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

Abstract

Multilingual topic models (MTMs) learn topics on documents in multiple languages. Past models align topics across languages by implicitly assuming the documents in different languages are highly comparable, often a false assumption. We introduce a new model that does not rely on this assumption, particularly useful in important low-resource language scenarios. Our MTM learns weighted topic links and connects cross-lingual topics only when the dominant words defining them are similar, outperforming LDA and previous MTMs in classification tasks using documents' topic posteriors as features. It also learns coherent topics on documents with low comparability.

1 Introduction

Topic models explain document collections at a high level (Boyd-Graber et al., 2017). Multilingual topic models (MTMs) uncover latent topics *across* languages and reveal commonalities and differences across languages and cultures (Ni et al., 2009; Shi et al., 2016; Gutiérrez et al., 2016). Existing models extend latent Dirichlet allocation (Blei et al., 2003, LDA) and learn *aligned* topics across languages (Mimno et al., 2009).

Prior models work well because they implicitly assume—even if not part of the model—parallel or highly comparable data with well-aligned topics. However, this assumption does not always comport with reality. Even documents from the same place and time can discuss very different things across languages: in multicultural London, Hindi tweets focus on a Bollywood actor's BBC appearance, French blogs fret about Brexit, and English articles focus on Tottenham's lineup. Generally, corpora have a range of "nonparallelness" (Fung, 2000). In less comparable settings, while some topics are shared, languages' emphasis may diverge and some topics may lack analogs.

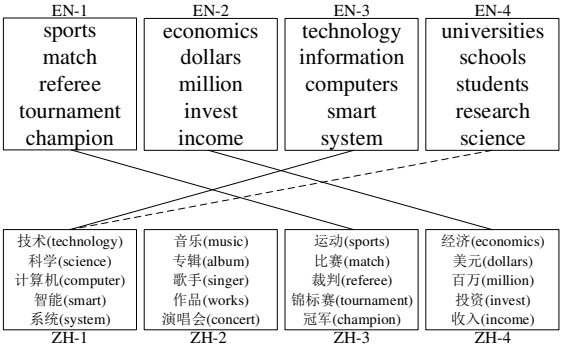

Figure 1: Topic pairs with many word translation pairs have high link weights, e.g., (EN-1, ZH-3) and (EN-2, ZH-4); topic pairs with partial overlap receive lower weights, e.g., (EN-4, ZH-1); a topic is unlinked if there is no corresponding topic in the other language (ZH-2).

We therefore introduce a new multilingual topic model that assumes each language has its own topic sets and jointly learns all topics, but does not force one-to-one alignment across languages. Instead, our MTM learns *weighted* topic links across languages and only assigns a high link weight to a topic pair whose top words have many direct translation pairs (Figure 1). Moreover, it allows unlinked topics if there is no matching topic in the other language. This makes the model robust for (more common) less-comparable data with topic misalignment. Joint inference also allows insights from high-resource languages to uncover low-resource language patterns. It is particularly useful in scenarios that involve modeling topics on low-resource languages in humanitarian assistance, peacekeeping, and/or infectious disease response, while limiting the additional cost to other steps that will also need to be taken, such as finding or creating a word translation dictionary.

We validate the MTM in two classification tasks using inferred topic posteriors as features. Our

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

MTM has higher F1 than other models in both intra- and cross-lingual evaluations, while discovering coherent topics and meaningful topic links.

2 Multilingual Topic Model for Connecting Cross-Lingual Topics

Yang et al. (2015) present a flexible framework for adding regularization to topic models. We extend this model to the multilingual setting by adding a potential function that links topics across languages. For simplicity of exposition, we focus on the bilingual case with languages S and T.

Unlike Yang et al. (2015) that encode monolingual information only, our potential function encodes multilingual knowledge parameterized by two matrices, $\boldsymbol{\rho_{S\to T}}$ and $\boldsymbol{\rho_{T\to S}}$, that transform topics between the two languages. Cells' values are between 0 and 1 and a cell $\rho_{S\to T,k_T,k_S}$ close to one is a strong connection of topics k_T and k_S in language T and S. Transformations $\boldsymbol{\rho}$ are learned from translation pairs' topic distributions.

