Crosslingual Topic Modeling with WikiPDA

Tiziano Piccardi
EPFL
tiziano.piccardi@epfl.ch

Robert West*
EPFL
robert.west@epfl.ch

Abstract

We present *Wikipedia-based Polyglot Dirichlet Allocation (WikiPDA),* a crosslingual topic model that learns to represent Wikipedia articles written in any language as distributions over a common set of language-independent topics. It leverages the fact that Wikipedia articles link to each other and are mapped to concepts in the Wikidata knowledge base, such that, when represented as bags of links, articles are inherently language-independent. WikiPDA works in two steps, by first densifying bags of links using matrix completion and then training a standard monolingual topic model. A human evaluation shows that WikiPDA produces more coherent topics than monolingual text-based LDA, thus offering crosslinguality at no cost. We demonstrate WikiPDA's utility in two applications: a study of topical biases in 28 Wikipedia editions, and crosslingual supervised classification. Finally, we highlight WikiPDA's capacity for zero-shot language transfer, where a model is reused for new languages without any fine-tuning.

1 Introduction

With 53 million articles written in 299 languages, Wikipedia is the largest encyclopedia in history. To leverage and analyze individual language editions, researchers have successfully used topic models (Singer et al., 2017). The goal of this paper is to move beyond individual language editions and develop a topic model that works for all language editions jointly. Our method, *Wikipedia-based Polyglot Dirichlet Allocation (WikiPDA),* learns to represent articles written in any language in terms of language-independent, interpretable semantic topics. This way, articles that cannot be directly compared in terms of the words they contain (as the words are from different vocabularies) can nevertheless be compared in terms of their topics.

Such a model is tremendously useful in practice. With close to a billion daily page views, Wikipedia plays an important role in everyday life, and it is equally important as a dataset and object of study for researchers across domains: Google Scholar returns about 2 million publications for the query "Wikipedia". Although English is but one of 299 language editions, it is currently by far the most studied by researchers, to an extent that goes well beyond what can be justified by size alone. The scarcity of easy-to-use crosslingual topic models contributes to this skew, affecting even the rare studies that go beyond English; e.g., it kept Lemmerich et al. (2019), who compared the usage of 14 language editions via survey data and browsing logs, from quantifying differences in users' topical interest across languages.

Although each language on its own can be readily handled via standard topic models, which are based on bags of words and thus straightforward to apply to any language with minimal preprocessing, such models are insufficient for comparing content across languages because in general the topics learned for one language do not have clearly corresponding topics in the other languages.

Prior solutions. To address this problem, researchers have extended monolingual topic models by mapping documents from separate monolingual spaces into a joint crosslingual topic space. This paradigm has been proposed under various names (e.g., crosslingual, multilingual, polylingual, bilingual topic models; cf. Sec. 2), but the basic idea is identical, namely to enhance the model by allowing for multiple languages while enforcing crosslingual alignment at the level of words or documents. For instance, in document-alignment models, a topic is modeled not as a single word distribution, but as a

*Robert West is a Wikimedia Foundation Research Fellow.

set of word distributions—one per language—, and different language versions of the same document are constrained to the same mix of topics during training. As Wikipedia articles are aligned across languages via the Wikidata knowledge base, Wikipedia has served as a prominent training dataset for models based on document alignment.

Proposed solution: WikiPDA. We leverage Wikipedia's crosslingual article alignment from a different angle, by recognizing that Wikipedia articles are not just plain text, but laced with links to other articles. An article's set of outgoing links (*"bag of links"*) is a concise summary of the article's key content. Crucially, since each linked article is itself associated with a language-independent Wikidata concept, bags of links immediately give rise to a crosslingual input representation "for free". Starting from this representation, WikiPDA works in two steps, by first densifying bags of links using matrix completion and then training a standard monolingual topic model. Whereas in previous methods, translating from mono- to crosslingual space constitutes the core computation, in WikiPDA it constitutes a mere preprocessing step. Put differently, whereas prior work has harnessed Wikipedia's crosslinguality to *increase model complexity,* we leverage it to *decrease data complexity.* This way, WikiPDA can leverage, as its core computation, standard monolingual topic models such as LDA (Blei et al., 2003), which have been vetted in practice, come with implementations on all platforms, and scale to massive datasets.

A human evaluation shows that WikiPDA topics extracted jointly from 28 language editions of Wikipedia are more coherent than those from monolingual text-based LDA, thus offering crosslinguality at no cost (Sec. 4). We demonstrate WikiPDA's practical utility in two applications (Sec. 5): a topical comparison of Wikipedia across 28 languages, and crosslingual supervised classification. Finally, we show WikiPDA's ability to operate in the challenging zero-shot setting (Sec. 6), where a model is applied to new languages without any fine-tuning.

