Shudong Hao, Michael J. Paul, and **Jordan Boyd-Graber**. **Lessons from the Bible on Modern Topics: Multilingual Topic Model Evaluation on Low-Resource Languages**. *North American Association for Computational Linguistics*, 2018, 9 pages.


```
    @inproceedings{Hao:Paul:Boyd-Graber-2018,
Title = {Lessons from the Bible on Modern Topics: Multilingual Topic Model Evaluation on Low-Resource Lang
Author = {Shudong Hao and Michael J. Paul and Jordan Boyd-Graber},
Booktitle = {North American Association for Computational Linguistics},
Year = {2018},
Location = {New Orleans, LA},
Url = {http://umiacs.umd.edu/~jbg//docs/2018_naacl_mltm_eval.pdf},
}
```


Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Lessons from the Bible on Modern Topics: Low-Resource Multilingual Topic Model Evaluation

Shudong Hao
Computer Science
University of Colorado
Boulder, CO
shudong@colorado.edu

Jordan Boyd-Graber
Computer Science, iSchool, LSC, and UMIACS,
University of Maryland
College Park, MD
jbg@umiacs.umd.edu

Michael J. Paul
Information Science
University of Colorado
Boulder, CO
mpaul@colorado.edu

Abstract

Multilingual topic models enable document analysis across languages through coherent multilingual summaries of the data. However, there is no standard and effective metric to evaluate the quality of multilingual topics. We introduce a new intrinsic evaluation of multilingual topic models that correlates well with human judgments of multilingual topic coherence as well as performance in downstream applications. Importantly, we also study evaluation for low-resource languages. Because standard metrics fail to accurately measure topic quality when robust external resources are unavailable, we propose an adaptation model that improves the accuracy and reliability of these metrics in low-resource settings.

1 Introduction

Topic models provide a high-level view of the main themes of a document collection (Boyd-Graber et al., 2017). Document collections, however, are often not in a single language, driving the development of **multilingual** topic models. These models discover topics that are consistent across languages, providing useful tools for multilingual text analysis (Vulić et al., 2015), such as detecting cultural differences (Gutiérrez et al., 2016) and bilingual dictionary extraction (Liu et al., 2015).

Monolingual topic models can be evaluated through likelihood (Wallach et al., 2009b) or coherence (Newman et al., 2010), but topic model evaluation is not well understood in multilingual settings. Our contributions are two-fold. We introduce an improved intrinsic evaluation metric for multilingual topic models, called Crosslingual Normalized Pointwise Mutual Information (CNPMI, Section 2). We explore the behaviors of CNPMI at both the model and topic levels with six language pairs and varying model specifications. This metric correlates well with human judgments and crosslingual classification results (Sections 5 and 6).

We also focus on evaluation in low-resource languages, which lack large parallel corpora, dictionaries, and other tools that are often used in learning and evaluating topic models. To adapt CNPMI to these settings, we create a coherence estimator (Section 3) that extrapolates statistics derived from antiquated, specialized texts like the Bible: often the only resource available for many languages.

2 Evaluating Multilingual Coherence

A multilingual topic contains one topic for each language. For a multilingual topic to be meaningful to humans (Figure 1), the meanings should be consistent across the languages, in addition to coherent within each language (*i.e.*, all words in a topic are related).

This section describes our approach to evaluating the quality of multilingual topics. After defining the multilingual topic model, we describe topic model evaluation extending standard monolingual approaches to multilingual settings.

2.1 Multilingual Topic Modeling

Probabilistic topic models associate each document in a corpus with a distribution over latent topics, while each topic is associated with a distribution over words in the vocabulary. The most widely used topic model, latent Dirichlet allocation (Blei et al., 2003, LDA), can be extended to connect languages. These extensions require additional knowledge to link languages together.

One common encoding of multilingual knowledge is **document links** (indicators that documents are parallel or comparable), used in polylingual topic models (Mimno et al., 2009; Ni et al., 2009). In these models, each document d indexes a tuple of parallel/comparable language-specific documents,

$d^{(\ell)}$, and the language-specific "views" of a document share the document-topic distribution θ_d. The generative story for the document-links model is:

for *each topic* k *and each language* ℓ **do**
  Draw a distribution over words $\phi_{\ell k} \sim \text{Dirichlet}(\beta)$;
for *each document tuple* $d = \left(d^{(1)}, \ldots, d^{(L)}\right)$ **do**
  Draw a distribution over topics $\theta_d \sim \text{Dirichlet}(\alpha)$;
  for *each language* $\ell = 1, \ldots, L$ **do**
    for *each token* $t \in d^{(\ell)}$ **do**
      Draw a topic $z_n \sim \theta_d$;
      Draw a word $w_n \sim \phi_{\ell z}$;

Alternatively, word translations (Jagarlamudi and Daumé III, 2010), concept links (Gutiérrez et al., 2016; Yang et al., 2017), and multi-level priors (Krstovski et al., 2016) can also provide multilingual knowledges. Since the polylingual topic model is the most common approach for building multilingual topic models (Vulić et al., 2013, 2015; Liu et al., 2015; Krstovski and Smith, 2016), our study will focus on this model.

