CONFERENCE PAPERS
Tamil Document Summarization Using Latent Dirichlet Allocation

N. Shreeya Sowmya\(^1\), T. Mala\(^2\)

\(^1\)Department of Computer Science and Engineering, Anna University
\(^2\)Department of Information Science and Technology, Anna University
Guindy, Chennai

\(^1\)shreeya.mel@gmail.com \(^2\)malanehru@annauniv.edu

Abstract

This paper proposes a summarization system for summarizing multiple tamil documents. This system utilizes a combination of statistical, semantic and heuristic methods to extract key sentences from multiple documents thereby eliminating redundancies, and maintaining the coherency of the selected sentences to generate the summary. In this paper, Latent Dirichlet Allocation (LDA) is used for topic modeling, which works on the idea of breaking down the collection of documents (i.e) clusters into topics; each cluster represented as a mixture of topics, has a probability distribution representing the importance of the topic for that cluster. The topics in turn are represented as a mixture of words, with a probability distribution representing the importance of the word for that topic. After redundancy elimination and sentence ordering, summary is generated in different perspectives based on the query.

Keywords- Latent Dirichlet Allocation, Topic modeling

I. Introduction

As more and more information is available on the web, the retrieval of too many documents, especially news articles, becomes a big problem for users. Multi-document summarization system not only shortens the source texts, but presents information organized around the key aspects. In multi-document summarization system, the objective is to generate a summary from multiple documents for a given query. In this paper, summary is generated from the multi-documents for a given query in different perspectives. In order to generate a meaningful summary, sentences analysis, and relevance analysis are included. Sentence analysis includes tagging of each document with keywords, named-entity and date. Relevance analysis calculates the similarity between the query and the sentences in the document set. In this paper, topic modeling is done for the query topics by modifying the Latent Dirichlet Allocation and finally generating the summary in different perspectives.

The rest of the paper is organized as follows. Section 2 discusses with the literature survey and the related work in multi-document summarization. Section 3 presents the overview of system design. Section 4 lists out the modules along with the algorithm. Section 5 shows the performance evaluation. Section 6 is about the conclusion and future work.
II. Literature Survey

Summarization approaches can be broadly divided into extractive and abstractive. A commonly used approach namely extractive approach was statistics-based sentence extraction. Statistical and linguistic features used in sentence extraction include frequent keywords, title keywords, cue phrases, sentence position, sentence length, and so on [3]. Cohesive links such as lexical chain, co-reference and word co-occurrence are also used to extract internally linked sentences and thus increase the cohesion of the summaries [2, 3]. Though extractive approaches are easy to implement, the drawback is that the resulting summaries often contain redundancy and lack cohesion and coherence. Maximal Marginal Relevance (MMR) metric [4] was used to minimize the redundancy and maximize the diversity among the extracted text passages (i.e. phrases, sentences, segments, or paragraphs).

There are several approaches used for summarizing multiple news articles. The main approaches include sentence extraction, template-based information extraction, and identification of similarities and differences among documents. Fisher et al [6] have used a range of word distribution statistics as features for supervised approach. In [5], qLDA model is used to simultaneously model the documents and the query. And based on the modeling results, they proposed an affinity propagation to automatically identify the key sentences from documents.

III. System Design

The overall system architecture is shown in the Fig. 1. The inputs to the multi-document summarization system are multi-documents which are crawled based on the urls given and the output given by the system is a summary of multiple documents.

![Image](image_url)

**Fig. 1 System Overview**

**System Description**

The description of each of the step is discussed in the following sections. The architecture of our system is as shown in Fig. 1.