2 Related work

Topic models (Blei, 2012) are unsupervised machine learning techniques that represent documents as low-dimensional vectors whose dimensions are interpretable as topics. Crosslingual topic models (Vulić et al., 2015) allow for documents to be written in different languages and represent them in terms of topics that are language-independent. Most crosslingual topic models are based on latent Dirichlet allocation (LDA) (Blei et al., 2003).

One set of methods uses document-aligned corpora, where documents that are loosely equivalent but written in different languages (e.g., Wikipedia articles about the same concept in different languages) are grouped and constrained to the same topic distribution during training (De Smet and Moens, 2009; Mimno et al., 2009; Ni et al., 2009; Fukumasu et al., 2012; Zhang et al., 2013).

Another set of methods does not use aligned corpora, but word alignments from bilingual dictionaries, modeling topics as distributions over crosslingual equivalence classes of words (Jagarlamudi and Daumé, 2010; Zhang et al., 2010; Hao and Paul, 2018). Boyd-Graber and Blei (2009) require neither an aligned corpus nor a dictionary.

Document-alignment-based methods make effective use of large aligned corpora, but are hampered by requiring that aligned documents be about the same topics, which is frequently not the case in practice and eliminates an important use case of crosslingual topic models *ab ovo,* namely quantifying how an identical concept is described in different languages (cf. Sec. 5.1). Word-alignment-based methods, on the contrary, do not suffer from this shortcoming, but are hampered by the scarcity of multilingual dictionaries beyond two languages.

WikiPDA marries the best of both worlds by leveraging the *document alignment* provided by Wikidata in the spirit of *word alignment* methods: representing articles as bags of links, rather than bags of words, may be seen as inducing a common vocabulary spanning 299 languages, without unnaturally forcing corresponding articles in different languages to have identical topic distributions.

3 WikiPDA: Wikipedia-based Polyglot Dirichlet Allocation

WikiPDA operates in two stages, link densification (Sec. 3.1) and topic modeling (Sec. 3.2).

3.1 Link densification

Wikipedia's *Manual of Style*[1] asks authors to add links that aid navigation and understanding. Key concepts are thus linked to their articles, allowing us to use bags of links, in lieu of bags of words, as concise article summaries. Crucially, bag-of-links

[1] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking

Table 1: Statistics of the 28 Wikipedia language editions used in this paper.

		Articles (in thousands)			Links (in millions)†			Disambiguation evaluation† (Sec. 4.1)				
			Num. w/	% of			Dens.	Ambig.	Accuracy for number of candidates in interval			
	Language	Num.	≥10 links*	total†	Sparse	Densified	ratio‡	anchors	$[1,\infty]$	$[2,\infty]$ (rand.)	$[1,10]$	$[2,10]$ (rand.)
ar	Arabic	987	507	2%	14.6	49.1	3.4	48%	0.86	0.70 (0.24)	0.88	0.75 (0.34)
ca	Catalan	611	468	2%	16.2	71.4	4.4	46%	0.88	0.74 (0.23)	0.90	0.78 (0.33)
cs	Czech	410	357	1%	14.3	52.0	3.6	48%	0.86	0.71 (0.26)	0.88	0.74 (0.34)
de	German	2043	1851	7%	70.1	304.1	4.3	44%	0.86	0.68 (0.22)	0.88	0.73 (0.33)
el	Greek	164	117	<1%	4.1	17.1	4.1	46%	0.87	0.71 (0.26)	0.89	0.75 (0.33)
en	English	5571	4511	18%	206.9	594.3	2.9	39%	0.88	0.68 (0.18)	0.90	0.74 (0.33)
es	Spanish	1461	1332	5%	56.8	179.9	3.2	37%	0.86	0.63 (0.19)	0.89	0.70 (0.33)
fa	Persian	674	341	1%	8.7	34.8	4.0	49%	0.86	0.71 (0.22)	0.89	0.78 (0.33)
fi	Finnish	451	348	1%	10.4	31.8	3.1	54%	0.86	0.75 (0.25)	0.89	0.80 (0.35)
fr	French	2013	1684	7%	81.5	247.3	3.0	42%	0.85	0.64 (0.19)	0.89	0.73 (0.32)
he	Hebrew	239	229	1%	12.3	52.7	4.3	47%	0.87	0.72 (0.25)	0.89	0.76 (0.33)
id	Indonesian	495	345	1%	8.9	33.8	3.8	52%	0.85	0.71 (0.23)	0.87	0.76 (0.33)
it	Italian	1458	1093	4%	54.0	193.6	3.6	45%	0.84	0.64 (0.20)	0.88	0.73 (0.33)
ja	Japanese	1097	1030	4%	60.4	80.8	1.3	44%	0.84	0.64 (0.22)	0.87	0.71 (0.33)
ko	Korean	418	307	1%	12.3	28.8	2.3	42%	0.84	0.63 (0.25)	0.89	0.74 (0.35)
nl	Dutch	1889	958	4%	33.2	116.2	3.5	55%	0.84	0.71 (0.22)	0.87	0.76 (0.33)
pl	Polish	1289	986	4%	35.3	105.4	3.0	43%	0.88	0.72 (0.21)	0.90	0.76 (0.31)
pt	Portuguese	964	742	3%	28.0	102.2	3.6	40%	0.86	0.64 (0.22)	0.88	0.70 (0.33)
ro	Romanian	378	240	1%	7.4	29.2	3.9	46%	0.90	0.78 (0.24)	0.90	0.79 (0.32)
ru	Russian	1406	1143	4%	47.7	172.9	3.6	37%	0.87	0.66 (0.21)	0.90	0.72 (0.31)
sq	Albanian	71	19	<1%	0.7	2.6	3.8	53%	0.89	0.79 (0.33)	0.89	0.79 (0.36)
sr	Serbian	579	424	2%	9.7	46.0	4.7	50%	0.87	0.75 (0.27)	0.89	0.79 (0.33)
sv	Swedish	3453	3178	12%	59.3	118.8	2.0	60%	0.91	0.85 (0.26)	0.92	0.87 (0.36)
tr	Turkish	319	227	1%	7.0	20.8	3.0	48%	0.86	0.71 (0.24)	0.89	0.78 (0.34)
uk	Ukrainian	905	742	3%	23.0	80.6	3.5	45%	0.88	0.73 (0.25)	0.90	0.78 (0.33)
vi	Vietnamese	1218	543	2%	15.0	71.8	4.8	59%	0.83	0.72 (0.30)	0.86	0.75 (0.39)
war	Waray	1251	1142	4%	15.6	29.8	1.9	99%	0.46	0.46 (0.37)	0.46	0.46 (0.37)
zh	Chinese	1028	576	2%	23.4	31.8	1.4	54%	0.85	0.72 (0.25)	0.88	0.77 (0.34)
	Average	1173	908		33.4	103.5	3.3	48%	0.85	0.69 (0.24)	0.87	0.74 (0.33)
	Total	32844	25437	100%	936.8	2900.0						

