Polylingual Topic Models

David Mimno Hanna M. Wallach Jason Naradowsky David A. Smith Andrew McCallum
University of Massachusetts, Amherst
Amherst, MA 01003
{mimno, wallach, narad, dasmith, mccallum}cs.umass.edu

Abstract

Topic models are a useful tool for analyzing large text collections, but have previously been applied in only monolingual, or at most bilingual, contexts. Meanwhile, massive collections of interlinked documents in dozens of languages, such as Wikipedia, are now widely available, calling for tools that can characterize content in many languages. We introduce a polylingual topic model that discovers topics aligned across multiple languages. We explore the model's characteristics using two large corpora, each with over ten different languages, and demonstrate its usefulness in supporting machine translation and tracking topic trends across languages.

1 Introduction

Statistical topic models have emerged as an increasingly useful analysis tool for large text collections. Topic models have been used for analyzing topic trends in research literature (Mann et al., 2006; Hall et al., 2008), inferring captions for images (Blei and Jordan, 2003), social network analysis in email (McCallum et al., 2005), and expanding queries with topically related words in information retrieval (Wei and Croft, 2006). Much of this work, however, has occurred in monolingual contexts. In an increasingly connected world, the ability to access documents in many languages has become both a strategic asset and a personally enriching experience. In this paper, we present the polylingual topic model (PLTM). We demonstrate its utility and explore its characteristics using two polylingual corpora: proceedings of the European parliament (in eleven languages) and a collection of Wikipedia articles (in twelve languages).

There are many potential applications for polylingual topic models. Although research literature is typically written in English, bibliographic databases often contain substantial quantities of work in other languages. To perform topic-based bibliometric analysis on these collections, it is necessary to have topic models that are aligned across languages. Such analysis could be significant in tracking international research trends, where language barriers slow the transfer of ideas.

Previous work on bilingual topic modeling has focused on machine translation applications, which rely on sentence-aligned parallel translations. However, the growth of the internet, and in particular Wikipedia, has made vast corpora of *topically* comparable texts—documents that are topically similar but are not direct translations of one another—considerably more abundant than ever before. We argue that topic modeling is both a useful and appropriate tool for leveraging correspondences between semantically comparable documents in multiple different languages.

In this paper, we use two polylingual corpora to answer various critical questions related to polylingual topic models. We employ a set of direct translations, the EuroParl corpus, to evaluate whether PLTM can accurately infer topics when documents genuinely contain the same content. We also explore how the characteristics of different languages affect topic model performance. The second corpus, Wikipedia articles in twelve languages, contains sets of documents that are not translations of one another, but are very likely to be about similar concepts. We use this corpus to explore the ability of the model both to infer similarities between vocabularies in different languages, and to detect differences in topic emphasis between languages. The internet makes it possible for people all over the world to access documents from different cultures, but readers will not be fluent in this wide variety of languages. By linking topics across languages, polylingual topic models can increase cross-cultural understanding by providing readers with the ability to characterize

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 880–889,
Singapore, 6-7 August 2009.

the contents of collections in unfamiliar languages and identify trends in topic prevalence.

2 Related Work

Bilingual topic models for parallel texts with word-to-word alignments have been studied previously using the HM-bitam model (Zhao and Xing, 2007). Tam, Lane and Schultz (Tam et al., 2007) also show improvements in machine translation using bilingual topic models. Both of these translation-focused topic models infer word-to-word alignments as part of their inference procedures, which would become exponentially more complex if additional languages were added. We take a simpler approach that is more suitable for topically similar document tuples (where documents are not direct translations of one another) in more than two languages. A recent extended abstract, developed concurrently by Ni et al. (Ni et al., 2009), discusses a multilingual topic model similar to the one presented here. However, they evaluate their model on only two languages (English and Chinese), and do not use the model to detect differences between languages. They also provide little analysis of the differences between polylingual and single-language topic models. Outside of the field of topic modeling, Kawaba et al. (Kawaba et al., 2008) use a Wikipedia-based model to perform sentiment analysis of blog posts. They find, for example, that English blog posts about the Nintendo Wii often relate to a hack, which cannot be mentioned in Japanese posts due to Japanese intellectual property law. Similarly, posts about whaling often use (positive) nationalist language in Japanese and (negative) environmentalist language in English.

