Modeling Word Relatedness in Latent Dirichlet Allocation

Xun Wang
Linguistic Intelligence
Research Group.
NTT Communication Science
Laboratories.
xunwang45@gmail.com

ABSTRACT

Standard LDA model suffers the problem that the topic assignment of each word is independent and word correlation hence is neglected. To address this problem, in this paper, we propose a model called Word Related Latent Dirichlet Allocation (WR-LDA) by incorporating word correlation into LDA topic models. This leads to new capabilities that standard LDA model does not have such as estimating infrequently occurring words or multi-language topic modeling. Experimental results demonstrate the effectiveness of our model compared with standard LDA.

1. INTRODUCTION

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has been widely used in many different fields for its ability in capturing latent semantics in textual and image data. In standard LDA, the generating probability is estimated from a document-word matrix using word frequency information. One main disadvantage of LDA is that topic assignment of words is conditionally independent from each other and the relevance between vocabularies (i.e semantic similarity) is totally neglected. This inability makes LDA fall short in many aspects. For example, it is quite plausible that synonymous words such as “use” and “utilize”, “Chinese” and “China” or “politics” and “politicians” should be generated from the same topic. However, standard LDA models treat them as independent units and if they do not co-appear in the same context, they can hardly be included in the same topic. An extreme case is cross-lingual topic modeling where words in different languages never co-occur with each other, though talking about same topics as they are.

Recently, a number of attempts have been made to address the limitation due to conditional-independence. In cross-lingual topic extraction, researchers try to match word pairs and discover the correspondence aligned either at sentence level or document level [18, 16, 11, 12]. Zhao and Xing [18] incorporated word − pair, sentence − pair and document − pair knowledge information into statistical models. One disadvantage is that these approaches usually require aligned text corpus based on machine translation techniques. Boyd-Graber and Blei [4] developed the unaligned topic approach called MUTO where topics are distributions over word pairs instead of just being distributions over terms. However, these methods focus on cross-lingual techniques where word correlation is only partly considered.

To the best of our knowledge, the closest existing works are the approach developed by Andrzejewski et al. [2] and the approach developed by Petterson et al. [14] where word relations are considered. In Andrzejewski et al.’s work [2], domain knowledge is used to model vocabulary relations such as Must-Link or Cannot-Link. By applying a novel Dirichlet Tree prior, Must-Link words are more likely to be generated from the same topic and Cannot-Link words are less likely. However, Andrzejewski et al.’s approach requires specific domain knowledge and cannot be extended to general word correlation. In Petterson et al.'s work [14], they use a sophisticated biased prior in Dirichlet distribution to consider the word correlations rather than the uniform one used in Standard LDA. One disadvantage is that their work considers word correlations only in the prior, but not in the algorithm.

In this paper, we propose an approach called Word Related Latent Dirichlet Allocation (WR-LDA) that incorporates word correlation into topic model. Given a collection of documents, standard LDA topic model wishes finding parameters by maximizing the marginal log likelihood of data. In our model, we sacrifice part of the maximization of log likelihood by incorporating vocabulary correlations based on the assumption that similar words tend to be generated from similar topics. We experiment our model on the different datasets and results indicate the effectiveness of model.

Section 2 presents some related work and Section 3 presents our model. The remaining is the experiments and the conclusion.

2. RELATED WORK

Topic models such as Latent Dirichlet Allocation (LDA) [3] and PLSA (Probabilistic Latent Semantic Analysis) [5] have been widely used in many different fields such as social analysis [6, 9, 10], event detection [1, 13], text classification [10, 15] or facet mining [7, 17], for their ability in discovering topics latent in the document collection. Standard topic models suffer the disadvantage that researchers only use word co-appearance frequency for topic modeling without considering word correlations. However similar or synonymous two terms are, if they do not co-appear in the document, they can hardly be classified into the same topic. This disadvantage is largely magnified in multi-lingual topic modeling where different languages never co-occur with each other.

