QCRI @ TREC 2011: Microblog Track

Ali El-Kahki, Kareem Darwish
Qatar Computing Research Institute, Qatar Foundation
ali_kahki@yahoo.com, kdarwish@qf.org.qa

Abstract

This paper briefly describes the Qatar Computing Research Institute (QCRI) participation in the TREC 2011 Microblog track. The focus of our TREC submissions was on using a generative graphic model to perform query expansion. We trained a model that attempted to predict appropriate hashtags to expand tweets as well as queries. In essence, we used hashtags to represent latent topics in tweets.

1. Introduction

Searching tweets, a popular type of microblogs, poses interesting research problems. Some of these problems include: 1) the short length of tweets limits the contexts that are available for search; and 2) the language of tweets typically contains non-standard abbreviations and colloquial expressions. We focused on solving the first problem that is related to the short length of tweets. In particular, we focused on bringing more contexts to tweets by expanding tweets and queries alike using appropriate hashtags. Essentially, we used hashtags to represent latent underlying topics in tweets. Massoudi et al. (2011) showed that Hashtags can be effective expansion terms in the context of search in microblogs. Though hashtags appear in less than 19% of all tweets[1] and popular hashtags are often used by spammers, there are sufficient numbers of tagged tweets to build effective hashtag models. We used a Latent Dirichlet Allocation (LDA) like graphical model to learn the relationship between words, latent topics, and hashtags. We assumed that the relationship between latent topics and hashtags to be m to n and that each tweet contains only one topic.

In remainder of the paper, we will describe: the preprocessing we performed on tweets (Sec. 2); the graphical model that we employed (Sec. 3); the experimental setup of the submitted runs (Sec. 4); and official TREC results for our runs (Sec. 5). We finally conclude the paper (Sec. 6).

2. Tweet Preprocessing

According to the track guidelines, only English tweets are considered relevant. Thus, we needed to extract the English from the approximately 16 million tweets in the collection. We used the language-detection open source Java library[2]. In all, we extracted roughly 4.8 million English tweets. We performed basic text tokenization where words were split on delimiters, except for “#” and “@” as they signify hashtags and user mentions respectively.

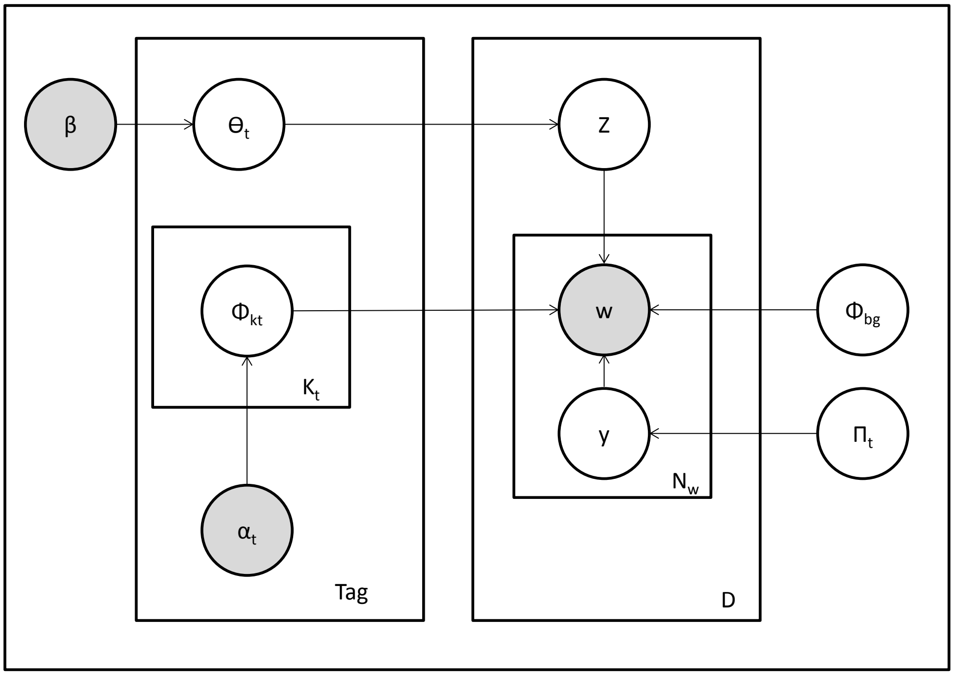
3. Graphic Model

We used an LDA-like graphical model. Figure 1 shows the plate representation of the model that we used. Formally, for each tag T, a set of documents D, Θ_t represents a distribution over different topics Z.

[1] Based on the English tweets in the TREC Microblog dataset
[2] http://code.google.com/p/language-detection/

For each tag also, there is a set of topics Φ_{kt} and distribution Π_t of for background and foreground probability. Each document is represented by its topic Z and a set of words. Words w are generated from either background distribution Φ_{bg} or from corresponding topic distribution $\Phi_{Z,t}$ based on whether the word is background or foreground which is determined by the binary variable y.

Figure 1. Generative process of proposed model.



The main differences between the standard LDA model and our model are

1. We assumed that each tweets is generated from only topic. This is a reasonable assumption for short documents like tweets and it has been used by Zhao et al. (2011).
2. We assumed that each tag contains a set of topics denoted by Θ_t to model different topics covered by the same tag. We limited the number of topics per tag to 3.
3. We used a global background distribution over all tags denoted by Φ_{bg}. Background word distribution has been used before in LDA models (Zhao et al., 2011). Our Model is different because we used separate background to foreground probability for each tag denoted by Π_t

We used the Factorie toolkit to describe the model and to perform inference. Factorie is an open source package that allows for factored graph construction, parameter estimation, and inference (McCullam et al., 2009).