These topic distributions come from the assignments of Gibbs sampling (Griffiths and Steyvers, 2004). Fortunately adding the potential function is equivalent to adding an additional term to Gibbs sampling for topic models (Yang et al., 2015). During sampling, each token is assigned to a topic, so we can compute a *post hoc* word distribution over topics. The probability of a topic k given a word w is $\Pr(k \mid w) \equiv \Omega_{w,k} \equiv N_{k,w}/N_w$, where $N_{k,w}$ is the number of times that word w is assigned to topic k and N_w is w's term frequency.

To find good topic links $\boldsymbol{\rho_{S\to T}}$, we use a dictionary. For instance, given the translation pair of "sports" and "运 动 (yùn dòng)", they should have similar topic distributions, so we want $\boldsymbol{\rho}_{\text{EN}\to\text{ZH}}\boldsymbol{\Omega}_{\text{sports}}$ to be close to $\boldsymbol{\Omega}_{\text{运动}}$ and vice versa. Moreover, the transformations should be symmetric: $\boldsymbol{\rho_{S\to T}\Omega_{w_S}}$ close to $\boldsymbol{\Omega_{w_T}}$, and vice versa. We encode this cross-lingual knowledge of topic transformations into the potential function Ψ which measures the difference of translation pairs' topic distributions after transformation:

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

where η_c is the statistical importance of the c-th translation pair to the corpus (Figure 2, full details in the Supplement).

While Yang et al. (2015) provide a blueprint for Gibbs sampling with potential functions without additional parameters, our model has additional parameters of $\boldsymbol{\rho_{S\to T}}$ and $\boldsymbol{\rho_{T\to S}}$ so we need to optimize them. Thus, we use stochastic EM (Celeux, 1985). The E-step updates tokens' topic assignments using Gibbs sampling, while holding the parameters of the topic link weight matrices $\boldsymbol{\rho}$ fixed. The M-step optimizes $\boldsymbol{\rho}$ while holding the topic assignments fixed. We optimize Ψ in log space using the objective function $J(\boldsymbol{\rho_{S\to T}})$ as

$$\sum_{c=1}^{C} \eta_c \log \|\boldsymbol{\Omega_{T,c}} - \boldsymbol{\rho_{S\to T,i_T}\Omega_{S,c}}\|_2 , \tag{2}$$

which is minimized by using L-BFGS (Liu and Nocedal, 1989), with the partial derivatives with respect to $\rho_{S\to T,k_T,k_S}$

$$-\sum_{c=1}^{C} \frac{\eta_c \Omega_{S,c,k_S} \left(\Omega_{T,c,k_T} - \boldsymbol{\rho_{S\to T,k_T}\Omega_{S,c}}\right)}{\|\boldsymbol{\Omega_{T,c}} - \boldsymbol{\rho_{S\to T,i_T}\Omega_{S,c}}\|_2^2}. \tag{3}$$

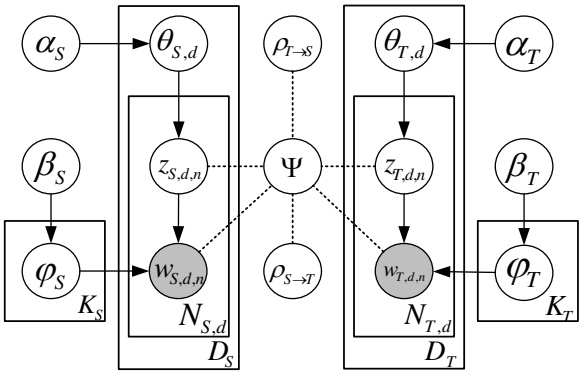

Figure 2: The graphical model of our multilingual topic model. The topic links $\boldsymbol{\rho}$, as instantiated by the function Ψ, encourage topics to encourage word translations to have consistent topics.

3 Experiments

We evaluate our model extrinsically on classification tasks, followed by intrinsic topic coherence.

3.1 Classification with Topic Posteriors

We use two datasets for classification: Wikipedia documents in English (EN) and Chinese (ZH) (Yuan et al., 2018) and an English-Sinhalese (SI) disaster response dataset (Strassel and Tracey, 2016).[1] Each dataset provides labeled documents and a dictionary. Yuan et al. (2018) extract the EN-ZH dictionary from MDBG, while Strassel and Tracey (2016) construct the EN-SI dictionary from online resources and manual annotation.[2] Each

[1] More dataset details in the Supplement.