*Links counted after link densification. †Considering only articles with ≥10 links after densification. ‡Densification ratio = Densified/Sparse.

elements—articles—are associated with language-independent Wikidata concepts, so in principle, the crosslingual article representations to be fed to the downstream topic model may be obtained simply by extracting links from articles.

In practice, however, human editors frequently fail to add all relevant links (West et al., 2009), and they are explicitly instructed to add links parsimoniously (e.g., by linking only the first mention of every concept). For topic modeling, such human-centric factors are of no concern; rather, we prefer semantically complete article summaries with information about the frequency of constituent concepts. Hence, the first phase of WikiPDA is link densification, where we link as many plain-text phrases as possible to the corresponding Wikidata concepts.

The difficulty arises from ambiguous phrases. Disambiguating phrases to the correct Wikidata concept is the so-called "wikification" task, with several existing solutions (Mihalcea and Csomai, 2007; Milne and Witten, 2008b; West et al., 2009; Noraset et al., 2014), any of which could be plugged in. Given the scale of our setting, we opted for a lightweight approach based on matrix completion: First, given a Wikipedia language edition, build the adjacency matrix A of the hyperlink graph, where both rows and columns represent Wikidata concepts, and entry a_{ij} is non-zero (details in Sec. 3.3) iff the article about concept i contains a link to that about concept j. Then, decompose $A \approx UV^\top$ using alternating least squares (Koren et al., 2009), such that both U and V are of low rank r. The rows of U are latent representations of articles when serving as link sources, and the rows of V, when serving as link targets, optimized such that, for existing links (i, j), we have $a_{ij} \approx u_i v_j^\top$ (where single subscripts are row indices). For non-existing links (i, j), the dot product $u_i v_j^\top$ provides a score that captures how well the new link (i, j) would be in line with the existing link structure.

Thus, the scores $u_i v_j^\top$ can be used to disambiguate the plain-text phrases p in article i: consider as the set C_p of candidate targets for p all articles j for which p occurs as an anchor at least once in the respective language edition of Wikipedia, and select the candidate with the largest score, i.e., link the phrase p in article i to $\arg\max_{j \in C_p} u_i v_j^\top$.

In principle, a decomposition computed for one language can be used to disambiguate links in any other language. For this paper, however, we computed a separate decomposition for each language.

3.2 Topic modeling

The bags of links resulting from link densification can be fed to any monolingual topic model based on bags of words, by using a vocabulary consisting of

Wikidata concepts rather words, and by using as the document corpus the union of all Wikipedia articles pooled across all languages considered. Concretely, we use LDA as the topic model. As usual, the number K of topics is set manually by the user.

