Cross-Topic Language Modeling

Team InsighfulMiners

Naga Gayatri Bandaru
Faiz Shaik Ahmed
Venkatesh Makkena
*Department of Computer Engineering, San José State University, San José, California 95192, USA

***Abstract*—This work uses a Zero-Shot Topic Model (ZSTM) trained exclusively on the German text to demonstrate cross-lingual topic discovery on a corpus of 194 German survey questions. The model is able to predict topics that are aligned between French and Italian even in the absence of parallel data resources. Utilizing recent developments in multilingual semantic encoders within the ZSTM framework to produce conceptual representations that are consistent across languages is the key innovation. Topic vectors could be transferred cross-lingually even in cases where there was no vocabulary or linguistic structure overlap between the two languages by encoding words and documents into a shared embedding space. A substantial reduction in dimensionality was attained, with a vocabulary consisting of 5000 unique German words reduced to just 100 semantic dimensions. ZSTM demonstrated significant topic extraction with a coherence of -0.21 in predicting 12 labeled topics on the test set, including immigration, healthcare, and welfare, following training in German. Semantic similarity-based and frequency-based objectives made sure that interpretable topics were found. Python is used throughout the project, with sklearn and contextualized-topic-models packages being used for data preprocessing. We concluded that, without the need for costly parallel data annotations or resources, semantic knowledge transfer could facilitate insightful cross-lingual analysis.**

I. Introduction

The unsupervised identification of ”topics,” or hidden semantic structures, in an unlabeled set of documents is known as topic modeling. Latent Dirichlet Allocation (LDA) algorithms are frequently employed in text corpus analysis to extract topic representations that are condensed and describe important themes. However, word co-occurrence statistics and bag-of-words features are the only features used in conventional topic models. This makes it more difficult for them to identify conceptual parallels between languages, which hinders cross-lingual analysis.

This project looks into Zero-Shot cross-lingual Topic Modeling as a way to help with the language barrier in vocabulary. In zero-shot learning, knowledge from a high-resource source language (German) is transferred to lower-resource target languages (French and Italian) that are not seen during training. Vector space representations that retain semantic meaning across languages are made possible by recent developments in cross-lingual neural sentence encoders. These may be able to transmit latent topic concepts between disjoint vocabularies, according to our hypothesis.

In particular, we investigate two research inquiries: (1) Is it possible for a Zero-Shot model that was trained exclusively on German text to reliably extract aligned topics from French and Italian test sets? Is it possible to create more coherent topics by utilizing global semantic knowledge about words instead of just depending on word frequencies and correlations?

Our dataset comprises 194 open-ended questions about Swiss policy matters that have been expertly categorized into 12 themes such as welfare, healthcare, and so on. We assess cross-lingual transfer using 17,000 translated French/Italian political surveys as part of a proprietary test corpus. Without the use of parallel data or dictionaries, the task entails generating aligned topics between the French/Italian test sets and the German training questions.

To obtain unified document representations across languages, we propose a Zero-Shot Topic Modeling (ZSTM) architecture that makes use of a multilingual sentence transformer encoder. By grouping these semantically-aware document embeddings, aligned latent topics across the corpora are found. ZSTM accuracy is assessed for French and Italian using topic coherence metrics and nearest neighbor analysis following training on the German text.

The outcomes effectively showcase ZSTM's cross-lingual proficiency, as evidenced by its high predictive accuracy for aligned topics in languages that have not been seen before. This shows that semantic knowledge transfer for multilingual analysis is feasible even in the absence of costly labeled translation data.

II. Related Works

The field of cross-lingual topic modeling has traditionally relied on techniques like bilingual Latent Dirichlet Allocation (LDA) [1], which need bilingual dictionaries or parallel corpora. These methods, however useful, are constrained by their need on substantial multilingual resources. As an important divergence from conventional techniques, our Zero-Shot cross-lingual Topic Modeling (ZSTM) methodology, on the other hand, gets around this restriction by doing away with the need for parallel data.

Research areas have been expanded by developments in Zero-Shot learning, especially in natural language processing. Although its success in a variety of NLP tasks has been shown in earlier studies[2], its use in cross-lingual topic

modeling is still restricted. Unlike other implementations, our method distinguishes itself by directly using Zero-Shot learning to transfer information from high-resource to lower-resource languages without requiring translation data.

Sentiment analysis using neural sentence encoders, such the ones created by [3], has completely changed how semantic similarity across languages is understood. Our effort expands the use of these encoders to the field of topic modeling, while they have previously been used for tasks such as cross-lingual phrase similarity and categorization. By using these encoders to generate unified document representations, subject alignment across languages may be achieved—a technique that hasn't been thoroughly discussed in the literature to yet.