2.2 Monolingual Evaluation

Most automatic topic model evaluation metrics use co-occurrence statistics of word pairs from a reference corpus to evaluate topic coherence, assuming that coherent topics contain words that often appear together (Newman et al., 2010). The most successful (Lau et al., 2014) is normalized pointwise mutual information (Bouma, 2009, NPMI). NPMI compares the joint probability of words appearing together $\Pr(w_i, w_j)$ to their probability assuming independence $\Pr(w_i)\Pr(w_j)$, normalized by the joint probability:

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{\Pr(w_i, w_j)}{\Pr(w_i)\Pr(w_j)}}{\log \Pr(w_i, w_j)}. \quad (1)$$

The word probabilities are calculated from a **reference corpus**, $\mathcal{R}$, typically a large corpus such as Wikipedia that can provide meaningful co-occurrence patterns that are independent of the target dataset.

The quality of topic k is the average NPMI of all word pairs (w_i, w_j) in the topic:

$$\text{NPMI}_k = \frac{-1}{\binom{C}{2}} \sum_{i \in \mathcal{W}(k,C)} \sum_{j \neq i} \text{NPMI}(w_i, w_j), \quad (2)$$

where $\mathcal{W}(k, C)$ are the C most probable words in the topic-word distribution ϕ_k (the number of words is the topic's **cardinality**). Higher NPMI_k means the topic's top words are more coupled.

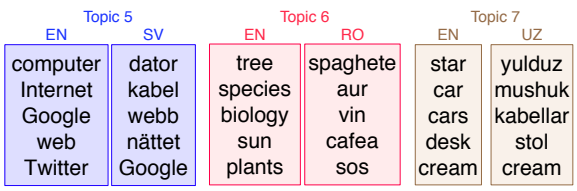

Figure 1: Topic 5 is multilingually coherent: both the English and Swedish topics are about technology. Topic 6 is about biology in English but food in Romanian, so it is low quality although coherent monolingually. Topic 7 is monolingually incoherent, so it is a low quality topic even if it contains word translations.

2.3 Existing Multilingual Evaluations

While automatic evaluation has been well-studied for monolingual topic models, there are no robust evaluations for multilingual topic models. We first consider two straightforward metrics that could be used for multilingual evaluation, both with limitations. We then propose an extension of NPMI that addresses these limitations.

Internal Coherence. A simple adaptation of NPMI is to calculate the monolingual NPMI score for each language independently and take the average. We refer this as internal NPMI (INPMI) as it evaluates coherence *within* a language. However, this metric does not consider whether the topic is coherent *across* languages—that is, whether a language-specific word distribution $\phi_{\ell_1 k}$ is related to the corresponding distribution in another language, $\phi_{\ell_2 k}$.

Crosslingual Consistency. Another straightforward measurement is Matching Translation Accuracy (Boyd-Graber and Blei, 2009, MTA), which counts the number of word translations in a topic between two languages using a bilingual dictionary. This metric can measure whether a topic is well-aligned across languages *literally*, but cannot capture non-literal more holistic similarities across languages.

2.4 New Metric: Crosslingual NPMI

We extend NPMI to multilingual models, with a metric we call crosslingual normalized pointwise mutual information (CNPMI). This metric will be the focus of our experiments.

A multilingually coherent topic means that if w_{i,ℓ_1} in language ℓ_1 and w_{j,ℓ_2} in language ℓ_2 are in the same topic, they should appear in similar contexts in comparable or parallel corpora $\mathcal{R}^{(\ell_1,\ell_2)}$.

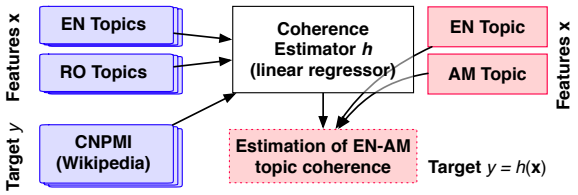

Figure 2: The coherence estimator takes multilingual topics and features from them then outputs an estimated topic coherence.

Our adaptation of NPMI is based on the same principles as the monolingual version, but focuses on the co-occurrences of *bilingual* word pairs. Given a bilingual word pair $(w_{i,\ell_1}, w_{j,\ell_2})$ the co-occurrence of this word pair is the event where word w_{i,ℓ_1} appears in a document in language ℓ_1 and the word w_{j,ℓ_2} appears in a comparable or parallel document in language ℓ_2.

The co-occurrence probability of each bilingual word pair is:

$$\Pr\left(w_{i,\ell_1}, w_{j,\ell_2}\right) \triangleq \frac{\left|\left\{\mathbf{d} : w_{i,\ell_1} \in d^{(\ell_1)}, w_{j,\ell_2} \in d^{(\ell_2)}\right\}\right|}{\left|\mathcal{R}^{(\ell_1,\ell_2)}\right|}, \quad (3)$$

where $\mathbf{d} = \left(d^{(\ell_1)}, d^{(\ell_2)}\right)$ is a pair of parallel/comparable documents in the reference corpus $\mathcal{R}^{(\ell_1,\ell_2)}$. When one or both words in a bilingual pair do not appear in the reference corpus, the co-occurrence score is zero.

Similar to monolingual settings, CNPMI for a bilingual topic k is the average of the NPMI scores of all C^2 bilingual word pairs,

$$\text{CNPMI}(\ell_1, \ell_2, k) = \frac{\sum_{i,j}^{C} \text{NPMI}\left(w_{i,\ell_1}, w_{j,\ell_2}\right)}{C^2}. \quad (4)$$

It is straightforward to generalize CNPMI from a language pair to multiple languages by averaging $\text{CNPMI}(\ell_i, \ell_j, k)$ over all language pairs (ℓ_i, ℓ_j).