294
1. Pre-processing

Pre-processing of documents involves removal of stop words and calculation of Term Frequency-Inverse Document Frequency. Each document is represented as feature vector, (ie.,) terms followed by the frequency. As shown in Fig. 1, the multi-documents are given as input for pre-processing, the documents are tokenized and the stop words are removed by having stop-word lists in a file. The relative importance of the word in the document is given by

\[ \text{Tfidf}(w) = \text{tf} \times (\log(N)/\text{df}(w)) \]

where, \( \text{tf}(w) \) – Term frequency (no. of word occurrences in a document)
\( \text{df}(w) \) – Document frequency (no. of documents containing the word)
N – No. of all documents

2. Document clustering

The pre-processed documents are given as input for clustering. By applying the k-means algorithm, the documents are clustered for the given k-value, and the output is the cluster of documents containing the clusters like cricket, football, tennis, etc., if the documents are taken from the sports domain.

3. Topic modeling

Topic models provide a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together. In this paper, Latent Dirichlet Allocation is used for discovering topics that occur in the document set. Basic Idea- Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

Sentence analysis

Multi-documents are split into sentences for analysis. It involves tagging of documents by extracting the keywords, named-entities and the date for each document. Summary generation in different perspectives can be done from the tagged document.

Query and Relevance analysis

The semantics of the query is found using Tamil Word Net. The relevant documents for the given query are retrieved. The relevance between the sentences and the query is calculated by measuring their similarity.

Query-oriented Topic modeling

In this paper, both topic modeling and entity modeling is combined [3]. Based on the query, the topic modeling is done by using Latent Dirichlet Allocation (LDA) algorithm. Query is given as prior to the LDA and hence topic modeling is done along with the query terms. Query may be topic or named-entity along with date i.e. certain period of time.
4. Sentence scoring

The relevant sentences are scored based on the topic modeling. For each cluster, the sum of the word’s score on each topic is calculated, the sentence with the word/topic of high probability are scored higher. This is done by using the cluster-topic distribution and the topic-word distribution which is the result of the Latent Dirichlet Allocation.

5. Summary generation

Summary generation involves the following two steps

5.1 Redundancy elimination

The sentences which are redundant are eliminated by using Maximal Marginal Relevance (MMR) technique. The use of MMR model is to have high relevance of the summary to the document topic, while keeping redundancy in the summary low.

5.2 Sentence ranking and ordering

Sentence ranking is done based on the score from the results of topic modeling. Coherence of the summary is obtained by ordering the information in different documents. Ordering is done based on the temporal data i.e. by the document id and the order in which the sentences occur in the document set.

IV. Results

Table 1 shows the topic distribution with number of topics as 5, the distribution includes the word, count, probability and z value. The topic distribution is for each cluster.

Table 1 Topic model with number of topics as 5

<table>
<thead>
<tr>
<th>TOPIC 0 (total count=1061)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>WORD ID</th>
<th>WORD</th>
<th>COUNT</th>
<th>PROB</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>645</td>
<td>அச்சாதக்கை</td>
<td>42</td>
<td>0.038</td>
<td>5.7</td>
</tr>
<tr>
<td>2806</td>
<td>துறுப்பிட்டி</td>
<td>38</td>
<td>0.035</td>
<td>5.4</td>
</tr>
<tr>
<td>2134</td>
<td>மங்குந்திகள்</td>
<td>36</td>
<td>0.033</td>
<td>5.2</td>
</tr>
<tr>
<td>589</td>
<td>அண்மூப்பிணி</td>
<td>32</td>
<td>0.029</td>
<td>4.8</td>
</tr>
<tr>
<td>1417</td>
<td>அங்கு</td>
<td>27</td>
<td>0.025</td>
<td>2.9</td>
</tr>
<tr>
<td>2371</td>
<td>மங்கொழுப்பு</td>
<td>27</td>
<td>0.025</td>
<td>4.5</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V. Conclusion and Future Work

In this paper, a system is proposed to generate summary for a query from the multi-documents using Latent Dirichlet Allocation. The multi-documents are pre-processed, clustered using k-means
algorithm. Topic modeling is done by using Latent Dirichlet Allocation. The relevant sentences are retrieved according to the query, by finding the similarity between the sentences and the query. Sentences are scored based on the topic modeling. Redundancy removal is done using MMR approach.

Topic modeling can be extended to find the relationship between the entities, i.e. the topics associated with the entity as a future work.

References
