METHODOLOGY OF TEXT SUMMARIZATION IN LOW-RESOURCE LANGUAGES LIKE THE PUNJABI LANGUAGE

Er. Rahul Garg[1] and Dr. Naresh Kumar Garg[2]

[1]Research Scholar, Department of CSE, GZSCCET, Maharaja Ranjit Singh Punjab Technical University, BATHINDA, Punjab, India
[2]Professor, Department of CSE, GZSCCET, Maharaja Ranjit Singh Punjab Technical University, BATHINDA, Punjab, India

ABSTRACT

Text summarization in last few years have witnessed significant advancements and showcased various approaches and techniques. While there were notable strengths, it is important to consider the limitations and challenges that emerged during this period. One of the strengths observed in text summarization literature in last few years was the continued progress in deep learning models. Transformer-based architectures, such as BERT and GPT, continued to dominate the field and achieved impressive results in summarization tasks. These models effectively captured the contextual information and semantic relationships in the text, leading to more accurate and coherent summaries. However, there is much less work done in low resource languages like Punjabi due to their complex contextual structure and low availability of standard dataset. In this paper, we have presented a methodology that can be followed to generate summaries of Punjabi Text.

KEYWORDS

Text Summarization, Punjab Language, Punjabi Text, Summaries, Machine Learning.

1. INTRODUCTION

Text Summarization is defined as the process of condensing the given text document while maintaining its meaning along with the context of the original document [1]. It is a crucial method for quickly and easily locating specific useful information in enormous amounts of text. It is the process of creating a shortened version of a text by extracting the most important information from the original source. This can be useful for condensing lengthy documents or articles, and for providing a quick overview of the main points. There are various methods for performing text summarization, including techniques that use natural language processing (NLP) to identify key phrases and sentences, and those that rely on algorithms to identify and extract important information. The goal of text summarization is to create a condensed version of the original text that retains the essential ideas and information.

The output of the summarizing system is particularly important because it is quite challenging for machines to decipher the text's true meaning and deliver it in a concise manner. The task of summarizing is divided into different categories based on how closely the summary produced by the machine and the human-made summary resemble each other:

 DOI:10.5121/ijci.2026.150128

• Extractive summarization: This technique involves sifting through the provided material to identify the most crucial phrases. The text provided in the text document is replicated in the summary that was thus prepared. Due to its simple implementation and minimal requirement for linguistic resources, this summary generation method is fairly well known among researchers[2]. The primary goals of extractive summarization are to increase the text's coherence, importance, and coverage as well as to cut down on repetition in the resulting summary.

• Abstractive summarization: This method of summary production entails rearranging sentences and presenting them so that the resulting summary resembles a summary produced by a human being more closely. This method is more challenging since it requires a thorough grasp of the concepts, extensive linguistic resources, and a representation of the actual context of summarization. The resulting summary is more similar to the desired summary.

1.1. Need of New Punjabi Text Summarization System

Punjabi is an official language of the Punjab state which encompasses northwest India and eastern Pakistan. It is the third most spoken language in the Indian subcontinent with more than 100 million native speakers around the world. Also Govt. of Punjab is promoting Punjabi Language. All the letters and notifications are released in Punjabi Language by Govt. of Punjab. Moreover, Punjabi is getting support on many social media platforms like Facebook, Instagram, Whatsapp etc. People posts/comments in Punjabi Language. Hence, there is a need of Punjabi Summarization System which can summarize the Punjabi Text.

