Geographical Latent Variable Models for Microblog Retrieval

Alexander Kotov[1], Vineeth Rakesh[1], Eugene Agichtein[2], and Chandan K. Reddy[1]

[1] Department of Computer Science, Wayne State University, Detroit MI 48226, USA
{kotov,ed3424}@wayne.edu,reddy@cs.wayne.edu
[2] Department of Mathematics and Computer Science, Emory University, Atlanta GA, 30322, USA
eugene@mathcs.emory.edu

Abstract. Although topic models designed for textual collections annotated with geographical meta-data have been previously shown to be effective at capturing vocabulary preferences of people living in different geographical regions, little is known about their utility for information retrieval in general or microblog retrieval in particular. In this work, we propose simple and scalable geographical latent variable generative models and a method to improve the accuracy of retrieval from collections of geo-tagged documents through document expansion that is based on the topics identified by the proposed models. In particular, we experimentally compare the retrieval effectiveness of four geographical latent variable models: two geographical variants of post-hoc LDA, latent variable model without hidden topics and a topic model that can separate background from geographically-specific topics. The experiments conducted on TREC microblog datasets demonstrate significant improvement in search accuracy of the proposed method over both the traditional probabilistic retrieval model and retrieval models utilizing geographical post-hoc variants of LDA.

Keywords: Microblog Retrieval, Latent Variable Models

1 Introduction

Collections of microblog documents pose difficult challenges and offer unique opportunities to retrieval systems at the same time. On one hand, microblog retrieval systems need to overcome severe vocabulary mismatch problem (i.e. how to retrieve very short documents, which might be conceptually relevant, but do not explicitly contain some or all of the query terms), while having to deal only with scarce relevance signals that can be derived from the text of the tweets alone. Furthermore, relevance in the context of microblog retrieval (MBR) is a multi-faceted phenomenon and involves many other factors besides content matching, such as recency, content quality, and geographical focus. On the other hand, social media documents in general and microblogs in particular naturally combine many different types of data besides textual content: timestamps, manually assigned topical tags (hashtags), geographical location of the users who created the tweets and their social networks (followers and followees), which can be leveraged in retrieval models as additional non-textual dimensions and indicators of relevance. As a result, combining lexical with non-lexical relevance signals, such as re-tweets [4] and timestamps [7] [6], has become a dominant theme across most of the recent developments in microblog retrieval.

While most such extrinsic dimensions of relevance (particularly, temporal) have recently received some degree of attention, geographical locations in textual form associated with Twitter user accounts is one important additional dimension and type of meta-data provided by Twitter, which remains relatively overlooked. The importance of accounting for geographical context can be illustrated by using the topic MB04 "Mexico drug war" from the 2011 TREC Microblog track query set as an example. The distribution of geographical locations of the authors of relevant tweets for this topic is shown in Figure 1.

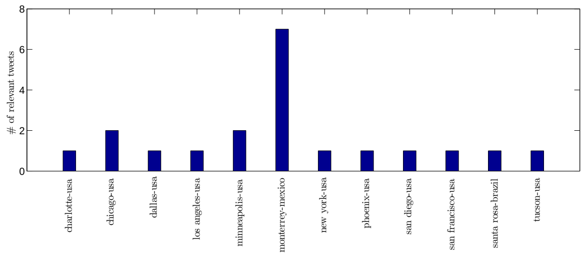


Fig. 1: **Distribution of geographical locations of the authors of relevant tweets for the topic MB04 "Mexico drug war"**

It is clear from Figure 1 that the majority of relevant tweets were authored by the users in a major city in Mexico as well as the cities in the United States, which are close to the Mexican border. Furthermore, from Table 1 it follows that the query terms "mexico" and "drug" individually occur in only about half of 111 relevant tweets for this topic, while the tweets in the other half include different, but conceptually related terms reflecting other aspect of this topic ("border", "catapult", "pot", "fire", "violence", "smuggler"). Only 16 relevant tweets contain include the query terms "mexico" and "drug" together and just 8 (less than 10%) relevant tweets contain all three query terms. Some relevant tweets, such as *"El Paso, Juarez Citizens Unite to Protest Border Violence: 'No Mas Sangre' http://amplify.com/u/..."*, do not include any query terms at all.