3.3 Implementation and corpus details

Link densification. We considered as potential anchors for new links all 1- to 4-grams, with preference given to longer n-grams. We did not consider n-grams whose occurrences are linked with a probability below the threshold of 6.5% (Milne and Witten, 2008b), since, like stop words, they usually do not represent semantically relevant content.

Decompositions of the adjacency matrix A used rank $r = 150$. Before the decomposition, A's entries were weighted in the spirit of inverse document frequency, giving more weight to links occurring in few articles: if i links to j, we set $a_{ij} = -\log(d_j/N)$, and $a_{ij} = 0$ otherwise, where d_j is j's in-degree, and N the number of articles in the respective Wikipedia (Milne and Witten, 2008a).

Topic modeling. Since LDA may perform poorly with short documents (Tang et al., 2014), we removed articles with fewer than 10 links after densification. Further, we ignored concepts appearing in fewer than 500 articles across all languages.

Corpus. We worked with 28 language editions of Wikipedia (details in Table 1), in their snapshots of 20 February 2020. Links and anchor texts were extracted from wiki markup. Redirects were resolved. After all preprocessing, the corpus encompassed 25m documents across all 28 languages, with a vocabulary of 437k unique Wikidata concepts.

Code and model availability. We release code and pre-trained models for a range of K between 20 and 200. For $K = 40$ and 100, topics were manually labeled with names. On a single machine (48 cores, 250 GB RAM), the full pipeline for all 28 languages with fixed hyperparameters ran in under 24 hours. As the code uses Apache Spark, parallelizing over many machines is straightforward.

4 Evaluation

4.1 Link densification

Densification increased the number of links substantially, by a factor of 3.3, to an effective 114 links per article (Table 1; means over languages).

The large fraction of ambiguous anchors (48%) underlines the importance of disambiguation. To evaluate disambiguation accuracy, we masked 5% of the entries of the adjacency matrix A before decomposing it (cf. Sec. 3.1). Each masked link is associated with a potentially ambiguous anchor text p. Given p, we generated all candidate targets j and ranked them by their score $u_i v_j^\top$. Disambiguation accuracy is then defined as the fraction of masked matrix entries for which the top-ranked candidate was correct. It is summarized, for all 28 languages, in the 4 rightmost columns of Table 1, where column "$[l,u]$" contains the accuracy for anchors with at least l and at most u candidates.

The column "$[1,\infty]$" shows the overall accuracy for all anchors (85% on average over all 28 languages). Since this column includes trivial, unambiguous anchors, the column "$[2,\infty]$" is more interesting. Although lower, these numbers are still satisfactorily high (69% on average), particularly when compared to the random baseline (24%).

Manual error analysis revealed that anchors with a large number of candidates tend to be inherently hard to disambiguate even for humans (e.g., "self-titled album" has 712 candidates). Hence, preferring precision over recall, we ignore phrases with more than 10 candidates, obtaining an average accuracy of 87% for the remaining anchors ("$[1,10]$").

The quality of disambiguated links is confirmed by the superior performance of densified, compared to raw, bags of links, as discussed next.

4.2 Topic modeling

We evaluated 4 model classes, trained on 4 corpora:

1. **WikiPDA, dense links, 28 languages:** full model as described in Sec. 3.
2. **WikiPDA, sparse links, 28 languages:** the same, but without link densification.
3. **WikiPDA, dense links, English:** trained on English only, rather than on all 28 languages.
4. **Text-based LDA, English:** bag-of-words LDA trained on English text (not links).

For each model class, we trained and evaluated models for 10 values of K, ranging from 20 to 200. In the following, "model" refers to an instance of a model class trained for a specific K.

Methodology: intruder detection. The evaluation of topic models is challenging. Traditionally, it has been based on automatic scores such as perplexity, capturing how "surprising" documents from a held-out corpus are, given the training set. Unfortunately, perplexity does not necessarily correlate

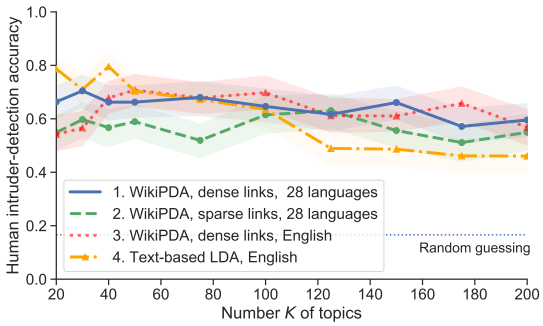


Figure 1: Evaluation of topic models. Topic coherence measured in terms of human intruder-detection accuracy (higher is better), with 95% confidence intervals.

with human judgment, and in some cases an inverse relation has even been reported (Chang et al., 2009). Since we are interested in interpretable models, we assessed the quality of topics in a human, rather than automatic, evaluation, using the *word intruder* framework proposed by Chang et al. (2009). Given a model to evaluate, we randomly selected $n = 20$ of the $K \geq 20$ topics and extracted the top 5 Wikidata concepts per topic. Then we selected an *intruder concept* for each topic: a concept that ranked low for that topic, but high for at least one other topic (in particular, the concept with the largest rank difference was selected). A shuffled list of the 6 concepts (described by their English names) was shown to a human evaluator, who was asked to spot the intruder. The more coherent a topic, the easier it is to spot the intruder, so human accuracy serves as a measure of topical coherence.