2. LITERATURE SURVEY

Punjabi is an official language of the Punjab state which encompasses northwest India and eastern Pakistan. It is the third most spoken language in the Indian subcontinent with more than 100 million native speakers around the world. It is the most widely spoken language in Pakistan, second/third language used by around 30 million people in India. In addition, Punjabi is a minority language in several other countries where Punjabi people have migrated, namely- United States of America, Australia, United Kingdom, and Canada. The Punjabi language comprises of canonical word order of Subject Object Verb (SOV), also contains postpositions; distinguishes gender- masculine/feminine, number- singular/plural, case- direct/oblique. The major writing system is the ਗੁਰਮੁਖੀ ‘Gurmukhi’, a Punjabi script. However, very fewer efforts are being made in the field of computer technology towards the development of the enriched Punjabi language. The language processing tasks for Punjabi language are as follows: (Gupta and Lehal [3]) have developed Named Entity Recognition (NER) system for Punjabi text to identify entities such as name of person, location and organizations using rule based and list look-up approach. (Gupta and Lehal [4]) have detailed a system that can automatically extracts and identified key phrases/keywords/key segments from the Punjabi text. The system works in phases such as removal of stopwords, nouns identification and stemming, computation of Term Frequency and Inverse Sentence Frequency (TF-ISF). The study has been done on more than 50 Punjabi documents from the Punjabi news corpus which gives precision of 80.40%, recall of 90.60% and F-measure 85.20% respectively. (Kaur and Gupta [5]) have implemented topic tracking using models like Vector Space Model (VSM), KNN classification, hierarchical clustering and others. (Kaur and Kaur [6]) have presented deadwood detection and elimination method to eliminate unwanted/unnecessary words or phrases from the given text which carried no meaning or have no relevance to the sentences. The summarising method based on hybrid ideas in the Punjabi language has been proposed by (Gupta and Kaur [7]). Their method uses SVM classifiers for distinct aspects in Punjabi, including concept-based, statistics-based, location-based, and linguistic-based features. It is compared with 10 baseline systems across 150 randomly chosen

articles from two Punjabi datasets. (Sarkar et al. [8]) have tried to eliminate sentence ambiguity by finding paraphrases in the Punjabi language using three similarity metrics and training probabilistic neural networks using feature sets. (Sharma et al. [9]) has worked upon the separation of dependent and independent clauses from a sentence in the given text in the Punjabi language using some rules and Part of Speech (POS). (Sharma et al. [10]) has developed syntactic analysis system in the Punjabi for language based compound sentences. (Ahmad et al. [11]) have developed first ever Named Entity Recognition (NER) corpus for the another script of Punjabi Language i.e. Shahmukhi. The corpus consists of 318,275 tokens and 16,300 named entities. They have compared the corpus specifications with the Gurmukhi counterpart using machine learning and deep learning techniques. Arti Jain et.al, 2021 [12] has proposed PSO algorithm for the summarization of Punjabi Language. PSO is based on intelligence that predicts which solution is best among the given set of solutions.

3. METHODOLOGY

3.1. Pre Processing Phase

In the pre-processing, an initial illustration over the textual data is marked through the given tasks. In this phase, data cleaning on Punjabi text is achieved using the following major operations- input restriction, sentence tokenization, removal of punctuation, word tokenization, stop-word elimination, word stemming, and normalization. Each of these is discussed in this section.

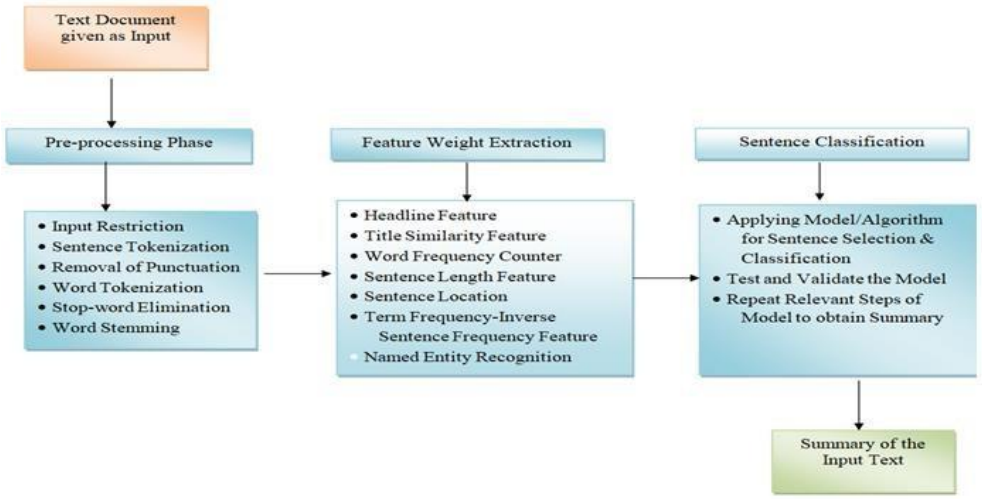

Figure1. Methodology of Text Summarization

3.1.1. Input Restriction

Text corpus must be majorly in the Punjabi language which comprises of total number of Punjabi characters as not lesser than 80% of the total number of corpus characters.

3.1.2. Sentence Tokenization

Existence of certain symbols such as ";", "।", "?", "!" are indicators of sentence boundary, particularly, the vertical bar "।" is used for the completion of the Punjabi sentence.

