Learning Polylingual Topic Models from Code-Switched Social Media Documents

Nanyun Peng **Yiming Wang** **Mark Dredze**
Human Language Technology Center of Excellence
Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD USA
{npeng1,freewym,mdredze}@jhu.edu

Abstract

Code-switched documents are common in social media, providing evidence for polylingual topic models to infer aligned topics across languages. We present Code-Switched LDA (csLDA), which infers language specific topic distributions based on code-switched documents to facilitate multi-lingual corpus analysis. We experiment on two code-switching corpora (English-Spanish Twitter data and English-Chinese Weibo data) and show that csLDA improves perplexity over LDA, and learns semantically coherent aligned topics as judged by human annotators.

1 Introduction

Topic models (Blei et al., 2003) have become standard tools for analyzing document collections, and topic analyses are quite common for social media (Paul and Dredze, 2011; Zhao et al., 2011; Hong and Davison, 2010; Ramage et al., 2010; Eisenstein et al., 2010). Their popularity owes in part to their data driven nature, allowing them to adapt to new corpora and languages. In social media especially, there is a large diversity in terms of both the topic and language, necessitating the modeling of multiple languages simultaneously. A good candidate for multi-lingual topic analyses are polylingual topic models (Mimno et al., 2009), which learn topics for multiple languages, creating tuples of language specific distributions over monolingual vocabularies for each topic. Polylingual topic models enable cross language analysis by grouping documents by topic regardless of language.

Training of polylingual topic models requires parallel or comparable corpora: document tuples from multiple languages that discuss the same topic. While additional non-aligned documents can be folded in during training, the "glue" documents are required to aid in the alignment across languages. However, the ever changing vocabulary and topics of social media (Eisenstein, 2013) make finding suitable comparable corpora difficult. Standard techniques – such as relying on machine translation parallel corpora or comparable documents extracted from Wikipedia in different languages – fail to capture the specific terminology of social media. Alternate methods that rely on bilingual lexicons (Jagarlamudi and Daumé, 2010) similarly fail to adapt to shifting vocabularies. The result: an inability to train polylingual models on social media.

User 1: ¡Don Samuel es un crack! #VamosMéxico #DaleTri
RT @User4: Arriba! Viva Mexico! Advanced to GOLD. medal match in "Football"!
User 2: @user1 rodo que tal el nuevo Mountain ?
User 3: @User1 @User4 wow this is something !! Ja ja ja Football well said

Figure 1: Three users discuss Mexico's football team advancing to the Gold medal game in the 2012 Olympics in code-switched Spanish and English.

In this paper, we offer a solution: utilize code-switched social media to discover correlations across languages. Social media is filled with examples of code-switching, where users switch between two or more languages, both in a conversation and even a single message (Ling et al., 2013). This mixture of languages in the same context suggests alignments between words across languages through the common topics discussed in the context.

We learn from code-switched social media by extending the polylingual topic model framework to infer the language of each token and then automatically processing the learned topics to identify aligned topics. Our model improves both in terms of perplexity and a human evaluation, and we provide some example analyses of social media that rely on our learned topics.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 674–679,
Baltimore, Maryland, USA, June 23-25 2014.

2 Code-Switching

Code-switched documents has received considerable attention in the NLP community. Several tasks have focused on identification and analysis, including mining translations in code-switched documents (Ling et al., 2013), predicting code-switched points (Solorio and Liu, 2008a), identifying code-switched tokens (Lignos and Marcus, 2013; Yu et al., 2012; Elfardy and Diab, 2012), adding code-switched support to language models (Li and Fung, 2012), linguistic processing of code switched data (Solorio and Liu, 2008b), corpus creation (Li et al., 2012; Diab and Kamboj, 2011), and computational linguistic analyses and theories of code-switching (Sankofl, 1998; Joshi, 1982).

Code-switching specifically in social media has also received some recent attention. Lignos and Marcus (2013) trained a supervised token level language identification system for Spanish and English code-switched social media to study code-switching behaviors. Ling et al. (2013) mined translation spans for Chinese and English in code-switched documents to improve a translation system, relying on an existing translation model to aid in the identification and extraction task. In contrast to this work, we take an unsupervised approach, relying only on readily available document level language ID systems to utilize code-switched data. Additionally, our focus is not on individual messages, rather we aim to train a model that can be used to analyze entire corpora.