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

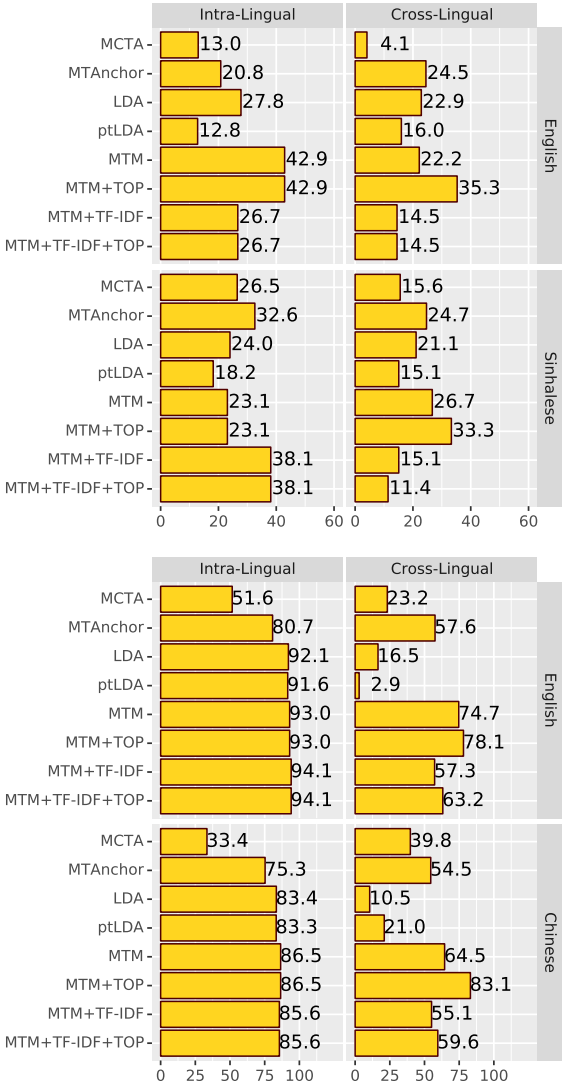

Figure 3: The F1 scores on disaster response (upper) and Wikipedia (lower) datasets. Our MTM outperforms all the baselines in intra- and cross-lingual evaluations.

Wikipedia document is labeled with one of the topics of *film*, *music*, *animals*, *politics*, *religion*, and *food*. A portion of the disaster response documents are labeled with one of eight types of needed rescue resources: *evacuation*, *food supply*, *search/rescue*, *utilities*, *infrastructure*, *medical assistance*, *shelter*, and *water supply*.

We follow Yuan et al. (2018) for preprocessing (such as lemmatization for English and segmentation for Chinese) and use a linear SVM for classification. For the Wikipedia dataset, we report micro-F1 scores on a six-way classification. For the disaster response dataset, our goal is binary classification of the need for *evacuation* versus other assistance. The classification uses features of topic posteriors: $\Pr(k \mid d) \equiv N_{d,k}/N_d$ which is the proportion of the tokens assigned to topic k in document d.

The baselines include polylingual tree LDA (Hu et al., 2014, ptLDA) which encodes the dictionary as a tree prior (Andrzejewski et al., 2009), Multilingual Topic Anchoring (Yuan et al., 2018, MTAnchor), and Multilingual Cultural-common Topic Analysis (Shi et al., 2016, MCTA). We also include LDA, which runs monolingually in each language. We use 20 topics and set hyperparameters $\alpha = 0.1$ and $\beta = 0.01$ (if applicable).

