

Research on Vietnamese multi-document summarization technique using graph models

Dao Thanh Chuyen

Abstract: This thesis investigates graph-based models for multi-document summarization in Vietnamese, addressing the challenges posed by the language's analytical structure and lack of inflection. The study proposes a framework that constructs a weighted graph from a cluster of related documents, where nodes represent sentences and edges denote lexical and semantic similarities. Key features include the integration of Vietnamese-specific preprocessing, such as syllable segmentation and part-of-speech filtering, to enhance graph connectivity. The system employs a ranking algorithm, adapted from TextRank, to identify salient sentences, which are then extracted to form a coherent summary. Experimental evaluations on a Vietnamese news corpus demonstrate that the graph-based approach outperforms baseline methods, including centroid-based and frequency-driven techniques, in terms of ROUGE scores. The findings highlight the efficacy of leveraging relational structures among sentences to capture cross-document redundancy and diversity, offering a robust solution for automatic summarization in Vietnamese. This work contributes to the limited research on Vietnamese text processing and provides a foundation for future studies on graph-based natural language generation.

Keywords: Vietnamese multi-document summarization, graph-based model, text summarization, Vietnamese NLP