Cross-Lingual Contextualized Topic Models with Zero-shot Learning on Same Script Languages

Divya Rustagi
Computer Science and Engineering
The Pennsylvania State University
State College, PA - 16801
dpr5375@psu.edu

Prithvi Rana
Computer Science and Engineering
The Pennsylvania State University
State College, PA - 16801
tpzr5169@psu.edu

Abstract

Topic models help us identify the abstract themes and overarching meaning of textual data by clustering words that frequently occur together. However, traditional topic modeling techniques have relied on language-specific vocabulary for training. A novel approach has been proposed to replace language-dependent bag-of-words (BoW) models with contextualized embeddings that allow us to train a model in any one language and apply it to unseen languages if we have available contextualized word embeddings. In this paper, we will extend the work of the original authors in cross-lingual contextualized topic models with zero-shot learning to test our hypothesis that cross-lingual contextualized TMs perform better on the same script languages. We use Indic languages for evaluation against two models trained in Hindi and English, respectively.

1 Introduction

Today, we are surrounded by large pools of texts in our always-online lives. We are more connected than ever, and as countries rapidly transform into digital societies, communication is fast, supporting languages across all regions [6]. Almost all textual data exists in multiple languages, from emails, messages, captions, and tweets to document archives, official reports, and even our search history [12]. To better manage this explosion of data, topic models were introduced to help us collect similar meaning documents for easy analysis.

Topic models in natural language processing (NLP) are probabilistic models that allow us to discover underlying semantic structures in unstructured text. In other words, topic models help us identify textual data's abstract themes and overarching meaning by clustering words that frequently occur together [3].

Traditional topic modeling techniques have relied on language-specific vocabulary for training [7]. To extract helpful word distributions, topic models rely on a bag-of-words (BoW) representation of the unstructured, language-specific texts. This means that once a topic model has been trained in one language, we cannot transfer it to another as the vocabulary would differ. The same model does not work on unknown words in the training language [5].

Next to topic modeling, sentiment analysis has been widely researched for cross-lingual purposes. Traditional approaches have included annotating monolingual data for building single-language models [9], packing several sentence embeddings in different languages, using machine translation to convert non-English documents into English, and developing generalized models for analysis [1]. However, these approaches have worked great for some high-resource languages or have proven computationally expensive and less optimized for others.

Other recent approaches have created code-mixed corpora for bilingual speech communities like Tamil-English [2] to enable popular NLP tasks in a multilingual or cross-lingual manner. There also have been novel additions in machine learning like LaBSE, which combines mask-language modeling pretraining to establish cross-lingual embedding

space for languages supported by M-BERT. Techniques for cross-lingual topic modeling have included Polylingual TM and multilingual dictionaries and embeddings [11]. As in traditional multilingual sentiment language models, these are language-dependent and share the same cons. However, a new approach has inspired our experimental research - Cross-lingual Contextualized Topic Models with Zero-shot Learning [5].

In this paper, we will extend the work of the original authors in cross-lingual contextualized topic models with zero-shot learning. We will address the following two aspects: (i) increase testing data to 1000-2000 samples instead of just 300 and (ii) show that contextualized topic models for zero-shot learning work better when we train and predict on the same script languages.

For training, we will focus on Indic languages that share the Devanagari script to test our hypothesis [4]. We will use 11-way parallel corpora, which covered 10 Indic languages and English and was submitted by the National Institute of Information and Communications Technology (NICT) and Kyoto University at WAT 2020 [8]. Our goal is to compare and benchmark the performance of a model trained in English against a model trained in Hindi.

We evaluate the models using NPMI scores with an existing s-BERT model as in the original paper. We will select the remaining same-script Indic languages for zero-shot learning and measure performance using matches, cosine similarity on centroid embeddings, and KL divergence to assess the data distribution.

Script We define 'script' as a writing system or an orthography consisting of visible marks, forms, or symbols called characters or graphs related to some structure in the linguistic system [14]. Latin is one of the most familiar scripts in our writing systems, as many Western languages like English, Italian, German, and more share it. Other widely used scripts include Cyrillic in Eurasia, Arabic in the Middle East, Canadian Aboriginal script in Indigenous Canadian communities, and Greek, Hebrew, and Chinese [10].