In this work we consider two types of code-switched documents: single messages and conversations, and two language pairs: Chinese-English and Spanish-English. Figure 1 shows an example of a code-switched Spanish-English *conversation*, in which three users discuss Mexico's football team advancing to the Gold medal game in the 2012 Summer Olympics. In this conversation, some tweets are code-switched and some are in a single language. By collecting the entire conversation into a single document we provide the topic model with additional content. An example of a Chinese-English code-switched messages is given by Ling et al. (2013):

> *watup Kenny Mayne!! - Kenny Mayne*
> 最近这么样啊*!!*

Here a user switches between languages in a single *message*. We empirically evaluate our model on both conversations and messages. In the model presentation we will refer to both as "documents."

3 csLDA

To train a polylingual topic model on social media, we make two modifications to the model of Mimno et al. (2009): add a token specific language variable, and a process for identifying aligned topics.

First, polylingual topic models require parallel or comparable corpora in which each document has an assigned language. In the case of code-switched social media data, we require a *per-token* language variable. However, while document level language identification (LID) systems are common place, very few languages have per-token LID systems (King and Abney, 2013; Lignos and Marcus, 2013).

To address the lack of available LID systems, we add a per-token latent language variable to the polylingual topic model. For documents that are not code-switched, we observe these variables to be the output of a document level LID system. In the case of code-switched documents, these variables are inferred during model inference.

Second, polylingual topic models assume the aligned topics are from parallel or comparable corpora, which implicitly assumes that a topics popularity is balanced across languages. Topics that show up in one language necessarily show up in another. However, in the case of social media, we can make no such assumption. The topics discussed are influenced by users, time, and location, all factors intertwined with choice of language. For example, English speakers will more likely discuss Olympic basketball while Spanish speakers football. There may be little or no documents on a given topic in one language, while they are plentiful in another. In this case, a polylingual topic model, which necessarily infers a topic-specific word distribution for each topic in each language, would learn two unrelated word distributions in two languages for a single topic. Therefore, naively using the produced topics as "aligned" across languages is ill-advised.

Our solution is to automatically identify aligned polylingual topics after learning by examining a topic's distribution across code-switched documents. Our metric relies on distributional properties of an inferred topic across the entire collection.

To summarize, based on the model of Mimno et al. (2009) we will learn:

- For each topic, a language specific word distribution.
- For each (code-switched) token, a language.
- For each topic, an identification as to whether the topic captures an alignment across languages.

The first two goals are achieved by incorporating new hidden variables in the traditional polylingual topic model. The third goal requires an automated post-processing step. We call the resulting model Code-Switched LDA (csLDA). The generative process is as follows:

- For each topic $z \in \mathcal{T}$
  - For each language $l \in \mathcal{L}$
    - Draw word distribution $\phi_z^l \sim Dir(\beta^l)$
- For each document $d \in \mathcal{D}$:
  - Draw a topic distribution $\theta_d \sim Dir(\alpha)$
  - Draw a language distribution $\psi_d \sim Dir(\gamma)$
  - For each token $i \in d$:
    - Draw a topic $z_i \sim \theta_d$
    - Draw a language $l_i \sim \psi_d$
    - Draw a word $w_i \sim \phi_z^l$

For monolingual documents, we fix l_i to the LID tag for all tokens. Additionally, we use a single background distribution for each language to capture stopwords; a control variable π, which follows a Dirichlet distribution with prior parameterized by δ, is introduced to decide the choice between background words and topic words following (Chemudugunta et al., 2006)[1]. We use asymmetric Dirichlet priors (Wallach et al., 2009), and let the optimization process learn the hyperparameters. The graphical model is shown in Figure 2.