This example illustrates the fact that queries are often geographically contextualized (i.e. in regions close to the border between the United States and Mexico, "war" is associated with different concepts than in its traditional definition). While many terms can be related to "war" in general, only a subset of these terms are relevant, given additional geographical ("mexico") and lexical ("drug") contexts. Therefore, the key intuition behind the methods proposed in this work is that accurately addressing vocabulary mismatch problem in MBR through expansion of microblog posts with conceptually related terms requires taking into account their geographical context.

In this work, geographical context is determined by projecting the tweets into lower-dimensional semantic space by leveraging the probabilistic machinery of latent variable generative models. In particular, we propose latent variable models (LVMs), which incorporate geographical locations as observed textual labels. Our work extends the line of information retrieval research, which ad-

Table 1: **Top 10 most frequently occurring terms in the relevant tweets for the query "Mexico drug war".**

term	**# rel. tweets**
mexico	58
drug	58
border	46
catapult	40
mexican	29
u.s.	22
pot	16
fire	15
found	15
smuggler	14

Table 2: **Top 10 geographical locations associated with the most number of tweets in 2011 TREC Microblog track corpus.**

location	**# tweets**
new york-usa	83,479
monterrey-mexico	65,473
sao paulo-brazil	62,243
rio branco-brazil	54,411
london-uk	48,496
los angeles-usa	36,096
caracas-venezuela	33,860
chicago-usa	33,394
jackarta-indonesia	29,795
san francisco-usa	23,642

dresses the problem of vocabulary mismatch through dimensionality reduction, by converting sparse and potentially noisy representation of documents as distributions over terms in the collection vocabulary into more compact representation as distributions over hidden topics, or clusters of semantically related terms. Although state-of-the-art methods to perform dimensionality reduction of document collections, such as Latent Dirichlet Allocation (LDA) [2] topic model, have been previously successfully applied to ad hoc information retrieval [18] [20], little is known about the utility of geographical topic models for information retrieval in general or MBR in particular. In this work, we propose latent variable models (LVMs) that utilize textual geographical locations in the profiles of Twitter users and document expansion methods that leverage the output of the proposed LVMs to address the vocabulary mismatch problem in MBR through *geographically-focused document expansion*.

The rest of this paper is organized as follows. In Section 2, we provide a brief overview of the previous work in closely related areas. Proposed geographical LVMs are discussed in detail in Section 4 and the method to derive document expansion LMs from the output of the proposed LVMs is presented in Section 4.5. Results of an experimental evaluation of the proposed methods are presented in Section 5 and our key contributions are summarized in Section 5.3.

2 Related work

Microblog retrieval. The main challenges in MBR, such as defining units of retrieval and relevance, factoring in quality, authority and timeliness of tweets as well as addressing the vocabulary mismatch problem are discussed in detail in [5], while [17] highlights the key differences between web search and microblog search. Most previously proposed methods for microblog IR focused on incorporating specific types of meta-data (e.g., temporal [7], [15], [3], [10], [1], [14] or social [11]) into retrieval models to address the issue of vocabulary mismatch.

Leveraging timestamps of tweets to model temporal relevance in pseudo-relevance feedback is one of the most well-explored directions in MBR up to date. In particular, Efron et al. proposed a document expansion method [7], in which each tweet is first submitted as a pseudo-query and then the retrieved tweets that are the closest to the timestamp of the original tweet are selected as expansion documents. Efron et al. also proposed a method [6] for re-ranking initial results by estimating the temporal density of relevant documents. A query expansion method proposed in [15] first obtains the initial results for a given query to construct its temporal profile as well as the temporal profiles for each top retrieved document. It then selects those expansion documents to construct the relevance model, for which the temporal profile is the closest to the query temporal profile as measured by Bhattacharyya distance. Amati [1] experimented with exponential, log-normal, log-logistic and Zipf-Mandelbrot distributions to model the freshness aspect of temporal relevance. Miyanishi et al. [14] proposed two methods to select query expansion terms based on analyzing temporal properties of queries and documents. The first method selects the expansion terms one by one from the top retrieved documents by constructing and comparing the temporal profiles for the original and expanded queries. The second method favors recency and selects the expansion terms for which the sample mean of the timestamps in the profile of the expanded query is close to the sample mean of the timestamps in the temporal profile of the original query.