Literature Review A small but dedicated section of the NLP research community focuses on sentiment analysis of low-resource languages. The common attribute shared by the works that arise from this community is the focus on building a general model trained on one primary language, which we can use on datasets of other languages. The easiest method of building models for low-resource languages would be to start by creating a good dataset for them that other researchers can leverage. One of the best-known multilingual corpora features is code-mixed [2]. Another approach that Google AI conducted made use of the Masked Language Model (MLM) and Translation Language Model (TLM) to create multilingual sentence embeddings [13]. This adaptation of multilingual BERT resulted in language-agnostic sentence embeddings for 109 languages

2 Contextualized Neural Topic Models

Cross-lingual contextualized TMs proposed moving away from language-dependent BoW. Instead, we replace BoW with contextualized embeddings that allow us to train a model in any language and apply it to unseen languages if we have available contextualized word embeddings. This made the models transferable and enabled zero-shot learning for cross-lingual topic modeling for the first time.

2.1 Replicating Parent Paper's Results

After conducting the experiments and tabulating the results using the original paper's implementation and multilingual representations from SBERT, we found the results to be reasonably consistent with that of the original article.

NPMI Coherence for each model showcased a slight improvement, and we can attribute this to an update in the off-the-shelf models used by the authors.

Model	τ (50)	τ (100)
ZeroShotTM	0.1692	0.1351
CombinedTM	**0.1792**	**0.1543**
Neural-ProdLDA	0.1710	0.1383
LDA	-0.0611	-0.1588

Table 1: NPMI Coherence on the W1 dataset.

Metrics The original model trained in English achieved, on average, a 70% topic coherence, an 85% similarity score showing that predicted topics

Lang	Mat25	KL25	CD25	Mat50	KL50	CD50
IT	80.67	0.13	0.85	68.00	0.17	0.76
FR	77.67	0.14	0.83	68.00	0.18	0.76
PT	81.00	0.12	0.85	70.00	0.15	0.78
DE	74.00	0.14	0.80	65.00	0.18	0.74
ZeroshotTM Avg	**78.33**	**0.13**	**0.83**	67.75	0.17	0.76

Table 2: Match, KL, and centroid similarity for 25 and 50 topics on various languages on W2 dataset.

were at least like English counterparts, and a 15% KL divergence proving that topic distributions are similar across the languages [5].

2.2 Evaluating Same Script Languages

We evaluate our hypothesis: *does contextualized TM tackle zero-shot cross-lingual topic modeling better on same script languages?*

We use the same extended Neural-ProdLDA approach as the original paper's authors. It consists of a neural network that maps the BoW representation of a document onto a continuous latent representation. We still need the BoW representation for training the models. To reconstruct the BoW representation, the latent representation is sampled from a Gaussian distribution parameterized by μ and σ2. Once we have trained the model, we can discard the BoW representation.

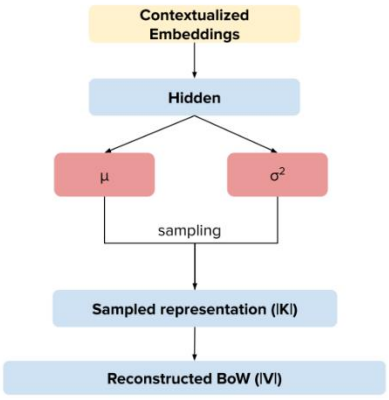

Figure 1: High-level schema of the architecture for the proposed contextualized neural topic model.

IndicBert We replace the BoW input in Neural-ProdLDA with pre-trained multilingual representations from IndicBert. The original paper used SBERT. But since our hypothesis focuses on the relationships between the same source script languages of the Indian subcontinent, we must use closely related representations.

Datasets We use the PMIndia dataset (4003 documents) as our primary data source. It contains parallel sentence text files which pair 13 major languages of India with English. There are up to 5600 sentences for each language pair. We only select speeches in Hindi and English with at least 500 tokens for training.

W1 contains 4003 randomly sampled speeches from the Hindi and English set of the PMIndia monolingual dataset.

W2 contains 800 randomly sampled test speeches in *Hindi and English with parallel instances in Assamese (as), Gujarati (gu), Kannada (kn), Malayalam (ml), Marathi (mr), Oriya (or), Punjabi (pa), and Tamil (ta), Telugu (te).*

2.3 Primary Language Models

Hindi-based ZeroshotTM model assessed our primary hypothesis and chose Hindi as our training language with the script of Brahmi scripts.

English-based ZeroshotTM model assessed our hypothesis against the standard method of using English as the training language with the script of Latin.