Previous work on multilingual topic models mostly require parallelism at either the sentence level or document level [2] and the approach developed by Petterson et al. [14]. Boyd-Graber and Blei (2009) proposed a multilingual topic approach which requires a list of word pairs. Topic is defined as distribution over word pairs. However, word correlation is only partly considered in multilingual
topic techniques.
Andrzejewski et al. [2] proposed the approach by considering domain knowledge to model vocabulary relations such as Must-Link or Cannot-Link by using a Dirichlet Tree prior in topic models. However, their model can not be extended to general word correlation. Petterson et al. [14] uses a biased Dirichlet prior by using a Logistic word smoother which takes accounts word relations. The disadvantage of their model is that their work take account of word correlations only in the prior, but not in the algorithm.

3. LDA TOPIC MODEL

The Latent Dirichlet allocation (LDA) model (Blei et al., 2003) defines the document likelihood using a hierarchical Bayesian scheme. Specifically, each document is presented as a mixture of topics which is drawn from a Dirichlet distribution and each topic is presented as a mixture of words. Document words are sampled from a topic-specific word distribution specified by a draw of the word-topic-assignment from the topic proportion vector. Let K be the number of topics; V be the number of the terms in a vocabulary. β is a K × V matrix and β_k is the distribution vector over V terms where \( β_{kj} = P(w = j | z_w = i) \) denotes the probability word w is generated from topic k. The generation for LDA is shown as follows:

1. For each document \( m \in [1, M] \):
   - Draw topic proportion \( θ_m | α \sim Dir(α) \)
2. For each word w
   - Draw topic \( z_w \sim Multinomial(θ_m) \)
   - Draw \( w \sim p(w | z_w, β) \)

LDA can be inferred from a collapsed Gibbs sampling (Heinrich 2005) or variational inference (Blei et al., 2003). Variational inference tries to find parameter α and β that maximize the (marginal) log likelihood of the data:

\[
L(α, β) = \sum_{d=1}^{M} \log(w_d | α, β)
\]

Since the posterior distribution of latent variables can not be computed efficiently, in variational inference of LDA, researchers use variational distribution \( q(θ, z | γ, φ) \) to approximate posterior distribution in each document.

\[
q(θ, z | γ, φ) = q(θ_m | γ_m) \prod_n q(z_n | φ_n)
\]

where \( φ_n = \{ φ_{n1}, φ_{n2}, ..., φ_{nK} \} \). \( φ_{nk} \) can be interpreted as the probability that word at \( n^{th} \) position in current document is generated from topic k. The next step is to set up an optimizing problem to determine the value of γ and φ where the desideratum of finding a tight lower bound on the log likelihood is transformed to the following optimization problem:

\[
(γ, φ) = \arg\min_{γ, φ} KL(q(θ, z | γ, φ) || p(θ, z | w, α, β))
\]

where \( KL(p || q) \) denotes the Kullback-Leibler (KL) divergence between two distributions.

\[
L(γ, φ; α, β) = E_q[log p(w | z, β)] + E_q[log p(z | θ)] + E_q[log p(θ | α)] - E_q[log q(θ)] - E_q[log q(z)] - E_q[log q(β)]
\]

{α, β} can be iteratively inferred from variational EM algorithm (Blei et al., 2003).

4. WR-LDA

4.1 Description

In this section, we present our WR-LDA model and show how it can incorporate word correlation into topic models. Let \( G = (V, E) \) denote the graph where \( V = \{ w_1, w_2, ..., w_n \} \) denotes word collection and \( E = \{ e_{w,w'} : w, w' \in V \} \). \( e_{w,w'} \) denotes the edge between node w and w'. \( \delta_{w,w'} \) denotes the edge weight which is the similarity between word w and w'. Based on the assumption that similar words should have similar probability generated by different topics, we introduce the Loss function \( R(β) \), inspired by the idea of graph harmonic function (Zhu et al., 2003; Mei et al., 2008).

\[
R(β) = \frac{1}{2} \sum_w \sum_{w'} \sum_k η_{w,w'} (β_{k,w} - β_{k,w'})^2
\]

Intuitively, \( R(β) \) measures the difference between \( p(w | β) \) and \( p(w' | β) \) for each pair \( (w, w') \). The more similar two words are, the larger penalty there would be for distribution difference. Clearly, we prefer the topic distribution with smaller \( R(β) \).