3.1.3. Removal of Punctuation

Punctuation symbols such as “;”, “,”, “.”, “:”, “-”, “”” etc. are necessarily be removed from the Punjabi sentences.

3.1.4. Word Tokenization

Every sentence is tokenized into words for the further operation such as stop-words removal and feature extraction.

3.1.5. Stop-word Elimination

Stop-words do not convey significant meaning to the sentence, and so, are removed from the Punjabi text. Eg. ਹੈ, ਉਹ

3.1.6. Word Stemming

The goal of stemming is to get the word into its basic form from its variant inflections and derivational forms. Eg. ਮੁੰਡਾ and ਮੁੰਡੇ, ਕੁੜੀ and ਕੁੜੀਆਂ.

3.1.7. Normalization

Words need to be normalized as there are several spelling mistakes in Punjabi as is shown in Table 1.

Table-1: Punjabi words and their spelling variations

Punjabi Words		Spelling Variations		English Translation
Word	**Transliteration**	**Variation**	**Transliteration**	
ਦੋਸ਼	Dōśa	ਦੋਸ	Dōsa	Accused
ਗੈਰਹਾਜ਼ਰੀ	Gairahāzarī	ਗੈਰਹਾਜਰੀ	Gairahājarī	Absence

3.2. Processing Phase

After the pre-processing phase, cleaned Punjabi text corpus is obtained over which the processing phase is applied to extract various statistical and linguistic features, for example- headline, similarity with title, length of sentence, TF-ISF, named entity recognition, cue phrase, and English-Punjabi common nouns. Each of these features is discussed in details here.

3.2.1. Headline Feature

The headline feature (hl) of a text document conveys the central theme of the document. For example:

Punjabi Text: ਭਾਰਤ ਅਤੇ ਪਾਕਕਸਤਾਨ ਕਿਚ ਅਲੇ ਅੱਜ ਹੋਿੇਗਾ ਕਕਿਕਟ ਮੈਚ
Transliteration: Bhārata atē pākisatāna vicālē aja hōvēgā krikaṭa maica
English Translation: Cricket match between India and Pakistan to be held today

3.2.2. Title Similarity Feature

The title similarity feature (ts) represents that the word(s) within a sentence if exist in the text title also, then the sentence is quite relevant to the Punjabi text document. The score of the title similarity is computed using Equation (1):

$$ts = \frac{\text{number of title words in sentence S}}{\text{Total number of words in title}}$$

3.2.3. Sentence Length Feature

The sentence length feature (sl) refers to the total number of words in a Punjabi sentence. Longer sentence has more likelihood to include vital information; however, extremely shorter sentence has lesser information and is usually not incorporated within summary. Sentence length is calculated using equation (2):

$$sl = \frac{\text{number of words in sentence S}}{\text{total number of words in longest sentence}}$$

3.2.4. Term Frequency-Inverse Sentence Frequency Feature

The **TF-ISF feature (ff)** extracts keywords from the Punjabi text using the following equation:
TF-ISF(x)=TF(x)×ISF(x)

Where:

- TF(x)\text{TF}(x)TF(x) = frequency of the word xxx within the Punjabi sentence
- ISF(x)=logNNi\text{ISF}(x) = \log \frac{N}{N_i}ISF(x)=logNiN
- NNN = total number of sentences in the Punjabi text
- NiN_iNi = number of sentences that contain the word xxx

3.2.5. Named Entity Recognition Feature

The **Named Entity Recognition (NE) feature** extracts **named entities (NEs)** from Punjabi text, such as **persons, locations, and organizations**. NEs in Punjabi are extracted using **rule-based approaches** and **gazetteer lists** (Gupta and Lehal [27]).

Consider the following example of a Punjabi sentence:

Punjabi:
ਕਿਲਗੇਟਸਨੇ PM ਮੋਦੀਨੂੰ ਆਦ਼ਾਦੀ ਕਦਿਸ ਦੀ ਕਦੱਤੀ ਿਧਾਈ, ਭਾਰਤ ਦੇ ਕਿਕਾਸ ਨੂੰ ਪਿਰਣਾ ਦਾਇਕ ਦੱਕਸਆ |

Transliteration:
Bila gēṭasa nē PM mōdī nū āzādī divasa dī ditī vadhā'ī, bhārata dē vikāsa nū prēranādā'ika dasi'ā

English Translation:

Bill Gates congratulated PM Modi on Independence Day and described India's development as inspiring.
In the above sentence, the following named entities are recognized, as shown in Table 2.