Geographical topic models. A series of recent studies [8] [9] [21] [13] have demonstrated that geography-aware topic models can capture lexical preferences and nuances of language use by people in different geographical locations. Unfortunately, these models are not usable for microblog IR, since they are computationally complex, only work with geographical coordinates and can only handle very small vocabularies. While previous studies [18] [20] have shown the effectiveness of basic topic models in improving retrieval accuracy in traditional ad-hoc IR scenario, with the exception of the preliminary work of Kotov et al. [12], who applied basic post-hoc geographical variant of LDA to MBR and reported promising results, no other work studied the utility of geographical topic models for MBR. In this work, we propose several new geography-aware topic models that use textual geographical locations rather than coordinates and compare their effectiveness for MBR.

3 Retrieval model

Our proposed methods are based on the query likelihood retrieval model, in which a document d is scored and ranked against a query q according to the likelihood of generating q from the language model (LM) of d:

$$P(q|d) = \prod_{w \in q} p(w|\Theta_d) \tag{1}$$

The maximum likelihood estimate $p_{ml}(w|\Theta_d) = \frac{c(w,d}{|d|}$ of document LM is normally smoothed to avoid zero probabilities for query terms that don't occur in d, for example using the Dirichlet prior smoothing:

$$p(w|\Theta_d) = \frac{|d|}{|d| + \mu} p_{ml}(w|\Theta_d) + \frac{\mu}{|d| + \mu} p(w|\mathbb{C}) \tag{2}$$

where $p_{ml}(w|\Theta_d)$ and $p(w|\mathbb{C})$ are the probabilities of w in the maximum-likelihood estimates of document LM and collection LM respectively, and $\mu \geq 0$ is the Dirichlet prior. A combination of query likelihood retrieval method with Dirichlet prior smoothing is used as one of the baselines in our experiments (denoted as **QL-DIR**).

Within the language modeling retrieval framework, the issue of vocabulary mismatch is typically addressed through expansion of either query or document LMs. We adopt the latter approach, in which a document expansion LM $\hat{\Theta}_d$ is first derived for each document by leveraging semantic terms associations mined either from external resources or the collection itself. Then the expanded document LM $p(w|\tilde{\Theta}_d)$ is obtained from the original document LM Θ_d through linear interpolation with a document expansion LM $\hat{\Theta}_d$ with the coefficient α:

$$p(w|\tilde{\Theta}_d) = \alpha p(w|\Theta_d) + (1-\alpha)p(w|\hat{\Theta}_d) \tag{3}$$

The key idea behind document expansion is to add more terms into the document LM that are conceptually relevant to the terms in the original document. In this work, we leverage geography-aware LVMs to identify clusters of semantically related terms within particular geographical regions. In the following sections, we present and discuss the details of the proposed LVMs.

4 Geographical latent variable models

4.1 Post-hoc geographical variants of LDA

We use retrieval methods based on two post-hoc geographical variants of Latent Dirichlet Allocation (LDA) [2], a popular topic model, as baselines. LDA considers each document d in the collection as a mixture of K multinomials (topics) ϕ_z drawn from a symmetric Dirichlet prior β.

Geo-specific topics can be mined from a geo-tagged document collection $\mathbb{C} = \{(d_1, l_{d_1}), \ldots, (d_M, l_{d_M})\}$, in which each document d is associated with textual location l_d from a set of L distinct locations, using standard LDA in a post-hoc way by grouping the documents labeled with each distinct geo-tag $l \in \mathcal{L}$ into sub-collections and running a separate instance of LDA on each sub-collection. The following two variants of this method are used as baselines in our experimental evaluation:

PH-GLDA: this variant uses the same number of geo-specific topics K^{loc} for each location sub-collection. The optimal number of local topics is determined by fitting LDAs with the same number of topics (starting with 2) for each location sub-collection to determine the setting that minimizes perplexity. Therefore, this method finds the optimal *global configuration* of post-hoc LDA and was used to obtain geo-specific topics in [12].

OPT-GLDA: this variant uses different number of geo-specific topics $K^{loc,l}$ for each location sub-collection. The optimal setting is determined by exhaustively trying different numbers of topics for each sub-collection LDA to find the setting that minimizes perplexity on the testing portion of each location sub-collection. This method finds the optimal *local configuration* of post-hoc LDA.