In WR-LDA, we try to maximize function \( O(α, β) \) which is the combination of \( L(α, β) \) and \(-R(β)\) instead of just optimizing log likelihood \( L(α, β) \) in standard LDA.

\[
O(α, β) = λL(α, β) - (1 - λ)R(β)
\]

where \( λ \) is the parameter that balances the likelihood and loss function. When \( λ = 1 \), our model degenerates into standard LDA model. Similar idea can also been found in Zhang et al. (2010) for bilingual topic extraction and Mei et al.(2008) for topic network construction.

4.2 Inference

In this subsection, we describe the variational inference for WR-LDA. We emphasize on the part that is different from LDA and briefly describe the part that is similar. We use the variational distribution \( q(θ, z | γ, φ) \) to approximate posterior distribution \( p(θ, z | w, α, β) \) and use variational EM algorithm for inference.

**E-step:** The E-step of WR-LDA tries to find the optimizing values of the variational parameters \( (γ, φ) \). Since the loss function \( R(β) \) does not involve γ and φ, E-step of WR-LDA is the same as that of standard LDA. We update \( γ, φ \) according to the following equations.

\[
φ_{nk} = β_{k,w} \cdot exp(E_q[log(θ_w | γ)])
\]

\[
γ_k = α_k + \sum_n φ_{nk}
\]

where \( E_q[log(θ_w | γ)] = Ψ(γ_k) - Ψ(∑_k γ_k) \) and \( Ψ(·) \) is the first derivative of \( log Γ \) function. The E-step of WR-LDA is shown in Figure 2 (Blei et al., 2003).

---

1The details of \( L(α, β) \) is shown in Appendix A.
shown in Equ.6 with respect to the model parameters $\sum$ according to the constraints that $LDA$.

If the size of vocabulary $V$ is $w$ instead of getting the local maximum of $\{\}$, so given $\beta$ is a $1 \times V$ vector. The update for $\beta_k$ in Newton-Raphson is as follows:

$\beta_{k+1} = \beta_k - H(\beta_k)^{-1} \nabla T(\beta_k)$

where $\nabla L(\beta_k)$ is the gradient of function $T(\alpha, \beta)$ with regard to $\beta_{k}$, $i \in [1, V]$. $H(\beta_k)$ is a Hessian matrix.

$$H(\beta_k)(i, j) = \frac{\partial^2 T(\alpha, \beta)}{\partial \beta_{ki} \partial \beta_{kj}} = -2(1 - \lambda)\kappa_{w_i, w_j}$$

If the size of vocabulary $V$ is $10^4$, $H(\beta_k)$ would be a $10^4 \times 10^4$ matrix, the inverse of which would be too expensive to be calculated directly. To solve this problem, we adopt the strategy proposed in Mei et al.(2008)’s work, where we only need to find the value of $\beta^n$ that makes $T(\alpha^n, \beta^n|\lambda^n, \phi^n, \gamma^n) > T(\alpha^{n-1}, \beta^{n-1}|\lambda^n, \phi^n, \gamma^n)$ instead of getting the local maximum of $T(\alpha^n, \beta^n|\lambda^n, \phi^n, \gamma^n)$ at each M-step. In each iteration of M-step, we firstly set the value of $\beta^n$ to the value which maximizes $L(\alpha, \beta)$, just as in Standard LDA.

$$\beta_{kw}^{n+1} = \lambda \beta_{kw}^{n} + (1 - \lambda) \frac{\sum_{i \in N_{kw}} \kappa_{w_i, w_k}}{\sum_{i \in N_{kw}} \kappa_{w_i, w_k}}$$