Table-2: Named entities for the sample Punjabi text

Punjabi Text	Transliteration	English Translation	Named Entity
ਨਰਿੰਦਰਮੋਦੀ	Naridaramōdī	NarendraModi	Person
ਭਾਰਤ	Bhārata	India	Location

3.3. Sentence selection, Classification and Summary Generation

After the preprocessing and feature weight extraction of the given text, machine leaning/deep learning techniques shall be applied for the text classification and final summary generation

4. CONCLUSIONS

In conclusion, text summarization literature witnessed notable advancements, particularly with the dominance of deep learning models and the exploration of multi-document summarization. However, challenges such as interpretability, information loss in extractive methods, evaluation standards, domain dependence, and scalability need to be addressed to further enhance the effectiveness and applicability of text summarization systems.

REFERENCES

[1] V. Gupta and G. S. Lehal, "Automatic text summarization system for Punjabi language,"Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 3, pp. 257–271, 2013.

[2] V. Gupta and N. Kaur, "A Novel Hybrid Text Summarization System for Punjabi Text," Cognitive Computation, vol. 8, no. 2, pp. 261–277, 2016.

[3] Kaur, K., & Gupta, V. "Topic Tracking for Punjabi Language", Computer Science Engineering International Journal, vol. 1, pp. 37–49, 2011.

[4] Gupta, V., & Lehal, G. S. "Named Entity Recognition for Punjabi Language Text Summarization", International Journal of Computers and Applications, vol. 33, no.3, pp. 28–32, 2011.

[5] Gupta, V., and Lehal, G. S. "Automatic Keywords Extraction for Punjabi Language", International Journal of Computer Science Issues, vol. 8, pp. 327-330, 2011.

[6] Kaur, M., & Kaur, J. "Deadwood Detection and Elimination in Text Summarization for Punjabi Language", International Journal of Engineering Science, vol. 8, pp. 51–59, 2013.

[7] Singh A. and Singh P., "Punjabi Dialects Conversion System for Malwai and Doabi Dialects," Indian Journal of Science and Technology, vol. 8, pp. 1-7, 2015.

[8] Sarkar, S., Saha, S., Bentham, J., Pakray, P., Das, D., & Gelbukh, A. F. "Language Independent Paraphrases Detection", FIRE (Working Notes), 256-259, 2016.

[9] Sharma, S. K. "Clauses Detection in Punjabi Language", International Journal of Innovations & Advancement in Computer Science, vol. 6(8), 2017.

[10] Sharma, S. K. "Sentence Reduction for Syntactic Analysis of Compound Sentences in Punjabi Language", EAI Endorsed Transactions on Scalable Information Systems, vol. 6(20), e4, 2019.

[11] Ahmad, M. T., Malik, M. K., Shahzad, K., Aslam, F., Iqbal, A., Nawaz, Z., & Bukhari, F. "Named Entity Recognition and Classification for Punjabi Shahmukhi", ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 19(4), pp. 1–13, 2020.

[12] Jain, A. "Named Entity Recognition for Hindi Language Using NLP Techniques (PhD Thesis)", Jaypee Institute of Information Technology. https://shodhganga.inflibnet.ac.in/handle/10603/241558, 2019.

[13] Kaur, R., Sharma, R. K., Preet, S., & Bhatia, P. "Punjabi WordNet Relations and Categorization of Synsets", 3rd National Workshop on IndoWordNet Under the Aegis of the 8th International Conference on Natural Language Processing, 2010.

[14] Narang, A., Sharma, R. K., & Kumar, P. "Development of Punjabi WordNet", CSI Transactions on ICT, vol. 1(4), pp. 349-354, 2013.

[15] Nidhi, V. G. "Domain Based Classification of Punjabi Text Documents", Proceedings of COLING, pp. 297-304, 2012.