3 Polylingual Topic Model

The polylingual topic model (PLTM) is an extension of latent Dirichlet allocation (LDA) (Blei et al., 2003) for modeling polylingual document tuples. 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 tuple share the same tuple-specific distribution over topics. This is unlike LDA, in which each document is assumed to have its own document-specific distribution over topics. Additionally, PLTM assumes that each "topic" consists of a *set* of discrete distributions

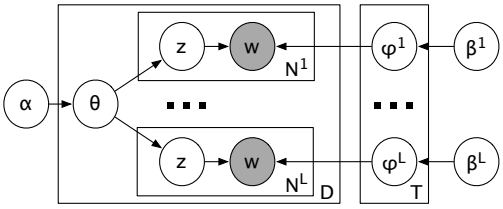

Figure 1: Graphical model for PLTM.

over words—one for each language $l = 1, \ldots, L$. In other words, rather than using a single set of topics $\Phi = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T\}$, as in LDA, there are L sets of language-specific topics, $\Phi^1, \ldots, \Phi^L$, each of which is drawn from a language-specific symmetric Dirichlet with concentration parameter β^l.

3.1 Generative Process

A new document tuple $\boldsymbol{w} = (\boldsymbol{w}^1, \ldots, \boldsymbol{w}^L)$ is generated by first drawing a tuple-specific topic distribution from an asymmetric Dirichlet prior with concentration parameter α and base measure $\boldsymbol{m}$:

$$\boldsymbol{\theta} \sim \text{Dir}\,(\boldsymbol{\theta}, \alpha \boldsymbol{m}). \tag{1}$$

Then, for each language l, a latent topic assignment is drawn for each token in that language:

$$\boldsymbol{z}^l \sim P(\boldsymbol{z}^l \,|\, \boldsymbol{\theta}) = \textstyle\prod_n \theta_{z_n^l}. \tag{2}$$

Finally, the observed tokens are themselves drawn using the language-specific topic parameters:

$$\boldsymbol{w}^l \sim P(\boldsymbol{w}^l \,|\, \boldsymbol{z}^l, \Phi^l) = \textstyle\prod_n \phi^l_{w_n^l | z_n^l}. \tag{3}$$

The graphical model is shown in figure 1.

3.2 Inference

Given a corpus of training and test document tuples—$\mathcal{W}$ and $\mathcal{W}'$, respectively—two possible inference tasks of interest are: computing the probability of the test tuples given the training tuples and inferring latent topic assignments for test documents. These tasks can either be accomplished by averaging over samples of $\Phi^1, \ldots, \Phi^L$ and $\alpha\boldsymbol{m}$ 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 held-out document tuples $\mathcal{W}'$ given training tuples $\mathcal{W}$ is then approximated by $P(\mathcal{W}' \,|\, \Phi^1, \ldots, \Phi^L, \alpha\boldsymbol{m})$.

Topic assignments for a test document tuple $\boldsymbol{w} = (\boldsymbol{w}^1, \ldots, \boldsymbol{w}^L)$ can be inferred using Gibbs

sampling. Gibbs sampling involves sequentially resampling each z_n^l from its conditional posterior:

$$P(z_n^l = t \mid \boldsymbol{w}, \boldsymbol{z}_{\setminus l,n}, \Phi^1, \ldots, \Phi^L, \alpha\boldsymbol{m}) \propto \phi^l_{w_n^l|t} \frac{(N_t)_{\setminus l,n} + \alpha m_t}{\sum_t N_t - 1 + \alpha}, \quad (4)$$

where $\boldsymbol{z}_{\setminus l,n}$ 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 z_n^l, the variable being resampled.

4 Results on Parallel Text

Our first set of experiments focuses on document tuples that are known to consist of direct translations. In this case, we can be confident that the topic distribution is genuinely shared across all languages. Although direct translations in multiple languages are relatively rare (in contrast with comparable documents), we use direct translations to explore the characteristics of the model.

4.1 Data Set

The EuroParl corpus consists of parallel texts in eleven western European languages: Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese and Swedish. These texts consist of roughly a decade of proceedings of the European parliament. For our purposes we use alignments at the speech level rather than the sentence level, as in many translation tasks using this corpus. We also remove the twenty-five most frequent word types for efficiency reasons. The remaining collection consists of over 121 million words. Details by language are shown in Table 1.

Table 1: Average document length, # documents, and unique word types per 10,000 tokens in the EuroParl corpus.

Lang.	Avg. leng.	# docs	types/10k
DA	160.153	65245	121.4
DE	178.689	66497	124.5
EL	171.289	46317	124.2
EN	176.450	69522	43.1
ES	170.536	65929	59.5
FI	161.293	60822	336.2
FR	186.742	67430	54.8
IT	187.451	66035	69.5
NL	176.114	66952	80.8
PT	183.410	65718	68.2
SV	154.605	58011	136.1

Models are trained using 1000 iterations of Gibbs sampling. Each language-specific topic–word concentration parameter β^l is set to 0.01.

