Supplement for Paper "A Multilingual Topic Model for Learning Weighted Topic Links Across Corpora with Low Comparability"

Weiwei Yang[*]
Computer Science

University of Maryland
wwyang@cs.umd.edu

Jordan Boyd-Graber[†]
Computer Science, iSchool, Language Science, UMIACS
University of Maryland
jbg@umiacs.umd.edu

Philip Resnik
Linguistics and UMIACS

University of Maryland
resnik@umd.edu

A Generative Details of our Multilingual Topic Model

Our MTM is a downstream model with a posterior regularizer. It has two LDA components that generate the documents in languages S and T. Then it generates the posterior regularizer Ψ which encodes the cross-lingual knowledge. The detailed generative process is as follows.

1. For each topic $k \in \{1, \dots, K_T\}$ in language T
   (a) Draw word distribution $\boldsymbol{\phi_{T,k}} \sim \text{Dirichlet}(\beta_T)$.
2. For each document $d \in \{1, \dots, D_T\}$ in language T
   (a) Draw topic distribution $\boldsymbol{\theta_{T,d}} \sim \text{Dirichlet}(\alpha_T)$.
   (b) For each token $t_{T,d,n}$ in document d
      i. Draw a topic $z_{T,d,n} \sim \text{Multinomial}(\boldsymbol{\theta_{T,d}})$.
      ii. Draw a word $w_{T,d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{T,z_{T,d,n}}})$.
3. For each topic $k \in \{1, \dots, K_S\}$ in language S
   (a) Draw word distribution $\boldsymbol{\phi_{S,k}} \sim \text{Dirichlet}(\beta_S)$.
4. For each document $d \in \{1, \dots, D_S\}$ in language S
   (a) Draw topic distribution $\boldsymbol{\theta_{S,d}} \sim \text{Dirichlet}(\alpha_S)$.
   (b) For each token $t_{S,d,n}$ in document d
      i. Draw a topic $z_{S,d,n} \sim \text{Multinomial}(\boldsymbol{\theta_{S,d}})$.
      ii. Draw a word $w_{S,d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{S,z_{S,d,n}}})$.
5. Draw the posterior regularizer $\Psi = \left(\prod_{c=1}^{C} \|\boldsymbol{\Omega_{S,c}} - \boldsymbol{\rho_{T\to S}}\boldsymbol{\Omega_{T,c}}\|_2^{\eta_c} \|\boldsymbol{\rho_{S\to T}}\boldsymbol{\Omega_{S,c}} - \boldsymbol{\Omega_{T,c}}\|_2^{\eta_c}\right)^{-1}$.

A.1 Posterior Inference

As we mentioned in the paper, the posterior inference is based on stochastic EM and consists of an E-step and an M-step [1]. In each iteration, the E-step updates every token's topic assignment using Gibbs sampling, while holding the values in the topic link weight matrices $\boldsymbol{\rho}$. The M-step optimizes $\boldsymbol{\rho}$ while holding the topic assignments.

[*]Now at Facebook
[†]Now at Google AI Zürich

A.1.1 E-Step: Topic Assignment Sampling

In addition to the usual word and topic dependencies, our MTM encourages topic assignments that maximize the posterior regularizer Ψ, thus make the related translation pairs' (transformed) topic distributions close. This is reflected in the Gibbs sampling equation to update $z_{T,d,n}$, the topic assignment of the n-th token of document d in language T:

$$\Pr(z_{T,d,n} = k \,|\, \boldsymbol{z_{-T,d,n}}, w_{T,d,n} = v, \boldsymbol{w_{-T,d,n}}, \boldsymbol{\rho}, \alpha_T, \beta_T) \propto \underbrace{\left(N_{T,d,k}^{-T,d,n} + \alpha_T\right) \frac{N_{T,k,v}^{-T,d,n} + \beta_T}{N_{T,k,\cdot}^{-T,d,n} + V_T \beta_T}}_{\text{LDA Sampling}}$$
$$\underbrace{\left(\prod_{v' \in \text{Dic}(v)} \left\| \boldsymbol{\Omega_{S,v'}} - \boldsymbol{\rho_{T \to S}} \boldsymbol{\Omega_{T,v}} \right\|_2^{\eta_{v',v}} \left\| \boldsymbol{\rho_{S \to T}} \boldsymbol{\Omega_{S,v'}} - \boldsymbol{\Omega_{T,v}} \right\|_2^{\eta_{v',v}} \right)^{-1}}_{\text{Minimizing the Topic Distribution Distances}}, \tag{1}$$