3 Adapting to Low-Resource Languages

CNPMI needs a reference corpus for co-occurrence statistics. Wikipedia, which has good coverage of topics and vocabularies is a common choice (Lau and Baldwin, 2016). Unfortunately, Wikipedia is often unavailable or not large enough for low-resource languages. It only covers 282 languages,[1] and only 249 languages have more than 1,000 pages: many of pages are short or unlinked to a high-resource language. Since CNPMI requires comparable documents, the usable reference corpus is defined by *paired* documents.

Another option for a parallel reference corpus is the Bible (Resnik et al., 1999), which is available in most world languages;[2] however, it is small and archaic. It is good at evaluating topics such as family and religion, but not "modern" topics like biology and Internet. Without reference co-occurrence statistics relevant to these topics, CNPMI will fail to judge topic coherence—it must give the ambiguous answer of zero. Such a score could mean a totally incoherent topic where each word pair never appears together (Topics 6 in Figure 1), or an unjudgeable topic (Topic 5).

Our goal is to obtain a reliable estimation of topic coherence for low-resource languages when the Bible is the only reference. We propose a model that can correct the drawbacks of a Bible-derived CNPMI. While we assume bilingual topics paired with English, our approach can be applied to any high-resource/low-resource language pair.

We take Wikipedia's CNPMI from high-resource languages as accurate estimations. We then build a coherence *estimator* on topics from high-resource languages, with the Wikipedia CNPMI as the target output. We use linear regression using the below features. Given a topic in low-resource language, the estimator produces an estimated coherence (Figure 2).

3.1 Estimator Features

The key to the estimator is to find features that capture whether we should trust the Bible. For generality, we focus on features independent of the available resources other than the Bible. This section describes the features, which we split into four groups.

Base Features (BASE) Our base features include information we can collect from the Bible and the topic model: cardinality C, CNPMI and INPMI, MTA, and topic word coverage (TWC), which counts the percentage of topic words in a topic that appear in a reference corpus.

Crosslingual Gap (GAP) A low CNPMI score could indicate a topic pair where each language has a monolingually coherent topic but that are not about the same theme (Topic 6 in Figure 1). Thus, we add two features to capture this information

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[2] The Bible is available in 2,530 languages.

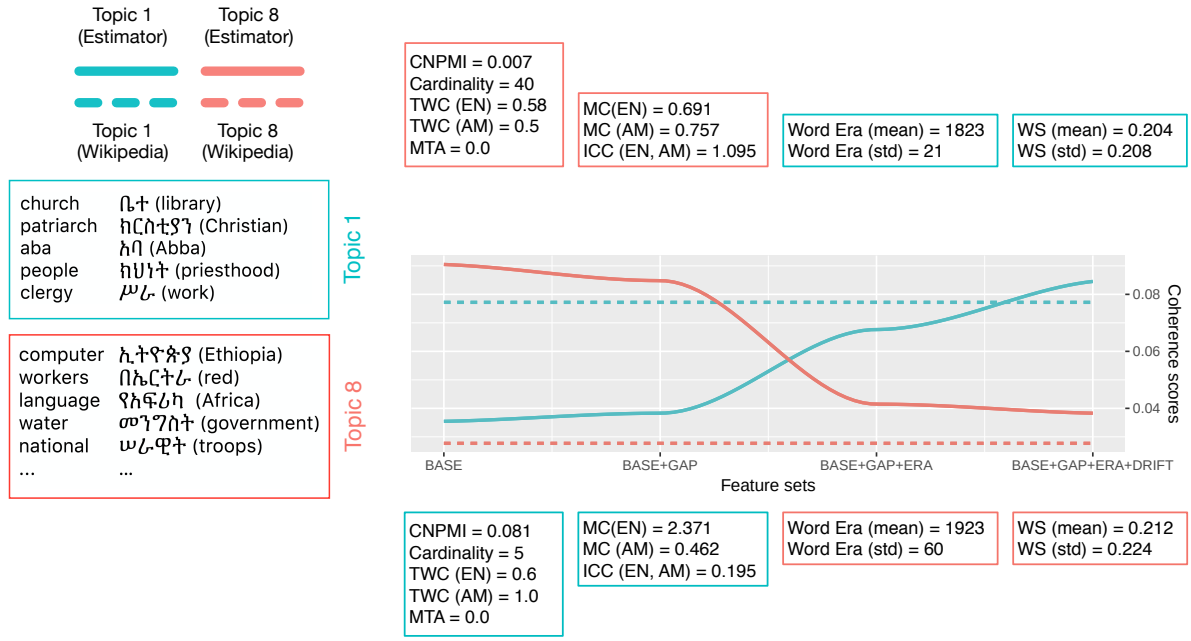

Figure 3: As the estimator adds additional features, the estimated topic coherence scores (solid lines) approach to Wikipedia CNPMI (dashed lines).

using the Bible: mismatch coefficients (MC) and internal comparison coefficients (ICC):

$$\text{MC}(\ell_1; \ell_2, k) = \frac{\text{CNPMI}(\ell_1, \ell_2, k)}{\text{INPMI}(\ell_1, k) + \alpha}, \quad (5)$$

$$\text{ICC}(\ell_1, \ell_2, k) = \frac{\text{INPMI}(\ell_1, k) + \alpha}{\text{INPMI}(\ell_2, k) + \alpha}, \quad (6)$$

where α is a smoothing factor ($\alpha = 0.001$ in our experiments). MC recognizes the gap between crosslingual and monolingual coherence, so a higher MC score indicates a gap between coherence within and across languages. Similarly, ICC compares monolingual coherence to tell if both languages are coherent: the closer to 1 the ICC is, the more comparable internal coherence both languages have.