Our evaluations are both intra- and cross-lingual. The intra-lingual evaluation trains and tests classifiers on the same language, while the cross-lingual evaluation trains classifiers on one language and tests on another. In cross-lingual evaluations, MTAnchor, MCTA, and ptLDA align topic spaces, so topic posterior transformation is not necessary. LDA cannot transform topic spaces, so we do not apply any transformation. For our MTM, we explore two transformation methods with $\boldsymbol{\rho}$. The first multiplies $\boldsymbol{\rho}$ with a language's document topic distributions, i.e., $\boldsymbol{\rho}_{\text{ZH}\rightarrow\text{EN}}\boldsymbol{\theta}_{\text{ZH}}$ and vice versa. The second (TOP), transfers each document topic's probability mass to the topic in the other language with the highest link weight.[3]

Our MTM has higher F1 both intra- and cross-lingually (Figure 3). TF-IDF weighting on translation pairs sometimes improves the intra-lingual F1, although it hurts the cross-lingual F1. Connecting the top linked topics (TOP) is better than directly using the topic link weight matrices. This indicates that $\boldsymbol{\rho}$'s values have some noise.

3.2 Looking at Learned Topics

Past MTMs align topics across languages but our MTM does not, so we compare the topics across models to see how they differ. We look at the Movies topics from the Wikipedia dataset (Table 1). For the Chinese MTM topics, we show the three English topics with the highest link weights.

The topics are about Movies, but the MCTA and MTAnchor topics do not rank "movie" or "电影 (diàn yǐng)" at the top. The ptLDA topics, although aligned well, incorrectly align some Chinese words. "胶片 (jiāo piàn)" means "*photographic* film", while "释放 (shì fàng)" means *release* as in "let something go", not movie distribution. ptLDA links words based on translations

[3] An example of TOP is available in the Supplement.

[4] In Tables 1 and 2, "[Q]" denotes the Chinese word is a counter for the following English word.

Lang.	Words
	MCTA
ZH	主演 (starring), 改编 (adapt), 本 (this), 小说 (novel), 拍摄 (shoot), 角色 (role)
EN	dog, san, movie, mexican, fighter, novel
	MTAnchor
ZH	主演 (starring), 改编 (adapt), 饰演 (act), 本片 (the movie), 演员 (actor), 编剧 (playwright)
EN	kong, hong, movie, official, martial, box
	LDA
ZH	电影 (movie), 部 ([Q] movie),[4] 美国 (USA), 上映 (release), 英语 (English), 剧情 (plot)
EN	film, star, direct, release, action, plot
	ptLDA
ZH	电影 (movie), 胶片 (film), 星 (star), 动作 (action), 释放 (release), 影片 (movie)
EN	film, star, direct, action, release, plot
	MTM
ZH	电影 (movie), 部 ([Q] movie), 上映 (release), 动画 (animation), 故事 (story), 作品 (works),
EN (.20)	film, direct, star, release, action, plot
EN (.12)	kill, find, death, attack, escape, return
EN (.11)	shrine, japanese, temple, japan, shinto, kami
	MTM + TF-IDF
ZH	电影 (movie), 部 ([Q] movie), 上映 (release), 美国 (USA), 英语 (English), 导演 (director)
EN (.32)	film, direct, star, action, release, plot
EN (.24)	film, kill, find, escape, attack, return
EN (.09)	character, series, star, game, trek, create

Table 1: The Movies topics given by models. For the Chinese (ZH) topics given by MTM, the top three English (EN) topics and their link weights are also given.

without looking at the context, which causes problems with multiple-sense words. The LDA and MTM topics are generally coherent.

The MTM's unique joint modeling of weighted topic links also recovers additional topical structure: after linking respective EN-ZH Movies topics, the next linked topics are Action Movies ("kill", "death", "attack", and "escape"). Further, the models capture a degree of connection between Movies and Computer Games (MTM + TF-IDF) and Japanese Animations (MTM).

3.3 Looking at Learned Topic Links

We give more examples of weighted MTM topic links in Table 2. High-weighted Biology (ZH-0, EN-12, and EN-19) and Music topics (EN-10, ZH-9, and ZH-17) are characterized by cross-lingual words in common. The model can also infer topic links beyond words, linking topics when the topical words have few direct translations but are related in senses. For instance, ZH-14 is about the "campaigns" against "government". Only "government" overlaps with EN-16 and EN-11, but