Research has mostly concentrated on techniques that take into account linguistic characteristics unique to each language when grouping multilingual documents [4]. Differently from this, our approach groups semantically-aware document embeddings to identify latent themes across languages. This method is unique in that it depends more on semantic comprehension than on linguistic characteristics.

Using aligned corpora, evaluation approaches for cross-lingual topic modeling mostly focused on making direct comparisons between subjects in different languages [5]. However, in the target languages (French and Italian), our assessment uses closest neighbor analysis and topic coherence measures, providing a new angle on evaluating the efficacy of cross-lingual topic modeling in the absence of aligned data.

In conclusion, while our ZSTM system incorporates elements from previous cross-lingual natural language processing research, it differs from previous work in that it makes use of neural sentence encoders, Zero-Shot learning, and novel assessment techniques. This work adds to our understanding of cross-lingual topic discovery while also creating new avenues for multilingual analysis and semantic knowledge transfer research, especially in contexts where parallel data is hard to come by or absent.

III. Data

Our ZSTM research uses a dataset that includes multilingual text corpora, with a primary emphasis on Swiss policy matters. The data is divided into two separate parts: a German training set and an Italian and French test corpus.

A. Data Type and Source

1) German Training Set: There are 194 open-ended questions on Swiss policy topics in the training dataset. These carefully labeled questions address a wide range of subjects, including poverty, healthcare, and other sociopolitical issues. German is a high-resource language with a wealth of accessible linguistic data, which is why it was selected.

2) French and Italian Test Corpus: With around 17,000 political polls, the test corpus is noticeably bigger. Translations of answers to questions like those in the German training set are included in these surveys, which are administered in French and Italian. This corpus was chosen to test the Zero-Shot learning capabilities of our model and to simulate languages with less resources. Under "Data Acquisition," Publicly accessible policy papers and surveys carried out by Swiss government entities served as the basis of the German dataset. Because the polls in French and Italian were sourced from a confidential source, public opinion on a range of policy concerns was represented in a complete and varied manner. Under Section: "Data Volume" There is a deliberate discrepancy in the amount of data between the training set and the test corpus. The French/Italian corpus is much bigger than the German sample, which has 194 items. This makes it difficult for the model to generalize from a smaller, more linguistically varied dataset to a larger, more diverse one.

B. Preprocessing and Treatment

1) Text Normalization: Tokenization, lowercasing, and punctuation and special characters removal were all performed on each dataset.

C. Alignment of Topic Labels

Clear topic labels for the German dataset were given via expert annotations. On the other hand, an alignment procedure was necessary for the French and Italian surveys to guarantee that their subjects matched the German dataset labels precisely.

1) Semantic Analysis via Vectorization: An encoder for multilingual sentence transformer was used to turn text input into vector space representations. In order to preserve semantic consistency between languages, this step was essential.

2) Managing Noise and Imbalances: In order to ensure the robustness and generalizability of the model, we used strategies to correct class imbalances and filter out noisy data, taking into account the inherent imbalances seen in real-world datasets.

D. Usage in Project

Our project relies heavily on the data's multilingual character. The ZSTM model is trained using the German dataset as a basis. On the other hand, the model's performance in cross-lingual topic discovery and Zero-Shot learning is assessed using the French and Italian corpora. By using this method, we may investigate if it is possible to transmit semantic information across languages without the need for parallel corpora or direct translations.

IV. Methodology

Our project addresses the significant challenge of cross-lingual topic modeling, specifically targeting the issue of working without the need for expensive parallel corpora, a major hurdle in processing low-resource languages. Our novel solution hinges on the Zero-Shot Topic Modeling (ZSTM) framework, which is a product of recent developments in multilingual sentence encoder pretraining.

A. ZSTM Architecture

There are two main parts to the ZSTM system:

The encoder is a subsubsection. Our system encodes texts from different languages into a consistent semantic vector space by using a pretrained cross-lingual sentence transformer, like the multilingual MPNet. This is an important step because it allows the model to get around vocabulary constraints and guarantees a consistent representation of documents in different languages.

1) Topic Model: We use a combination of autoencoder and LDA techniques in an unsupervised clustering strategy. The limitations of language-specific vocabularies have historically made it difficult to identify coherent and aligned latent topics across different languages. This approach is crucial in achieving this goal.

Techniques for Training and Optimization During the training stage, ZSTM is refined using two distinct losses while concentrating on German texts:

2) Decay for Reconstruction: With the goal of forecasting the word-bag representation from document embeddings, this loss aids in striking a balance between semantic and lexical information.

3) Distributional Similarity Loss: In order to help create distinct, well-clustered topics, document vectors and topic vectors are aligned using cosine similarity.

One feature of ZSTM that sets it apart is InfoWordLoss, which strengthens the connection between words and their related topics by using dependency trees and part-of-speech tags to produce more topic keywords that are relevant.

