International Journal of Advanced Trends in Engineering and Technology (IJATET)
Impact Factor: 5.965, ISSN (Online): 2456 - 4664
(www.dvpublication.com) Volume 3, Issue 1, 2018

COMPARISON OF TEXT SUMMARIZER IN INDIAN LANGUAGES

D. K. Kanitha*, D. Muhammad Noorul Mubarak & S. A. Shanavas***

* Department of Linguistics, University of Kerala, Kariyavattom, Thiruvananthapuram, Kerala

** Department of Computer Science, University of Kerala, Kariyavattom, Thiruvananthapuram, Kerala

Cite This Article: D. K. Kanitha, D. Muhammad Noorul Mubarak & S. A. Shanavas, "Comparison of Text Summarizer in Indian Languages", International Journal of Advanced Trends in Engineering and Technology, Volume 3, Issue 1, Page Number 79-82, 2018.

Abstract:

Text summarization is the process of extracting the relevant information from a source text keeps the significant information. Mainly two types of text summarization methods such as abstractive and extractive. The extractive summarization ranks all sentences and high scored sentences are selected as summary. The abstractive summarization understands the content of a document and re-state in few words. This paper discusses about various text summarization methods followed by the Indian languages. The existing algorithms are explained and then the merits and demerits are discussed. This paper also investigates which method is suitable for summarizing documents in Indian languages.

Key Words: Text Summarization, Extractive Methods, Abstractive Methods & Natural Language Processing

1. Introduction:

The new era of digital world abundance of text materials are available in Internet in any topic. The developments of Natural Language Processing many documents are available in Natural Languages. Read every page and find the relevant information takes lot of time and effort. At the same time text summarizers generate summary of a text within a limited time. So user can recognize whether it is relevant or not.

Text summarization is a technique that creates summary of a text. The summarization systems begin in early 1950s. The earlier works are focused on word frequency and phrase frequency. Then different machine learning algorithms are used for text summarization. Now the statistical, algebraic methods and language processing tools are used for create a summary.

Mainly two types of text summarization such as extractive and abstractive. Extractive summarization extracts the important topics and creates a subset of main text. Abstractive summarization generates a new text from the source text. Linguistics techniques and theories are required for summarization. Today most of the systems follow the extractive based summarization. The extractive summarization systems require less memory space and easy implement. Microsoft word auto summarizer is an example of simple extractive based summarizer.

The rest of the paper is organized as follows. Section2 describes the various algorithms used by the Indian languages. Section 3 describes the merits and demerits of various algorithms. Section 4 shows the comparative study of various models. Finally Section 5 describes the conclusion of the results.

2. Text Summarizer for Indian Languages:

Numerous automatic text summarization systems are available in all languages especially in English and other foreign languages. The Indian languages, summarization models are very few. Some of the text summarizer for Indian languages is discussed below.

2.1 Auto Summarizer for Urdu [1]: The authors proposed a simple sentence scoring algorithm to rank the sentences.

Algorithm:

- ✓ Step 1: Calculate total words within the document.
- ✓ Step 2: Remove the stop words from the input text
- ✓ Step 3: Calculate the content words in each sentence.

  Content words=Total words-Stop words
- ✓ Step 4: Calculate the sentence weight

  Sentence weight= (Content words/Total words) *100
- ✓ Sort the sentences.
- ✓ Generate the summary

2.2 Auto Summarizer for Panjabi [2]: Numerous algorithms are used for Punjabi text summarization. The summarizer systems follow abstractive and extractive methods. The abstractive method requires the linguistics tools for recognize the source document. Named Entity Recognizer, WordNet, Semantic Parser etc are required for summarization. The extractive based summarization use simple sentence weight learning method.

Algorithm simple sentence weight learning method:

- ✓ Step 1: Split the text into sentences and words.

- ✓ Step 2: Find the number of words in each sentence
- ✓ Step 3: Find the number of words in maximum lengthy sentence
- ✓ Step 4: Calculate the sentence score.

