alignment model, to obtain $P(t|w)$.

3.3 Tag recommendation using Topic-specific translation probabilities

3.3.1 Topic identification

Suppose given an unlabeled dataset $\mathbf{W}^* = \{w^*_d\}^U_{d=1}$ with U tweets, where the dth tweet $w^*_d = \{w^*_{dn}\}^{L_d}_{n=1}$ consists of L_d words. $z^*_d = \{z^*_{dn}\}^{L_d}_{n=1}$ denotes topics of words in dth tweet and $\mathbf{Z}^* = \{z^*_d\}^U_{d=1}$. we first identify topics for each tweet using the standard LDA model. The collapsed Gibbs sampling is also applied for inference. After the topics of each word become stable, we can estimate the distribution of topic choice for hashtags of the dth tweet in unlabeled data by: $\eta^*_{dk} = \frac{N^d_k + \gamma}{N^d_{(.)} + \gamma K}$, where N^d_k is a count of words that are assigned topic k in the dth tweet of unlabeled dataset.

3.3.2 Tag recommendation

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

$$P(t^*_{dm}|w^*_d, \eta^*_d, \mathbf{B}^*) = \sum_{c^*_{dm}=1}^{K} \sum_{n=1}^{L_d} P(t^*_{dm}|c^*_{dm}, w^*_{dn}, \mathbf{B}^*) P(c^*_{dm}|\eta^*_d) P(w^*_{dn}|w^*_d)$$

Where $P(w^*_{dn}|w^*_d)$ indicates the importance of the words in the tweet. Here, we used IDF to compute this importance score. According to the ranking scores, we can suggest the top-ranked hashtags for each tweet to users.

4 Experiments

4.1 Data collection and analysis

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. It was launched in August, 2009 and quickly become the most popular microblogging service in China. We collected a dataset with totally **10,320,768** tweets. Among them, there are **551,479** tweets including hashtags annotated by users. We extracted these annotated tweets for training and evaluation. Some detailed statistical information is shown in Table 3. We divided them into a training set of 446,909 tweets and a test set of 104,570 tweets. The training set is applied for building topic-specific translation model, while the test set is for evaluation. We use hashtags annotated by users as the golden set.

#tweet	W	T	$\bar{N_w}$	$\bar{N_t}$
551,479	244,027	116,958	19.97	1.24

Table 3: 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.

4.2 Evaluation metrics and settings

We use Precision(P), Recall(R), and F-value(F) to evaluate the performance of hashtag recommendation methods. We ran topic-specific translation model with 1000 iterations of Gibbs sampling. After trying a few different numbers of topics, we empirically set the number of topics to 100. We use $\alpha = 50.0/K$ and $\beta = 0.1$ as (Griffiths and Steyvers, 2004) suggested. Parameter γ is also set to 0.1. We use IDF to indicate the importance of a word and set smoothing parameter λ to 0.8 which gives the best performance. The influence of smoothing to our model can be found in Section 4.5.

4.3 Comparison with other methods

In this subsection, we implement several methods for comparison, where Naive Bayes(NB) is a representative classification method, while LDA (Krestel et al., 2009) is selected to represent generative model for tag suggestion, IBM model-1 (Liu et al., 2011) is a novel translation-based model.

[1] http://weibo.com/

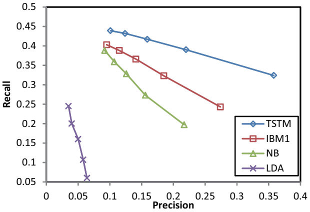


Figure 4: Performance comparison between NB, LDA-based, IBM1 and TSTM.

Method	Precision	Recall	F-measure
NB	0.217	0.197	0.203
LDA	0.064	0.060	0.062
IBM1	0.271	0.241	0.249
TSTM	**0.358**	**0.324**	**0.334**

Table 4: Comparison results of NB, LDA-based, IBM1 and TSTM when suggesting top-1 hashtag.

In Figure 4, we show the Precision-Recall curves of NB, LDA, IBM1 and TSTM 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)TSTM outperforms all the baselines. This indicates the robustness and effectiveness of our approach for hashtag recommendation. (2)IBM1 underperforms TSTM, because IBM1 relies solely on word-tag co-occurrence statistics. And contextual topical information can help to disambiguate word-alignment choices in TSTM. (3)LDA performs so poor, because it ranks the candidate hashtags by the hashtag distribution for each topic. So it can only suggest general hashtags.

To further demonstrate the performance of TSTM and other baseline methods, in Table 4, we show the Precision, Recall and F-measure of NB, LDA, IBM1 and TSTM suggesting top-1 hashtag, because the number is near the average number of hashtags in dataset. We find that the F-measure of TSTM comes to 0.334, outperforming all the baselines more than 8%.

4.4 Example

In Table 5, we show top-8 hashtags suggested by NB, LDA, IBM1 and TSTM for the tweet in Table 1[2]. The number in brackets after the name of each method is the count of correctly suggested hashtags. The correctly suggested hashtags are marked in bold face.