Word Era (ERA) Because the Bible's vocabulary is unable to evaluate modern topics, we must tell the model what the modern words are. The **word era** features are the earliest usage year [3] for each word in a topic. We use both the mean and standard deviation as features.

Meaning Drift (DRIFT). The meaning of a word can expand and drift over time. For example, in the Bible, "web" appears in Isaiah 59:5:

> They hatch cockatrice' eggs, and weave the spider's **web**.

[3] https://oxforddictionaries.com/

The word "web" could be evaluated correctly in an animal topic. For modern topics, however, Bible fails to capture modern meanings of "web", as in Topic 5 (Figure 1).

To address this **meaning drift**, we use a method similar to Hamilton et al. (2016). For each English word, we calculate the context vector from Bible and from Wikipedia with a window size of five and calculate the cosine similarity between them as **word similarity**. Similar context vectors mean that the usage in the Bible is consistent with Wikipedia. We calculate word similarities for all the English topic words in a topic and use the average and standard deviation as features.

3.2 Example

In Figure 3, Topic 1 is coherent while Topic 8 is not. From left to right, we incrementally add new feature sets, and show how the estimated topic coherence scores (dashed lines) approach the ideal CNPMI (dotted lines). When only using the BASE features, the estimator gives a higher prediction to Topic 8 than to Topic 1. Their low MTA and TWC prevent accurate evaluations. Adding GAP does not help much. However, ICC(EN, AM, $k = 1$) is much smaller, which might indicate a large gap of internal coherence between the two languages.

Adding ERA makes the estimated scores flip between the two topics. Topic 1 has word era of 1823, much older than Topic 8's word era of 1923, in-

Pair	Training	Reference		
		Wikipedia	The Bible	Wiktionary
EN-RO	1,272	8,126	1,189	29,836
EN-SV	3,378	9,067	1,189	42,953
EN-AM	421	1,581	1,189	1,091
EN-TL	542	4,166	1,189	10,970
EN-TR	874	5,524	1,189	16,853
EN-ZH	874	10,000	1,189	22,946

Table 1: Number of document pairs in the training and reference datasets and number of dictionary entries for each language pair.

dicating that Topic 8 includes modern words the Bible lacks (*e.g.*, "computer"). Using all the features, the estimator gives more accurate topic coherence evaluations.

4 Experiments: Bible to Wikipedia

We experiment on six languages (Table 1) from three corpora: Romanian (RO) and Swedish (SV) from EuroParl as representative of well-studied and rich-resource languages (Koehn, 2005); Amharic (AM) and Tagalog (TL) from collected news, as low-resource languages (Huang et al., 2002a,b); and Chinese (ZH) and Turkish (TR) from TED Talks 2013 (Tiedemann, 2012), adding language variety to our experiments. Each language is paired with English as a bilingual corpus.

Typical preprocessing methods (stemming, stop word removal, *etc.*) are often unavailable for low-resource languages. For a meaningful comparison across languages, we do not apply any stemming or lemmatization strategies, including English, except removing digit numbers and symbols. However, we remove words that appear in more than 30% of documents for each language.

Each language pair is separately trained using the `MALLET` (McCallum, 2002) implementation of the polylingual topic model. Each experiment runs five Gibbs sampling chains with 1,000 iterations per chain with twenty topics. The hyperparameters are set to the default values ($\alpha = 0.1$, $\beta = 0.01$), and are optimized every 50 iterations in `MALLET` using slice sampling (Wallach et al., 2009a).

4.1 Evaluating Multilingual Topics

We use Wikipedia and the Bible as reference corpora for calculating co-occurrence statistics. Different numbers of Wikipedia articles are available for each language pair (Table 1), while the Bible contains a complete set of 1,189 chapters for all of its translations (Christodoulopoulos and Steedman, 2015). We use Wiktionary as the dictionary to calculate MTA.

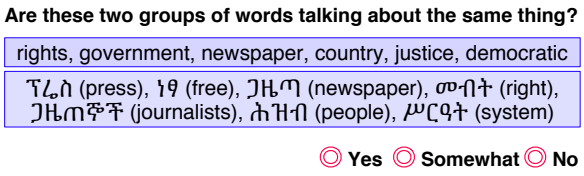

Figure 4: The interface for topic quality judgments. Users read the topic first, and make a judgment on whether the words in this pair are talking about the same thing. The translations are here for illustration; they are not shown to the users.

	Wikipedia		The Bible		MTA
	CNPMI	INPMI	CNPMI	INPMI	
EN-RO	0.490	0.118	−0.096	0.031	**0.592**
EN-SV	**0.453**	−0.295	0.164	−0.351	0.248
EN-AM	0.110	0.019	**0.289**	0.249	0.172
EN-TL	**0.512**	0.277	0.166	0.002	0.289
EN-TR	0.664	0.243	0.209	−0.246	**0.677**
EN-ZH	**0.436**	0.297	0.274	0.157	0.411

Table 2: Pearson correlations between human judgments and CNPMI are higher than INPMI, while MTA correlations are comparable to CNPMI.

4.2 Training the Estimator

In addition to experimenting on Wikipedia-based CNPMI, we also re-evaluate the topics' Bible coherence using our estimator. In the following experiments, we use an AdaBoost regressor with linear regression as the coherence estimator (Friedman, 2002; Collins et al., 2000). The estimator takes a topic and low-quality CNPMI score as input and outputs (hopefully) an improved CNPMI score.