  Sentence score= Number of words/Number of words in maximum length sentence
- ✓ Step 5: Rank the sentences and highest ranking sentences as summary

2.3 Auto Summarizer for Kannada [3] [4] [6]: Some of the methods are used for Kannada text summarization are sentence weight score [3], tf-idf weighting scheme [4], template based abstractive summary method [5] etc. Sentence weight score- This system uses adjectives, adverbs and nouns as key terms.

Algorithm for Sentence Weighting [3]:

- ✓ Step 1: Split the text into sentences and words.
- ✓ Step 2: Calculate the position score of each sentences. The first sentences in each document got highest score.
- ✓ Step 3: Add additional score to numeric held sentences.
- ✓ Step 4: Assign score to the keywords
- ✓ Step 5: Calculate sentence score. Sentence score is the sum of words in the sentences.
- ✓ Step 6: Assign feature weight to the sentences. Sum all the feature score and extract the important sentences

The tf-idf method [4] gives the weight to the sentences and highest scoring sentences are extracted as summary sentences. The abstraction methods [5] follow the Natural Language Generation Techniques. Kansum [6] follow statististical based methods for summarization.

2.4 Bengali Language [6] [7] [8] [9]: Islam and Masum (2004) developed a summarizer for Bengali language called 'Bhasa'. The sentences are scored on the basis of matching between query vector and sentence vector. Das and Bandyopadhyay (2010) developed a Bengali opinion text summarizer based on sentiment analysis of the text. Kamal Sarkar (2012) proposed a model based on tf-idf score, sentence length and position score.

Algorithm:

- ✓ Step 1: Segment the text into words and sentences
- ✓ Step 2: Calculate the value of thematic terms using tf-idf score.
- ✓ Step 3: Calculate the position score (1/position of sentence).
- ✓ Step 4: Calculate the length of sentences.
- ✓ Step 5: Calculate the total score and extract the highest scoring sentences in proper order.

2.5 Auto Summarizer for Tamil [10]: M. Banu, C. Karthika, P. Sudarmani and T. V. Geetha, proposed an abstractive based summarization system based on semantic graph method. Kumar and Devi (2011) proposed Tamil language summarization system for scoring of sentences in summary using graph theoretic scoring technique. This system uses statistics of frequency of words and a term positional and weight-age calculation by string pattern for scoring of sentences.

Algorithm (Graph Theoretic Method):

- ✓ Step 1: Split the text into words and sentences.
- ✓ Step 2: Construct graph and represents each vertex as sentences and edges shows the occurrence of words in the sentences.
- ✓ Step 3: Calculate the total number of words.
- ✓ Step 4: Find the affinity weight (aw) of word

  Aw=document frequency of a word/total words.
- ✓ Step 5: Calculate the sentence weight. It is the sum of affinity weight.
- ✓ Step 6: Calculate the Levenshtein similarity weight. It is difference between maximum length of two sentences and Levenshtein distance (LD) of two sentences then it is divided by maximum length of two sentences. Levenshtein distance is the distance between two words.
- ✓ Step 7: Calculate the vertex weight.
- ✓ Step 8: Rank the sentences on the basis of similarity weight and vertex weight.

M. K. Keyan and K.G. Srinivasagan, (2012) [21] proposed a model based on vector space model and Neural Networks. The method is suitable for multi documents and it does summarize the news articles in English and Hindi. Vector space model is used for rank the single document and it is the input to the multi document. Neural network extract the significant sentences.

2.6 Auto Summarizer for Hindi [16] [18] [19]: The authors proposed a statistical, linguistics and heuristics method is used for summarization. Chetan Thaokar and Latesh Malik [18] an extractive approach to genetic algorithms for ranking the sentence. K Vimal Kumar and Divakar Yadav [19] proposed an extractive approach to summaries the text here scores of the sentence is calculated based on occurrence of word in the sentences.

2.7 Auto Summarizer for Malayalam [13] [14]:

Algorithm [13]:

- ✓ Step 1: Split the text into words and sentences
- ✓ Step 2: Calculate the position score of sentences.