From Table 5, we observe that classification model NB suggests some unrelated hashtags. While LDA, as generative models, tends to suggest general hashtags, such as "Information News", "mobile phone" and "Technology leaders", and fail to generate the specific hashtags "WWDC", "MAC OS Lion". IBM1 method will suggest some topic-unrelated hashtags. For instance, "2012 Jinshan Inc cloud computing" and "2012 spring and summer men's week" are triggered by the word "2012". On the contrary, TSTM succeeds to suggest specific hashtags, and most of them are topic-related to the tweet.

4.5 Influences of smoothing

To validate the power of smoothing in TSTM on different sizes of datasets, the experiments were conducted on two datasets, including a small dataset(a training set of 100,000 tweets

[2] Hashtags are translated from Chinese

NB(+1): **MAC OS Lion**, 2012 wishes, OS, Smiles to the world, 2012 salary report, 2012 Jinshan Inc cloud computing, Lion, Noah's ark 2012
LDA(+1): Android, Information news, Japan earthquake, mobile phone, **Apple Inc**, Cloud computing, Tablet PC, Technology leaders
IBM1(+2): **WWDC**, Android, 2012 Jinshan Inc cloud computing, **Apple Inc**, 2012 spring and summer men's week, 2012, mobile phone OS, Information news
TSTM(+3): **Mac OS Lion**, **WWDC**, MAC, **Apple Inc**, Baidu union conference, Microsoft, Android, iphone

Table 5: Top-8 hashtags suggested by NB, LDA, IBM1 and TSTM.

and a test set of 10,000 tweets) and a large dataset(100% training set and 100% test set). Figure 5 and Figure 6 show the performance on both of the datasets when λ ranges from 0.0 to 1.0. We find that TSTM achieves the best performance when $\lambda = 0.8$ in both of the two Figures. Furthermore, the model cannot perform well without smoothing (when $\lambda = 1$) on the small data set. That indicates smoothing is more powerful on the small data set. While the model can still perform well without smoothing on the large data set. This is reasonable because large data set can help to solve the problem of data sparsity to some extent.

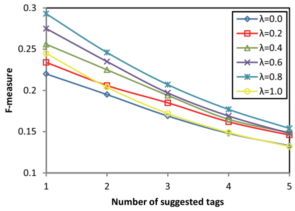


Figure 5: F-measure of TSTM on the small data set when smoothing parameter λ ranges from 0.0 to 1.0.

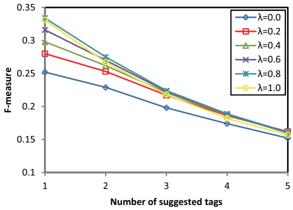


Figure 6: F-measure of TSTM on the large data set when smoothing parameter λ ranges from 0.0 to 1.0.

Conclusions

In this paper, we address the issue of suggesting hashtags for microblogs. The existing methods cannot be directly applied to this task due to the following challenges. (1) tweets are much shorter than traditional documents. (2) topics are more diverse in microblogs than other media. To solve these problems, we proposed a topic-specific translation model, which combines the advantages of both topic model and translation model. Experimental result on tweets crawled from real world service demonstrates that the proposed method can outperforms some state-of-the-art methods.

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), and "Chen Guang" project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (11CG05).

References

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

Bernhard, D. and Gurevych, I. (2009). Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceeding of ACL*, pages 728–736.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The machematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):263–311.

Davidov, D., Tsur, O., and Rappoport, A. (2010). 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. Association for Computational Linguistics.

Efron, M. (2010). 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. ACM.

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

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53.

Heymann, P., Ramage, D., and Garcia-Molina, H. (2008). 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. ACM.

Hulth., A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *proceedings of EMNLP*.

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

Krestel, R., Fankhauser, P., and Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. In *RecSys*.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

Liu, Z., Chen, X., and Sun, M. (2011). A simple word trigger method for social tag suggestion. In *Proceedings of EMNLP*.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. In *Proceedings of EMNLP*.

Ohkura, T., Kiyota, Y., and Nakagawa, H. (2006). Browsing system for weblog articles based on automated folksonomy. *Workshop on the Weblogging Ecosystem Aggregation Analysis and Dynamics at WWW*.

Rendle, S., Balby Marinho, L., Nanopoulos, A., and Schmidt-Thieme, L. (2009). 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. ACM.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2(4):303–336.

Turney, P. D. (2003). Coherent keyphrase extraction via web mining. In *proceedings of IJCAI*.

Wan, X. and Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of AAAI*.

Wang, X., Wei, F., Liu, X., Zhou, M., and Zhang, M. (2011). 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. ACM.

Zhou, G., Cai, L., Zhao, J., and Liu, K. (2011). Phrase-based translation model for question retrieval in community question answer archives. In *Proceeding of ACL*, pages 653–662.