To make our testing scenario more realistic, we treat one language as our estimator's test language and train on multilingual topics from the other languages. We use three-fold cross-validation over languages to select the best hyperparameters, including the learning rate and loss function in `AdaBoost.R2` (Drucker, 1997).

5 Topic-Level Evaluation

We first study CNPMI at the topic level: does a particular topic make sense? An effective evaluation should be consistent with human judgment of the topics (Chang et al., 2009). In this section, we measure gold-standard human interpretability of multilingual topics to establish which automatic measures of topic interpretability work best.

Test	Bible	Train		
		RO+SV	ZH+TR	RO+SV+ZH+TR
AM	−0.015	0.332	0.315	0.333
TL	−0.309	0.767	0.631	0.705
		AM+TL	ZH+TR	AM+TL+ZH+TR
RO	−0.269	0.736	0.681	0.713
SV	0.000	0.787	0.645	0.683
		RO+SV	AM+TL	RO+SV+AM+TL
ZH	0.217	0.751	0.732	0.741
TR	0.113	0.680	0.642	0.666

Table 3: Correlations between the Wikipedia-based CNPMI and the Bible-based CNPMI, before and after using the coherence estimator, at the topic level. Strong correlations indicate that the estimator improves CNPMI estimates.

5.1 Task Design

Following monolingual coherence evaluations (Lau et al., 2014), we present topic pairs to bilingual CrowdFlower users. Each task is a topic pair with the top ten topic words ($C = 10$) for each language. We ask if both languages' top words in a multilingual topic are talking about the same concept (Figure 4), and make a judgment on a three-point scale—coherent (2 points), somewhat coherent (1 point), and incoherent (0 points). To ensure the users have adequate language competency, we insert several topics that are easily identifiable as incoherent as a qualification test.

We randomly select sixty topics from each language pair (360 topics total), and each topic is judged by five users. We take the average of the judgment points and calculate Pearson correlations with the proposed evaluation metrics (Table 2). NPMI-based scores are separately calculated from each reference corpus.

5.2 Agreement with Human Judgments

CNPMI (the extended metric) has higher correlations with human judgments than INPMI (the naive adaptation of monolingual NPMI), while MTA (matching translation accuracy) correlations are comparable to CNPMI.

Unsurprisingly, when using Wikipedia as the reference, the correlations are usually higher than when using the Bible. The Bible's archaic content limits its ability to estimate human judgments in modern corpora (Section 3).

Next, we compare CNPMI to two baselines: INPMI and MTA. As expected, CNPMI outperforms INPMI regardless of reference corpus overall, because INPMI only considers monolingual coherence. MTA has higher correlations than CNPMI scores from the Bible, because the Bible fails to give accurate estimates due to limited topic coverage. MTA, on the other hand, only depends on dictionaries, which are more comprehensive than the Bible. It is also possible that users are judging coherence based on translations across a topic pair, rather than the overall coherence, which would closely correlate with MTA.

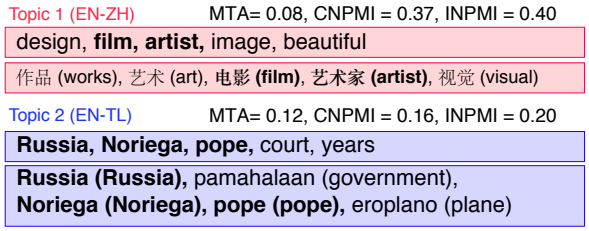

Figure 5: MTA fails to capture semantically related words (Topic 1) and only looks at translation pairs regardless of internal coherence (Topic 2).

5.3 Re-Estimating Topic-Level Coherence

The Bible—by itself—produces CNPMI values that do not correlate well with human judgments (Table 2). After training an estimator (Section 4.2), we calculate Pearson's correlation between Wikipedia's CNPMI and the estimated topic coherence score (Table 3). A higher correlation with Wikipedia's CNPMI means more accurate coherence.

As a baseline, the correlation of Bible-based CNPMI without adaptation has negative and near-zero correlations with Wikipedia;[4] it does not capture coherence. After training the estimator, the correlations become stronger, indicating the estimated scores are closer to Wikipedia's CNPMI.

5.4 When MTA Falls Short

We analyze MTA from two aspects—the inability to capture semantically-related *non-translation* topic words, and insensitivity to cardinality—to show why MTA is not an ideal measurement, even though it correlates well with human judgments.

Semantics We take two examples with EN-ZH (Topic 1) and EN-TL (Topic 2) in Figure 5. Topic 1 has fewer translation pairs than Topic 2, which leads to a lower MTA score for Topic 1. However, all words in Topic 1 talk about art, while it is hard to interpret Topic 2. Wikipedia CNPMI scores reveals

[4] Normally one would not estimate CNPMI on rich-resource languages using low-resource languages. For completeness, however, we also include these situations.

Topic 1 is more coherent. Because our experiments are on datasets with little divergence between the themes discussed across languages, this is uncommon for us but could appear in noisier datasets.

Cardinality Increasing cardinality diminishes a topic's coherence (Lau and Baldwin, 2016). We vary the cardinality of topics from ten to fifty at intervals of ten (Figure 6). As cardinality increases, more low-probability and irrelevant words appear the topic, which lowers CNPMI scores. However, MTA stays stable or increases with increasing cardinality. Thus, MTA fails to fulfill a critical property of topic model evaluation.

