Cross - Language based Multi-Document Summarization Model using Machine Learning Technique

Ms. P. Mahalakshmi [1,2], Dr. N. Sabiyath Fatima[1] Vishaal Saravanan[1] Mohamed Arshad SS[1]

Department of Computer Science and Engineering, B.S.Abdur Rahman Crescent Institute of Science & Technology, Chennai

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur

Article History:Received:11 november 2020; Accepted: 27 December 2020; Published online: 05 April 2021

ABSTRACT : Cross-Language Multi-document summarization (CLMDS) process produces a summary generated from multiple documents in which the summary language is different from the source document language. The CLMDS model allows the user to provide query in a particular language (e.g., Tamil) and generates a summary in the same language from different language source documents. The proposed model enables the user to provide a query in Tamil language, generate a summary from multiple English documents, and finally translate the summary into Tamil language. The proposed model makes use of naïve Bayes classifier (NBC) model for the CLMDS. An extensive set of experimentation analysis was performed and the results are investigated under distinct aspects. The resultant experimental values ensured the supremacy of the presented CLMDS model.

Keywords: Cross-lingual information retrieval, Document summarization, Machine learning

1. Introduction

Text Summarization (TS) is the task of purifying the essential data from the initial document to offer abbreviated version for particular operation. An important objective of this work is to establish a multi document text summarization framework. Multi Document Summarization (MDS) is one of the well-known and automated process where the required data has been extracted from several input documents. Several representations of models that has been evolved on generating the summary from single and multiple documents. Both single and multiple DS frameworks have experienced massive challenges. Followed by, the primary role in MDS is to collect several resources from the data extraction point as it is comprised of risk with maximum redundancy when it is compared with single document. Furthermore, the sequence of gained data within the coherent text for making a coherent summary is highly complex operation.

Summarization is implemented in the form of abstractive or extractive. Initially, abstractive summarization usually needs data unification, sentence compression as well as reformulation. Secondly, extractive summarization is operated by identifying the prominent factors to the statements of documents. Here, the extracted sentence will have the maximum score derived from consequent summary. In recent days, developers have focused on automated TS called extractive summarization.

Cross-Language Text Summarization (CLTS) is determined as the procedure of examining the document in a language to learn the prominent factors which in turn generatesa short, suitable and accurate summary of the document in specific language. The schemes used for CLTS are divided into TS application which depends upon the extractive [1]. Mostly, the advanced approaches for CLTS make use of extractive class. Nowadays, the systems have employed compressive as well as abstractive frameworks to maximize the usefulness and grammatical supremacy of summaries. But these models need special resources for a language [2] and unification of diverse models which restricts the applicability of these approaches in summary generation in various languages. In addition, the need for specific set of resources [12] would be of great impact in developing the applications.

This paper presents a novel query optimization with Cross-Language Multi-document summarization model. The proposed model enables the user to provide a query in Tamil language, generate a summary from multiple English documents, and finally translate the summary into Tamil language. The proposed model involves naïve Bayes classifier (NBC) based document summarization. Extensive simulation analysis is performed to highlight the effective outcome of the presented model.

2. Related works

Cross-Language Automatic TS (CLATS) resolves to produce a summary in which the language varies from document language. The cross-lingual based set of documents [13] have been retrieved for all the queries, where the expansion terms are identified by term selection value. Traditionally, CLDS examined the data in single language [3]. The cross-lingual based set of documents has been retrieved for all the queries where the expansion terms are identified by term selection value. Typically, 2 CLATS methods were applied namely, early

[*] Corresponding author: Ms. P. Mahalakshmi
Department of Computer Science and Engineering, B.S.Abdur Rahman Crescent Institute of Science & Technology, Chennai

and late translations. Initially, the source documents are converted into the target language and summarize the converted documents under the application of data in changed sentences. Secondly, the document is summarized with the help of abstractive, compressive, and extractive approaches and then convert into summary of desired language. Some of the current models have increased the supremacy of cross-lingual summarization by translation quality value as well as data of documents in all languages [4,5].[7] applied a Support Vector Machine (SVM) regression approach for predicting the translation quality of pair of English-Chinese sentences from fundamental features like sentence length, sub-sentence value, proportion of nouns and adjective, as well as parse features.[8] trained ϵ-Support Vector Regression (ϵ-SVR) for the purpose of predicting the score of translation quality according to the automatic NIST measure as quality indicator. It generates the translated English documents to French under the application of Google Translate and examined the features for estimating the conversion excellence of a sentence. Based on the phrase-relied translation schemes, [9] developed a phrase-based approach for computing the sentence scoring, extraction as well as compression. Followed by, a scoring model has been deployed for CLATS task relied on a submodular norm of reduced sentences.