4.2 GLTA

Geographic Latent Term Allocation (**GLTA**) associates a latent variable with each word, which determines its type (whether a word is generated from a background or geo-specific LM) instead of topical assignment. It considers each document d labeled with geo-tag l_d as a mixture of the *background LM* ϕ^{bg}, which is drawn from β^{bg} (all Dirichlet priors in this work are symmetric and have a single hyper-parameter), and *location-specific LM* ϕ^{loc,l_d}, which is drawn from β^{loc}. GLTA models document generation according to the following probabilistic process:

1. draw $\lambda_d \sim Beta(\gamma)$, a binomial distribution controlling the mixture of local and a background LMs for d
2. for each word position i of N_d in d:
   (a) draw Bernoulli switching variable $m_{d,i} \sim \lambda_d$
   (b) if $m_{d,i} = bg$:
      i. draw a word $w_{d,i} \sim \phi^{bg}$
   (c) if $m_{d,i} = loc$:
      i. draw a word $w_{d,i} \sim \phi^{loc,l_d}$

Figure 2a shows the graphical model of GLTA in plate notation. GLTA is a probabilistic extension of the geography-aware Naïve Bayes method proposed in [19].

4.3 GLDA

Geographical LDA (**GLDA**) considers each document d labeled with geo-tag l_d as a mixture of the *background topic* ϕ^{bg} drawn from β^{bg} and K^{loc} *location-specific* topics ϕ^{loc,l_d} drawn from β^{loc} and models document generation according to the following probabilistic process:

1. draw $\lambda_d \sim Beta(\gamma)$, a binomial distribution controlling the mixture of local topics and a background topic for d
2. draw $\Theta_d^{loc,l_d} \sim Dir(\alpha^{loc})$
3. for each word position i of N_d in d:
   (a) draw Bernoulli switching variable $m_{d,i} \sim \lambda_d$
   (b) if $m_{d,i} = bg$:
      i. draw a word $w_{d,i} \sim \phi^{bg}$
   (c) if $m_{d,i} = loc$:
      i. draw a topic $z_{d,i} \sim \Theta_d^{loc,l_d}$
      ii. draw a word $w_{d,i} \sim \phi_{z_{d,i}}^{loc,l_d}$

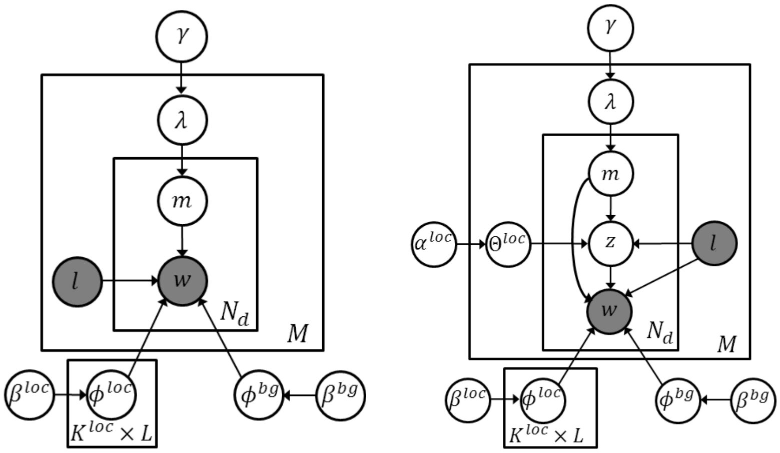
The graphical model of GLDA in plate notation is presented in Figure 2b.

4.4 Posterior inference

Posterior inference for **GLTA** is done using Gibbs sampler, which at each iteration selects the topic type $m_{d,i}$ for a word at every position i in each document in $\mathbb{C}$ based on the following formulas:

$$p(m_{d,i} = bg|\boldsymbol{m}_{\neg i}) \propto \frac{n(d,bg)_{\neg i} + \gamma}{n_d + 2\gamma - 1} \times \frac{n(w_{d,i},bg)_{\neg i} + \beta^{bg}}{\sum_{j=1}^{N} n(w_j,bg) + N\beta^{bg} - 1} \tag{4}$$

$$p(m_{d,i} = loc|\boldsymbol{m}_{\neg i}) \propto \frac{n(d,loc)_{\neg i} + \gamma}{n_d + 2\gamma - 1} \times \frac{n(w_{d,i},loc)_{\neg i} + \beta^{loc}}{\sum_{j=1}^{N} n(w_j,loc) + N\beta^{loc} - 1} \tag{5}$$



