Urdu Text Summarization Module Prototype

Progress Report

Contents

1. Introduction

Text summarization is the process of abridging the text in a way that the summarized form retains the original message of the text. There are two techniques of summarization: extractive and abstractive. The extractive techniques work by selecting the sentences from the original that are strong candidates to be appeared in the summary. The other technique is abstractive that uses natural language generation techniques for generating the summaries.

2. Literature Survey

This section presents the review of work that has been conducted in the area of text summarization. The work in [1] presents an extractive text summarization system for Bahasa Indonesia. The proposed system produces a summary of a given input document on the basis of identification and extraction of important sentences in the document. The system also counts frequency of verbs and nouns because they are supposed to be most representative to the content of the text. The system also uses features such as title of the article and position of the sentence in the document. The subject evaluation method has been used for evaluating the quality of the generated summaries and a performance accuracy of 83.3% has been reported. The objective evaluation has resulted in average f-measure of 78%.

The paper[2] presents a clustering based approach for text summarization. The system first preprocesses the text by applying tokenization, punctuation and stop removal. After that the TF-IDF scores for each term is calculated. The sentences are scored by summing the TF-IDF scores of their words. The sentences are then represented as vectors in one-d space. The k-means algorithm is applied on the sentences. The sentences of the dense cluster are selected as summary sentences.

The system in [3] generates extractive summary of the text by using supervised machine learning approaches. The manual summaries are generated for the selected corpus. The features used for training the models are Mean-TF-ISF, Sentence Length Sentence Position, Similarity to Title, Similarity to Keywords Sentence-to-Sentence Cohesion, Sentence-to-Centroid Cohesion. Two supervised machine learning models used are C4.5 and Naive Bayes. The sentences of the documents are marked as summary and non summary sentence using the information from manually generated summaries. The classifiers are trained for two class problem. The best percision of 37.50% has been achieved by naive bayes classifier for 20% compression rate.

An extractive summarization system for Punjabi language has been proposed in [4]. The features used for scoring the sentences are Sentence length, Keywords, numeric data, NER, nouns, proper nouns and cue phrase. Each feature is assigned a weight and then the weighted features are added

to calculate score of sentences. The weights of the features are calculated using mathematical regression.

The paper[5] presents a text summarization work for Bengali language. The features used in the proposed methodology are TF-IDF, Sentence Position, Sentence Length. The weights are assigned to each sentence and then top sentences are selected as candidates of summary. The corpus of 38 documents has been used and Average Unigram based Recall Score of 41.2% has been achieved.

The work in [6] presents a genetic algorithm based approach for text summarization. The features used for the methodology are: sentence position, positive keywords, Negative keywords, senetence centerality, presence of numerical data, presence of brackets and commas, sentence length, presence of acronym. The 40 documents of English manually summarized documents were used. The precision measure is used for performance evaluation.

A neural network based text summarization model has been presented in [7]. The features for training the network are: Paragraph follows title, Paragraph location in document, Sentence location in paragraph, First sentence in paragraph, Sentence length, Number of thematic words in the sentence and Number of title words in the sentence. The 85 news articles from the Internet with various topics such as technology, sports, and world news to train the network. Each article consists of 19 to 56 sentences with an average of 34 sentences. The entire set consists of 2,'835 sentences. The average accuracy of the real-valued neural network (NI) was 93%, the average accuracy of the discretized real-values into intervals neural network (NJ was 96%, .and the average accuracy of the discretized real-values into single values neural network (N3) was 99% when compared with the human reader’s summaries.

The paper [8] presents a work for Urdu text summarization module in MS Word. The algorithm selects sentences that contain highest percentage of content words. The algorithm first removes all the stop words from the content. The sentences are scored on the basis of percentage of their content words. The top n sentences are selected as summary sentences.

A Latent semantic Analysis based approach for text summarization is presented in [9]. Different LSA-based summarization algorithms are explained in the paper. The author has also proposed two new algorithms. The system was evaluated on different datasets containing English and Turkish documents. The ROUGE measure has been used for evaluating the performance of the system.

A sentence clustering based approach for summarization has been presented in [10]. The sentences are first clustered using k-means and then ranked. The process of summary sentences selection starts from selecting high ranked sentences from each cluster in descending orders of cluster density.

3. Architecture Diagram

The architecture of the Urdu text summarization module is presented in Figure 1 below. The input document is first passed through the pre-processing sub module. The pre-processed text is the processed using the summarization algorithm for generating the document summary.