3.1 Inference

Inference for csLDA follows directly from LDA. A Gibbs sampler learns the word distributions ϕ_z^l for each language and topic. We use a block Gibbs sampler to jointly sample topic and language variables for each token. As is customary, we collapse out ϕ, θ and ψ. The sampling posterior is:

$$P(z_i, l_i | \mathbf{w}, \mathbf{z}_{-i}, \mathbf{l}_{-i}, \alpha, \beta, \gamma) \propto \frac{(n_{w_i}^{l,z})_{-i} + \beta}{n_{-i}^{l,z} + \mathcal{W}\beta} \times \frac{m_{-i}^{z,d} + \alpha}{m_{-i}^{d} + \mathcal{T}\alpha} \times \frac{o_{-i}^{l,d} + \gamma}{o_{-i}^{d} + \mathcal{L}\gamma} \quad (1)$$

where $(n_{w_i}^{l,z})_{-i}$ is the number of times the type for word w_i assigned to topic z and language l (excluding current word w_i), $m_{-i}^{z,d}$ is the number of tokens assigned to topic z in document d (excluding current word w_i), $o_{-i}^{l,d}$ is the number of tokens assigned to language l in document d (excluding current word w_i), and these variables with superscripts or subscripts omitted are totals across all values for the variable. $\mathcal{W}$ is the number of words in the corpus. All counts omit words assigned to the background. During sampling, words are first assigned to the background/topic distribution and then topic and language are sampled for non-background words.

[1] Omitted from the generative process but shown in Fig. 2.

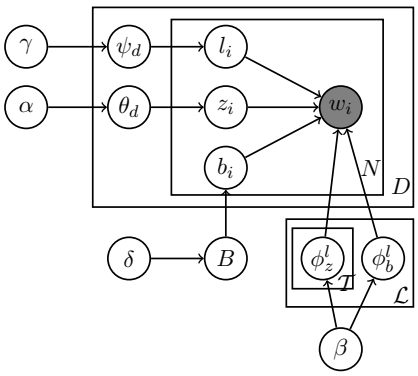


Figure 2: The graphical model for csLDA.

We optimize the hyperparameters α, β, γ and δ by interleaving sampling iterations with a Newton-Raphson update to obtain the MLE estimate for the hyperparameters. Taking α as an example, one step of the Newton-Raphson update is:

$$\alpha^{new} = \alpha^{old} - \mathbf{H}^{-1} \frac{\partial \mathcal{L}}{\partial \alpha} \quad (2)$$

where $\mathbf{H}$ is the Hessian matrix and $\frac{\partial \mathcal{L}}{\partial \alpha}$ is the gradient of the likelihood function with respect to the optimizing hyperparameter. We interleave 200 sampling iterations with one Newton-Raphson update.

3.2 Selecting Aligned Topics

We next identify learned topics (a set of related word-distributions) that truly represent an aligned topic across languages, as opposed to an unrelated set of distributions for which there is no supporting alignment evidence in the corpus. We begin by measuring how often each topic occurs in code-switched documents. If a topic never occurs in a code-switched document, then there can be no evidence to support alignment across languages. For the topics that appear at least once in a code-switched document, we estimate their probability

in the code-switched documents by a MAP estimate of θ. Topics appearing in at least one code-switched document with probability greater than a threshold p are selected as candidates for true cross-language topics.

4 Data

We used two datasets: a Sina Weibo Chinese-English corpus (Ling et al., 2013) and a Spanish-English Twitter corpus.

Weibo Ling et al. (2013) extracted over 1m Chinese-English parallel segments from Sina Weibo, which are code-switched messages. We randomly sampled 29,705 code-switched messages along with 42,116 Chinese and 42,116 English messages from the the same time frame. We used these data for training. We then sampled an additional 2475 code-switched messages, 4221 English and 4211 Chinese messages as test data.

Olympics We collected tweets from July 27, 2012 to August 12, 2012, and identified 302,775 tweets about the Olympics based on related hashtags and keywords (e.g. olympics, #london2012, etc.) We identified code-switched tweets using the Chromium Language Detector[2]. This system provides the top three possible languages for a given document with confidence scores; we identify a tweet as code-switched if two predicted languages each have confidence greater than 33%. We then used the tagger of Lignos and Marcus (2013) to obtain token level LID tags, and only tweets with tokens in both Spanish and English are used as code-switched tweets. In total we identified 822 Spanish-English code-switched tweets. We further expanded the mined tweets to full conversations, yielding 1055 Spanish-English code-switched documents (including both tweets and conversations), along with 4007 English and 4421 Spanish tweets composes our data set. We reserve 10% of the data for testing.