[16] N. Desai and P. Shah, “Automatic text summarization using supervised machine learning technique for Hindi langauge,” International Journal of Research in Engineering and Technology, vol. 05, no. 06, pp. 361–367, 2016.
[17] G. Pareek, D. Modi, and A. Athaiya, “A Meticulous Approach for Extractive based Hindi Text Summarization using Genetic Algorithm,” International Journal of Innovations & Advancement in Computer Science, vol. 6, no. 8, pp. 264-273, 2017.
[18] P. Sethi, S. Sonawane, S. Khanwalker, and R. B. Keskar, “Automatic text summarization of news articles,” in Proc. International Conference on Big Data, IoT and Data Science, pp. 23–29, 2017.
[19] K. Shetty and J. S. Kallimani, “Automatic extractive text summarization using K-means clustering,” in Proc. International Conference on Electrical, Electronics, Communication Computer Technologies and Optimization Techniques, pp. 881–890, 2017.
[20] R. Vale, R. Lins, and R. Ferreira, “Assessing sentence simplification methods applied to text summarization,” in Proc. - 2018 Brazilian Conference on Intelligent System, pp. 49–54, 2018.
[21] Sahoo, Ashutosh Bhoi, Rakesh Chandra Balabantaray, "Hybrid Approach To Abstractive Summarization", Procedia Computer Science, vol. 132, pp. 1228-1237, 2018.
[22] N. Alami, N. En-nahnahi, S. A. Ouatik, and M. Meknassi, “Using Unsupervised Deep Learning for Automatic Summarization of Arabic Documents,” Arabian Journal for Science and Engineering, vol. 43, no. 12, pp. 7803–7815, 2018.
[23] K. Al-Sabahi, Z. Zhang, J. Long, and K. Alwesabi, “An Enhanced Latent Semantic Analysis Approach for Arabic Document Summarization,” Arabian Journal of Science and Engineering, vol. 43, no. 12, pp. 8079–8094, 2018.
[24] Anh, D. T., & Trang, N. T. T. “Abstractive text summarization using pointer generator networks with pre-trained word embedding”, In Proceedings of the tenth international symposium on information and communication technology , pp. 473-478, 2019.
[25] C. Mallick, A. K. Das, M. Dutta, A. K. Das, and A. Sarkar, “Graph-based text summarization using modified TextRank,” in Advances in Intelligent Systems and Computing, vol. 758, Springer Verlag, pp. 137–146, 2019.
[26] X. Mao, H. Yang, S. Huang, Y. Liu, and R. Li, “Extractive summarization using supervised and unsupervised learning,” Expert Systems with Applications, vol. 133, pp.173–181, 2019.
[27] P. Verma and H. Om, “A novel approach for text summarization using optimal combination of sentence scoring methods,” Sadhana - Academy Proceedings in Engineering Sciences, vol. 44, no. 5, pp. 1-17, 2019.
[28] A. Khan, M. A. Gul, M. Zareei, R. R. Biswal, A. Zeb. M. Naeem, Y. Saeed and N. Salim,“Movie Review Summarization Using Supervised Learning and Graph-Based Ranking Algorithm,” Computational Intelligence and Neuro Science, pp. 1–14, 2020.
[29] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “EdgeSumm: Graphbased framework for automatic text summarization,” Information Processing and Management, vol. 57, no. 6, pp. 1–21, 2020.
[30] Gupta, A., Chugh, D., Anjum, Katarya, R. (2022). Automated News Summarization Using Transformers. In: Aurelia, S., Hiremath, S.S., Subramanian, K., Biswas, S.K. (eds) Sustainable Advanced Computing. Lecture Notes in Electrical Engineering, vol 840. Springer, Singapore. https://doi.org/10.1007/978-981-16-9012-9_21.
[31] Mohsin, Muhammad & Latif, Shazad & Haneef, Muhammad & Tariq, Usman & Khan, Muhammad & Kadry, Seifedine & Yong, Hwan-Seung & Choi, Jungin. “Improved Text Summarization of News Articles Using GA-HC and PSO-HC”, Applied Sciences, 2021.
[32] Kumari, Namrata & Singh, Pardeep “Hindi Text Summarization using Sequence to Sequence Neural Network”. 10.21203/rs.3.rs-2036546/v1, Research Square, 2022.
[33] Arti Jain & Divakar Yadav & Anuja Arora "Particle Swarm Optimization for Punjabi Text Summarization," International Journal of Operations Research and Information Systems (IJORIS), IGI Global, vol. 12, no. 3, pp. 1-17, 2021.
[34] Jain, A. “Automatic Text Summarization for Hindi Using Real Coded Genetic Algorithm”, Applied Sciences, vol. 12, no.13, 2022.
[35] Gupta, J. P., Tayal, D. K., & Gupta, A.” A TENGRAM Method Based Part-of-Speech Tagging of Multi-Category Words in Hindi Language”, Expert Systems with Applications, vol. 38, no.12, pp. 15084–15093, 2011.