DA centralbank europæiske ecb s lån centralbanks
DE zentralbank ezb bank europäischen investitionsbank darlehen
EL τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN **bank central ecb banks european monetary**
ES banco central europeo bce bancos centrales
FI keskuspankin ekp n euroopan keskuspankki eip
FR banque centrale bce européenne banques monétaire
IT banca centrale bce europea banche prestiti
NL bank centrale ecb europese banken leningen
PT banco central europeu bce bancos empréstimos
SV centralbanken europeiska ecb centralbankens s lån

DA børn familie udnyttelse børns børnene seksuel
DE kinder kindern familie ausbeutung familien eltern
EL παιδιά παιδιών οικογένεια οικογένειας γονείς παιδικής
EN **children family child sexual families exploitation**
ES niños familia hijos sexual infantil menores
FI lasten lapsia lapset perheen lapsen lapsiin
FR enfants famille enfant parents exploitation familles
IT bambini famiglia figli minori sessuale sfruttamento
NL kinderen kind gezin seksuele ouders familie
PT crianças família filhos sexual criança infantil
SV barn barnen familjen sexuellt familj utnyttjande

DA mål nå målsætninger målet målsætning opnå
DE ziel ziele erreichen zielen erreicht zielsetzungen
EL στόχους στόχο στόχος στόχων στόχοι επίτευξη
EN **objective objectives achieve aim ambitious set**
ES objetivo objetivos alcanzar conseguir lograr estos
FI tavoite tavoitteet tavoitteena tavoitteiden tavoitteita tavoitteen
FR objectif objectifs atteindre but cet ambitieux
IT obiettivo obiettivi raggiungere degli scopo quello
NL doelstellingen doel doelstelling bereiken bereikt doelen
PT objectivo objectivos alcançar atingir ambicioso conseguir
SV mål målet uppnå målen målsättningar målsättning

DA andre anden side ene andet øvrige
DE anderen andere einen wie andererseits anderer
EL άλλες άλλα άλλη άλλων άλλους όπως
EN **other one hand others another there**
ES otros otras otro otra parte demás
FI muiden toisaalta muita muut muihin muun
FR autres autre part côté ailleurs même
IT altri altre altro altra dall parte
NL andere anderzijds anderen ander als kant
PT outros outras outro lado outra noutros
SV andra sidan å annat ena annan

Figure 2: EuroParl topics (T=400)

The concentration parameter α for the prior over document-specific topic distributions is initialized to $0.01\,T$, while the base measure $\boldsymbol{m}$ is initialized to the uniform distribution. Hyperparameters $\alpha\boldsymbol{m}$ are re-estimated every 10 Gibbs iterations.

4.2 Analysis of Trained Models

Figure 2 shows the most probable words in all languages for four example topics, from PLTM with 400 topics. The first topic contains words relating to the European Central Bank. This topic provides an illustration of the variation in technical terminology captured by PLTM, including the wide array of acronyms used by different languages. The second topic, concerning children, demonstrates the variability of everyday terminology: although the four Romance languages are closely

related, they use etymologically unrelated words for children. (Interestingly, all languages except Greek and Finnish use closely related words for "youth" or "young" in a separate topic.) The third topic demonstrates differences in inflectional variation. English and the Romance languages use only singular and plural versions of "objective." The other Germanic languages include compound words, while Greek and Finnish are dominated by inflected variants of the same lexical item. The final topic demonstrates that PLTM effectively clusters "syntactic" words, as well as more semantically specific nouns, adjectives and verbs.

Although the topics in figure 2 seem highly focused, it is interesting to ask whether the model is genuinely learning mixtures of topics or simply assigning entire document tuples to single topics. To answer this question, we compute the posterior probability of each topic in each tuple under the trained model. If the model assigns all tokens in a tuple to a single topic, the maximum posterior topic probability for that tuple will be near to 1.0. If the model assigns topics uniformly, the maximum topic probability will be near $1/T$. We compute histograms of these maximum topic probabilities for $T \in \{50, 100, 200, 400, 800\}$. For clarity, rather than overlaying five histograms, figure 3 shows the histograms converted into smooth curves using a kernel density estimator.[1] Although there is a small bump around 1.0 (for extremely short documents, e.g., "Applause"), values are generally closer to, but greater than, $1/T$.

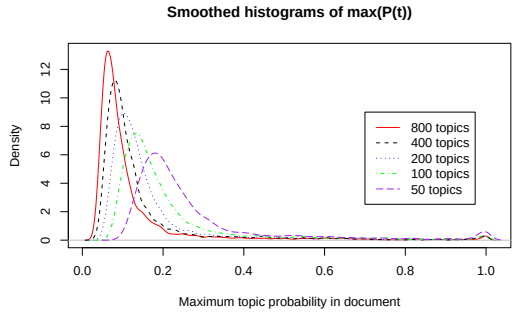

Figure 3: Smoothed histograms of the probability of the most probable topic in a document tuple.