Lang.	Words
ZH-0	学名 (scientific name), 它们 (they), 呈 (show), 白色 (white), 长 (long), 黑色 (black)
EN-12 (.57)	specie, bird, eagle, genus, white, owl
EN-19 (.13)	breed, chicken, white, goose, bird, black
ZH-14	主义 (-ism), 组织 (organization), 美国 (USA), 革命 (revolution), 运动 (campaign), 政府 (government)
EN-16 (.32)	sex, law, act, sexual, marriage, court
EN-11 (.17)	traffic, victim, government, trafficking, child, force
EN-10	album, release, record, music, song, single
ZH-9 (.30)	专辑 (album), 张 ([Q] album), 发行 (release), 音乐 (music), 首 ([Q] song), 唱片 (record)
ZH-17 (.20)	音乐 (music), 乐团 (musical group), 艺术 (art), 创作 (create), 奖 (award), 演出 (perform)

Table 2: Topics are linked because they have overlap in topical words. Our MTM can also infer the topic relations beyond words, e.g., ZH-14 and EN-16.

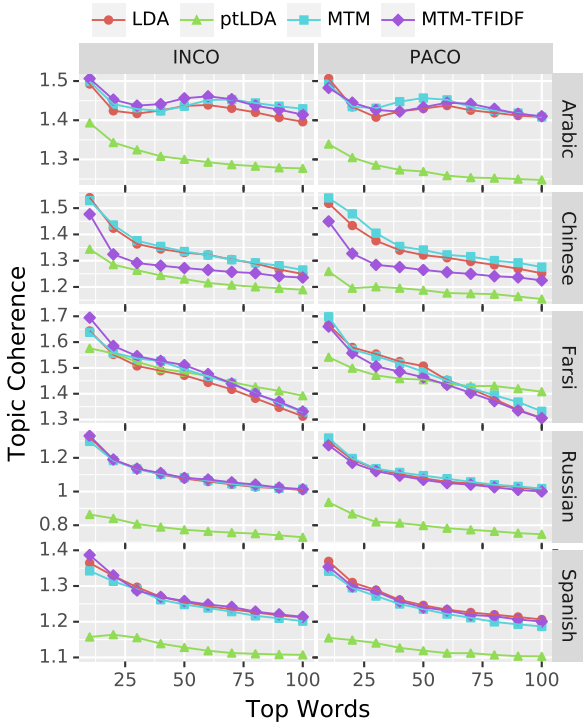

Figure 4: Topic coherence on INCO and PACO datasets with the number of top words in each topic.

MTM identifies the two English topics as the top linked topics for ZH-14: EN-16 is about the "campaign" in Sexual Rights, while EN-11 is about the Crime of human trafficking. This shows that our MTM can incorporate word translations and infer more cross-lingual word and topic relationships.

3.4 Evaluating Topic Coherence

We intrinsically evaluate models' topic coherence on two sets of preprocessed bilingual Wikipedia corpora (Hao and Paul, 2018) that vary in "non-

parallelness". Both pair English with Arabic, Chinese, Spanish, Farsi, and Russian. In PACO, 30% of documents have direct translations across languages, and in INCO none has direct translations. Dictionaries are extracted fromn Wiktionary.[5] Standard preprocessing has already been applied to the datasets, including stemming, stop word removal, and high-frequency word removal.

We use an intra-lingual metric to evaluate topic coherence (Lau et al., 2014): for every topic, we compute its top N words' average pairwise PMI score on a disjoint subset of Wikipedia documents (Hao and Paul, 2018). We report average coherence with N from 10 to 100 with a step size of 10 (five-fold cross-validation). We use the same translation pair weighting options as in our classification tasks and also compare against monolingual LDA and ptLDA (Hu et al., 2014).

MTM is no worse than LDA and sometimes slightly better (Figure 4). TF-IDF weighting on translation pairs sometimes further improves coherence (e.g., Arabic, Farsi, Russian, and Spanish on INCO) but occasionally hurts (e.g., Chinese). ptLDA mostly works poorly, except on Farsi with a high number of top words. ptLDA aligns topic spaces, which is hard for low-comparability data, thus sacrificing coherence for alignment; in contrast MTM only connects topics when they align well in senses.

3.5 Topic Coherence vs. Target Language Corpora Sizes

We next vary the size of target language (non-English languages in PACO and INCO) corpora: how much can MTM help topic coherence for low-resource languages? We start from 10% of the randomly-selected documents in target languages and incrementally add more target language documents at a step size of 10% until it reaches 100%. Meanwhile, we always use 100% of the English documents. We train monolingual LDA, ptLDA, and MTMs with and without TF-IDF weighting on translation pairs on each setting and evaluate the topic coherence on the same reference corpora using the top thirty words of each topic (Figure 5).