Finally, MTA requires a comprehensive multilingual dictionary, which may be unavailable for low-resource languages. Additionally, most languages often only have one dictionary, which makes it problematic to use the same resource (a language's single multilingual dictionary) for training and evaluating models that use a dictionary to build multilingual topics (Hu et al., 2014). Given these concerns, we continue the paper's focus on CNPMI as a data-driven alternative to MTA. However, for many applications MTA may suffice as a simple, adequate evaluation metric.

6 Model-Level Evaluation

While the previous section looked at individual topics, we also care about how well CNPMI characterizes the quality of *models* through an average of a model's constituent topics.

6.1 Training Knowledge

Adding more knowledge to multilingual topic models improves topics (Hu et al., 2014), so an effective evaluation should reflect this improvement as knowlege is added to the model. For polylingual topic models, this knowledge takes the form of the *number* of linked documents.

We start by experimenting with no multilingual knowledge: no document pairs share a topic distribution θ_d (but the documents are in the collection as unlinked documents). We then increase the number of document pairs that share θ_d from 20% of the corpus to 100%. Fixing the topic cardinality at ten, CNPMI captures the improvements in models (Figure 7) through a higher coherence score.

6.2 Agreement with Machines

Topic models are often used as a feature extraction technique for downstream machine learning applications, and topic model evaluations should reflect whether these features are useful (Ramage et al., 2009). For each model, we apply a document classifier trained on the model parameters to test whether CNPMI is consistent with classification accuracy.

Test	Bible	Train		
		RO+SV	ZH+TR	RO+SV+ZH+TR
AM	0.607	0.677	0.707	0.694
TL	0.796	0.875	0.924	0.918
		AM+TL	ZH+TR	AM+TL+ZH+TR
RO	0.631	0.912	0.919	0.931
SV	0.797	0.959	0.848	0.878
		RO+SV	AM+TL	RO+SV+AM+TL
ZH	0.907	0.918	0.951	0.939
TR	0.911	0.862	0.898	0.887

Table 4: At the model level, the estimator improves correlations between CNPMI and downstream classification for all languages except for Turkish.

Specifically, we want our classifier to transfer information from training on one language to testing on another (Smet et al., 2011; Heyman et al., 2016). We train a classifier on one language's documents, where each document's feature vector is the document-topic distribution θ_d. We apply this to TED Talks, where each document is labeled with multiple categories. We choose the most frequent seven categories across the corpus as labels,[5] and only have labeled documents in one side of a bilingual topic model. CNPMI has very strong correlations with classification results, though using the Bible as the reference corpus gives slightly lower correlation—with higher variance—than Wikipedia (Figure 8).

6.3 Re-Estimating Model-Level Coherence

In Section 5.3, we improve Bible-based CNPMI scores for individual topics. Here, we show the estimator also improves model-level coherence. We apply the estimator on the models created in Section 6.2 and calculate the correlation between estimated scores and Wikipedia's CNPMI (Table 4).

The coherence estimator substantially improves scores except for Turkish: the correlation is better *before* applying the estimator (0.911). We suspect a lack of overlap between topics between Turkish and languages other than Chinese is to blame (Figure 9); the features used by the estimator do not generalize well to other kinds of features; training on many languages pairs would hopefully solve this

[5] design, global issues, art, science, technology, business, and culture

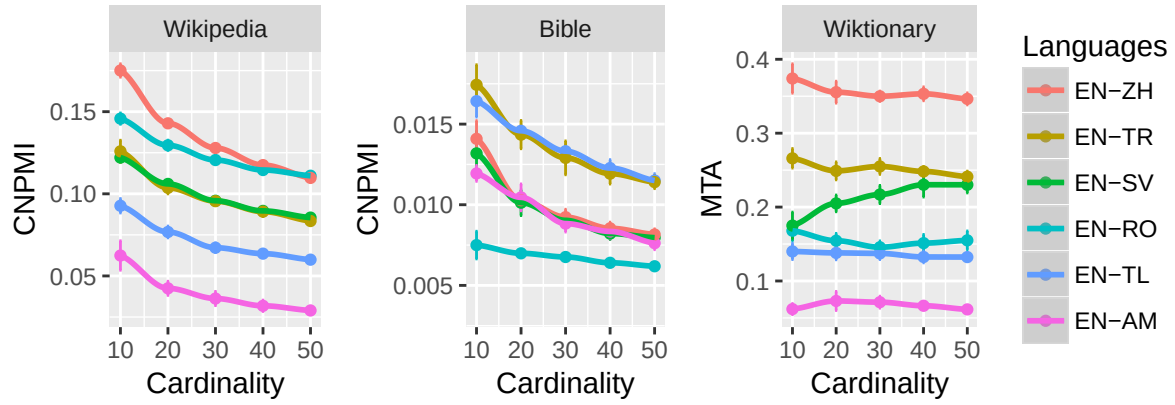

Figure 6: Increasing cardinality of topic pairs makes it harder to judge the coherence. Decreasing CNPMI scores reflect the diminished interpretability of topics, while MTA scores do not.

