Scene-Based Movie Search using Multi-Modal Embeddings

Arun Karthik Sengottuvel, Ashwinkumar Venkatnarayanan, Rohil Wattal,
Nitin Sairaj Paruchuri, Arun Prakash Jayakanthan
University of Southern California

Movie identification based on plot is challenging yet fascinating in Natural Language Processing. This report outlines the methodology and results of a project aimed at identifying movies based on their plot descriptions. To achieve its objectives, the project leveraged web scraping, data augmentation, scene separation, similarity checks, and ranking techniques.

1 Tasks Performed

1.1 Dataset Preparation

Two datasets were used for training the model. The first dataset was obtained from TOMT (Tip Of My Tongue) (Arguello et al., 2021) and consisted of plot descriptions. The second dataset was created by web scraping information from the plots of highest-grossing movies. The plots were tokenized using sentence tokenization from the NLTK package. Additional information was obtained from Wikipedia using Beautiful Soup. Preprocessing techniques such as lemmatization and dropping NaN values were applied to the data.

Data augmentation was performed using the Pegasus model for summarizing plots. Since the size of the plots was huge, we summarized them using Pegasus-LARGE (Zhang et al., 2020). Meanwhile, the Parrot Paraphraser generated query sentences from the plots.

1.2 Training

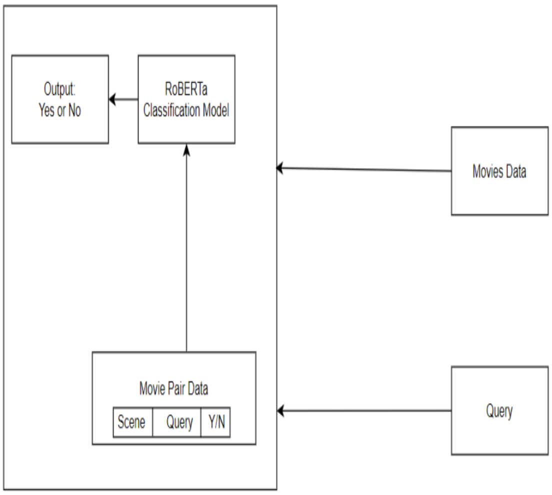
Our approach reframed the similarity between plot and query as a binary classification problem using RoBERTa (Liao et al., 2021). RoBERTa was fine-tuned for sentence pair classification using a pairwise dataset created through manual annotation and weak supervision.

1.3 Similarity Check

Three approaches were used for similarity checks. The first approach was using RoBERTa and the second was with BERT embeddings with cosine similarity on the embeddings produced. The third approach was using BERTScore between the rephrased plots and the query.

1.4 Ranking

Two metrics were used for ranking movies. The first metric was the similarity score between an input and any query in the dataset using BERT embeddings. The second metric was the percentage of positive classifications using RoBERTa. To produce a ranking, the model was fed with the query and each scene of every movie in the dataset, which returned a "Y" (match) or "N" (no match) label. The percentage of positive classifications was then used to produce a "most likely movie" ranking.



1.5 Results

The model achieved an accuracy of 52.56% over the test split using RoBERTa for sentence pair classification. The ranking algorithm produced a list of movies based on the percentage of positive classifications, indicating the likelihood of a match between the query and the movie plot.

1.6 Video Synopsis Generation

In addition to our text-based model development, we have initiated the integration of video embeddings to enrich our scene-based movie search system. Video embeddings involve extracting visual features from movie scenes and encoding them into numerical representations to enhance the accuracy and relevance of our movie identification framework.

Our approach involves the development of a video synopsis generation pipeline, encompassing scene extraction, text generation, and scene-query matching. This pipeline seamlessly integrates with our movie identification framework, enabling enhanced scene understanding and more accurate search results.

2 Risks and Challenges

a) Paraphrasing: The use of paraphrasing techniques may not effectively generate queries representative of typical end-users due to the presence of proper nouns and specific terminology in film media that may not be accurately paraphrased.

b) Model Bias and Performance: There is a risk of bias in the classification approach, particularly towards movies with fewer scenes and multiple similar scenes. This bias could lead to a higher recall scenario, affecting the accuracy of scene-based movie searches.

c) Handling Vague Queries: Answering vague queries presents a challenge, as these queries often lack specificity or contain ambiguous terms. Such queries make it difficult for the model to accurately identify relevant movie scenes, impacting the overall effectiveness of the scene-based movie search system.

d) Computational Resource Constraints for Video Synopsis Generation: Tasks such as scene extraction, feature extraction, and text generation require significant computational resources, and limitations in computing power or memory can lead to longer processing times and reduced system performance.

3 Plan to Mitigate Risks and Address Challenges

a) Paraphrasing: Our strategy involves integrating domain-specific knowledge or resources to improve the accuracy of paraphrased queries. We also plan to implement a feedback mechanism to iteratively refine the paraphrasing process based on user feedback and query relevance.

b) Model Bias and Performance: We intend to benchmark the model against existing approaches and evaluate its performance on diverse datasets to ensure unbiased and robust scene-based movie searches. By exploring ensemble learning techniques to combine multiple models, we can reduce bias in predictions.

c) Handling Vague Queries: We aim to incorporate entity recognition and semantic parsing to extract relevant information from vague queries and enhance query understanding.

d) Computational Resource Constraints for Video Synopsis Generation: We plan to implement algorithm optimization techniques, such as parallelization using frameworks like OpenMP or CUDA, to reduce processing time and enhance scalability.

4 Individual Contributions

Data Collection through online web scraping and accessing existing datasets - Arun Prakash and Nitin Sairaj

Data Augmentation through paraphrasing and summarization - Arun Prakash and Nitin Sairaj

Scene Separation for segmenting movie plots into distinct scenes - Ashwinkumar and Arun Karthik

Training Models - Arun Karthik

Evaluating Similarity Check to assess scene-query resemblances - Ashwinkumar

Development of a multi-modal approach focusing on visual data - Rohil

Creating preliminary design of the embedding model - Rohil

References

Arguello, J., Ferguson, A., Fine, E., Mitra, B., Zamani, H. and Diaz, F., 2021, March. Tip of the tongue known-item retrieval: A case study in movie identification. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (pp. 5-14).

Zhang, J., Zhao, Y., Saleh, M. and Liu, P., 2020, November. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International conference on machine learning (pp. 11328-11339). PMLR.

Liao, W., Zeng, B., Yin, X. et al. An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. Appl Intell 51, 3522–3533 (2021).