Although the posterior distribution over topics for each tuple is not concentrated on one topic, it is worth checking that this is not simply because the model is assigning a single topic to the tokens in each of the languages. Although the model does not distinguish between topic assignment variables within a given document tuple (so it is technically incorrect to speak of different posterior distributions over topics for different documents in a given tuple), we can nevertheless divide topic assignment variables between languages and use them to estimate a Dirichlet-multinomial posterior distribution for each language in each tuple. For each tuple we can then calculate the Jensen-Shannon divergence (the average of the KL divergences between each distribution and a mean distribution) between these distributions. Figure 4 shows the density of these divergences for different numbers of topics. As with the previous figure, there are a small number of documents that contain only one topic in all languages, and thus have zero divergence. These tend to be very short, formulaic parliamentary responses, however. The vast majority of divergences are relatively low (1.0 indicates no overlap in topics between languages in a given document tuple) indicating that, for each tuple, the model is not simply assigning all tokens in a particular language to a single topic. As the number of topics increases, greater variability in topic distributions causes divergence to increase.

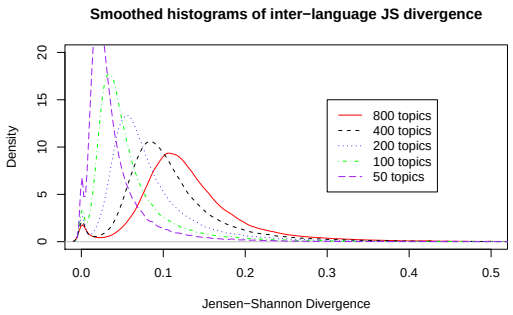

Figure 4: Smoothed histograms of the Jensen-Shannon divergences between the posterior probability of topics between languages.

4.3 Language Model Evaluation

A topic model specifies a probability distribution over documents, or in the case of PLTM, document tuples. Given a set of training document tuples, PLTM can be used to obtain posterior estimates of $\Phi^1, \ldots, \Phi^L$ and $\alpha\boldsymbol{m}$. The probability of previously unseen held-out document tuples given these estimates can then be computed. The higher the probability of the held-out document tuples, the better the generalization ability of the model.

[1] We use the R `density` function.

Analytically calculating the probability of a set of held-out document tuples given $\Phi^1, \ldots, \Phi^L$ and $\alpha\boldsymbol{m}$ is intractable, due to the summation over an exponential number of topic assignments for these held-out documents. However, recently developed methods provide efficient, accurate estimates of this probability. We use the "left-to-right" method of (Wallach et al., 2009). We perform five estimation runs for each document and then calculate standard errors using a bootstrap method.

Table 2 shows the log probability of held-out data in nats per word for PLTM and LDA, both trained with 200 topics. There is substantial variation between languages. Additionally, the predictive ability of PLTM is consistently slightly worse than that of (monolingual) LDA. It is important to note, however, that these results do not imply that LDA should be preferred over PLTM—that choice depends upon the needs of the modeler. Rather, these results are intended as a quantitative analysis of the difference between the two models.

Table 2: Held-out log probability in nats/word. (Smaller magnitude implies better language modeling performance.) PLTM does slightly worse than monolingual LDA models, but the variation between languages is much larger.

Lang	PLTM	sd	LDA	sd
DA	-8.11	0.00067	-8.02	0.00066
DE	-8.17	0.00057	-8.08	0.00072
EL	-8.44	0.00079	-8.36	0.00087
EN	-7.51	0.00064	-7.42	0.00069
ES	-7.98	0.00073	-7.87	0.00070
FI	-9.25	0.00089	-9.21	0.00065
FR	-8.26	0.00072	-8.19	0.00058
IT	-8.11	0.00071	-8.02	0.00058
NL	-7.84	0.00067	-7.75	0.00099
PT	-7.87	0.00085	-7.80	0.00060
SV	-8.25	0.00091	-8.16	0.00086

As the number of topics is increased, the word counts per topic become very sparse in monolingual LDA models, proportional to the size of the vocabulary. Figure 5 shows the proportion of all tokens in English and Finnish assigned to each topic under LDA and PLTM with 800 topics. More than 350 topics in the Finnish LDA model have zero tokens assigned to them, and almost all tokens are assigned to the largest 200 topics. English has a larger tail, with non-zero counts in all but 16 topics. In contrast, PLTM assigns a significant number of tokens to almost all 800 topics, in very similar proportions in both languages. PLTM topics therefore have a higher granularity – i.e., they are more specific. This result is important: informally, we have found that increasing the granularity of topics correlates strongly with user perceptions of the utility of a topic model.