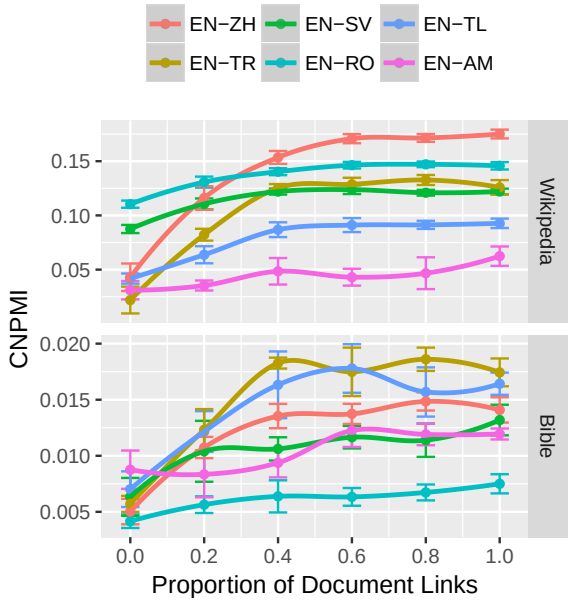

Figure 7: Adding more document links to the model produces more multilingually coherent topics. CNPMI captures this improvement.

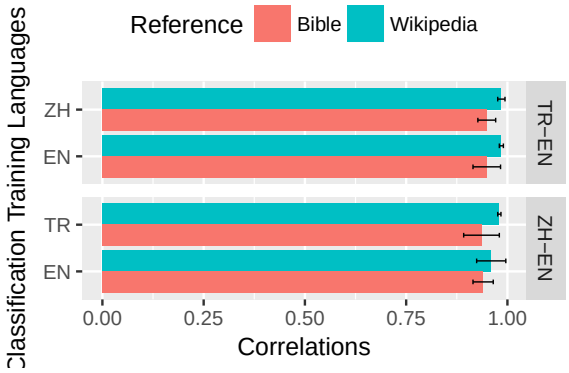

Figure 8: Pearson correlation between classification F1 scores and CNPMI: both CNPMI data sources predict whether a classifier using topic features will work well, but Wikipedia has slightly higher correlation with lower variance.

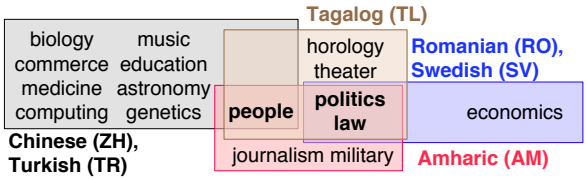

Figure 9: The overlap of topics and domain: only one out of nine Turkish and Chinese topics have domain overlap with Tagalog and Amharic topics. This hinders the Turkish estimator from capturing model-level properties.

issue. Turkish is also morphologically rich, and our preprocessing completely ignores morphology.

6.4 Reference Size

One challenge with low-resource languages is that even if Wikipedia is available, it may have too few documents to accurately calculate coherence. As a final analysis, we examine how the reliability of CNPMI degrades with a smaller reference corpus.

We randomly sample 20% to 100% of document pairs from the reference corpora and evaluate the polylingual topic model with all document links (Figure 10), again fixing the cardinality as 10.

CNPMI is stable across different amounts of reference documents, as long as the number of reference documents is sufficiently large. If there are too few reference documents (for example, 20% of Amharic Wikipedia is only 316 documents), then CNPMI degrades.

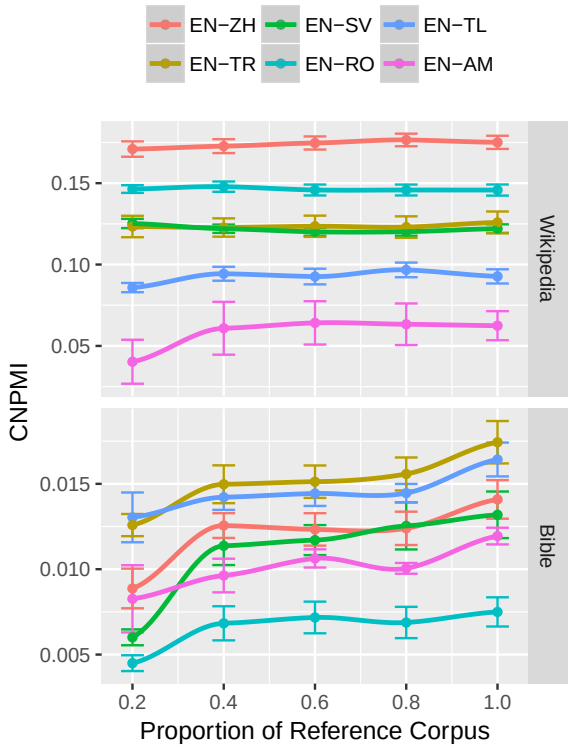

Figure 10: CNPMI is stable once the number of reference documents is large enough (around five thousand documents).

7 Related Work

Topic Coherence Many coherence metrics based on co-occurrence statistics have been proposed besides NPMI. Similar metrics—such as asymmetrical word pair metrics (Mimno et al., 2011) and combinations of existing measurements (Lau et al., 2014; Röder et al., 2015)—correlate well with human judgments. NPMI has been the current gold standard for evaluation and improvements of monolingual topic models (Pecina, 2010; Newman et al., 2011).

External Tasks Another approach is to use a model for predictive tasks: the better the results are on external tasks, the better a topic model is assumed to be. A common task is held-out likelihood (Wallach et al., 2009b; Jagarlamudi and Daumé III, 2010; Fukumasu et al., 2012), but as Chang et al. (2009) show, this does not always reflect human interpretability. Other specific tasks have also been used, such as bilingual dictionary extraction (Liu et al., 2015; Ma and Nasukawa, 2017), cultural difference deteciton (Gutiérrez et al., 2016), and crosslingual document clustering (Vulić et al., 2015).