Crowdsourcing setup. For each model, human accuracy was estimated based on $12n = 240$ workers' guesses obtained from 12 independent rounds of the above procedure on Amazon Mechanical Turk. Workers solved 16 intruder-detection tasks per batch. To not reveal a pattern, we used each model and each intruder at most once per batch.

Results. Fig. 1 shows that, with the full WikiPDA model (model class 1), human intruder-detection accuracy was 60–70%, depending on K, far above random guessing (16.7%).

Comparing model classes 1 and 2, we see that the dense WikiPDA model yielded results consistently above the sparse model (by up to 15 percentage points), showing the utility of link densification.

Comparing model classes 1 and 3, we find that the dense WikiPDA model for 28 languages performed indistinguishably from the dense model for English only; i.e., adding more languages did not make the topics less coherent. This outcome is noteworthy, since on other crosslingual tasks (e.g., document retrieval), performance on a fixed testing language decreased when adding languages to the training set (Josifoski et al., 2019).

Comparing model classes 3 and 4 (both English only) shows that, whereas the performance of text-based LDA degrades with growing K, WikiPDA is more stable. While text-based LDA is slightly better for small $K \leq 50$, WikiPDA prevails for $K \geq 75$. This suggests that the link-based models are more customizable to use cases where the problem requires a specific K.

Note that the text-based LDA model is not language-independent and thus not truly a competitor with crosslinguality in mind. Rather, it should *a priori* be considered a strong ceiling: text-based LDA is the de-facto standard for analyzing the content of Wikipedia articles in monolingual settings (Singer et al., 2017; Lemmerich et al., 2019). Thus, by surpassing the topical coherence of text-based models, WikiPDA offers crosslinguality "for free".

5 Applications

WikiPDA enables a wide range of applications, some of which we spotlight next.

5.1 Comparing Wikipedia across languages

Understanding the differences in content coverage across language editions constitutes a major topic for Wikipedia researchers (Bao et al., 2012).

Topical bias. Using WikiPDA, we studied the topical bias of 28 language editions (cf. Table 1) via regression analysis. For each language L, we randomly sampled 20k articles as positive examples and $20\text{k}/27 = 740$ from each of the 27 other languages as negative examples (80/20 train/test split), and trained a one-vs.-all logistic regression classifier to predict whether an article is from language L, given the article's topic distribution. On average the 28 classifiers achieved an area under the ROC curve (AUC) of 78% for $K = 40$, or 84% for $K = 200$, significantly above the random baseline of 50%, indicating major differences across language editions. Inspecting the fitted coefficients ($K = 40$), depicted in Fig. 2, revealed the specificities of individual languages. First and foremost, country- or region-specific topics appeared among the most discriminative topics. Additionally, several more surprising associations emerged: e.g., COMICS is the topic most indicative of English and Dutch,

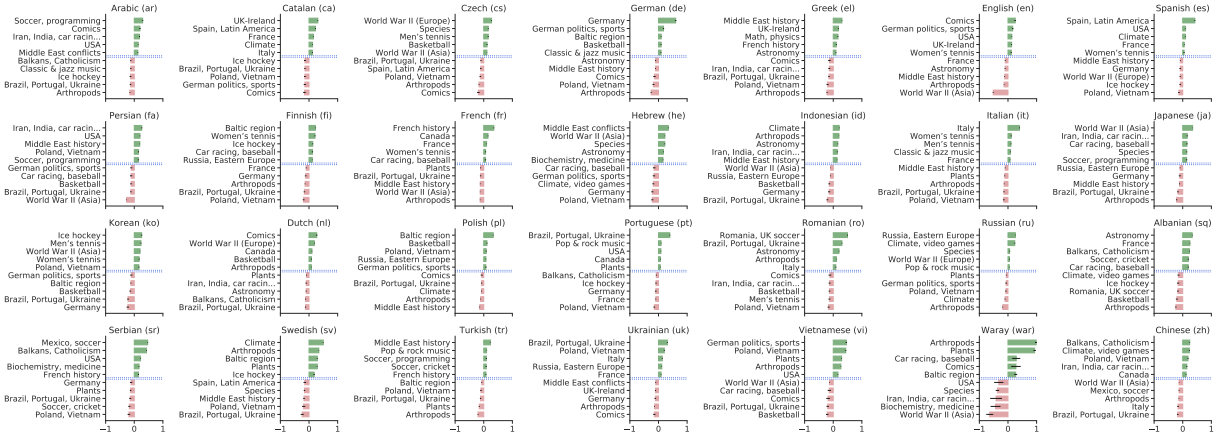
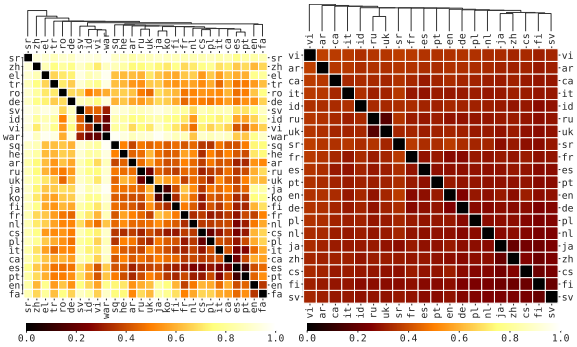