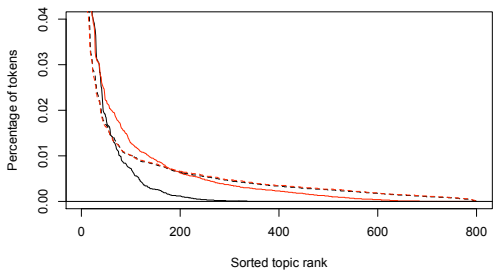

Figure 5: Topics sorted by number of words assigned. Finnish is in black, English is in red; LDA is solid, PLTM is dashed. LDA in Finnish essentially learns a 200 topic model when given 800 topics, while PLTM uses all 800 topics.

4.4 Partly Comparable Corpora

An important application for polylingual topic modeling is to use small numbers of comparable document tuples to link topics in larger collections of distinct, non-comparable documents in multiple languages. For example, a journal might publish papers in English, French, German and Italian. No paper is exactly comparable to any other paper, but they are all roughly topically similar. If we wish to perform topic-based bibliometric analysis, it is vital to be able to track the same topics across all languages. One simple way to achieve this topic alignment is to add a small set of comparable document tuples that provide sufficient "glue" to bind the topics together. Continuing with the example above, one might extract a set of connected Wikipedia articles related to the focus of the journal and then train PLTM on a joint corpus consisting of journal papers and Wikipedia articles.

In order to simulate this scenario we create a set of variations of the EuroParl corpus by treating some documents as if they have no parallel/comparable texts – i.e., we put each of these documents in a single-document tuple. To do this, we divide the corpus $\mathcal{W}$ into two sets of document tuples: a "glue" set $\mathcal{G}$ and a "separate" set $\mathcal{S}$ such that $|\mathcal{G}| \,/\, |\mathcal{W}| = p$. In other words, the proportion of tuples in the corpus that are treated as "glue" (i.e., placed in $\mathcal{G}$) is p. For every tuple in $\mathcal{S}$, we assign each document in that tuple to a new single-document tuple. By doing this, every document in $\mathcal{S}$ has its own distribution over topics, independent of any other documents. Ideally, the "glue" doc-

uments in $\mathcal{G}$ will be sufficient to align the topics across languages, and will cause comparable documents in $\mathcal{S}$ to have similar distributions over topics even though they are modeled independently.

Table 3: The effect of the proportion p of "glue" tuples on mean Jensen-Shannon divergence in estimated topic distributions for pairs of documents in $\mathcal{S}$ that were originally part of a document tuple. Lower divergence means the topic distributions distributions are more similar to each other.

p	Mean JS	# of pairs	Std. Err.
0.01	0.83755	487670	0.00018
0.05	0.79144	467288	0.00021
0.1	0.70228	443753	0.00026
0.25	0.38480	369608	0.00029
0.5	0.29712	246380	0.00030

Table 4: Topics are meaningful within languages but diverge between languages when only 1% of tuples are treated as "glue" tuples. With 25% "glue" tuples, topics are aligned.

lang	Topics at $p = 0.01$
DE	rußland russland russischen tschetschenien sicherheit
EN	china rights human country s burma
FR	russie tchétchénie union avec russe région
IT	ho presidente mi perché relazione votato

lang	Topics at $p = 0.25$
DE	rußland russland russischen tschetschenien ukraine
EN	russia russian chechnya cooperation region belarus
FR	russie tchétchénie avec russe russes situation
IT	russia unione cooperazione cecenia regione russa

We train PLTM with 100 topics on corpora with $p \in \{0.01, 0.05, 0.1, 0.25, 0.5\}$. We use 1000 iterations of Gibbs sampling with $\beta = 0.01$. Hyperparameters $\alpha\boldsymbol{m}$ are re-estimated every 10 iterations. We calculate the Jensen-Shannon divergence between the topic distributions for each pair of individual documents in $\mathcal{S}$ that were originally part of the same tuple prior to separation. The lower the divergence, the more similar the distributions are to each other. From the results in figure 4, we know that leaving all document tuples intact should result in a mean JS divergence of less than 0.1. Table 3 shows mean JS divergences for each value of p. As expected, JS divergence is greater than that obtained when all tuples are left intact. Divergence drops significantly when the proportion of "glue" tuples increases from 0.01 to 0.25. Example topics for $p = 0.01$ and $p = 0.25$ are shown in table 4. At $p = 0.01$ (1% "glue" documents), German and French both include words relating to Russia, while the English and Italian word distributions appear locally consistent but unrelated to Russia. At $p = 0.25$, the top words for all four languages are related to Russia.

These results demonstrate that PLTM is appropriate for aligning topics in corpora that have only a small subset of comparable documents. One area for future work is to explore whether initialization techniques or better representations of topic co-occurrence might result in alignment of topics with a smaller proportion of comparable texts.