Representation Learning Topic models are one example of a broad class of techniques of learning representations of documents (Bengio et al., 2013). Other approaches learn respresentations at the word (Klementiev et al., 2012; Vyas and Carpuat, 2016), paragraph (Mogadala and Rettinger, 2016), or corpus level (Søgaard et al., 2015). However, neural representation learning approaches are often data hungry and not adaptable to low-resource languages. The approaches here could help improve the evaluation of all multilingual representation learning algorithms (Schnabel et al., 2015).

8 Conclusion

We have provided a comprehensive analysis of topic model evaluation in multilingual settings, including for low-resource languages. While evaluation is an important area of topic model research, no previous work has studied evaluation of multilingual topic models. Our work provided two primary contributions to this area, including a new intrinsic evaluation metric, CNPMI, as well as a model for adapting this metric to low-resource languages without large reference corpora.

As the first study on evaluation for multilingual topic models, there is still room for improvement and further applications. For example, human judgment is more difficult to measure than in monolingual settings, and it is still an open question on how to design a reliable and accurate survey for multilingual quality judgments. As a measurement of multilingual coherence, we plan to extend CNPMI to high-dimensional representations, *e.g.,* multilingual word embeddings, particularly in low-resource languages (Ruder et al., 2017).

Acknowledgement

We thank the anonymous reviewers for their insightful and constructive comments. Hao has been supported under subcontract to Raytheon BBN Technologies, by DARPA award HR0011-15-C-0113. Boyd-Graber and Paul were supported by NSF grant IIS-1564275. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new

perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(8):1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

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

Gerlof Bouma. 2009. Normalized (oointwise) mutual information in collocation dxtraction. In *Proceedings of the German Society for Computational Linguistics and Language Technology Conference*.

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

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. *Applications of Topic Models*, volume 11 of *Foundations and Trends in Information Retrieval*. NOW Publishers. http://www.nowpublishers.com/article/Details/INR-030.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation* 49(2):375–395.

Michael Collins, Robert E. Schapire, and Yoram Singer. 2000. Logistic regression, AdaBoost and Bregman distances. In *Proceedings of Conference on Learning Theory*.

Harris Drucker. 1997. Improving regressors using boosting techniques. In *Proceedings of the International Conference of Machine Learning*.

Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38(4):367–378.

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

E. Dario Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics* 4:47–60.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of Empirical Methods in Natural Language Processing*.

Geert Heyman, Ivan Vulic, and Marie-Francine Moens. 2016. C-BiLDA: Extracting cross-lingual topics from non-parallel texts by distinguishing shared from unshared content. *Data Mining and Knowledge Discovery* 30(5).

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

Shudong Huang, David Graff, and George Doddington. 2002a. NMSU Amharic language pack from REFLEX V1.1. Web download file. Philadelphia: Linguistic Data Consortium.

Shudong Huang, David Graff, and George Doddington. 2002b. Tagalog language pack from reflex V1.1. Web download file. Philadelphia: Linguistic Data Consortium.

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

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of International Conference on Computational Linguistics*.

Philipp Koehn. 2005. Europarl: a parallel corpus for statistical machine translation.

Kriste Krstovski and David A. Smith. 2016. Bootstrapping translation detection and sentence extraction from comparable corpora. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Kriste Krstovski, David A. Smith, and Michael J. Kurtz. 2016. Online multilingual topic models with multilevel hyperpriors. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Jey Han Lau and Timothy Baldwin. 2016. The sensitivity of topic coherence evaluation to topic cardinality. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

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

Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. 2015. Multilingual topic models for bilingual dictionary extraction. *ACM Transactions on Asian & Low-Resource Language Information Processing* 14(3):11:1–11:22.

Tengfei Ma and Tetsuya Nasukawa. 2017. Inverted bilingual topic models for lexicon extraction from non-parallel data. In *International Joint Conference on Artificial Intelligence*.

Andrew Kachites McCallum. 2002. MALLET: a machine learning for language toolkit. `http://mallet.cs.umass.edu`.

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

David M. Mimno, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of Empirical Methods in Natural Language Processing*.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

David Newman, Edwin V. Bonilla, and Wray L. Buntine. 2011. Improving topic coherence with regularized topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

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

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Proceedings of the Language Resources and Evaluation Conference* 44(1-2).

Daniel Ramage, David Leo Wright Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of Empirical Methods in Natural Language Processing*.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a parallel corpus: annotating the 'book of 2000 tongues'. *Computers and the Humanities* 33(1/2):129–153.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of ACM International Conference on Web Search and Data Mining*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *CoRR* abs/1706.04902.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of Empirical Methods in Natural Language Processing*. pages 298–307.

Wim De Smet, Jie Tang, and Marie-Francine Moens. 2011. Knowledge transfer across multilingual corpora via latent topics. In *Pacific-Asia Advances in Knowledge Discovery and Data Mining*.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual nlp. In *Proceedings of the Association for Computational Linguistics*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Language Resources and Evaluation Conference*.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2013. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval* 16(3):331–368.

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.

Yogarshi Vyas and Marine Carpuat. 2016. Sparse bilingual word representations for cross-lingual lexical entailment. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Hanna Wallach, David Mimno, and Andrew McCallum. 2009a. Rethinking LDA: Why priors matter. In *Proceedings of Advances in Neural Information Processing Systems*.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David M. Mimno. 2009b. Evaluation methods for topic models. In *Proceedings of the International Conference of Machine Learning*.

Weiwei Yang, Jordan L. Boyd-Graber, and Philip Resnik. 2017. Adapting topic models using lexical associations with tree priors. In *Proceedings of Empirical Methods in Natural Language Processing*.