a) GLTA **b) GLDA**

Fig. 2: **Graphical models of the proposed LVMs in plate notation**

The Gibbs sampler for **GLDA** at each iteration selects both the topic type $m_{d,i}$ and topical assignment $z_{d,i}$ for each word based on the following formulas:

$$p(m_{d,i} = bg|\boldsymbol{z}_{\neg i}, \boldsymbol{m}_{\neg i}) \propto \frac{n(d, bg)_{\neg i} + \gamma}{n_d + 2\gamma - 1} \times \frac{n(w_{d,i}, bg)_{\neg i} + \beta^{bg}}{\sum_{j=1}^{N} n(w_j, bg) + N\beta^{bg} - 1} \tag{6}$$

$$p(z_{d,i}^{loc,l_d}, m_{d,i} = loc|\boldsymbol{z}_{\neg i}, \boldsymbol{m}_{\neg i}) \propto \frac{n(d, loc)_{\neg i} + \gamma}{n_d + 2\gamma - 1} \times \frac{n(w_{d,i}, z_{d,i}^{loc,l_d})_{\neg i} + \beta^{loc}}{\sum_{j=1}^{N} n(w_j, z_{d,i}^{loc,l_d}) + N\beta^{loc} - 1} \times \frac{n(d, z_{d,i}^{loc,l_d})_{\neg i} + \alpha^{loc}}{\sum_{k=1}^{K^{loc}} n(d, z_k^{loc,l_d}) + K^{loc}\alpha^{loc} - 1} \tag{7}$$

where $n(w, z)_{\neg i}$ is the number of times a term w is assigned to a topic z in the entire collection and $n(d, z)_{\neg i}$ $n(d, bg)$, $n(d, loc)$ are the number of terms in document d that are assigned to topic z, background or geo-specific topics (all counts exclude the current assignments of topic category m and topic z to the word at position i in document d).

4.5 Constructing document expansion LMs

Background and geo-specific topics, per document topic type mixtures and topic distributions obtained by the LVMs presented above can be used to derive a document expansion LM $p(w|\hat{\Theta}_d)$ for each d. In case of **GLDA**, $p(w|\hat{\Theta}_d)$ is obtained using the following formula:

$$p(w|\hat{\Theta}_d) = p(bg|\lambda_d)p(w|\phi^{bg}) + p(loc|\lambda_d)\sum_{k=1}^{K^{loc}} p(w|\phi_k^{loc,l_d}) \times p(z_k^{loc,l_d}|\Theta_d^{loc}) \quad (8)$$

5 Experiments

We used the 2011 TREC Microblog track [16] corpus, which is a 1% sample of Twitter over a period of 2 weeks, as the base dataset for all experiments in this work. The query set for 2011 TREC Microblog track was used to tune the parameters of LVMs and retrieval model, while the 2012 query set was used for the final comparison of retrieval performance. To avoid the sparsity issue (only $1 \sim 2\%$ of microblog posts have geographic coordinates [9]), all microblog posts were labeled with the location of their authors extracted from their Twitter account. Potential noise that may be introduced by a fraction of tweets that are not about the user's primary location can be tolerated by the proposed LVMs, since they identify *major* topical patterns in large volumes of textual data created by many users (there are about 600,000 unique users in TREC dataset). Since the proposed retrieval method requires geographical meta-data, which is not available in the original TREC corpus, we performed additional data collection and pre-processing steps. Firstly, we post-processed the corpus by filtering out all non-English tweets (tweets that do not include any words from the English dictionary of the spell-checking program *aspell*). Secondly, we determined all unique users, who authored the tweets in TREC dataset, extracted their locations from their Twitter profiles and normalized those location to the common "city-country" format using a manually compiled dictionary of suburbs and popular name variants of major cities (e.g. ny, nyc, brooklyn, bronx were all converted to "new york-usa") and Google Geocoding API [3]. Then we selected the top 150 locations (top 10 of which are in Table 2) and used only the documents labeled with those locations to train the proposed LVMs.

Although all retrieval runs are based on using the original TREC corpus, for the purpose of unbiased evaluation of all retrieval models, we only considered the relevant tweets which are covered by the geo-coded subset of the original dataset.