3 Results

NPMI Coherence results for the Hindi model were exceptionally low compared to the English model. Because the Hindi topics lacked coherence, they did not generate a meaningful vector space.

There could also have been errors when we tried to fit ZeroShotTM for Hindi text.

We made two observations: (i) Pre-processing with Hindi leads to disjointed topic predictions. (ii) Tokens also affected the performance of the two models. The Hindi model performs better with more, while the English model performs better with fewer tokens. We tried using 200, 350, and 500 tokens in our experiments.

Model	τ (50)	τ (100)
ZeroShotTM for Hindi	-0.0138	-0.0093
ZeroShotTM for English	0.1391	0.1345

Table 3: NPMI Coherence for 25 and 50 topic models in Hindi and English trained on ~4000 PMIndia speeches.

Metrics The results show that the English model performs much better than the Hindi model. The main reason we see this result is that when we use off-the-shelf libraries to create BoW for languages that use Devanagari script, the BoW cannot capture the accents, which are a big part of these Indic languages. This problem does not exist in its English model counterpart.

Across the board, the topic matches (MAT), Topic similarity (CD), and Topic distribution (KL divergence) were worse for the Hindi model when compared to the English model.

Interestingly, the Hindi model MAT benefits from comparing more topics as the MAT50 is greater than MAT25 for most languages. But this does not hold for the English model.

Lang	Mat25	KL25	CD25	Mat50	KL50	CD50
as	8.37	0.55	0.0	19.00	0.43	0.0
bn	7.75	0.62	0.0	12.75	0.51	0.0
en	9.63	8.06	0.0	3.63	8.26	0.0
gu	37.88	0.29	0.0	37.50	0.26	0.0
kn	12.63	0.52	0.0	18.50	0.41	0.0
ml	5.38	0.66	0.0	5.50	0.62	0.0
mr	18.63	0.49	0.0	23.88	0.38	0.0
or	14.25	0.49	0.0	14.75	0.57	0.0
pa	40.88	0.34	0.0	32.75	0.33	0.0
ta	4.88	0.68	0.0	2.63	0.68	0.0
te	8.75	0.70	0.0	8.23	0.56	0.0
ZeroshotTM Avg	15.36	1.22	0.0	**16.28**	1.18	0.0

Table 3: Hindi-based ZeroshotTM on Match, KL, and centroid similarity for 25 and 50 topics on W2.

Lang	Mat25	KL25	CD25	Mat50	KL50	CD50
as	43.50	0.33	0.59	29.13	0.41	0.48
bn	38.75	0.41	0.57	21.12	0.48	0.43
gu	62.88	0.27	0.73	47.00	0.26	0.64
kn	40.00	0.40	0.56	32.63	0.40	0.51
ml	31.00	0.53	0.48	16.00	0.57	0.38
mr	39.88	0.42	0.56	31.11	0.43	0.51
or	46.50	0.40	0.62	26.40	0.48	0.48
pa	55.63	0.25	0.67	51.50	0.24	0.67
ta	30.01	0.63	0.48	13.13	0.67	0.36
te	28.00	0.48	0.45	10.75	0.63	0.33
hi	0.0	0.34	0.69	25.34	0.30	0.0
ZeroshotTM Avg	**37.84**	0.40	0.58	25.34	0.44	0.43

Table 4: English-based ZeroshotTM on Match, KL, and centroid similarity for 25 and 50 topics on W2.

4 Conclusions

We investigate the performance of cross-lingual contextualized topic models with zero-shot learning on same script languages, with a focus on Indic languages sharing the Devanagari script. We extend the work of the original authors by increasing the testing data and comparing models trained in English and Hindi. We hypothesize that contextualized topic models perform better when trained and predicted on same script languages.

However, our results do not support this hypothesis. The Hindi model, based on the Devanagari script, performs poorly compared to the English model, which uses the Latin script. We attribute these results to the limitations of the off-the-shelf libraries when creating BoW representations for languages that use the Devanagari script, as they fail to capture the accents and diacritics crucial to these Indic languages.