4.5 Machine Translation

Although the PLTM is clearly not a substitute for a machine translation system—it has no way to represent syntax or even multi-word phrases—it is clear from the examples in figure 2 that the sets of high probability words in different languages for a given topic are likely to include translations. We therefore evaluate the ability of the PLTM to generate bilingual lexica, similar to other work in unsupervised translation modeling (Haghighi et al., 2008). In the early statistical translation model work at IBM, these representations were called "cepts," short for concepts (Brown et al., 1993).

We evaluate sets of high-probability words in each topic and multilingual "synsets" by comparing them to entries in human-constructed bilingual dictionaries, as done by Haghighi et al. (2008). Unlike previous work (Koehn and Knight, 2002), we evaluate all words, not just nouns. We collected bilingual lexica mapping English words to German, Greek, Spanish, French, Italian, Dutch and Swedish. Each lexicon is a set of pairs consisting of an English word and a translated word, $\{w_e, w_\ell\}$. We do not consider multi-word terms. We expect that simple analysis of topic assignments for sequential words would yield such collocations, but we leave this for future work.

For every topic t we select a small number K of the most probable words in English (e) and in each "translation" language (ℓ): $\mathcal{W}_{te}$ and $\mathcal{W}_{t\ell}$, respectively. We then add the Cartesian product of these sets for every topic to a set of candidate translations $\mathcal{C}$. We report the number of elements of $\mathcal{C}$ that appear in the reference lexica. Results for $K = 1$, that is, considering only the single most probable word for each language, are shown in figure 6. Precision at this level is relatively high, above 50% for Spanish, French and Italian with $T = 400$ and 800. Many of the candidate pairs that were not in the bilingual lexica were valid translations (e.g. EN "comitology" and IT

"comitalogia") that simply were not in the lexica. We also do not count morphological variants: the model finds EN "rules" and DE "vorschriften," but the lexicon contains only "rule" and "vorschrift." Results remain strong as we increase K. With $K = 3, T = 800$, 1349 of the 7200 candidate pairs for Spanish appeared in the lexicon.

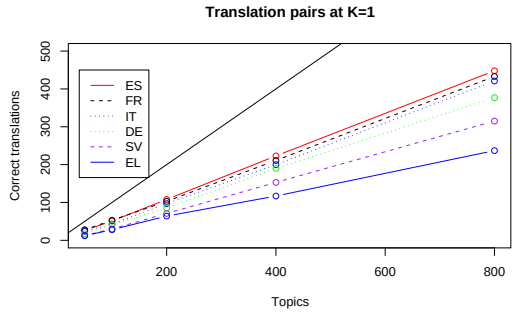

Figure 6: Are the single most probable words for a given topic in different languages translations of each other? The number of such pairs that appear in bilingual lexica is shown on the y-axis. For $T = 800$, the top English and Spanish words in 448 topics were exact translations of one another.

4.6 Finding Translations

In addition to enhancing lexicons by aligning topic-specific vocabulary, PLTM may also be useful for adapting machine translation systems to new domains by finding translations or near translations in an unstructured corpus. These aligned document pairs could then be fed into standard machine translation systems as training data. To evaluate this scenario, we train PLTM on a set of document tuples from EuroParl, infer topic distributions for a set of held-out documents, and then measure our ability to align documents in one language with their translations in another language.

It is not necessarily clear that PLTM will be effective at identifying translations. In finding a low-dimensional semantic representation, topic models deliberately smooth over much of the variation present in language. We are therefore interested in determining whether the information in the document-specific topic distributions is sufficient to identify semantically identical documents.

We begin by dividing the data into a training set of 69,550 document tuples and a test set of 17,435 document tuples. In order to make the task more difficult, we train a relatively coarse-grained PLTM with 50 topics on the training set. We then use this model to infer topic distributions for each

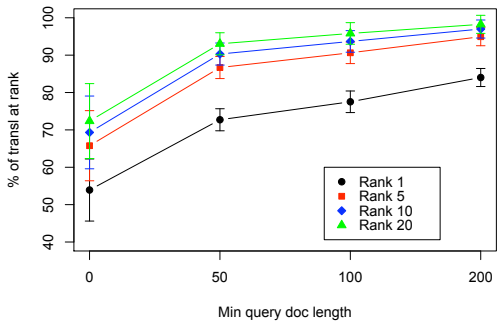

Figure 7: Percent of query language documents for which the target language translation is ranked at or above 1, 5, 10 or 20 by JS divergence, averaged over all language pairs.

of the 11 documents in each of the held-out document tuples using a method similar to that used to calculate held-out probabilities (Wallach et al., 2009). Finally, for each pair of languages ("query" and "target") we calculate the difference between the topic distribution for each held-out document in the query language and the topic distribution for each held-out document in the target language. We use both Jensen-Shannon divergence and cosine distance. For each document in the query language we rank all documents in the target language and record the rank of the actual translation.