- ✓ Step 3: Calculate length of score of each sentence.
- ✓ Step 4: Calculate term frequency of each word and find the weight of sentences.
- ✓ Step 5: A combination function is used to score the sentences and extract the high scored sentences.

Maximum Marginal Approach [14] is used to find the relevance of each term using a word Dictionary. On the basis of relevance score high score sentences are selected as summary.

2.8 Auto Summarizer for Odia [10]:

Algorithm:

- ✓ Step 1: Calculate the total number of terms within the text.
- ✓ Step 2: Calculate the weight of each term.

Weight of a term= frequency of a term/total number of terms

- ✓ Step 3: Calculate the weight of sentences

Weight of sentence=Sum of all the weighted term in a sentence/number of terms in the sentence.

3. Advantages and Disadvantages of Indian Summarizers:

An abstractive summarization method requires heavy language processing tools. The tools in Natural language processing are developing stage. So building this tool requires large memory space and heavy machineries. At the same extractive based tools are easy to implement. The literature most of the works in Indian languages are extractive based sentence ranking methods. These algorithms are mainly follows two steps pre processing step and processing steps. The pre-processing step the unstructured text is converted into structured. The stop words are eliminated and extract the content words. The processing steps extract the feature of sentences and give feature score to the sentences. The highest ranking sentences are selected as summary. The advantages and disadvantages are shown in Table1:

Table 1: Advantages and disadvantages of methods

Languages	Methods	Advantages	Disadvantages
Urdu	Extraction-Sentence weighting scheme	Easy to implement	Doesn't consider the semantics of sentences
Panjabi	Simple weight learning	Simple	Less semantics
	Liguistics based	More semantics	Requires language related resources
Kannada	Feature weight algorithm	Accurate	Complex algorithm
	Tf-idf	Keywords with similar sentences are extracted as summary	Less semantics
Bengali	Vector space term matching	Extract query held sentences	Determine the thematic term is difficult.
	Sentiment analysis	More semantics	Difficult to compute
	Tf-idf, position, sentence length	Accurate	Duplication in summary
Tamil	Graph based	Suitable for agglutinative languages	Difficult to implement
Hindi	Extractive-Linguistics theories	Semantically related sentences	Difficult to compute and domain dependant
Malayalam	Extractive and abstractive	More semantics	Abstractive requires training data
Odia	word and sentence weighting algorithm	Simple	Less semantics

4. Comparison of Summarization Models in Various Indian Languages:

The methods used by the summarizers are extracting the important sentences and show the overall idea about the topic. The Indian language summarizers are follows simple extractive methods. The implementation of machine learning algorithms requires the trained data for summarization. The statistical and linguistic methods increase the accuracy of the summarizer and extract the summary without the use of trained data. The pre-process of text is done by linguistics theories and ranking is performs by statistical method increase accuracy of summary. The statistical and algebraic method Latent Semantic analysis [16] extracts the semantically similar sentences. It extract the semantically similar sentences on the basis of word co-occur in the sentences. LSA based model is proposed in Kannada text summarization. The LSA based and graph based method extract semantically similar sentences and it is suitable for Indian languages.

5. Conclusions:

This paper explains some of the algorithms used by the Indian language summarizers. Most of the systems in earlier stages are followed the sentence scoring methods. These methods are domain independent and it is suitable for all languages. The summarizers are not producing the complete summary about the document but it give an idea about the document. Recently some systems follow the linguistics, statistical and heuristic

techniques for summarization. It is very useful to the user because it consider the semantic of sentences extract the relevant sentences as summary. It is efficient for single and multi document summarization.