Figure 2: Topical bias of 28 Wikipedia language editions. For each language L, a logistic regression was trained to predict if an article belongs to L, using the article's distribution over WikiPDA topics (labeled manually with names) as predictors. Most predictive positive and negative coefficients are shown, with 95% confidence intervals.



(a) All articles (b) Common articles

Figure 3: Cosine distance between Wikipedia language editions. **(a)** 28 languages, each represented via average topic vector of all articles. **(b)** 20 top languages, considering only the 16k articles included in all 20.

and it is most counter-indicative of Ukrainian and Catalan; GEOPOLITICS is prominently featured in Hebrew; ICE HOCKEY and TENNIS, in Korean; etc.

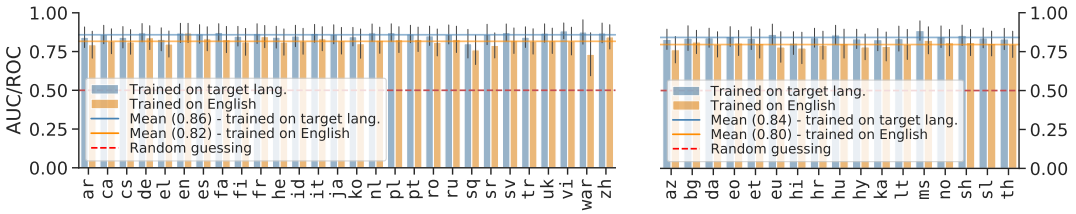
Distance between language editions. Next, we computed pairwise distances for all language editions via the cosine distance between the languages' mean topic vectors ($K = 200$). Fig. 3a shows the distance matrix. Clear topical similarities (darker colors) emerge for languages of countries with historical or geographical ties, including Japanese/Korean, Russian/Ukrainian, Czech/Polish, and Portuguese/Spanish. Waray (spoken in the Philippines) clusters with Indonesian, Vietnamese, and—more surprisingly—Swedish, a language that, linguistically speaking, could not be more distant. Investigating the reasons, we found that Swedish and Waray are among the three Wikipedias (the third being Cebuano) in which Lsjbot was active, a bot that created 80–99% of the articles in those languages. Fig. 2 suggests that Lsjbot created particularly many biology-related articles, a finding not even mentioned on the Wikipedia page about Lsjbot itself.[2] Also, it seems that the bot, which was created by a Swede, gave Waray Wikipedia a Swedish bias, as indicated by Waray's large coefficient for the topic BALTIC REGION in Fig. 2. These nuggets exemplify how WikiPDA enables the cross-cultural study of Wikipedia.

The above-noted differences may be due to different language editions containing articles about different concepts. An equally interesting question asks to what extent the languages differ in how they discuss identical concepts. To quantify this, we found the 16k articles in the intersection of the 20 largest language editions and computed, for each language pair and each common article, the cosine distance of the two languages' topic vectors for the article. Averaging the 16k distances yields Fig. 3b, which paints a more uniform picture than Fig. 3a, with no important differences remaining between languages. Note, however, that Russian/Ukrainian, Finnish/Swedish, and Chinese/Japanese cover the same concepts in particularly similar ways.

5.2 Supervised topic classification

WikiPDA is an unsupervised technique. It discovers whatever topics are best suited for summarizing the data. Sometimes, however, researchers may want to exert more control by fixing a set of topics ahead of time and classifying documents into those in a supervised manner. For instance, with

[2] https://en.wikipedia.org/w/index.php?title=Lsjbot&oldid=949492392



(a) 28 languages seen during WikiPDA training

(b) 17 zero-shot languages

Figure 4: Performance on supervised topic classification, using unsupervised WikiPDA topics as features. For each language L, two models were evaluated: trained on L (blue); trained on English (orange). Error bars: standard deviation over 64 binary classification tasks (one per supervised topic label). Similarity of blue and orange shows that classifier works on languages not seen during classifier training. Similarity between (a) and (b) shows that classifier and WikiPDA models work on languages not seen during WikiPDA training.

the ORES library,[3] Wikimedia provides a supervised classifier for categorizing English articles into a manually constructed taxonomy of 64 topics, based on features derived from the articles' English text (Asthana and Halfaker, 2018). We explore if WikiPDA topic vectors can be used as features instead, giving rise to a language-independent model, whereas the ORES model is language-specific.