5.1 Optimization of topic models

In order to determine the optimal number of local topics for **PH-GLDA** and **GLDA**, we trained both models on 90% of the documents in each location subcollection and estimated the perplexity on the remaining 10% of documents. During both training and testing the Gibbs sampler was run for 1000 iterations for all models. We found out that **GLDA** achieves significantly lower perplexity than both post-hoc baselines (**OPT-GLDA** and **PH-GLDA**) and **GLTA**, which we attribute to the inclusion of an additional hidden variable, which determines the topic type. Our experiments indicated that the optimal number of geo-specific topics per location for **GLDA** is 30. Examples of the topics discovered by **PH-GLDA**, **OPT-GLDA**, **GLTA** and **GLDA** are provided in Table 3.

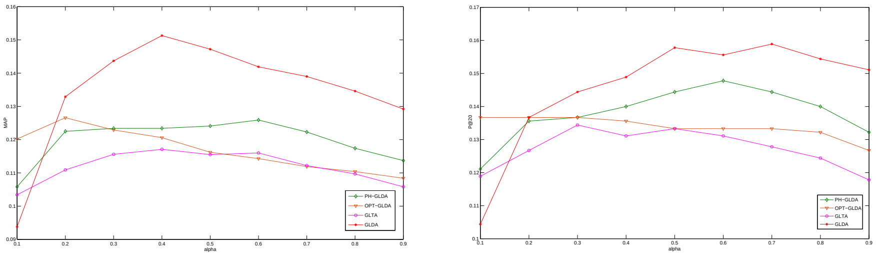
[3] data is available at `http://www.cs.wayne.edu/kotov/code.html#geombr`

Table 3: **Sample geographically-specific topics and LMs extracted by the proposed LVMs.**

PH-GLDA						**OPT-GLDA**					
chicago-usa			cairo-egypt			chicago-usa			cairo-egypt		
topic 1	topic 2	topic 3	topic 1	topic 2	topic 3	topic 1	topic 2	topic 3	topic 1	topic 2	topic 3
snow	tax	new	protest	tahrir	sunni	come	chicago	get	egypt	tahrir	will
take	nice	day	night	old	regime	snomg	mayor	bulls	revolut	protest	people
close	idea	game	fun	light	problem	blizzard	story	beat	police	egypt	mubarak
weather	rahm	race	intern	square	bandar	snow	rahm	point	please	cairo	peace
inch	chi	bears	airport	govt	mobile	stuck	today	rose	thug	freedom	kill
GLTA						**GLDA**					
bkg	chicago-usa		cairo-egypt			bkg	chicago-usa		cairo-egypt		
	geo LM		geo LM				topic 1	topic 2	topic 1	topic 2	topic 3
new	chicago	blizzard	egypt	cairo	regime	rt	snow	court	egypt	muslim	amin
rt	snow	home	thug	people	arab	go	storm	rahm	tahrir	silent	shahira
love	day	good	tahrir	square	arrest	new	blizzard	mayor	square	brotherhood	mustafa
time	bears	snow	mubarak	police	army	love	inch	emanuel	protest	aljazeera	heenim
video	game	work	protest	revolut	afp	time	shovel	ballot	mubarak	moham	sabah

5.2 Optimization of parameters and training performance summary

We used 2011 TREC Microblog track query set to optimize the parameters of different document expansion-based retrieval models proposed in this work with respect to precision at 20 (P@20). First, we optimized the value of Dirichlet prior μ in **QL-DIR** and achieved the best performance when $\mu = 50$. After that we optimized the value of interpolation coefficient α. Sensitivity of retrieval performance of the document expansion methods in terms of mean average precision (MAP) and P@20 on different settings of interpolation coefficient α is shown in Figures 3a and 3b, respectively. Retrieval performance of the proposed methods and the baselines on the training query set is summarized in Table 4.



a) MAP **b) P@20**

Fig. 3: **Performance of document expansion methods based on different LVMs by varying the interpolation coefficient α**

As follows from Table 4, document expansion methods leveraging the output of geography-aware LVMs all improve over the baseline retrieval model (**QL-DIR**), while **GLDA** consistently achieves the best performance across all metrics and outperforms a state-of-the-art baseline (**PH-GLDA**).