In most cases, the topic coherence improves with larger target corpora, except Arabic and Russian on PACO. This confirms our intuition that more data yield a better topic model. MTM is helpful in cases when the target language corpora sizes are small, e.g., Chinese and Russian with 10% or 20% of the corpora. TF-IDF weighting is not consistently better or worse than equal weights.

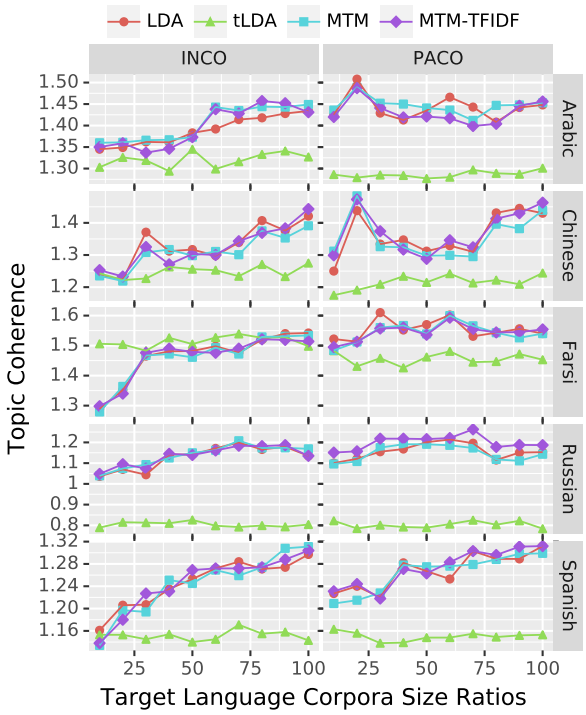

Figure 5: The models' topic coherence on INCO and PACO datasets when the sizes of target language corpora grow from 10% to 100%, with a step size of 10%.

The ptLDA with tree priors based on dictionaries performs poorly in topic coherence, except Farsi in INCO. In most cases, its topic coherence is substantially below others' and improves little when the target corpora grow.

4 Conclusions and Future Work

We introduce a novel multilingual topic model (MTM) that learns weighted topic links across languages by minimizing the Euclidean distances of translation pairs' (transformed) topic distributions, where translation pairs can be weighted, e.g., by TF-IDF. This connects topics in different languages *only* when necessary and is more robust on low-comparability corpora. The MTM outperforms baselines substantially in both intra- and cross-lingual classification tasks, while achieving no worse or slightly better topic coherence than monolingual LDA on low-comparability data.

We plan to explore weighting methods to better evaluate the importance of translation pairs. We will also study how to improve topic transformation with the topic link weight matrices.

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

Acknowledgements

We thank Shudong Hao and Michelle Yuan for providing their datasets. We thank the anonymous reviewers for their insightful and constructive comments. This research has been supported under subcontract to Raytheon BBN Technologies, by DARPA award HR0011-15-C-0113. Boyd-Graber is also supported by NSF grant IIS-1409287. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the International Conference of Machine Learning*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. *Applications of Topic Models*, volume 11 of *Foundations and Trends in Information Retrieval*. NOW Publishers.

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

Pascale Fung. 2000. A statistical view on bilingual lexicon extraction. In *Parallel Text Processing*, pages 219–236.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, pages 5228–5235.

E. Dario Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, pages 47–60.

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

Yuening Hu, Ke Zhai, Vlad Eidelman, and Jordan Boyd-Graber. 2014. Polylingual tree-based topic models for translation domain adaptation. In *Proceedings of the Association for Computational Linguistics*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the Association for Computational Linguistics*.

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

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of Empirical Methods in Natural Language Processing*.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of the World Wide Web Conference*.

Bei Shi, Wai Lam, Lidong Bing, and Yinqing Xu. 2016. Detecting common discussion topics across culture from news reader comments. In *Proceedings of the Association for Computational Linguistics*.

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

Yi Yang, Doug Downey, and Jordan Boyd-Graber. 2015. Efficient methods for incorporating knowledge into topic models. In *Proceedings of Empirical Methods in Natural Language Processing*.

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