where the first two terms are the same as LDA: $N_{T,d,k}$ denotes the number of tokens in document d assigned to topic k; $N_{T,k,v}$ denotes the number of times word v is assigned to topic k; $\cdot$ denotes marginal counts; $^{-T,d,n}$ means the count excludes the token. The final term corresponds to the posterior regularizer: $\text{Dic}(v)$ is word v's translation word set in language S; The values of $\boldsymbol{\Omega_{T,v}}$, the topic distribution of word v, *assume* topic k is chosen as follows:

$$\Omega_{T,v,k'} = \frac{N_{T,k',v}^{-T,d,n} + \mathbb{1}(k' = k)}{N_{T,v}}, \tag{2}$$

where $\mathbb{1}(\cdot)$ is an indicator function.

Symmetrically, the Gibbs sampling equation to update $z_{S,d,n}$, the topic assignment of the n-th token of document d in language S, is

$$\Pr(z_{S,d,n} = k \,|\, \boldsymbol{z_{-S,d,n}}, w_{S,d,n} = v, \boldsymbol{w_{-S,d,n}}, \boldsymbol{\rho}, \alpha_S, \beta_S) \propto \left(N_{S,d,k}^{-S,d,n} + \alpha_S\right) \frac{N_{S,k,v}^{-S,d,n} + \beta_S}{N_{S,k,\cdot}^{-S,d,n} + V_S \beta_S}$$
$$\left(\prod_{v' \in \text{Dic}(v)} \left\| \boldsymbol{\Omega_{S,v}} - \boldsymbol{\rho_{T \to S}} \boldsymbol{\Omega_{T,v'}} \right\|_2^{\eta_{v,v'}} \left\| \boldsymbol{\rho_{S \to T}} \boldsymbol{\Omega_{S,v}} - \boldsymbol{\Omega_{T,v'}} \right\|_2^{\eta_{v,v'}} \right)^{-1}. \tag{3}$$

The values of $\boldsymbol{\Omega_{S,v}}$, assuming topic k is chosen, are

$$\Omega_{S,v,k'} = \frac{N_{S,k',v}^{-S,d,n} + \mathbb{1}(k' = k)}{N_{S,v}}. \tag{4}$$

A.1.2 M-Step: Parameter Optimization

Here we optimize the topic link weight matrices $\boldsymbol{\rho}$ while fixing the topic assignments. As Ψ is the product over all translation pairs, we modify Ψ to obtain the objective functions $J(\boldsymbol{\rho_{T \to S}})$ and $J(\boldsymbol{\rho_{S \to T}})$ as the weighted logarithmic sum[1]

$$J(\boldsymbol{\rho_{T \to S}}) = \sum_{c=1}^{C} \eta_c \log \left\| \boldsymbol{\Omega_{S,c}} - \boldsymbol{\rho_{T \to S, i_S}} \boldsymbol{\Omega_{T,c}} \right\|_2 \tag{5}$$

$$J(\boldsymbol{\rho_{S \to T}}) = \sum_{c=1}^{C} \eta_c \log \left\| \boldsymbol{\Omega_{T,c}} - \boldsymbol{\rho_{S \to T, i_T}} \boldsymbol{\Omega_{S,c}} \right\|_2. \tag{6}$$

[1]It makes sense to add regularization on $\boldsymbol{\rho}$'s to prevent overfitting, but the data already adds a strong constraint on $\boldsymbol{\rho}$'s—each word's $\boldsymbol{\Omega}$ values should add up to one.

The objective functions are then minimized by using L-BFGS [2] and the partial derivatives with respect to $\rho_{T\to S,k_S,k_T}$ and $\rho_{S\to T,k_T,k_S}$ as

$$\frac{\partial J(\boldsymbol{\rho_{T\to S}})}{\rho_{T\to S,k_S,k_T}} = -\sum_{c=1}^{C} \frac{\eta_c \Omega_{T,c,k_T} \left(\Omega_{S,c,k_S} - \boldsymbol{\rho_{T\to S,k_S}\Omega_{T,c}}\right)}{\|\boldsymbol{\Omega_{S,c}} - \boldsymbol{\rho_{T\to S,i_S}\Omega_{T,c}}\|_2^2} \tag{7}$$

$$\frac{\partial J(\boldsymbol{\rho_{S\to T}})}{\rho_{S\to T,k_T,k_S}} = -\sum_{c=1}^{C} \frac{\eta_c \Omega_{S,c,k_S} \left(\Omega_{T,c,k_T} - \boldsymbol{\rho_{S\to T,k_T}\Omega_{S,c}}\right)}{\|\boldsymbol{\Omega_{T,c}} - \boldsymbol{\rho_{S\to T,i_T}\Omega_{S,c}}\|_2^2}. \tag{8}$$

B Corpora Statistics

There are four datasets used in our experiments. The first one is collected from Wikipedia for document category classification in English (EN) and Chinese (ZH) [3]. There are six document categories for Wikipedia documents: *film*, *music*, *animals*, *politics*, *religion*, and *food*. The dictionary comes from MDBG, a website for learning Chinese.[2]

The second dataset is about disaster response in English and Sinhalese (SI) [4, 5]. A subset of the documents are annotated with one of eight need types: *evacuation*, *food supply*, *search/rescue*, *utilities*, *infrastructure*, *medical assistance*, *shelter*, and *water supply*. The dictionary is provided by the dataset authors.

The last two are also collected from Wikipedia, one is partially comparable (PACO) and the other one is incomparable (INCO) [6]. Either one contains five bilingual corpora. Among these bilingual corpora, one of the language is always English (EN), while the other language is one of Arabic (AR), Chinese (ZH), Spanish (ES), Farsi (FA), and Russian (RU). The dictionaries are extracted from Wiktionary.[3]

The detailed corpora statistics are in Table 1.

Dataset	Language Pair	Language	#Docs	#Tokens	#Vocabuary	#Translations
Wikipedia	EN-ZH	EN	11,043	1,906,142	13,200	6,812
		ZH	10,135	1,169,056	13,972	
Disaster Response	EN-SI	EN	1,100	32,714	6,920	6,330
		SI	4,790	168,082	31,629	
PACO	EN-AR	EN	1,999	622,955	47,790	4,384
		AR	1,999	107,434	19,900	
	EN-ZH	EN	2,000	405,976	39,847	8,691
		ZH	1,997	86,585	30,481	
	EN-ES	EN	2,000	238,092	30,278	18,221
		ES	2,000	188,469	27,465	
	EN-FA	EN	2,000	513,855	41,685	4,419
		FA	1,814	37,158	9,987	
	EN-RU	EN	1,999	296,148	34,618	2,981
		RU	1,999	101,922	24,341	
INCO	EN-AR	EN	2,000	581,473	45,444	4,380
		AR	1,999	107,434	19,900	
	EN-ZH	EN	2,000	432,442	38,369	8,766
		ZH	1,997	86,585	30,481	
	EN-ES	EN	1,999	557,602	46,161	20,954
		ES	2,000	188,469	27,465	
	EN-FA	EN	2,000	324,858	34,278	4,280
		FA	1,814	37,158	9,987	
	EN-RU	EN	2,000	547,748	47,167	3,345
		RU	1,999	101,922	24,341	

Table 1: Corpora Statistics.

[2] https://www.mdbg.net/chinese/dictionary?page=cc-cedict

[3] https://dumps.wikimedia.org/enwiktionary/

C An Example of Topic Space Transformation with Top-Linked Topics

We give an example to illustrate how we transform topic space with top-linked topics. Suppose that we have a Chinese topic with a probability mass of 0.2 in a document and its topic link weight to English Topics 0–4 are 0.1, 0.4, 0.2, 0.1, 0.2. Given that English Topic 1 has the highest link weight with the Chinese topic, when transforming the document's topic distribution into English, the probability mass of the Chinese topic is transferred to English Topic 1.

References

[1] Gilles Celeux. The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, pages 73–82, 1985.

[2] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, pages 503–528, 1989.

[3] Michelle Yuan, Benjamin Van Durme, and Jordan Boyd-Graber. Multilingual anchoring: Interactive topic modeling and alignment across languages. In *Proceedings of Advances in Neural Information Processing Systems*, 2018.

[4] Stephanie M. Strassel and Jennifer Tracey. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Language Resources and Evaluation Conference*, 2016.

[5] Stephanie M. Strassel, Ann Bies, and Jennifer Tracey. Situational awareness for low resource languages: The LORELEI situation frame annotation task. In *Proceedings of the European Conference on Information Retrieval*, 2017.

[6] Shudong Hao and Michael J. Paul. Learning multilingual topics from incomparable corpora. In *Proceedings of International Conference on Computational Linguistics*, 2018.