Clearly, $\sum_{k} \beta_{kw} = 1$ and $\beta_{kw} \geq 0$ always hold in Equ.14. Equ.14 can be interpreted as follows: when $\rho$ is set to 0, it means that the updating value of $\beta_{kw}$ is totally decided by its neighbors. In graph harmonic algorithm (zhu2003semi), Equ.14 is just the optimization of $Loss function R(\beta)$ when $\rho$ is set to 0. So we try to optimize $O(\alpha, \beta)$ by firstly decreasing the value of $Loss function R(\beta)$ until $O(\alpha, \beta)$ decreases.

update $\alpha$: The first derivative of $T(\alpha, \beta)$ in regard with $\alpha_k$ also depends on $\alpha_k$, $k \neq k'$, we also use Newton-Raphson method for optimization. Since the Hessian matrix $H(\alpha)$ is the form of:

$$H(\alpha) = diag(h) + 11^T$$

where $h_k = -M \Psi'(\alpha_k)$ and $z = M \Psi'(\alpha_k)$ the inverse of $H^{-1}(\alpha) = diag(h)^{-1} - \frac{\sum_{k} h_k}{z - 1}$

$\alpha$ can be updated as follows:

$$\alpha^{t+1} = \alpha^t - H(\alpha)^{-1} \nabla T(\alpha)$$

Multiplying by the gradient, we easily obtain the $i$th component of matrix $H(\alpha)^{-1} \nabla T(\alpha)$ (Blei et al., 2003).

$$[H(\alpha)^{-1} \nabla T(\alpha)]_i = \frac{[\nabla T(\alpha)]_i - c}{h_i}$$

where $c = \sum_{k} [\nabla T(\alpha)]_k / h_k$

5. EXPERIMENTS

In this section, we compare WR-LDA with standard LDA and other baselines in multiple applications.

5.1 Text Modeling

We firstly compare text modeling of the WR-LDA with standard LDA on the 20 Newsgroups data set with a standard list of 598 stop words removed.

We fit the dataset to a 110 topics. In WR-LDA, we need to tune $\lambda$. To evaluate the performances of WR-LDA with different $\lambda$, we have the following function.

$$M(\hat{\gamma}) = \frac{1}{C_t} \sum_{i=1}^{20} \sum_{d \in G_i} \sum_{j \in G_i} \zeta_2(KL(\hat{\gamma}_d^{\hat{\gamma}}||\gamma_d^{\hat{\gamma}}))$$

$$+ \frac{1}{C_t} \sum_{i=1}^{20} \sum_{j \neq d} \sum_{d' \in G_j} \sum_{j \in G_j} \zeta_1(KL(\hat{\gamma}_d^{\hat{\gamma}}||\gamma_d^{\hat{\gamma}}))$$

http://qwone.com/jason/20Newsgroups/
and contributions. In our experiments, we set \(\lambda\) in an interval of \(10^5\). The results are shown in Fig 4. In the following experiments, we set \(\lambda = 3.6 \times 10^5\).

Figure 7 shows the 2D embedding of the expected topic proportions of WR-LDA \((\lambda = 3.6 \times 10^5)\) and LDA using the t-SNE stochastic neighborhood embedding (van der Maaten and Hinton, 2008).

5.2 Regression

We evaluate supervised version of WR-LDA on the hotel reviews dataset from TripAdvisor\(^1\). As in Zhu and Xing (2010), we take logs of the response values to make them approximately normal. Each review is associated with an overall rating score and five aspect rating scores for the aspects: Value, Rooms, Location, Cleanliness, and Service. In this paper, we focus on predicting the overall rating scores for reviews.

We compare the results from the following approaches: LDA+SVR, sLDA, WR-LDA+SVR and sWR-LDA. For LDA+SVR and WC-LDA+SVR, we use their low dimensional representation of documents as input features to a linear SVR and denote this method. The evaluation criterion is predictive \(R^2\) (pR2) as defined in the work of Blei and McAuliffe (2007).