Table 4: **Comparison of the best performance of different document expansion-based methods optimized with respect to P@20 on the training query set**

method	**MAP**	**GMAP**	**P@20**	**Bpref**
QL-DIR	0.1015	0.0333	0.1189	0.6223
PH-GLDA	0.1259	0.0464	0.1478	0.6264
OPT-GLDA	0.1266	0.0397	0.1367	0.6170
GLTA	0.1156	0.0356	0.1344	0.6225
GLDA	**0.1390**	**0.0503**	**0.1589**	**0.6445**

5.3 Testing performance summary

Table 5 summarizes and compares with the baselines the retrieval effectiveness of document expansion methods that use the output of the proposed latent variable models on 2012 TREC Microblog query set, which we use as a testing set in this work. Both the proposed methods and the baselines are used with the optimal parameters determined on the training set as described in Section 5.2.

Table 5: **Summary of retrieval performance of document expansion methods based on the output of the proposed latent variable models on testing query set using the optimal parameters determined on the training query set. The magnitude of improvement (↑) or degradation (↓) in percentage relative to QL-DIR baseline is shown in parenthesis. ▲ indicates statistically significant improvement according to the paired *t*-test ($p < 0.05$).**

method	**MAP**	**GMAP**	**P@20**	**Bpref**
QL-DIR	0.0849	0.0469	0.106	0.6135
PH-GLDA	0.1167 (↑37.45%)	0.0662 (↑41.15%)	0.1664 (↑56.98%)	0.6053 (↓1.34%)
OPT-GLDA	0.1123 (↑32.27%)	0.0493 (↑5.11%)	0.1603 (↑51.23%)	0.571 (↓6.93%)
GLTA	0.1123 (↑32.27%)	0.0575 (↑22.6%)	0.1466 (↑38.3%)	0.5976 (↓2.59%)
GLDA	**0.1289 (↑51.83%$^▲$)**	**0.0745 (↑58.85%$^▲$)**	**0.1698 (↑60.19%$^▲$)**	**0.6323 (↑3.06%)**

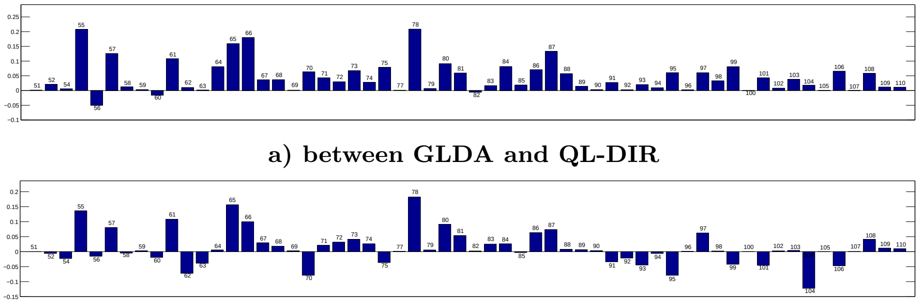
The results in Table 5 indicate that document expansion based on the post-hoc geographical variants of LDA (**PH-GLDA** and **OPT-GLDA**) and our proposed latent variable models (**GLTA** and **GLDA**) all result in significant improvement over **QL-DIR** baseline according to MAP, GMAP and P@20. Remarkably, retrieval accuracy in terms of Bpref measure is improved only when document expansion based on **GLDA** is used and gets worse in case of both post-hoc LDA variants and **GLTA**. Hence, **GLDA**-based document expansion is able to not only retrieve more relevant documents at higher ranks, but also consistently ranks relevant documents above non-relevant ones. Furthermore, **GLDA**-based document expansion results in the highest improvement of retrieval accuracy relative to **QL-DIR** baseline across all metrics.

Table 6 compares the retrieval accuracy of the proposed latent variable models and post-hoc LDA variants relative to the state-of-the-art baseline (**PH-GLDA**). As follows from Table 6, only **GLDA** was able to consistently outperform **PH-GLDA**. The improvement is substantial (over 10%) in terms of MAP and GMAP and statistically significant across most metrics.

Figure 4 shows per-topic differences in average precision between the best preforming document expansion method (based on **GLDA** model) and **QL-DIR**