[15] adapted the analysis of 3-gram for multi-sentence compression in order to cluster similar sentences. [16,17] has enhanced the analysis of the relevance related to the coherence of words and the keywords on the sentence-based compressions.[14] associated the sentence and multi-sentence compression procedure for the generation of cross-lingual summaries. The cohesion metrics is utilized for producing clusters of similar sentences in multi-sentence compression procedure.

3. NBC based Multi-document summarization

Figure 1 represents the block diagram for multi-document summarization using NBC. The documents are processed and scored the sentences based on the centroids. The given query has been translated to the English language, and based on the classifier score the sentences are summarized and converted to the query language.

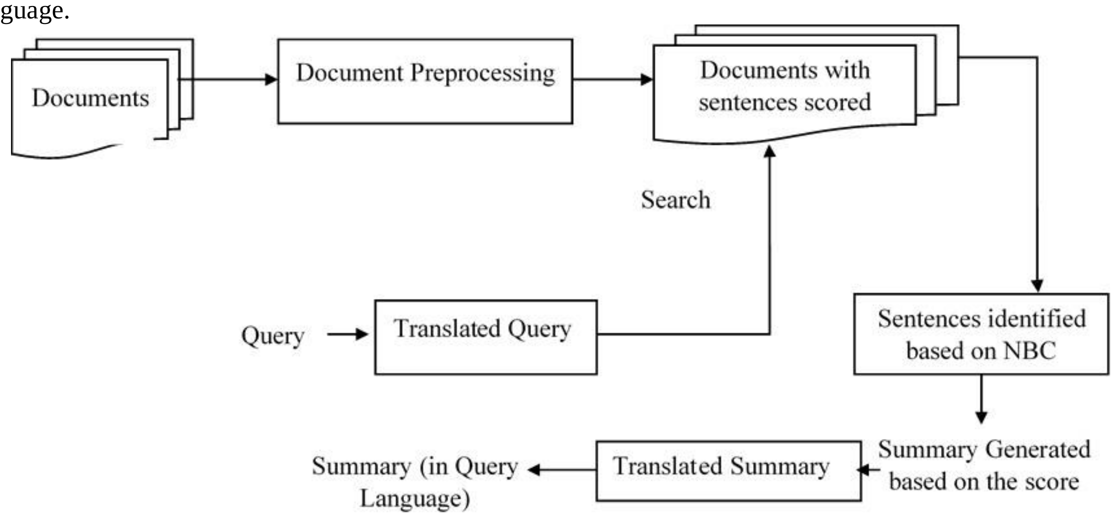
Documents
Document Preprocessing
Documents with sentences scored
Search
Query
Translated Query
Sentences identified based on NBC
Summary Generated based on the score
Translated Summary
Summary (in Query Language)

Figure 1 Multi-Document Summarization using NBC

NBC model is applied for summarizing multiple documents. It is used to extract the highly related details from many documents. In addition, scoring mechanism is utilized for the calculation of the word scoring of the words to attain word frequency. Actually, keywords are defined as significant tools that offer limited summary of a text document. In this work, an automatic keyword extraction method has been developed. It is used for identifying the keywords from training set which depends upon the frequencies and location. To extract the summary from prominently applied documents, Naïve Bayesian (NB) classification model is applied for the supervised learning. Assume the input is the collection of text documents [6,10]. The major objective of this work is to filter the prominently utilized documents and produce small paragraph that saves the significant information available in the source document. There are 7 multistage compression phases as shown in the following:

- The collection of documents is subjected to computation.
- From the group of documents, regularly applied documents are decided by a system for further computation.
- Preprocessing is carried out and the sentences are divided as to words.
- The score is measured for a word with the help of Bayesian classifier.
- For a sentence group, a sentence level representation is chosen.
- The sentence level representation is created as linear sentence and compressed whenever required.

Initially, the compression is applied to reduce the complexity in representation with diverse approaches. The score value is evaluated on the basis of given expression:

$$Score\ (S_\mathrm{i}) = \sum \left(wcC_{\mathrm{i},k} + wpP_{i,k} + wfF_{i,k}\right), \qquad (10)$$

Where $C_{i,k}$ indicates the centroid value estimated from (11). $P_{i,k}$ denotes the positional value measured from (12). $F_{\mathrm{i},k}$ refers first sentence overlap value derived from (13). wc, wp, wf represents the weights. The metric of centroids for sentence $S_{i,k}$ is calculated from the normalized sum of centroid elements and is given as:

$$C_{i,k} = \sum \frac{\left(\mathrm{TF}(\alpha_i) * \mathrm{IDF}(\alpha_i)\right)}{|\mathcal{R}|}, \alpha \in S_{i,k}, \qquad (11)$$