6. References:

1. Aqil Burney, Badar Sami, Nadeem Mahmood, Zain Abbas and Kashif Rizwan, (2012), "Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors" Pakistan International Journal of Computer Applications (0975 – 8887) Volume 46– No.19.
2. Vishal Gupta, Gurpreet Singh Lehal, "Features Selection and Weight learning for Punjabi Text Summarization", International Journal of Engineering Trends and Technology- Volume2 Issue2- 2011.
3. Jagadish S Kallimani., Srinivasa K, G., (2010) "Information Retrieval by Text Summarization for an Indian Regional Language", IEEE.
4. Jayashree.R. Srikanta Murthy. & Sunny.K. (2011). "Document summarization in kannada using keyword extraction", CS & IT-CSCP, pp. 121-127.
5. Embar, V.R., Deshpande, S.R., &Vaishnavi, A.K., (2013). "sArAmsha- A Kannada Text Summarizer", Advances in computing, ICACCI, International Conferencence on 22-25 Aug. 540-544, IEEE.
6. J. S. Kallimani, K. G. Srinivasa and B. R. Eswara, "Summarizing News Paper Articles: Experiments with Ontology Based, Customized, Extractive Text Summary and Word Scoring", Journal of Cybernetics and Information Technologies, Bulgarian Academy of Sciences, vol. 12, pp. 34-50, 2012.
7. T. Islam and S. M. A. Masum, (2004). "Bhasa: A Corpus Based Information Retrieval and Summarizer for Bengali Text," Macquarie University, Sydney, Australia.
8. K. Sarkar, (2012)."Bengali text summarization by sentence extraction," In Proceedings of International Conference on Business and Information Management (ICBIM-2012), NIT Durgapur, pp. 233-245.
9. K. Sarkar, (2012). "An approach to summarizing Bengali news documents," In proceedings of the International Conference on Advances in Computing, Communications and Informatics, ACM, pp. 857-862.
10. A. Das and S. Bandyopadhyay, (2010). "Topic-Based Bengali Opinion Summarization", International Conference COILING '10, Beijing, pp. 232–240.
11. Sankar K, Vijay Sundar Ram R and Sobha Lalitha Devi, (2011). Text Extraction for an Agglutinative Language, Problems of Parsing in Indian Languages, Special Volume.
12. R. C. Balabantaray, B. Sahoo, D. K. Sahoo, M. Swain, (2012). Odia Text Summarization using Stemmer, International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868, Volume 1– No.3, 2012.
13. Nilofar Mulla, Shital K. Dhamal, (2016). "A Survey of Text Summarization Techniques for Different Indian Regional Languages", International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 8.
14. Dhanya, P. M., and M. Jathavedan, (2013). NCILC seminar proceedings.
15. Renjith. S. R, Sony.P, (2015). "An automatic text summarization for Malayalam using sentence extraction", IRF International Conference, ISBN: 978-93-85465-35-2.
16. Ajmal E.B, Posna P Haron, (2015) "Summarization of Malayalam Document Using Relevance of Sentences" International Journal of Latest Research in Engineering and Technology, Volume I Issue 6 pp 08-13.
17. Anjana T G (2017). "Summarizing Malayalam News Articles Using Topic Modeling" International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 6, Issue 10.
18. Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T. (1983). Statistical semantics: Analysis of the potential performance of key-word information systems. Bell System Technical Journal, 62(6), 1753-1806.
19. U. Garain, A. K. Datta, U Bhattacharya and S.K. Parui, (2006), Summarization of JBIG2 Compressed Indian Textual Images,‖ Proceeding of 18th International Conference onPattern Recognition (ICPR'06), IEEE, Kolkata, India, Vol. 3, Pp. 344-347, 2006.
20. Chetan Thaokar And Latesh Malik " Test Model for Summarizing Hindi Text using Extraction Method"In Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013
21. Kumar, K. Vimal, and Divakar Yadav. "An Improvised Extractive Approach to Hindi Text Summarization." In Information Systems Design and Intelligent Applications, pp. 291-300, Springer India, 2015
22. Keyan, M.K., & Srinivasagan,K.G., "Multi-Document and Multi-Lingual Summarization using Neural Networks", Proceedings of International Conference on Recent Trends pp. 11-14, 2012.
23. M. Banu, C. Karthika, P. Sudarmani and T.V. Geetha, "Tamil Document Summarization Using Semantic Graph Method", International Conference on Computational Intelligence and Multimedia Applications, IEEE, pp. 128-134, 2007.