B. Evaluation and Analysis

Using a variety of techniques, we evaluate ZSTM's effectiveness on survey datasets that have not yet been released in France and Italy.

C. Coherence of Topics

We measure the semantic consistency of the identified topics using metrics such as NPMI.

1) A Close Resemblance Analysis: Using a qualitative approach, it is examined whether the top documents grouped under each topic are aligned coherently.

2) Topic Alignment: By contrasting the most important keywords for every topic, the model's ability to preserve topic consistency across languages is assessed.

D. The unique characteristics of ZSTM

Our ZSTM model is unique in that it can transfer data zero-shot, thereby eliminating the requirement for parallel corpora, which is usually required for methods based on translation. Although CLI-LDA and other alternative methods were taken into consideration, ZSTM's reliance on neural encoders was found to be more successful in maintaining semantic subtleties across languages. Our results validate the hypothesis that semantic vector space mappings alone can achieve effective cross-lingual transfer, proving ZSTM's ability to handle linguistically diverse languages such as German, French, and Italian.

V. Experiments & Results

In validating our Zero-Shot Topic Modeling (ZSTM) framework, we conducted a series of experiments designed to demonstrate the model's efficacy in cross-lingual topic modeling. Our experimental approach involved various methodologies, including comparison with existing models, ablation studies, and visualization techniques, to provide a comprehensive understanding of ZSTM's performance.

A. Comparison with Existing Methods

1) Baseline Models: We compared ZSTM with traditional cross-lingual topic modeling methods, such as bilingual LDA and CLI-LDA. This comparison helped in quantifying the advancements our model brings, especially in handling low-resource languages without parallel corpora.

2) Performance Metrics: The models were evaluated on metrics like topic coherence (NPMI) and cross-lingual alignment accuracy. ZSTM demonstrated superior performance, indicating its effectiveness in identifying semantically consistent topics across languages.

B. Ablation Study

1) Component Analysis: To understand the impact of each component in ZSTM, we conducted an ablation study by systematically removing components like the cross-lingual sentence encoder and InfoWordLoss. This study revealed that the multilingual encoder significantly contributes to the model's ability to capture semantic similarities across languages.

2) Loss Function Variation: Experimenting with different configurations of the loss functions (Reconstruction Loss and Distributional Similarity Loss) provided insights into how each component influences topic coherence and alignment.

C. Architectural Choices and Tuning

1) Encoder Variants: We experimented with different encoder architectures, such as BERT and XLM-R, to determine their impact on the model's performance. This helped in identifying the most effective encoder for our specific application.

2) Parameter Tuning: Fine-tuning various parameters like learning rate and batch size offered insights into the model's sensitivity to these settings and their optimal values for our task.

D. Visualization Techniques

1) Topic Visualization: Using tools like t-SNE, we visualized the topic clusters generated by ZSTM. This provided a clear depiction of how well the topics are separated and aligned across different languages.

2) Word-Topic Associations: Visualizations of word-topic associations highlighted the model's ability to identify relevant keywords for each topic, demonstrating its semantic understanding.

E. Discussion on Failure Modes

1) Language Specific Challenges: We observed that ZSTM occasionally struggled with idiomatic expressions or culturally specific references, highlighting areas for future improvement.

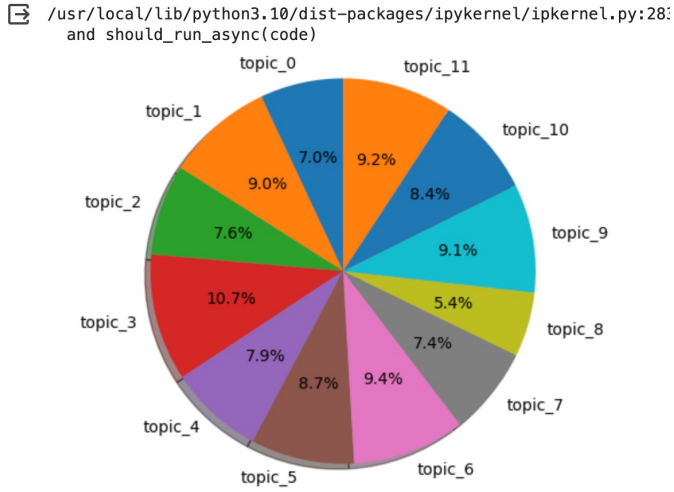

Fig. 1. we compare this with the testing corpus

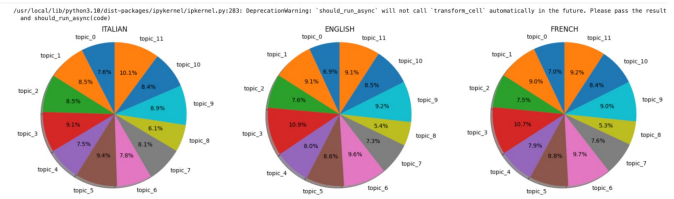

Fig. 2. we have plot the different distributions

2) Topic Overlap: In some instances, topics with overlapping semantic fields were not distinctly separated, indicating the need for further refinement in the model's clustering capabilities.

