Assessing multilingual news article similarity

Introduction: Assessing the similarity between multilingual news articles is a crucial task in various domains such as journalism, information retrieval, and natural language processing. With the increasing availability of news articles in multiple languages, it becomes essential to develop effective techniques to measure the similarity between articles written in different languages. By accurately assessing the similarity, we can facilitate tasks like cross-lingual information retrieval, summarization, and topic clustering.

Objective: The objective of this project is to develop a system for assessing the similarity between multilingual news articles. The system will utilize state-of-the-art natural language processing techniques and machine learning algorithms to capture semantic and contextual information from the articles and provide a measure of their similarity.

Background:

Assessing multilingual news article similarity is a challenging task that requires expertise in natural language processing, machine learning, and cross-lingual information retrieval. Previous work has focused on techniques for representing multilingual text, cross-lingual information retrieval methods, similarity measures, supervised and unsupervised learning approaches, and evaluation metrics. These studies have contributed to bridging the language gap and developing effective models for comparing and clustering news articles in different languages. By building upon this background work, we can develop an algorithm that accurately assesses the similarity between multilingual news articles.

Benefits:

Cross-lingual information retrieval: The developed system will enable users to search for news articles in one language and retrieve similar articles in different languages, facilitating access to information across language barriers.

News clustering and organization: Similarity assessment can help in clustering news articles across different languages, allowing users to explore related articles regardless of the language they are written in.

News summarization: By identifying similar articles across languages, the system can aid in the summarization process by selecting representative articles from different languages to provide a comprehensive overview of a news event.

Methodology:

Data collection: Gather a diverse dataset of multilingual news articles from various sources. The dataset should cover different languages and a wide range of topics.

Preprocessing: Clean and preprocess the articles by removing noise, stopwords, and performing language-specific normalization techniques.

Feature extraction: Utilize language-specific and language-independent feature extraction techniques to capture the semantic and contextual information from the articles. These features can include TF-IDF, word embeddings, topic modeling, and syntactic analysis.

Similarity measurement: Apply appropriate similarity measures such as cosine similarity, Jaccard similarity, or language-specific similarity measures to calculate the similarity between pairs of articles.

Model development: Build a machine learning model (e.g., supervised or unsupervised) to learn the relationship between the extracted features and the ground truth similarity scores, using a subset of the dataset with labeled similarity scores.

Evaluation: Evaluate the performance of the developed model using appropriate evaluation metrics such as precision, recall, F1 score, or mean average precision. Use cross-validation techniques to ensure robustness and generalizability.

System implementation: Develop a user-friendly interface or API that allows users to input news articles in different languages and obtain similarity scores or rankings.

Project Plan: Week 1:

Familiarize with the problem domain and existing literature on multilingual text similarity.

Gather and preprocess a diverse dataset of multilingual news articles.

Week 2:

Explore and implement language-specific and language-independent feature extraction techniques.

Investigate and implement various similarity measurement methods.

Develop a machine learning model for similarity prediction.

Evaluate the model's performance and fine-tune the parameters.

Week 3:

Implement a user-friendly interface or API for the system.

Test the system using sample inputs and evaluate its effectiveness. Write the project report, including the methodology, experimental results, and conclusion.

Revise the report based on feedback from the instructor or peers.

Participant: Venkata Sumanth Nagabhairu (20012395)

By successfully developing a system for assessing multilingual news article similarity, this project will contribute to the advancement of cross-lingual information retrieval and facilitate effective organization and summarization of news content across different languages.