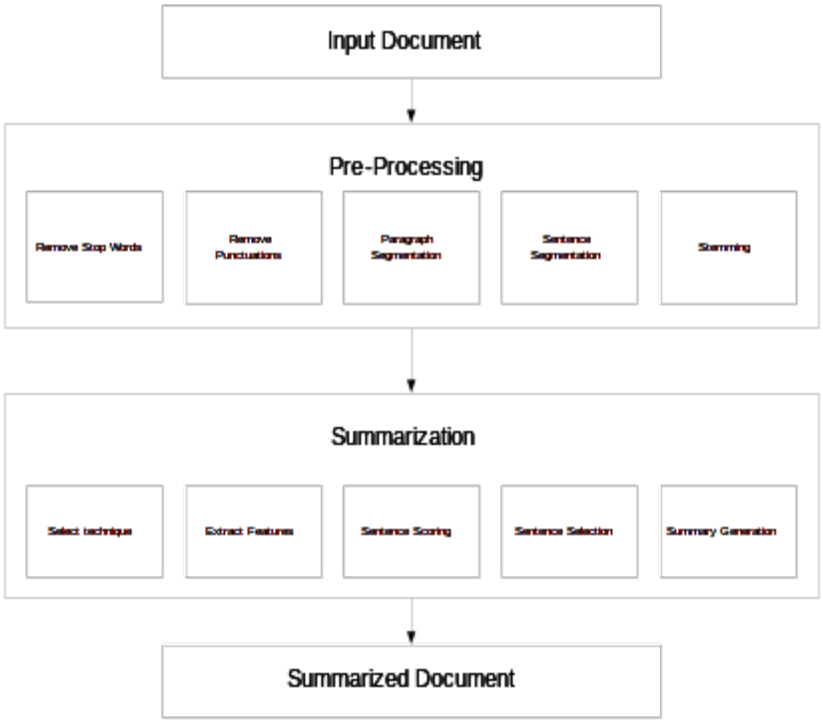

Figure 1. Text summarization module architecture

4. Proposed Methodology

This section explains the methodology that has been followed to implement the text summarization module for Urdu search engine. The Figure 1 below is depicting the steps involved in the procedure of summarization module development.

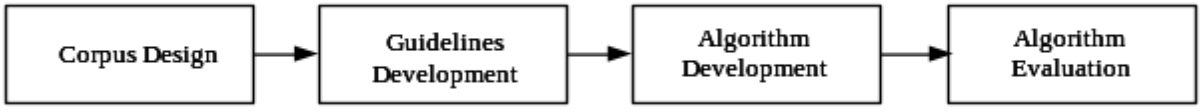

Figure 2. Methodology for text summarization module

4.1. Corpus Design

The first step in implementing the text summarization module for Urdu is the design of text corpus. The CLE Urdu Digest Corpus of 500k [11] has been used for selecting the documents for generating the manual summaries. The 100 documents are selected from the corpus from different domains. The corpus contains the text from the following 15 fields as shown in Table 1 below.

Table 1. Selected corpus domains

1	Book Reviews	9	Press
2	Culture	10	Religion
3	Education	11	Science
4	Entertainment	12	Short Stories
5	Health	13	Sports
6	Interviews	14	Technology
7	Letters	15	Translation
8	Novels		

Seven different files have been selected from each single domain for creating a balanced corpus of documents for generating manual summaries. The summaries of the selected documents have been generated by expert Urdu linguistic. The summaries of three sentences, five sentences and 30% compression have been generated.

4.2. Guidelines development

The next step is development of the guidelines for the human for selecting the summary sentences from the original document. The developed guidelines are presented below.

1. Read the provided text carefully and be sure you understand it.
2. Prepare outline of the text and note down the major points.
3. Select sentences from the text that you think are relevant to central and main ideas of the text.
4. Generate three different summaries of the document in the following manner.
    4.1. Select three sentences for first summary.
    4.2. Select five sentences for second summary.
    4.3. Select 1/3 sentences for third summary.
5. Assign a score to the selected sentences on a scale of 1 to 10.

6. Copy the selected sentences in the summary file in same order in which they appear in original text.

7. The summary should retain the original theme of the text

8. The summary should be coherent. It should not sound like a list of loosely-related sentences.

4.3. Selected Algorithm

We have used the technique presented in [8] for implementing prototype of text summarization module. The working mechanism of this is presented in the algorithm below.

1. Calculate total words of the document.
2. Find all stop words from the document.
3. Remove stop words from the vocabulary.
4. Score the sentences using the equation

*Sentence score=((content words in sentence)/(total content words of document))*100*

5. Sort sentence in descending order of their scores.
6. Pick top N sentences.
7. Arrange the selected sentences in same order as they appear in original document.

The flow of the technique is presented in Figure 2 below.