Table 6: **Improvement (↑) or degradation (↓) of retrieval performance of document expansion LMs derived from the proposed latent variable models relative to PH-GLDA baseline. • indicates statistically significant improvement ($p < 0.05$).**

method	**MAP**	**GMAP**	**P@20**	**Bpref**
OPT-GLDA	↓3.77%	↓25.53%	↓3.67%	↓5.67%
GLTA	↓3.77%	↓13.14%	↓11.9%	↓1.27%
GLDA	↑10.45%•	↑12.54%•	↑2.04%	↑4.46%•



a) between GLDA and QL-DIR

b) between GLDA and PH-GLDA

Fig. 4: **Per-topic difference in average precision between GLDA and the baselines**

and **PH-GLDA** baselines. As follows from both Figure 4a and 4b, improvement in retrieval accuracy varies by the topic. The highest improving queries are shared between both baselines, which indicates that they are both benefiting from accounting for the same phenomena in retrieval, and include: *MB57 "Chicago blizzard"*, *MB61 "Hu Jintao visit to the United States"*, *MB65 "Michelle Obama's obesity campaign"*, *MB66 "Journalists' treatment in Egypt"*, *MB78 "McDonalds food"*, *MB86 "Joanna Yeates murder"*.

In contrast, the topics, for which applying geographically focused document expansion results in decreased retrieval performance (e.g. *MB62 "Starbucks Trenta cup"*, *MB70 "farmers markets opinions"*, *MB70 "texting and driving"*) are broad queries that are not tied to any particular geographical location.

Summary and conclusions

The main contribution of the present work are as follows:

– we proposed new geography-aware LVMs that work with *textual* geographical labels;
– we proposed a method to derive document expansion LMs that leverages the output of the proposed LVMs and compared the retrieval effectiveness of the proposed LVMs on standard TREC datasets for MBR evaluation. Unlike most of the previously proposed methods for microblog IR, our approach does not rely on pseudo-relevance feedback, and hence is more robust and efficient.

Our work has implications beyond microblog retrieval. In particular, the proposed methods can be applied to any geo-tagged document collections other than microblogs. We believe that an interesting direction for future research would be to consider an interplay between different dimensions of relevance in microblog retrieval, such as geographic and temporal.

References

1. G. Amati, G. Amodeo, and C. Gaibisso. Survival analysis for freshness in microblogging search. In *Proceedings of CIKM'12*, pages 2483–2486, 2012.
2. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
3. J. Choi and W. B. Croft. Temporal models for microblogs. In *Proceedings of CIKM'12*, pages 2491–2494, 2012.
4. J. Choi, W. B. Croft, and J. Y. Kim. Quality models for microblog retrieval. In *Proceedings of CIKM'12*, pages 1834–1838, 2012.
5. M. Efron. Information search and retrieval in microblogs. *ASIS&T*, 62(6):996–1008, 2011.
6. M. Efron, J. Lin, J. He, and A. de Vries. Temporal feedback for tweet search with non-parametric density estimation. In *Proceedings of SIGIR'14*, pages 33–42, 2014.
7. M. Efron, P. Organisciak, and K. Fenlon. Improving retrieval of short texts through document expansion. In *Proceedings of SIGIR'12*, pages 911–920, 2012.
8. J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of EMNLP'10*, pages 1277–1287, 2010.
9. L. Hong, A. Ahmed, S. Gurumurthy, A. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of WWW'12*, pages 769–778, 2012.
10. M. Keikha, S. Gerani, and F. Crestani. Time-based relevance models. In *Proceedings of SIGIR'11*, pages 1087–1088, 2011.
11. A. Kotov and E. Agichtein. The importance of being socially-savvy: Quantifying the influence of social networks on microblog retrieval. In *Proceedings of CIKM'13*, pages 1905–1908, 2013.
12. A. Kotov, Y. Wang, and E. Agichtein. Leveraging geographical metadata to improve search over social media. In *Proceedings of WWW'13*, pages 151–152, 2013.
13. Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of WWW'06*, pages 533–542, 2006.
14. T. Miyanishi, K. Seki, and K. Uehara. Combining recency and topic-dependent temporal variation for microblog search. In *Proceedings of ECIR'13*, pages 331–343, 2013.
15. T. Miyanishi, K. Seki, and K. Uehara. Improving pseudo-relevance feedback via tweet selection. In *Proceedings of CIKM'13*, pages 439–448, 2013.
16. I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the trec-2011 microblog track. In *Proceedings of TREC'11*, 2011.
17. J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: A comparison of microblog search and web search. In *Proceedings of ACM WSDM'11*, pages 35–44, 2011.
18. X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of ACM SIGIR'06*, pages 178–185, 2006.
19. B. P. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the ACL'11*, pages 955–964, 2011.
20. X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *Proceedings of ECIR'09*, pages 29–41, 2009.
21. Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topics dsicovery and comparison. In *Proceedings of WWW'11*, pages 247–256, 2011.