Where $C_{\mathrm{i},k}$ denotes the optimized sum of centroid measures. The positional measure for a sentence is measured from the given expression:

$$P_{i,k} = \frac{n-i+1}{n} * C_{\max}, \qquad (12)$$

In which $C_{\max}$ denotes the maximum centroid value of a sentence. The overlap score can be calculated as the product of sentence vectors in recent sentence i and initial sentence of a document. The first sentence overlap is determined from the applied function:

$$F_{i,k} = \overrightarrow{S_{\imath,k}} * \overrightarrow{S_{1,k}}. \qquad (13)$$

The ML approaches consider keyword extraction as classification issue. The procedure follows the word in a document that comes under the class of normal words. Bayesian decision theory is one of the commonly used statistical methods that depending upon the tradeoffs among the classification decisions with the help of cost as well as probability. The words involved in a document has to be classified using keyword features for examining the word and declare as a key. The term frequency is evaluated on the basis of count of iterations in a word. Basically, prepositions are not comprised of value as it is a keyword. When the word has maximum frequency when compared with alternate documents then it is decided whether it is a keyword or not. The integration of these features results in metric TF $*$ IDF measures. TF denotes the frequency and IDF indicates the inverse document frequency. It can be also a benchmark measure applied for data extraction.

Word α present in a document $\mathcal{R}$ is calculated in the following:

$$TF\ *\ IDF\ (P, \mathcal{R}) = P\ (word\ in\ \mathcal{R}\ is\ \alpha)\ *(-\ log\ P\ (\alpha\ in\ a\ document)). \quad (14)$$

(i) It can be measured by computing the iterations of words suited in a document and classify them into the overall number of words.

(ii) It can be determined by calculating the overall count of documents in a training set, in which the word exists except $\mathcal{R}$, and classifying the measure by documents in a training set.

Actually, the keyword is concentrated in both starting and endpoints of the text. At this point, the required keywords are found initial point of a sentence. Bayes theorem has been employed for computing the possibility of a word and described in the following:

$$\frac{P\ (T|\ \mathrm{key})\ * P(\mathcal{R}|\ \mathrm{key})\ * P\ (\mathrm{PT}\ |\ \mathrm{key})\ * P\ (\mathrm{PS}\ |\ \mathrm{key})}{P(\mathrm{key}\ |T, \mathcal{R}, \mathrm{PT}, \mathrm{PS})} \qquad (15)$$

Once the NB is completed, the summary is generated according to the Score and applying timestamp. The last summary is produced on the basis of a score. When the documents are computed by NB, sentences are organized on the basis of a score. It is pointed that sentence from initial document has to be placed prior to the sentences decided from upcoming document. Thus, series of sentences from a document summary is not reasonable in the incidence. This problem can be resolved using timestamp strategy.

4. Performance Validation

This section validates the results analysis of the presented model on distinct aspects. Table 1 and Fig. 1 investigates the results of document summarization models in terms of different measures [11]. On examining the summarization outcome in terms of precision, the presented NBC model has resulted in a maximum precision of 85.78% whereas an average precision of 78% has been achieved by the CT-LC model. Likewise, on investigative summarization results with respect to recall, the proposed NBC method has resulted in a superior recall of 82.84% while an average recall of 77.7% has been obtained by the CT-LC manner. Similarly, on determining the summarization outcome in terms of F-measure, the projected NBC technique has resulted in a higher F-measure of 91.64% but an average F-measure of 86% has been reached by the CT-LC methodology.

Table 1 Result Analysis of Various Method on Document Summarization in terms of Precision, Recall and F-Measure