Figure 7 shows the predictive \(R^2\) scores of different models. First, since supervised topic models can leverage the side information (e.g., rating scores) to discover latent topic representations, they generally outperform the decoupled two-step procedure as adopted in unsupervised topic models. Secondly, we can see that sWR-LDA outperforms sLDA and WR-LDA+SVR outperforms LDA+SVR.

From the top words of different topics shown at Fig8, we can easily explain why sWR-LDA is better than sLDA. For sWR-LDA, part of topics show a regular positiveness/negativeness pattern. Topic 3 is on the negative aspects of a hotel while Topic 7 and Topic 9 describe positive aspects. This is because word correlation is considered in WR-LDA. For example, positive vocabularies such as great, good and wonderful, they have similar semantic meanings and thus high edge weight. So they are very likely to be generated from the same topic. The topics discovered by LDA do not show a regular pattern on positiveness or negativeness, resulting in low predictive \(R^2\). Moreover, we can see that topics are more coherent in WR-LDA. Topic 2 talks about sea sceneries, Topic 5 talks about food, Topic 6 talks about conditions of bathroom or swimming pool and Topic 10 talks about room objects. However, topics in LDA do not show such clear pattern.

6. CROSS-LINGUAL TOPIC EXTRACTION

6.1 Dataset Construction

To qualitatively compare our approach with the baseline method, we test different models on Cross-Lingual Topic Extraction task in different respects. The data set we used in this experiment is selected from a Chinese news website Sina\(^2\) and Google News\(^3\). Selected news talks about 5 topics categorized by news websites such as "Sports(体育)", "Entertainments(娱乐)", "World(国际)", "Science(科学)", and "Business(财经)". Each Chinese news passage is associated with an English news passage which talks about the same event (i.e. Passage entitled "土土耳其民众对政府改造广场抗议蔓延至全国" vs Passage entitled "Protests no Turkish Spring", says PM Erdogan"). We select 1000 English passages and 1000 Chinese passages during the period from March. 8th, 2001 to June. 23th, 2013, including [Sports:360, WordNews:296, Business:92, Science:68, and Entertainments:184].

6.2 Details

Let \(C_c\) denote Chinese news collection, where \(C_c = \{d_1^c, d_2^c, ..., d_N^c\}\), \(N = 1000\). \(C_e\) is the English news collection \(C_e = \{d_1^e, d_2^e, ..., d_N^e\}\), and \(d_i^e\) is the correspondent passage of \(d_i^c\). We firstly removed infrequent words \([\text{infrequent words}]\) and frequent words \([\text{frequent words}]\). A bilingual Chinese-English dictionary would give us a many-to-many mapping between the vocabularies of the two languages. If one word can be potentially translated into another word, the two words would be connected with an edge. Specifically, let \(w_c\) denote an English word and \(w_e\) denotes a Chinese word. \(C_{w_c}\) denotes the set of Chinese words that are translated from \(w_e\), \(C_{w_e} = \{w_c | e(w_c, w_e) \neq 0\}\)

\[
\kappa(w_c, w_e) = \begin{cases} 
1 & \text{if } w_c \in C_{w_e} \\
0 & \text{if } w_c \not\in C_{w_e}
\end{cases}
\]  

(18)

6.3 Evaluation

Part of the results are presented in Table 1. Since commonly used measures for LDA topic model such as perplexity can not capture whether topics are coherent or not, we use the measures in Pettersen et al. (2011)’s work.

We compare the topic distributions of each Chinese news passage with its correspondent English pair based on the following measures:

- Mean \(l_2\) Distance (L2-D):
  \[
  \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{N} \sum_{k=1}^{N} \gamma_k^c d_i^c - \gamma_k^e d_i^e\right)^2 \right)^{1/2}
  \]
- Mean Hellinger Distance (H-D):
  \[
  \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} \left(\frac{1}{N} \sum_{k=1}^{N} \gamma_k^c d_i^c - \gamma_k^e d_i^e\right)^2 \right)^{1/2}
  \]

\(^1\)http://www.tripadvisor.com
\(^2\)http://www.sina.com.cn
\(^3\)https://news.google.com

\(^4\)Words that occurred less than 3 times in the corpus.