4. Experimental Setup

In inspecting the hashtags that were used in the tweets, we found that there were 223,145 unique hashtags, of which 203,065 were used 5 times or fewer. Hashtags that appear very few times may represent hashtags that were not adopted by other users for reasons beyond the scope of this paper. Some of the hashtags that were used just once include: #ilovejakewolf, #federalism, #whencanistart, #grungy, #andywho, and #promisingnight.

Tags that are used tens of times (not hundreds of times) tend to belong some of the following categories:

- Too specific such as #veronicamarsmovie (used 55 times)
- Unpopular tags such as #greatmovie (used 14 times) and #cinema (used 30 times), where there are more popular tags such #movies and #movie (used 400 and 220 times respectively)
- Topics with waning interest such as #tudors, #thetudors, #tudors4, and #tudor (used 7, 3, 3, 1 times respectively)
- Nondescript tags such as #days and #all (used 37 and 70 times respectively)
- Tags of narrow interest such as #luton and #medicaljobs (used 38 and 36 times respectively)

More frequent tags, that are used hundreds of times, typically have broad interest. Some examples with high count include #xbox (used 240 times), #breakingnews (used 170 times), #blackhistorymonth (used 869 times), #packers (used 2,226 times), and #jan25 (used 17,810 times).

We chose tags that appeared at least 100 times in the tweets, limiting the number of hashtags to 1,208. There were two main reasons for this choice: 1) we wanted tags that cover broad or more popular interests, and 2) we were constrained by the computational capabilities[3].

Then given all tweets in the collection (regardless of whether or not they had hashtags), we used our model to generate the most likely 5 hashtags for each tweet and for each query. The hashtags were appended to the query or the tweet if the probability of the inferred hashtag was greater than 0.001. Hashtag prediction had mixed results with variable success. Table 1 shows some the inferred hashtags for some tweets and queries. Some of the successfully inferred hashtags are those for T2 and Q10. Partial success was achieved for T1 and Q2, and no success was achieved for T3 and Q50. Further analysis is required to approximate the accuracy of hashtag prediction. However, it is noteworthy that consistency of prediction could be more important than correctness of prediction. For example, for the two tweets: “upset at my dad because he bought sugarfree apple juice” and “eating an apple”, the hashtag #ipad was inferred. Though it is incorrect for both, perhaps eating apples is related to drinking apple juice and the inferred hashtag can help connect them. Again, further error analysis is required to ascertain the effect of such errors.

Table 1: Sample tweets and queries and the inferred hashtags

No.	Tweet/Query	Inferred hashtags
T1	saying no to carbs is saying no to fat loss. what follows is a dull mind, tired body & frustrated soul.	#health, #fitness, #gemini, #sagittarius, #knockitoff
T2	with journalists hounded out of tahrir square the crackdown on protestors could be worse overnight, with few cameras to catch it	#tahrir, #cairo, #jan25, #egypt, #aljazeera
T3	"great things are not done by impulse, but by a series of small things brought together." vincent van gogh	#blackparentquotes, #cricket, #skins, #arsenal, #neversaynever
Q2	2022 FIFA soccer	#football, #sports
Q10	Egyptian protesters attack museum	#jan25, #egypt, #mubarak, #cairo, #tahrir
Q50	war prisoners hatch act	#ipod, #iphone, #ipad

[3] Even with the filtering of tags based on the number of times they appeared, the training of the graphical model used 40 G of RAM.

We used Indri to index all tweets twice: once in their original form, and a second time with inferred hashtags. Likewise, we submitted two runs: first without the inferred hashtags; and second with the inferred hashtags. When inferred hashtags were included in queries, the original query was assigned 75% of the weight of the query and inferred hashtags were given 25% of the weight. The inferred hashtags were treated as weighted synonyms using the Indri #wsyn operator. Since the top 30 tweets were to be judged by NIST, we limited the number of results to 30 per query.

5. Experimental Results

Table 2 shows the results of the two runs that we submitted to TREC. As can be seen from the results, the inclusion of the inferred hashtags had a very slight positive effect on retrieval effectiveness, but all differences were not statistically significant. We used a paired two-tailed t-test with $p < 0.05$ to ascertain statistical significance. The reasons why inferring hashtags had little effect need further investigation. We suspect that there is an error in the run with inferred hashtags.

Table 2. Official TREC Results

	Original Queries	Original + Inferred Hashtags
P5	0.388	0.396
P10	0.384	0.380
P15	0.352	0.367
P20	0.347	0.347
P30	0.318	0.318

6. Conclusion

In this paper we described the QCRI submissions to the TREC Microblog track. The essence of the work revolved around expanding tweets and queries using hashtags. We used a generative graphical model for inferring hashtags. Further investigation is required to estimate the accuracy of hashtag prediction. The effect of expansion using hashtags was very limited. We need to debug our experiments.

7. References

Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, Wouter Weerkamp. (2011). Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts. ECIR-2011, LNCS 6611/2011, pp. 362-367, 2011.

Andrew McCallum, Karl Schultz, Sameer Singh. (2009). FACTORIE: Probabilistic Programming via Imperatively Defined Factor Graphs. In Advances on Neural Information Processing Systems (NIPS), 2009.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, Xiaoming Li. (2011). Comparing Twitter and Traditional Media Using Topic Models. ECIR-2011, LNCS 6611/2011, pp. 338-349.