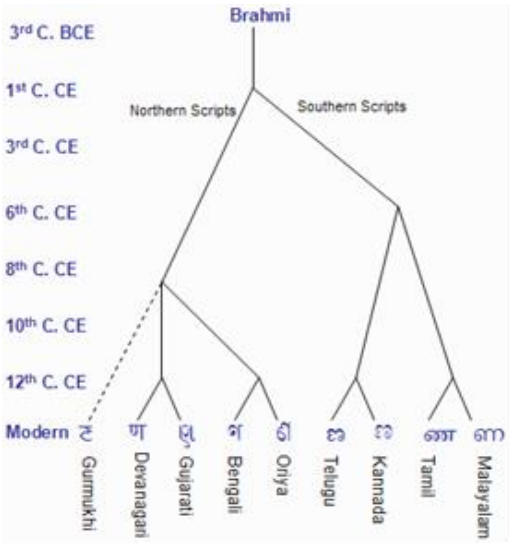

Figure 2: Derivation of the character *NNA*, showing how all the different forms arose from a common source (Brahmi) for the modern scripts. The diagram shows an early divergence between north and south Indian scripts.

We also highlight the challenges when working with languages sharing the same script, particularly those with complex orthographies like Devanagari. The script of a language significantly changes text preprocessing in NLP, as noted by W3C (2002) [15], with different scripts requiring specialized handling and processing techniques. Future research should focus on improving the preprocessing and handling of these languages to enhance the performance of cross-lingual contextualized topic models. Additionally, further exploration into other languages and scripts may provide valuable insights into the generalizability of these models across different language families.

The viability of using the ZeroShotTM method for topic modeling on Indic languages requires modifying state-of-the-art NLP resources, which were primarily built for and on high-resource languages like English. To address the unique challenges posed by the Devanagari script and other low-resource languages, investment in time and resources is necessary for developing processing techniques specifically tailored to these languages and scripts.

5 Acknowledgements

Authors of the original paper and support from class DS-340W.

Related Works A novel approach to creating a corpus for every low-resource language was formulated using an RNN-based framework [1]. This method could train sentiment analysis model training on one language like English, for which much data is available. Then, the dataset of the target language would be translated into English so that the model could be used on it. However, this approach creates the overhead of machine translation. Furthermore, how well the target language was translated into the primary language would impact the model's accuracy.

References

[1] Can, E.F, Can, A.E, Can, F. (2018). Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data. arXiv. https://arxiv.org/abs/1806.04511

[2] Chakravarthi, B.R., Muralidaran, V., Priyadharshini, R., & McCrae, J.P. (2020). Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. Proceedings of the 1st Joint Workshop on SLTU and CCURL Workshop. LREC 2020. https://aclanthology.org/2020.sltu-1.28/

[3] David M. Blei. 2012. Probabilistic topic models. Commun. ACM 55, 4 (April 2012), 77–84. DOI: https://doi.org/10.1145/2133806.2133826

[4] Devanagari. (2001, December 1). In Wikipedia. https://en.wikipedia.org/wiki/Devanagari

[5] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1676–1683, Online. Association for Computational Linguistics.

[6] Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. Journal of Economic Literature, 57(3), 535–574. https://doi.org/10.1257/jel.20181020

[7] Krstovski, K., Kurtz, M. J., Smith, D. A., & Accomazzi, A. (2017). Multilingual Topic Models. arXiv preprint arXiv:1712.06704.

[8] Kyoto University & National Institute of Information and Communications Technology. (2021, January 4). Indic Languages Multilingual Parallel Corpus. MultiIndicMT: An Indic Language Multilingual Task. http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.html

[9] Larkey, L. S., Feng, F., Connell, M., & Lavrenko, V. (2004, July). Language-specific models in multilingual topic tracking. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 402-409).

[10] List of writing systems. (2003, October 1). In Wikipedia. https://en.wikipedia.org/wiki/List_of_writing_systems

[11] Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., & McCallum, A. (2009, August). Polylingual topic models. In Proceedings of the 2009 conference on empirical methods in natural language processing (pp. 880-889).

[12] Statista, & Johnson, J. J. (2022, January 26). Most common languages used on the Internet 2020. Statista. Retrieved March 15, 2022, from https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/

[13] Yang, Y., Feng, F., Cer, D., Arivazhagan, N., Wang, W. (2020). Language-agnostic BERT Sentence Embedding. arXiv. https://arxiv.org/abs/2007.01852

[14] Writing - Types of writing systems. (n.d.). Encyclopedia Britannica. https://www.britannica.com/topic/writing/Types-of-writing-systems

[15] “Indic script blocks.” https://www.w3.org/2002/Talks/09-ri-indic/indic-paper.html