Setup. For training and testing (80/20 split), we used a dataset of 5.1m English articles manually labeled with 64 binary labels,[4] specifying for each of the 64 topical classes defined by ORES whether the article belongs to the class. Each article can belong to multiple classes, so we trained an independent binary logistic regression classifier per class, on a balanced training set where negative examples were sampled evenly from the 63 other classes. Performance was found to increase with K, so we used $K = 200$. For each language L, we performed two evaluations: first, with a model trained on articles from L (after transferring labels from the English dataset via the alignment given by Wikidata) and second, with a model trained on English.

Results. In Fig. 4a, we show two AUC values (macro-averages over the 64 classes) for each language L: one when testing the classifier trained on L itself (blue); the other, when testing the classifier trained on English (orange). Performance is high across all languages, with an average AUC of 86% for the language-specific classifiers. The single, fixed classifier trained on English performed only slightly worse when evaluated on the other languages, with an average AUC of 82%.

Note that the primary goal of these experiments was not to achieve maximum classification performance by all means. Indeed, exploratory results showed that simply switching from logistic regression to gradient-boosted trees immediately boosted the AUC by 2–3 percentage points. Rather, the main take-aways are twofold: (1) WikiPDA's unsupervised topics can be readily translated to a different set of manually defined topics, which demonstrates their utility as a general low-dimensional representation that captures the topical essence of a document. (2) Due to the crosslingual nature of WikiPDA topics, a supervised model trained on one language (here: English) can be transferred to any other language not seen during supervised training, achieving high performance without any fine-tuning.

In our final set of experiments, described in the next section, we push the language-transfer paradigm even further, by moving to a setting where the target language was absent not only during training of the supervised classifier, but also during unsupervised training of the WikiPDA topics that the supervised classifier uses as features.

6 Zero-shot language transfer

The bags of links by which WikiPDA represents input documents are composed of language-independent Wikidata concepts. No matter in what language an article is written, its bag of links can be immediately compared to the bags of links extracted from any other language. This way, a WikiPDA model trained on a certain set of languages can be used to infer topics for articles from any new language not seen during WikiPDA training. In other words, WikiPDA inherently enables *zero-shot language transfer.* This capability is particularly convenient for low-resource languages, where the available data might not suffice to learn meaningful topics, and it sets WikiPDA apart from

[3] https://ores.wikimedia.org/

[4] Code: https://github.com/wikimedia/drafttopic

all the previously proposed crosslingual topics models discussed in Sec. 2, which need to be retrained whenever a new language is added.

To showcase WikiPDA's zero-shot capability, we used the model trained on the 28 languages of Table 1 to infer topics for all articles in 17 more languages (cf. Fig. 4b) and repeated the supervised topic classification experiments (Sec. 5.2) for these languages. As in Sec. 5.2, we evaluated two supervised topic classifiers for each language: one trained on the respective language, the other trained on English. Note that in neither case were the 17 new languages included in topic model training; rather, the topic vectors that served as input to the supervised classifier were inferred based on a WikiPDA model trained only on the 28 old languages. Despite this, the mean AUC on the 17 new languages (Fig. 4b) nearly reached that on the 28 old ones, both for the language-specific classifiers (84% for the new languages, vs. 86% for the old ones) and for the English classifier (80% vs. 82%).

Finally, to further validate the applicability of WikiPDA topics in the zero-shot setting, we repeated the analysis of Fig. 2, fitting logistic regression models to predict the language of an article given its topic vector. Classification performance was as high on the 17 new languages as on the 28 languages seen during topic model training (mean AUC 79% for $K = 40$; 84% for $K = 200$), indicating that the topic vectors capture the peculiarities of the 17 new languages well, even though the languages were not seen during topic model training.

7 Discussion and conclusion

We presented Wikipedia-based Polyglot Dirichlet Allocation (WikiPDA), a novel crosslingual topic model for Wikipedia. Our human evaluation showed that the topics learned from 28 languages are as coherent as those learned from English alone, and more coherent than those from text-based LDA on English, a noteworthy finding, given that other crosslingual tasks have suffered by adding languages to the training set (Josifoski et al., 2019). We demonstrated WikiPDA's practical utility in several example applications and highlighted its capability for zero-shot language transfer.

The key insight underpinning WikiPDA is that, when represented as bags of links, Wikipedia articles are crosslingual from the get-go, leading to two big advantages, interpretability and scalability:

Interpretability. As WikiPDA's vocabulary consists of Wikidata concepts, which have names in all languages, bags of links and learned topics (distributions over the vocabulary), can be interpreted even without understanding the corpus languages.