5 Experiments

We evaluated csLDA on the two datasets and evaluated each model using perplexity on held out data and human judgements. While our goal is to learn polylingual topics, we cannot compare to previous polylingual models since they require comparable data, which we lack. Instead, we constructed a baseline from LDA run on the entire dataset (no language information.) For each model, we measured the document completion perplexity (Rosen-Zvi et al., 2004) on the held out data. We experimented with different numbers of topics ($\mathcal{T}$). Since csLDA duplicates topic distributions ($\mathcal{T} \times \mathcal{L}$) we used twice as many topics for LDA.

Figure 3 shows test perplexity for varying $\mathcal{T}$ and perplexity for the best setting of csLDA ($\mathcal{T}$ =60) and LDA ($\mathcal{T}$=120). The table lists both monolingual and code-switched test data; csLDA improves over LDA in almost every case, and across all values of $\mathcal{T}$. The background distribution (-bg) has mixed results for LDA, whereas for csLDA it shows consistent improvement. Table 4 shows some csLDA topics. While there are some mistakes, overall the topics are coherent and aligned.

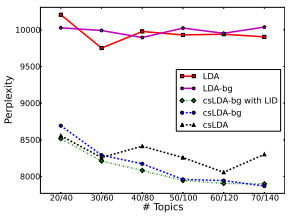
We use the available per-token LID system (Lignos and Marcus, 2013) for Spanish/English to justify csLDA's ability to infer the hidden language variables. We ran csLDA-bg with l_i set to the value provided by the LID system for code-switched documents (csLDA-bg with LID), which gives csLDA high quality LID labels. While we see gains for the code-switched data, overall the results for csLDA-bg and csLDA-bg with LID are similar, suggesting that the model can operate effectively even without a supervised per-token LID system.

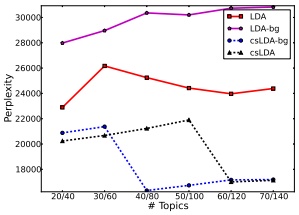
5.1 Human Evaluation

We evaluate topic alignment quality through a human judgements (Chang et al., 2009). For each aligned topic, we show an annotator the 20 most frequent words from the foreign language topic (Chinese or Spanish) with the 20 most frequent words from the aligned English topic and two random English topics. The annotators are asked to select the most related English topic among the three; the one with the most votes is considered the aligned topic. We count how often the model's alignments agree.

LDA may learn comparable topics in different languages but gives no explicit alignments. We create alignments by classifying each LDA topic by language using the KL-divergence between the topic's words distribution and a word distribution for the English/foreign language inferred from the monolingual documents. Language is assigned to a topic by taking the minimum KL. For Weibo data, this was not effective since the vocabularies of each language are highly unbalanced. Instead,

[2] https://code.google.com/p/chromium-compact-language-detector/





$\mathcal{T}$=60/120	Olympics			Weibo		
	En	Es	CS	En	Cn	CS
LDA	11.32	9.44	6.97	29.19	23.06	11.69
LDA-bg	11.35	9.51	6.79	40.87	27.56	10.91
csLDA	8.72	7.94	6.17	18.20	17.31	12.72
csLDA-bg	8.72	7.73	6.04	18.25	17.74	12.46
csLDA-bg with LID	8.73	7.93	4.91	-	-	-

Figure 3: Plots show perplexity for different $\mathcal{T}$ (Olympics left, Weibo right). Perplexity in the table are in magnitude of 1×10^3.