\(^5\)Words that occurred more than M/10 times in the corpus, where M is the total number of documents.
Figure 4: Tuning $\lambda$ (20News-Group data)

Figure 5: t-SNE 2D embedding (a)LDA (b)WR-LDA

| class                  | Top Popular Words                                                                 |
|------------------------|-----------------------------------------------------------------------------------|
| comp.graphics          | T17 jpeg, gif, pictures, pigment frames, color, bit, pixels, symbols, points    |
|                        | T31 images, image, graphics, software, file, db, format, files, programs, systems |
|                        | T62 file, information, vga, tar, files, photograph, chip, images, windows, files  |
| talk.politics.mideast  | T53 Israel, Israeli, people, Arab, Jews, Arabs, American, America, Jewish, Muslim |
|                        | T77 attacks, soldiers, villages, peace, war, rights, government, writes, Israel, citizen |
|                        | T101 Israel, Lebanese, Lebanese, people, Turkish, turkey, Jewish, Armenians, Turks, Arab |
| comp.electrics         | T55 circuit, current, voltage, power, signal, wires, electrical, resistance, ground, wire |
|                        | T82 input, output, signal, audio, high, low, time, chip, circuit, loop            |
| comp.sys.mac.hardware   | T22 drive, apple, mac, hardware, memory, Macintosh, system, laptop, etc, macs    |
|                        | T22 monitor, scsi, hardware, CPU, drive, processor, quadras, mb, memory, chip     |
| talk.politics.guns     | T43 gun, guns, fire, people, weapons, firearms, militia, pistol, weapon, government|
| misc.forsale            | T78 sail, price, sell, shipping, shops, offer, buy, email, send, credit           |
| sci.space              | T25 earth, space, moon, astronomy, nasa, shuttle, satellite, lunar, orbit, stars  |
| rec.sport.baseball     | T22 games, baseball, scoring, players, game, runs, team, hit, player, season     |
| rec.autos              | T93 cars, price, drive, car, engine, oil, steering, pay, driving, speed          |
| soc.religion.christian | T86 Christians, christ, bible, church, god, Catholic, christian, Jews, Jesus, soul |
| sci.med                | T55 disease, doctor, medical, pain, health, care, patients, cancer, treatment, remedy |

Figure 6: Top words in selected classes from 20NewsGroup by WR-LDA

| WC-LDA                                                                 |
|---------------------------------------------------------------|
| Topic1 | Topic2 | Topic3 | Topic4 | Topic5 | Topic6 | Topic7 | Topic8 | Topic9 | Topic10 |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|
| room   | beach  | bad    | place  | restaurant | bathroom | great  | room   | like   | door    |
| hotel  | island | worst  | room   | room restaurants | clean | take   | loved  | TV      |
| floor  | ocean  | poor   | hotel  | food restaurants | food   | stay   | meet   | room    |
| block  | trip   | dislike| rooms  | dinner | take   | children | love   | bed     |
| location | area | time  | staff | time | wonderful | kids | left   | enjoy   |
| stay   | weather | dislike | staff | time | leave | kids | door | curtain |
| house  | sea    | hate   | staff | pool | best | perfect | enjoyed | bedroom |
| rooms  | day    | dirty  | staff | swimming | swimming | terrible | recommend | beds |
| day    | warm   | morning | service | swimming | excellent | decided | walk | beds |
| stay   | sea    | night  | staff | water | children | check | trip | have |
| room   | back   | night  | staff | cookies | wash   | check | table | have |