Scalability. In bag-of-links space, the corpus can be treated as monolingual, such that standard topic models apply, for which highly efficient algorithms exist; e.g., online algorithms for LDA can handle massive amounts of text (Hoffman et al., 2010) and have been implemented for high-performance machine learning libraries (e.g., Vowpal Wabbit) and massively parallel big data platforms (e.g., Apache Spark). Scaling WikiPDA to all of Wikipedia's 299 languages is thus fully within reach, whereas previous methods have usually been deployed on small numbers of languages only. That said, given WikiPDA's zero-shot capability (Sec. 6), training on all 299 languages may not even be necessary, since a model trained on a few languages can be immediately applied to unseen languages "for free".

Limitations. Finally, we discuss two potential concerns: language imbalance and link sparsity.

First, Wikipedia's language editions vary considerably in size, so the learned topics are dominated by larger languages. Whether this is desirable or not depends on the specific use case. Future work should explore the effects of upweighting smaller languages, e.g., by downsampling large languages, upsampling small languages, or incorporating weights into LDA's objective function.

Second, compared to text-based models, our link-based model works with sparser inputs, even after link densification. While advantageous computationally, this raises the question if the method can handle very "short" documents, i.e., articles with very few outgoing links. To understand this aspect, we trained and tested the supervised topic classification model (Sec. 5.2) for English again, this time only on articles with fewer than 10 links (19% of all articles). The model still performed well (AUC 85%), only 3 percentage points lower than when using all articles, including those with many links, indicating that WikiPDA is not hampered in important ways by articles with few links.

Conclusion. WikiPDA offers researchers a practical tool for studying the content of all of Wikipedia's 299 language editions in a unified framework, thus better reflecting Wikipedia's real diversity. We look forward to seeing it deployed in practice.

Acknowledgments. We thank Ahmad Ajalloeian and Blagoj Mitrevski for their valuable early contributions to the project, and we gratefully acknowledge funding from the Swiss National Science Foundation (grant 200021_185043) and from Facebook, Google, and Microsoft.

References

Sumit Asthana and Aaron Halfaker. 2018. With few eyes, all hoaxes are deep. In *Proc. ACM on Human-Computer Interaction*.

Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: bridging the wikipedia language gap. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

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

Jordan Boyd-Graber and David M Blei. 2009. Multilingual topic models for unaligned text. In *Proc. Conference on Uncertainty in Artificial Intelligence*.

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 22*, pages 288–296.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the Web using interlingual topic modelling. In *Proc. ACM Workshop on Social Web Search and Mining*.

Kosuke Fukumasu, Koji Eguchi, and Eric P Xing. 2012. Symmetric correspondence topic models for multilingual text analysis. In *Proc. Advances in Neural Information Processing Systems*.

Shudong Hao and Michael J Paul. 2018. Learning multilingual topics from incomparable corpus. *arXiv preprint arXiv:1806.04270*.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent Dirichlet allocation. In *Proc. Advances in Neural Information Processing Systems*.

Jagadeesh Jagarlamudi and Hal Daumé. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proc. European Conference on Information Retrieval*.

Martin Josifoski, Ivan S Paskov, Hristo S Paskov, Martin Jaggi, and Robert West. 2019. Crosslingual document embedding as reduced-rank ridge regression. In *Proc. ACM International Conference on Web Search and Data Mining*.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the world reads wikipedia: Beyond english speakers. In *Proc. ACM International Conference on Web Search and Data Mining*.

Rada Mihalcea and Andras Csomai. 2007. Wikify! Linking documents to encyclopedic knowledge. In *Proc. ACM Conference on Information and Knowledge Management*.

David Milne and Ian H. Witten. 2008a. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proc. AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*.

David Milne and Ian H. Witten. 2008b. Learning to link with wikipedia. In *Proc. ACM Conference on Information and Knowledge Management*.

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

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *Proc. International Conference on World Wide Web*.

Thanapon Noraset, Chandra Bhagavatula, and Doug Downey. 2014. Adding high-precision links to Wikipedia. In *Proc. Conference on Empirical Methods in Natural Language Processing*.

Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read Wikipedia. In *Proc. International Conference on the World Wide Web*.

Jian Tang, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proc. International Conference on Machine Learning*.

Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, 51(1):111–147.

Robert West, Doina Precup, and Joelle Pineau. 2009. Completing wikipedia's hyperlink structure through dimensionality reduction. In *Proc. ACM Conference on Information and Knowledge Management*.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proc. Annual Meeting of the Association for Computational Linguistics*.

Tao Zhang, Kang Liu, and Jun Zhao. 2013. Cross lingual entity linking with bilingual topic model. In *Proc. International Joint Conference on Artificial Intelligence*.