Results averaged over all query/target language pairs are shown in figure 7 for Jensen-Shannon divergence. Cosine-based rankings are significantly worse. It is important to note that the length of documents matters. As noted before, many of the documents in the EuroParl collection consist of short, formulaic sentences. Restricting the query/target pairs to only those with query and target documents that are both longer than 50 words results in significant improvement and reduced variance: the average proportion of query documents for which the true translation is ranked highest goes from 53.9% to 72.7%. Performance continues to improve with longer documents, most likely due to better topic inference. Results vary by language. Table 5 shows results for all target languages with English as a query language. Again, English generally performs better with Romance languages than Germanic languages.

5 Results on Comparable Texts

Directly parallel translations are rare in many languages and can be extremely expensive to produce. However, the growth of the web, and in particular Wikipedia, has made *comparable* text cor-

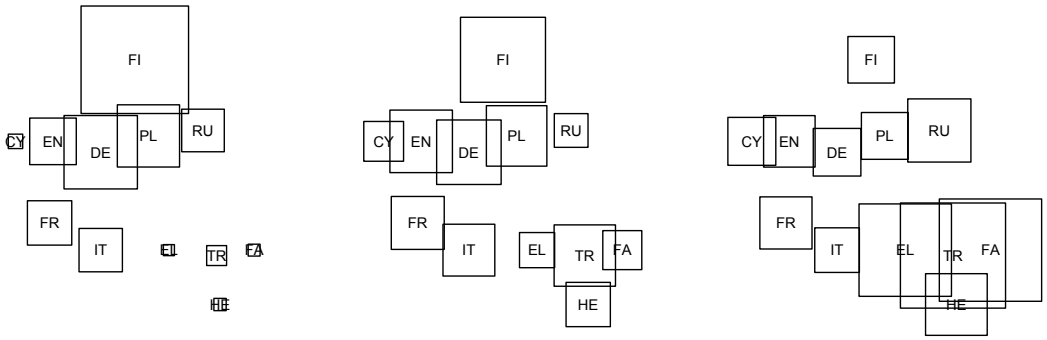

Figure 8: Squares represent the proportion of tokens in each language assigned to a topic. The left topic, *world ski km won*, centers around Nordic counties. The center topic, *actor role television actress*, is relatively uniform. The right topic, *ottoman empire khan byzantine*, is popular in all languages but especially in regions near Istanbul.

Table 5: Percent of English query documents for which the translation was in the top $n \in \{1, 5, 10, 20\}$ documents by JS divergence between topic distributions. To reduce the effect of short documents we consider only document pairs where the query and target documents are longer than 100 words.

Lang	1	5	10	20
DA	78.0	90.7	93.8	95.8
DE	76.6	90.0	93.4	95.5
EL	77.1	90.4	93.3	95.2
ES	81.2	92.3	94.8	96.7
FI	76.7	91.0	94.0	96.3
FR	80.1	91.7	94.3	96.2
IT	79.1	91.2	94.1	96.2
NL	76.6	90.1	93.4	95.5
PT	80.8	92.0	94.7	96.5
SV	80.4	92.1	94.9	96.5

pora – documents that are topically similar but are not direct translations of one another – considerably more abundant than true parallel corpora.

In this section, we explore two questions relating to comparable text corpora and polylingual topic modeling. First, we explore whether comparable document tuples support the alignment of fine-grained topics, as demonstrated earlier using parallel documents. This property is useful for building machine translation systems as well as for human readers who are either learning new languages or analyzing texts in languages they do not know. Second, because comparable texts may not use exactly the same topics, it becomes crucially important to be able to characterize differences in topic prevalence at the document level (do different languages have different perspectives on the same article?) and at the language-wide level (which topics do particular languages focus on?).

5.1 Data Set

We downloaded XML copies of all Wikipedia articles in twelve different languages: Welsh, German, Greek, English, Farsi, Finnish, French, Hebrew, Italian, Polish, Russian and Turkish. These versions of Wikipedia were selected to provide a diverse range of language families, geographic areas, and quantities of text. We preprocessed the data by removing tables, references, images and info-boxes. We dropped all articles in non-English languages that did not link to an English article. In the English version of Wikipedia we dropped all articles that were not linked to by any other language in our set. For efficiency, we truncated each article to the nearest word after 1000 characters and dropped the 50 most common word types in each language. Even with these restrictions, the size of the corpus is 148.5 million words.

We present results for a PLTM with 400 topics. 1000 Gibbs sampling iterations took roughly four days on one CPU with current hardware.