| LDA                                                                 |
|---------------------------------------------------------------|
| Topic1 | Topic2 | Topic3 | Topic4 | Topic5 | Topic6 | Topic7 | Topic8 | Topic9 | Topic10 |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|
| hotel  | beach  | rooms  | day    | hotel  | great  | food   | restaurant | hotel | resort |
| location | hotel  | rooms  | asked  | stay   | children | shower  | hotel | pool | resort |
| room   | holiday | rooms  | door   | stay   | time   | spend  | bathroom | asked | beach |
| good   | nice   | rooms  | door   | stay   | time   | spend  | breakfast | work  | walk |
| service | hotel  | rooms  | desk   | desk   | time   | spend  | like | beach | reception |
| staff  | pool   | rooms  | front  | front  | front | trip | enjoy | area | stay |
| rooms  | nice   | rooms  | back   | trip   | front | trip | enjoy | area | view |
| time   | nice   | rooms  | stay   | place | dinner | experience | service | view | wonderful |
| time   | small  | rooms  | nice   | place | dinner | experience | water | service | wondere |
| place  | hotel  | floor  | sea    | check | kids | terrific | trip | view | trip |
| place  | booked | hotel  | right | sea    | check | terrific | trip | view | trip |

Figure 8: WR-LDA & LDA on TripAdvisor dataset
| Topic 1                  | Topic 5                  | Topic 6                  | Topic 11                  |
|-------------------------|--------------------------|--------------------------|---------------------------|
| James Korea             | [Image] Korea north      | Xi Jinping (Xi Jinping)   | Barcelona (Barcelona)     |
| James (James)           | North Korea              | China Champions          |                           |
| Kobe (Kobe)             | North Korea              | Ferguson                 |                           |
| Heat (Jong-un)          | Kim (Kim Jong-un)        | Xi                        |                           |
| rebound (Kim Jong-un)   | [Image]李克强 (Li Keqiang)| Semi final               |                           |
| Heat (Heat)             | China (China)            | China                    |                           |
| Lakers                  | United                   | Barcelona                |                           |
| basketball              | [Image]篮球 (basketball) | [Image]中国 (China)      |                           |
| 湖人 (Lakers)           | South (South Korea)      | [Image]电子 (Chinese)    |                           |
| Durant                  | Obama (Obama)            | [Image]主席 (chairman)   |                           |
| Durant (Durant)         | [Image]导弹 (missile)    | 马克 (representative)    |                           |
| score                   | United                   | [Image]德国 (German)     |                           |
| 耐 (Thailand)            | 军事 (military)           | [Image]国家 (country)    |                           |
| 雷霆 (Thunder)          | [Image]外交部 (Foreign Ministry) | [Image]中国 (China) |                           |
| season                  | military                 | [Image]人民 (People)    |                           |
| 洛杉矶 (Los Angeles)    |                         | [Image]多特蒙德 (Dortmund) |                           |

Table 1: Top Word in selected topic extracted from WR-LDA in Cross-Lingual task

- Agreements on first topic: (A-1)
  \[ \frac{1}{N} \sum_{i=1}^{N} I(\arg \max_k \gamma_{d_i}^{c_i} = \arg \max_k \gamma_{d_i}^{e_i}) \]

- Mean number of agreements in top 5 topics (A-5):
  \[ \frac{1}{N} \sum_{i=1}^{N} \text{agreement}(d_i^c, d_i^e), \text{where agreements} \text{agreement}(d_i^c, d_i^e) \]
  is the cardinality of the intersection of the 5 most likely topics of \( d_i^c, d_i^e \).

Clearly, we prefer smaller values of L2-D and H-D and larger values of A-1 and A-5. We compare the performances from following approaches:

- WR-LDA1: Only Chinese-English correlations are considered and the weights of edges between English-English words, Chinese-Chinese words are 0.
- WR-LDA2: All word correlations are considered.
- DC model: Approach proposed by Petterson et al.(2011), where word correlations are incorporated into a Beta prior using a logistic smooth function.
- LDA: the standard LDA topic model.

From Figure 9, LDA achieves the worst results as expected because it can hardly detect the topics shared by bilingual documents due to the reason that vocabularies from different languages hardly co-appear in the same passages. WR-LDA1 is better than the DC model, which considers word correlations only in the prior construction, rather than in the algorithm. WR-LDA2, which consider word correlations fully, achieves better results than WR-LDA1, which partly considers word correlations.