Measures	**NBC**	**CT-LC**
Precision	85.78	78.00

Recall	82.84	77.70
F-measure	91.64	86.00

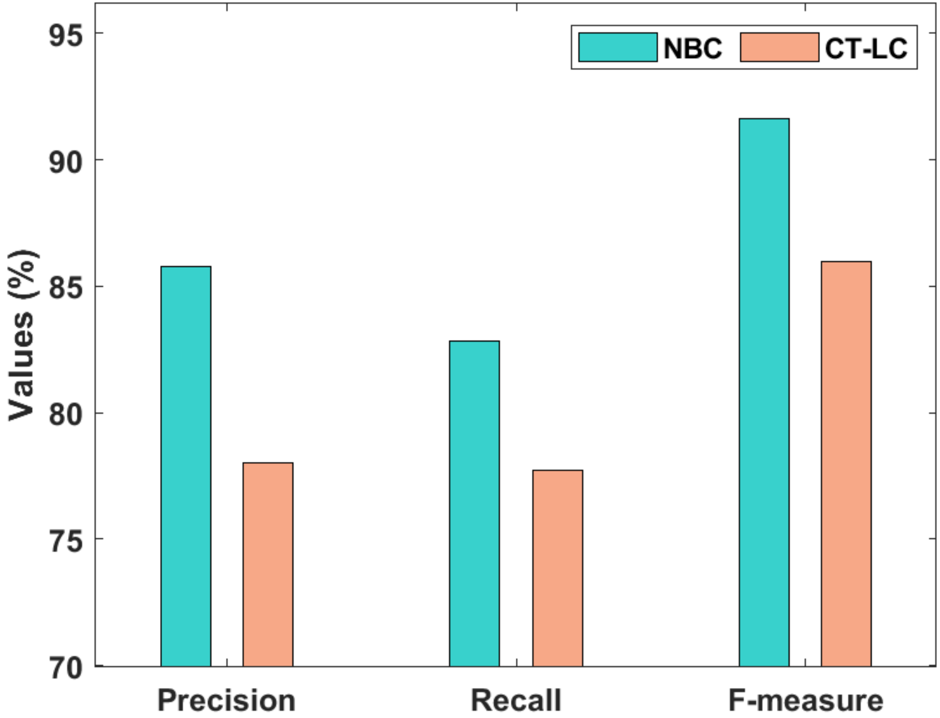

Fig.1. Comparative analysis of Document Summarization with different measures

5. Conclusion

The newly developed approach is used of enabling the user to apply a query in Tamil language, produce a summary from several English documents, and convert the summary into Tamil language. The proposed system generates the summarization of numerous documents in various languages using NBC mechanism. An extensive set of experimentation analyses is carried out and simulation outcomes are examined under various factors. Finally, the experimental values make sure the quality of newly developed CLMDS approach.

6.References

L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, H. Li, Generative adversarial network for abstractive text summarization, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018.

Orasan, C., Chiorean, O.A.: Evaluation of a Cross-lingual Romanian-English Multi-document Summariser. In: 6th International Conference on Language Resources and Evaluation (LREC) (2008)

J. Zhang, Y. Zhou, C. Zong, Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing, IEEE/ACM Trans. Audio Speech Lang. Process. 24 (10) (2016) 1842–1853.

Zhang, J., Zhou, Y., Zong, C.: Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. IEEE/ACM Trans. Audio, Speech & Language Processing 24(10), 1842–1853 (2016)

Wan, X.: Using bilingual information for cross-language document summarization. In: ACL. pp. 1546–1555 (2011)

Mahalakshmi, P., Fatima, N.S. Ensembling of text and images using Deep Convolutional Neural Networks for Intelligent Information Retrieval. Wireless Pers Commun (2021)

Wan, X., Li, H., Xiao, J.: Cross-language document summarization based on machine translation quality prediction. In: ACL. pp. 917–926 (2010)

Boudin, F., Huet, S., Torres-Moreno, J.: A graph-based approach to cross-language multi-document summarization. Polibits 43, 113–118 (2011)

Yao, J., Wan, X., Xiao, J.: Phrase-based compressive cross-language summarization. In: EMNLP. pp. 118–127 (2015)

Ramanujam, N. and Kaliappan, M., 2016. An automatic multidocument text summarization approach based on Naive Bayesian classifier using timestamp strategy. *The Scientific World Journal*, *2016*.

S. Saraswathi and R. Arti, “Multi-document text summarization using clustering techniques and lexical chaining,” ICTACT Journal on Soft Computing, vol. 1, no. 1, pp. 23–29, 2010.

P.Mahalakshmi and N.Sabiyath Fatima, “An art of review on Conceptual based Information Retrieval”, Webology Journal, volume 18, issue no. 1, pp. 51-61, 2021.

Chandra, G. and Dwivedi, S.K., 2020. Query expansion based on term selection for Hindi–English cross lingual IR. Journal of King Saud University-Computer and Information Sciences, 32(3), pp.310-319.

ElvysLinhares Pontes, Stéphane Huet, Juan-Manuel Torres-Moreno and Andréa Carneiro Linhares, “Compressive approaches for cross-language multi-document summarization”, Data & Knowledge Engineering, vol, 125, 2020.

Linhares Pontes E., Huet S., Gouveia da Silva T., Linhares A.C., Torres-Moreno J.-M, “Multi-sentence compression with word vertex-labeled graphs and integer linear programming”, Proceedings of TextGraphs-12: The Workshop on Graph-Based Methods for Natural Language Processing, Association for Computational Linguistics, 2018.

P. Mahalakshmi and N. S. Fatima, "Collaborative Text and Image based Information Retrieval Model using BiLSTM and Residual Networks," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 958-964, 2020

Wan X., Luo F., Sun X., Huang S., Yao J.-g, “Cross-language document summarization via extraction and ranking of multiple summaries”, Knowledge Information System,2018.