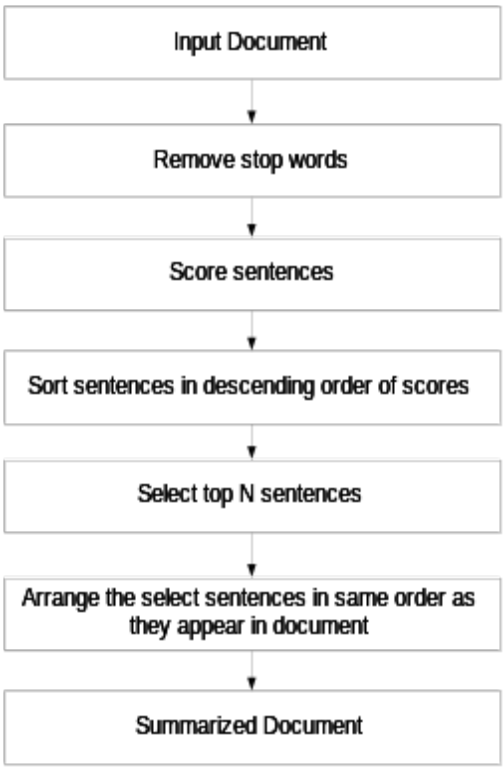

Figure 2. Flowchart of the algorithm

4.4. Evaluation

For evaluating the performance of the algorithm, the summaries of 100 documents generated by the algorithm is compared with the manual summaries of the documents. The following metrics have been used for evaluating the performance of the system.

1. Percentage overlap of sentences between manual and algorithm generated summary.
2. ROUGE [12]

The results algorithm evaluation are presented in Table 2 below.

Table 2. Evaluation results

	Summary Size		
Evaluation Metric	**3 Sentences**	**5 Sentences**	**One third Sentences**
Percentage Overlap	19.47	29.4	**48.29**
ROUGE-2	19.5	31.4	**50.1**

5.Usage of the developed module

The developed module provides functionality of Urdu text summarization to computer systems, mobile phones and SMS.

5.1.Summarization for Computer Systems

The computer systems are used for reading text materials from online resources like web sites as well as from offline resources like stored text documents on user machine. The developed module provides summarization facility to the computer system users by abridging the lengthy contents from different resources. The users are provided with the options of generating summaries of three sentences, five sentences and 30% compression ratio. The module is integrated with the humkinar Urdu search engine. The users can visit humkinar search engine, search for the material and generate the summaries of the desired length.

5.2.Summarization for Mobiles and SMS

The module also facilitates users of mobile phones by providing them summaries of the articles they want to read. The mobile users can search Urdu articles using humkinar search engine and generate summaries. The summaries are sent to the users via SMS. The maximum length of the is 160 characters. So, the most relevant two sentences are sent to mobile users as summary of the article.

6. Conclusion and Future Work

The details of implementing the Urdu text summarization system for Urdu search engine has been presented in this document. The developed system has been evaluated and performance accuracy of 50% has been achieved. The next step will be the performance improvement of the developed module. This improvement will be done by testing other algorithms with different variations and indentifying the areas of existing system that need improvement.

References

1. Fachrurrozi, M., Yusliani, N., & Yoanita, R. U. (2013). Frequent Term based Text Summarization for Bahasa Indonesia.
2. Agrawal, A., & Gupta, U. (2014). Extraction based approach for text summarization using k-means clustering. *International Journal of Scientific and Research Publications*, *4*(11), 1.
3. Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002, November). Automatic text summarization using a machine learning approach. In *Brazilian Symposium on Artificial Intelligence* (pp. 205-215). Springer Berlin Heidelberg.
4. LEHAL, V. G. G. S. (2012, December). Automatic Punjabi text extractive summarization system. In *24th International Conference on Computational Linguistics* (p. 191).

5. Sarkar, K. (2012). Bengali text summarization by sentence extraction. *arXiv preprint arXiv:1201.2240*.
6. Anil Kumar, Jyoti Yadav, Seema Rani (2015).Automatic Text Summarization Using Regression Model (GA).
7. Kaikhah, K. (2004, June). Automatic text summarization with neural networks. In *Intelligent Systems, 2004. Proceedings. 2004 2nd International IEEE Conference* (Vol. 1, pp. 40-44). IEEE.
8. Aqil Burney, B. S., Mahmood, N., Abbas, Z., & Rizwan, K. Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors.
9. Ozsoy, M. G., Alpaslan, F. N., & Cicekli, I. (2011). Text summarization using latent semantic analysis. *Journal of Information Science*, *37*(4), 405-417.
10. Bhole, P., & Agrawal, A. J. Single Document Text Summarization Using Clustering Approach Implementing for News Article.
11. cle.org.pk
12. https://en.wikipedia.org/wiki/ROUGE_(metric)