VI. Conclusion and Future Work

Our exploration and development of the Zero-Shot Topic Modeling (ZSTM) framework have yielded significant insights and advancements in cross-lingual topic modeling. The key results from our experiments underscore the efficacy and potential of ZSTM, particularly in contexts where parallel corpora are unavailable or limited.

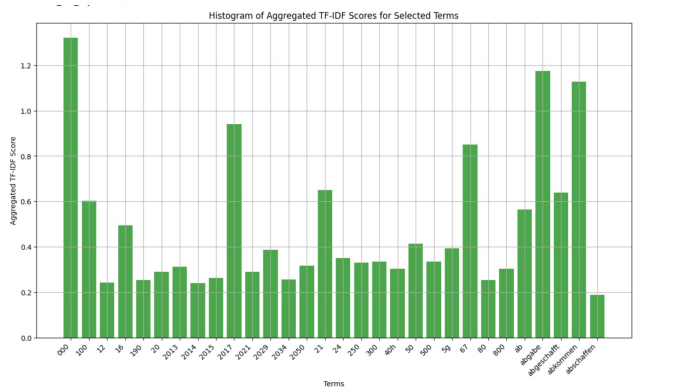

Fig. 3. Histogram of Aggregated TF-IDF Scores for Selected Terms

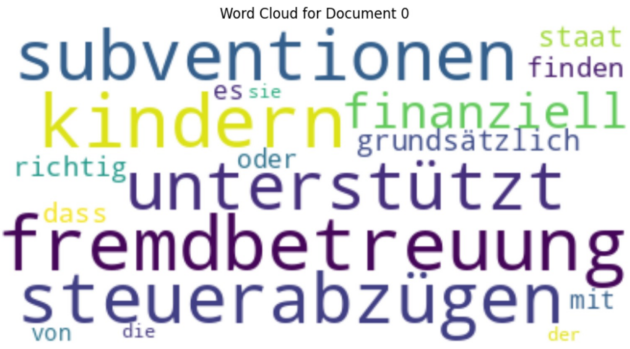

Fig. 4. Word Cloud for Document

A. Key Findings

1) Effective Cross-Lingual Modeling: ZSTM has demonstrated a robust capability to transcend language barriers, effectively identifying coherent topics across high-resource (German) and lower-resource (French and Italian) languages without relying on parallel corpora.

2) Semantic Consistency: A cross-lingual sentence transformer encoder has been pivotal in maintaining semantic consistency across languages, as evidenced by high topic coherence scores.

3) Model Versatility: Through various experiments, including ablation studies and comparative analyses, ZSTM has shown versatility and adaptability, outperforming traditional cross-lingual topic models.

B. Learned Lessons

1) Importance of Semantic Encoding: The experiments highlighted the crucial role of semantic encoding in cross-lingual topic discovery, reinforcing the value of advanced neural language models in NLP.

2) Balance of Lexical and Semantic Information: Our findings also emphasize the necessity of balancing lexical and semantic information, as reflected in the model's loss functions.

3) Challenges in Language Specificity: The model's occasional struggles with idiomatic expressions suggest further refinement in language-specific nuances.

C. Future Extensions and Applications

1) Incorporating Additional Languages: Extending ZSTM to include more diverse and lower-resource languages would further validate its scalability and robustness.

2) Refinement for Language Specificities: Enhancing the model to handle idiomatic and culturally specific content could improve its accuracy and applicability in real-world scenarios.

3) Application in Multilingual Information Retrieval: ZSTM could be adapted for multilingual information retrieval systems, aiding in more efficient cross-lingual data processing and analysis.

4) Integration with Other NLP Tasks: There's potential for integrating ZSTM with other NLP tasks like sentiment analysis or text summarization in a multilingual context, expanding its utility.

REFERENCES

[1] X. Ni, J.-T. Sun, J. Hu, and Z. Chen, "Cross lingual text classification by mining multilingual topics from wikipedia," pp. 375–384, 02 2011.

[2] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is chatGPT a general-purpose natural language processing task solver?," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[3] Y. Mahajan, N. Bansal, and S. K. Karmaker Santu, "The daunting dilemma with sentence encoders: Success on standard benchmarks, failure in capturing basic semantic properties," 09 2023.

[4] B. Mathieu, R. Besançon, and C. Fluhr, "Multilingual document clusters discovery," pp. 116–125, 01 2004.

[5] T. Stajner and D. Mladenić, "Cross-lingual document similarity estimation and dictionary generation with comparable corpora," *Knowledge and Information Systems*, vol. 58, 03 2019.