5.2 Which Languages Have High Topic Divergence?

As with EuroParl, we can calculate the Jensen-Shannon divergence between pairs of documents within a comparable document tuple. We can then average over all such document-document divergences for each pair of languages to get an overall "disagreement" score between languages. Interestingly, we find that almost all languages in our corpus, including several pairs that have historically been in conflict, show average JS divergences of between approximately 0.08 and 0.12 for $T = 400$, consistent with our findings for EuroParl translations. Subtle differences of sentiment may be below the granularity of the model.

CY sadwrn blaned gallair at lloeren mytholeg
DE space nasa sojus flug mission
EL διαστημικό sts nasa αγγλ small
EN **space mission launch satellite nasa spacecraft**
FA فضایی ماموریت ناسا مدار فضانورد ماهواره
FI sojuz nasa apollo ensimmäinen space lento
FR spatiale mission orbite mars satellite spatial
HE החלל הארץ חלל כדור א תוכנית
IT spaziale missione programma space sojuz stazione
PL misja kosmicznej stacji misji space nasa
RU космический союз космического спутник станции
TR uzay soyuz ay uzaya salyut sovyetler

CY sbaen madrid el la josé sbaeneg
DE de spanischer spanischen spanien madrid la
EL ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN **de spanish spain la madrid y**
FA اسپانیا اسپانیایی کوبا مادرید de ترین
FI espanja de espanjan madrid la real
FR espagnol espagne madrid espagnole juan y
HE ספרד ספרדית דה מדריד הספרדית קובה
IT de spagna spagnolo spagnola madrid el
PL de hiszpański hiszpanii la juan y
RU де мадрид испании испания испанский de
TR ispanya ispanyol madrid la küba real

CY bardd gerddi iaith beirdd fardd gymraeg
DE dichter schriftsteller literatur gedichte gedicht werk
EL ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN **poet poetry literature literary poems poem**
FA شاعر شعر ادبیات فارسی ادبی آثار
FI runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR poète écrivain littérature poésie littéraire ses
HE משורר ספרות שירה סופר שירים המשורר
IT poeta letteratura poesia opere versi poema
PL poeta literatury poezji pisarz in jego
RU поэт его писатель литературы поэзии драматург
TR şair edebiyat şiir yazar edebiyatı adlı

Figure 9: Wikipedia topics (T=400).

Overall, these scores indicate that although individual pages may show disagreement, Wikipedia is on average consistent between languages.

5.3 Are Topics Emphasized Differently Between Languages?

Although we find that if Wikipedia contains an article on a particular subject in some language, the article will tend to be topically similar to the articles about that subject in other languages, we also find that across the whole collection different languages emphasize topics to different extents. To demonstrate the wide variation in topics, we calculated the proportion of tokens in each language assigned to each topic. Figure 8 represents the estimated probabilities of topics given a specific language. Competitive cross-country skiing (left) accounts for a significant proportion of the text in Finnish, but barely exists in Welsh and the languages in the Southeastern region. Meanwhile, interest in actors and actresses (center) is consistent across all languages. Finally, historical topics, such as the Byzantine and Ottoman empires (right) are strong in all languages, but show geographical variation: interest centers around the empires.

6 Conclusions

We introduced a polylingual topic model (PLTM) that discovers topics aligned across multiple languages. We analyzed the characteristics of PLTM in comparison to monolingual LDA, and demonstrated that it is possible to discover aligned topics. We also demonstrated that relatively small numbers of topically comparable document tuples are sufficient to align topics between languages in non-comparable corpora. Additionally, PLTM can support the creation of bilingual lexica for low resource language pairs, providing candidate translations for more computationally intense alignment processes without the sentence-aligned translations typically used in such tasks. When applied to comparable document collections such as Wikipedia, PLTM supports data-driven analysis of differences and similarities across *all* languages for readers who understand *any one* language.

7 Acknowledgments

The authors thank Limin Yao, who was involved in early stages of this project. This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant number IIS-0326249, and in part by Army prime contract number W911NF-07-1-0216 and University of Pennsylvania subaward number 103-548106, and in part by National Science Foundation under NSF grant #CNS-0619337. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

David Blei and Michael Jordan. 2003. Modeling annotated data. In *SIGIR*.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *JMLR*.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *CL*, 19(2):263–311.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, pages 771–779.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *EMNLP*.

Mariko Kawaba, Hiroyuki Nakasaki, Takehito Utsuro, and Tomohiro Fukuhara. 2008. Cross-lingual blog analysis based on multilingual blog distillation from multilingual Wikipedia entries. In *ICWSM*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*.

Gideon Mann, David Mimno, and Andrew McCallum. 2006. Bibliometric impact measures leveraging topic analysis. In *JCDL*.

Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. 2005. Topic and role discovery in social networks. In *IJCAI*.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *WWW*.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. 2007. Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 28:187–207.

Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *ICML*.

Xing Wei and Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *SIGIR*.

Bing Zhao and Eric P. Xing. 2007. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *NIPS*.