For further illustration, we randomly choose two pairs of words, "Messi" and "梅西" (Messi in Chinese), which is the name of a famous Argentine soccer player, and "China" and "中国" (China in Chinese). Since each pair of words have the same meaning, we prefer that they have the similar probabilities generated from the same topic. Figure 8 shows the probability that "Messi" and "梅西" generated from different topics and Figure 9 shows "China" and "中国". The x-axis is the index to the topic and y-axis is the log-likelihood of probability. We can see that in LDA, words within
each pair are always negative correlated. This can be easily explained by the fact that Chinese words and English words never co-appear in a document. So if a Chinese word has large probability generated by one topic, it means this top is a Chinese dominated topic and most English words would be excluded. Word pairs in DC model are correlated but not as strong as that in WR-LDA. Since DC model models word correlation only in the prior, such influence can be diluted while Gibbs sampling goes on. In WR-LDA, word correlations are considered in the algorithm and during each iteration, algorithm will fix the gap between correlated words. That is why WR-LDA outperforms DC model.

7. CONCLUSION

In this paper, we present WR-LDA, a revised version of LDA topic model that incorporates word correlations. Experiments on text modeling, review rating prediction and cross-lingual topic modeling demonstrate the effectiveness of our model. There are two disadvantages of WR-LDA when compared with LDA. (1) The value of parameter lambda involved in the model is hard to tune while LDA there is no additional parameter involved. (2) Due to the introduction of Penalty function in WR-LDA, we have to keep record of parameter $\phi$ in varational inference for all documents, which largely increase the cost of both memory and time.

8. REFERENCES

[1] C. C. Aggarwal and C. Zhai. Mining text data. Springer, 2012.
[2] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 25–32. ACM, 2009.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.
[4] J. Boyd-Graber and D. M. Blei. Multilingual topic models for unaligned text. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pages 75–82. AUAI Press, 2009.
[5] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57. ACM, 1999.
[6] J. Li and C. Cardie. Timeline generation: tracking individuals on twitter. In Proceedings of the 23rd international conference on World wide web, pages 643–652. International World Wide Web Conferences Steering Committee, 2014.
[7] J. Li, M. Ott, and C. Cardie. Identifying manipulated offerings on review portals. In EMNLP, pages 1933–1942, 2013.
[8] J. Li, A. Ritter, C. Cardie, and E. Hovy. Major life event extraction from twitter based on congratulations/condolences speech acts. 2014.
[9] J. Li, A. Ritter, and E. Hovy. Weakly supervised user profile extraction from twitter. ACL, 2014.
[10] J. D. McCalline and D. M. Blei. Supervised topic models. In Advances in neural information processing systems, pages 121–128, 2008.
[11] D. Memno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polyligual topic models. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pages 880–889. Association for Computational Linguistics, 2009.
[12] X. Ni, J.-T. Sun, J. Hu, and Z. Chen. Mining multilingual topics from wikipedia. In Proceedings of the 18th international conference on World wide web, pages 1155–1156. ACM, 2009.
[13] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. International journal of computer vision, 79(3):299–318, 2008.
[14] J. Petterson, W. Buntine, S. M. Narayananurthy, T. S. Caetano, and A. J. Smola. Word features for latent dirichlet
allocation. In Advances in Neural Information Processing Systems, pages 1921–1929, 2010.

[15] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, pages 248–256. Association for Computational Linguistics, 2009.

[16] Y.-C. Tam and T. Schultz. Bilingual lda-based translation lexicon adaptation for spoken language translation. In INTERSPEECH, pages 2461–2464, 2007.

[17] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In Proceedings of the 17th international conference on World Wide Web, pages 111–120. ACM, 2008.

[18] B. Zhao and E. P. Xing. Bitam: Bilingual topic admixture models for word alignment. In Proceedings of the COLING/ACL on Main conference poster sessions, pages 969–976. Association for Computational Linguistics, 2006.