Cross-Lingual Contextualized Topic Models With Zero-Shot Learning

Prince Titiya, Mitul Dudhat, Ayush Patel and Hiten Gondaliya

Abstract

ZeroShotTM is a scalable and efficient framework for multilingual topic modeling, enabling cross-lingual topic inference without additional training in unseen languages. By leveraging SBERT embeddings for semantic representation and a Variational Autoencoder (VAE) to learn latent topic distributions, the model overcomes challenges like vocabulary sparsity, scalability, and dependency on language-specific training. Tested on multilingual datasets, ZeroShotTM generalizes topics across languages, achieving competitive results in zero-shot scenarios and paving the way for robust multilingual natural language understanding.

1. Problem Statement

Traditional topic modeling techniques, such as Latent Dirichlet Allocation (LDA), rely heavily on word count representations (Bag-of-Words) and often require labeled training data. These approaches face significant challenges when applied to multilingual or cross-lingual datasets:

- **Vocabulary Dependence**: They rely on language-specific BoW representations, making cross-lingual application impractical.
- **Scalability Issues**: Training on multilingual corpora leads to a vast, sparse vocabulary, causing overfitting and computational inefficiency

2. Dataset

1. **Training Dataset**:

   • **Source**: 99,700 English documents from DBpedia, covering diverse topics such as arts, sports, and technology.
   • **Preprocessing**: Text is preprocessed to remove punctuation, stopwords, and low-frequency words, and converted into SBERT embeddings.

2. **Testing Dataset**:

   • **Languages**: 300 documents each in Portuguese, French, German, and Italian, with parallel topics across languages.
   • **Evaluation Context**: Testing focuses on the ability to generalize topics learned from English to these unseen languages.

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India
202318010@daiict.ac.in (P. Titiya); 202318024@daiict.ac.in (M. Dudhat); 202318036@daiict.ac.in (A. Patel); 202318063@daiict.ac.in (H. Gondaliya)
202318010 (P. Titiya); 202318024 (M. Dudhat); 202318036 (A. Patel); 202318063 (H. Gondaliya)

3. Evaluation Metrics

The effectiveness of ZeroShotTM is evaluated using the following metrics:

- **Match Score (Mat↑):** Measures the proportion of documents where the most probable topic in the source language matches the most probable topic in the target language.
- **KL Divergence (KL↓):** Quantifies the divergence between the topic distributions of comparable documents in source and target languages. Lower values indicate better alignment.
- **Centroid Distance (CD↑):** Evaluates the similarity between the topic-word distributions of source and target languages. Higher values indicate better alignment.

4. Results

The model achieves competitive results across languages:

Table 1
Performance Metrics for 25 and 50 Topics Across Languages

Lang	Mat25↑	KL25↓	CD25↑	Mat50↑	KL50↓	CD50↑
IT	0.51	0.21	0.73	0.44	0.22	0.71
FR	0.57	0.22	0.71	0.48	0.21	0.72
DE	0.56	0.21	0.72	0.49	0.21	0.69
PT	0.59	0.19	0.76	0.52	0.19	0.74

These results demonstrate ZeroShotTM's ability to generalize topics across languages without additional training.

Given results in research paper.

Lang	Mat25↑	KL25↓	CD25↑	Mat50↑	KL50↓	CD50↑
IT	75.67	0.16	0.84	62.00	0.21	0.75
FR	79.00	0.14	0.86	63.33	0.19	0.77
PT	78.00	0.14	0.85	68.00	0.19	0.79
DE	79.33	0.15	0.85	64.33	0.20	0.77

5. Key Challenges and Learnings

1. **Vocabulary Sparsity**
   Challenge: Multilingual data brings a vast vocabulary, leading to sparsity and inefficiency during computation. Managing such diverse lexical content is critical to ensure smooth processing and meaningful topic extraction.
   Solution: Implementing preprocessing steps, such as stopword removal, vocabulary pruning, and utilizing Sentence-BERT (SBERT) embeddings, significantly reduced vocabulary size while preserving semantic information.

2. **Named Entity Bias**
   Challenge: Frequently occurring named entities, such as "Sachin Tendulkar," dominated certain topics, causing imbalances and masking other valuable insights.
   Solution: Leveraging SBERT embeddings helped abstract named entities into their semantic meanings, ensuring that the model focused on overall context rather than entity frequency.

3. **Evaluation Alignment**
   Challenge: Aligning topic distributions across source and target languages proved difficult due to variations in semantic structures and linguistic nuances.
   Solution: Metrics like KL Divergence and Centroid Distance provided robust evaluation methods, ensuring meaningful comparisons and alignment between multilingual topics.

4. **Lower Match Score Compared to Baselines**
   Challenge: The match scores in our results were lower compared to state-of-the-art research papers, highlighting potential limitations in topic modeling robustness.
   Key Observation: Match scores for topics in ZeroShotTM were 51–59% (25 topics) and 44–52% (50 topics), which, while competitive, revealed room for improvement in capturing topic relevance compared to benchmark studies.
   Solution: Enhancing match scores may involve fine-tuning the VAE architecture, incorporating advanced topic coherence metrics, or exploring hybrid embeddings for better semantic representation.

5. **Cross-Lingual Generalization**
   Key Learning: Pretrained embeddings like SBERT were crucial for enabling ZeroShotTM's cross-lingual generalization. These embeddings captured shared semantic structures across languages, facilitating zero-shot topic inference with minimal reliance on additional training data.

6. Conclusion

ZeroShotTM presents a scalable and efficient approach to multilingual topic modeling, addressing critical challenges like vocabulary sparsity, named entity bias, evaluation alignment, and match score discrepancies. By leveraging SBERT embeddings and a Variational Autoencoder (VAE) architecture, the model achieves competitive results even in zero-shot scenarios. This study highlights the potential for more robust cross-lingual topic analysis, paving the way for future advancements in multilingual natural language understanding.

References

[1] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, E. Fersini, 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, Association for Computational Linguistics, Online, 2021, pp. 1676–1683. URL: https://arxiv.org/abs/2004.07737.

[1]