Football		Basketball	
English	Spanish	English	Spanish
mexico	mucho	game	españa
brazil	argentina	basketball	baloncesto
soccer	méxico	year	basketball
vs	brasil	finals	bronce
womens	ganará	gonna	china
football	tri	nba	final
mens	yahel_castillo	obama	rusia
final	delpo	lebron	española

Social Media		Transportation	
English	Chinese	English	Chinese
twitter	啊啊啊	car	汽车
bitly	微博	drive	这个
facebook	更新	road	真真
check	下载	line	明年
use	转发	train	自行车
blog	视频	harry	车型
free	pm	汽车	奔驰
post	推特	bus	大众

Figure 4: Examples of aligned topics from Olympics (left) and Weibo (right).

we manually labeled the topics by language. We then pair topics across languages using the cosine similarity of their co-occurrence statistics in code-switched documents. Topic pairs with similarity above t are considered aligned topics. We also used a threshold p (§3.2) to select aligned topics in csLDA. To ensure a fair comparison, we select the same number of aligned topics for LDA and csLDA.[3]. We used the best performing setting: csLDA $\mathcal{T}$=60, LDA $\mathcal{T}$=120, which produced 12 alignments from Olympics and 28 from Weibo.

Using Mechanical Turk we collected multiple judgements per alignment. For Spanish, we removed workers who disagreed with the majority more than 50% of the time (83 deletions), leaving 6.5 annotations for each alignment (85.47% inter-annotator agreement.) For Chinese, since quality of general Chinese turkers is low (Pavlick et al., 2014) we invited specific workers and obtained 9.3 annotations per alignment (78.72% inter-annotator agreement.) For Olympics, LDA alignments matched the judgements 25% of the time, while csLDA matched 50% of the time. While csLDA found 12 alignments and LDA 29, the 12 topics evaluated from both models show that csLDA's alignments are higher quality. For the Weibo data, LDA matched judgements 71.4%, while csLDA matched 75%. Both obtained high quality alignments – likely due both to the fact that the code-switched data is curated to find translations and we hand labeled topic language – but csLDA found many more alignments: 60 as compared to 28. These results confirm our automated results: csLDA finds higher quality topics that span both languages.

[3]We used thresholds $p = 0.2$ and $t = 0.0001$. We limited the model with more alignments to match the one with less.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*.

Mona Diab and Ankit Kamboj. 2011. Feasibility of leveraging crowd sourcing for the creation of a large scale annotated resource for Hindi English code switched data: A pilot annotation. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 36–40, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model

for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *NAACL*.

Heba Elfardy and Mona Diab. 2012. Token level identification of linguistic code switching. In *Proceedings of COLING 2012: Posters*, pages 287–296, Mumbai, India, December. The COLING 2012 Organizing Committee.

Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM.

Jagadeesh Jagarlamudi and Hal Daumé. 2010. Extracting multilingual topics from unaligned comparable corpora. *Advances in Information Retrieval*, pages 444–456.

Aravind K Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th Conference on Computational linguistics (COLING)*, pages 145–150.

Ben King and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *NAACL*.

Ying Li and Pascale Fung. 2012. Code-switch language model with inversion constraints for mixed language speech recognition. In *Proceedings of COLING 2012*, pages 1671–1680, Mumbai, India, December. The COLING 2012 Organizing Committee.

Ying Li, Yue Yu, and Pascale Fung. 2012. A mandarin-english code-switching corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2515–2519, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1573.

Constantine Lignos and Mitch Marcus. 2013. Toward web-scale analysis of codeswitching. In *Annual Meeting of the Linguistic Society of America*.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, ACL '13. Association for Computational Linguistics.

David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual 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.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM*.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2(Feb):79–92.

Daniel Ramage, Susan T Dumais, and Daniel J Liebling. 2010. Characterizing microblogs with topic models. In *ICWSM*.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

David Sankofl. 1998. The production of code-mixed discourse. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 8–21, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981, Honolulu, Hawaii, October. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008b. Part-of-Speech tagging for English-Spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060, Honolulu, Hawaii, October. Association for Computational Linguistics.

Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking lda: Why priors matter. In *NIPS*, volume 22, pages 1973–1981.

Liang-Chih Yu, Wei-Cheng He, and Wei-Nan Chien. 2012. A language modeling approach to identifying code-switched sentences and words. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 3–8, Tianjin, China, December. Association for Computational Linguistics.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer.