Polylingual Topic Models

David Mimno **Hanna M. Wallach** **Limin Yao** **Jason Naradowsky** **Andrew McCallum**

Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
{mimno,wallach,lmyao,narad,mccallum}@cs.umass.edu

Statistical topic models are a useful tool for analyzing large, unstructured document collections [1, 2]. Such collections are increasingly available in multiple languages. Previous work on bilingual topic modeling [4] has focused on aligning pairs of translated sentences. In contrast, we consider "loosely parallel" corpora, in which tuples of documents in different languages are not direct translations, but are known to be about similar topics. We introduce a polylingual topic model (PLTM) and demonstrate it on a collection of Wikipedia articles in ten languages.

Model: PLTM is an extension of LDA for modeling polylingual document tuples, where each tuple is a set of documents that are loosely equivalent to each other, but written in different languages, e.g., corresponding Wikipedia articles in French, English and German. PLTM assumes that the documents in a single tuple share the same distribution over topics. This is unlike LDA, where each document is assumed to have its own document-specific topic distribution. In addition to this, PLTM assumes that each "topic" t consists of a *set* of discrete distributions over word—one for each language. A new document tuple $(\boldsymbol{w}^1, \ldots, \boldsymbol{w}^L)$ is generated by drawing

$$\boldsymbol{\theta} \sim \mathrm{Dir}\,(\boldsymbol{\theta}; \alpha \boldsymbol{m}) \qquad \text{a tuple specific distribution over topics,}$$

and then, for each language l,

$$\boldsymbol{z}^l \sim P(\boldsymbol{z}^l \,|\, \boldsymbol{\theta}) = \textstyle\prod_n \theta_{z_n^l} \qquad \text{a latent topic assignment for each token,}$$
$$\boldsymbol{w}^l \sim P(\boldsymbol{w}^l \,|\, \boldsymbol{z}^l, \Phi^l) = \textstyle\prod_n \phi^l_{w_n^l | z_n^l} \qquad \text{and finally the observed tokens.}$$

The language-specific "topic" parameters Φ^l are drawn from a language-specific symmetric Dirichlet with concentration parameter β^l. The graphical model for PLTM is shown in figure 1(a).

Inference: Given a corpus of training and test document tuples—$\mathcal{W}$ and $\mathcal{W}'$, respectively—two possible tasks of interest are: computing the probability of the test data given the training data, and inferring latent topic assignments for the test data. These tasks can either be accomplished by averaging over samples from $P(\Phi^1, \ldots, \Phi^L, \alpha\boldsymbol{m} \,|\, \mathcal{W}', \beta)$ or by evaluating a point estimate. We take the latter approach, and use the MAP estimate for $\alpha\boldsymbol{m}$ and the predictive distributions over words for $\Phi^1, \ldots, \Phi^L$. The probability of $\mathcal{W}$ given $\mathcal{W}'$ is then approximated by $P(\mathcal{W} \,|\, \Phi^1, \ldots, \Phi^L, \alpha\boldsymbol{m})$.

Topic assignments for a test document tuple $(\boldsymbol{w}^1, \ldots, \boldsymbol{w}^L)$ can be inferred using Gibbs sampling. Gibbs sampling involves sequentially resampling each z_n^l from its conditional posterior:

$$P(z_n^l = t \,|\, (\boldsymbol{w}^1, \ldots, \boldsymbol{w}^L), (\boldsymbol{z}^1, \ldots, \boldsymbol{z}^L)_{\setminus l,n}, \Phi^1, \ldots, \Phi^L, \alpha\boldsymbol{m}) \propto \phi^l_{w_n^l | t} \frac{\{N_t\}_{\setminus l,n} + \alpha m_t}{\sum_t N_t - 1 + \alpha}, \qquad (1)$$

where $(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^L)_{\setminus l,n}$ is the current set of topic assignments for all other tokens in the tuple, while $\{N_t\}_{\setminus l,n}$ is the number of occurrences of topic t in the tuple, excluding the variable of interest.

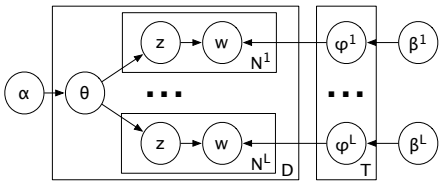
Presenter: HMW **Category:** graphical models **Preference:** Oral

Model	Training Data	Average Log Probability	St. Dev.
LDA	French	-248568.65	16.33
PLTM	French/German	-248492.77	9.67
PLTM	French/English	-249031.59	10.19
PLTM	French/English/German	-249924.29	10.37

Table 1: Average log probability of 74 French test documents, with bootstrap-based error bars.

Generalization Ability: To evaluate generalization ability, we gathered 975 English documents, 796 French documents and 727 German documents from Wikipedia. (The larger number of English documents meant that some document tuples consisted of an English document only.) PLTM was then run on three training corpora of Wikipedia documents: a) French and German, b) French and English, and c) French, English and German. For each corpus, estimates of Φ^F and $\alpha\boldsymbol{m}$ were obtained. As a baseline, LDA was run on French documents only. 74 French test documents were then used to compare generalization ability. For every test document $\boldsymbol{w}^F$, the probability $P(\boldsymbol{w}^F \mid \Phi^F, \alpha\boldsymbol{m})$ was computed using each of the four sets of parameters estimated by PLTM and LDA and a "left-to-right" evaluation algorithm [3]. Log probabilities are shown in table 1. These results demonstrate that additional documents in other languages can indeed help generalization ability of topic models. PLTM trained on French and German documents performed significantly better than LDA (which was trained on only French documents). PLTM trained on the other corpora (all of which included the English documents) performed significantly worse than LDA, however, indicating that simply adding more documents in another language is insufficient to improve performance—the additional documents must be at least roughly semantically equivalent.

Inferred Topics: Figure 1(b) shows the most probable words in each of ten languages for a single topic related to the Ottoman empire. We did not need to remove stop words (e.g. "the", "and", etc.) as the model effectively isolates the syntactic words in all languages in a small number of topics (not shown). Since Wikipedia documents are specifically not translations of one another, it is interesting to explore differences in focus between the languages. We therefore report the percentage of tokens in each language that are assigned to the topic, indicating that Greek, Polish and Russian are relatively more focused on the Ottomans than Finnish and Italian. PLTM also facilitates inference of fine-grained topics for relatively resource-poor languages like Welsh.



(a) Graphical model for PLTM.

Lang	%	Most probable words
Welsh	.20	ei greu swltan ottoman strwythr bymtheg
German	.18	osmanischen osmanische osmanen sultan konstantinopel truppen
Greek	.29	οθωμανική πόλης αυτοκρατορία μωάμεθ κωνσταντινούπολη
English	.15	ottoman the empire turkish ottomans constantinople
Finnish	.06	balkanin turkin muureja kaupungin toukokuuta tuottanut
French	.10	lempire ottoman sultan turcs ottomans éd
Italian	.07	turchi ottomano ottomani limpero sultano veneziano
Polish	.31	turcy turków murów rogu sułtan mury
Portug.	.18	turcos sultão constantinopla ataque otomano muralhas
Russian	.57	османской турки империи турок султан султана

(b) Top words for a single topic in ten languages, along with the percentage of each corpus assigned to this topic.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

[2] T. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1), 2004.

[3] H. M. Wallach. *Structured topic models for language*. PhD thesis, University of Cambridge, 2008.

[4] B. Zhao and E. P. Xing. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *NIPS*, 2007.