International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-09, Issue-11, Feb 2024

A Survey on Text Summarization for Urdu Language using Deep Learning Techniques

Sumayya Afreen, Research Scholar, Dept. of Computer Science and Engineering, University College of Engineering (OU), sumayyaafreen21@gmail.com

Prof. S. Sameen Fatima, Retd. Professor, Dept. of Computer Science and Engineering, University College of Engineering (OU), sameenf@gmail.com

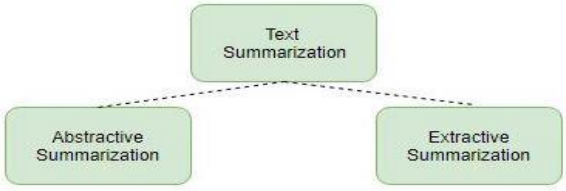
Abstract: Urdu is a language that has developed over time through a combination of various linguistic influences like Arabic, Persian, Turkian languages and Sanskrit. Urdu is a widely spoken language primarily in South Asia. It is spoken by approximately 100 million people as a first language worldwide. Urdu is one of the 22 officially recognized languages of India and is spoken in various parts of the country. In India, it is often spoken alongside Hindi. Urdu is a low-resource language in the context of natural language processing (NLP) and machine learning. This means that there is relatively limited data and resources available for developing NLP applications and models in Urdu compared to more widely spoken languages like English or Spanish. The availability of resources for a language can significantly impact the development and performance of NLP systems in that language. Some of the challenges faced by low-resource languages like Urdu include, Limited Text Data: There is a scarcity of large, high-quality text corpora in Urdu. This makes it challenging to train robust language models, as they require extensive text data for pre-training. Language Models: Pre-trained language models like GPT-3 and BERT have been successful in English and a few other languages, but they may not perform as well in Urdu due to the lack of large-scale pre-training data. Language-Specific Challenges: Urdu has its unique linguistic features, including script, grammar, and vocabulary, which require specialized models and resources. There are various applications of NLP like Sentiment Analysis, Chat Bot, Text Summarization etc. In this paper we try to present the survey of text summarization and provide the list of available datasets for Urdu Summaries.

Keywords — Abstractive Text Summarization, Extractive Text Summarization, Large Language Models,

I. INTRODUCTION

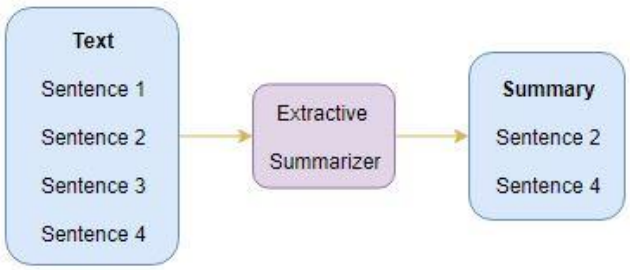
Text summarization refers to the technique of condensing a lengthy text document into a succinct and well-written summary that captures the essential information and main ideas of the original text, achieved by highlighting the significant points of the document.

There are broadly two different approaches that are used for text summarization: Extractive Summarization and Abstractive Summarization



A. *Extractive Summarization*

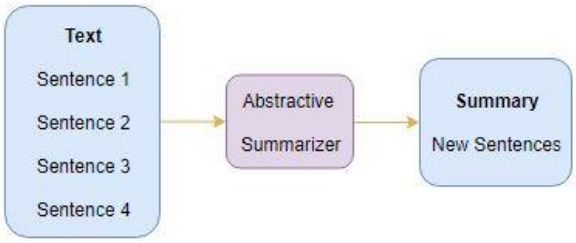
The name gives away what this approach does. We identify the important sentences or phrases from the original text and extract only those from the text. Those extracted sentences would be our summary. The below diagram illustrates extractive summarization:



B. *Abstractive Summarization*

This is a very interesting approach. Here, we generate new sentences from the original text. This contrasts with the extractive approach we saw earlier where we used only the sentences that were present. The sentences generated through abstractive summarization might not be present in the original text:

 DOI : 10.35291/2454-9150.2024.0040



Urdu abstractive summarization involves developing algorithms and models that can automatically generate concise and coherent summaries of longer Urdu text while preserving the core meaning and important details. Abstractive summarization is a subfield of natural language processing (NLP) that focuses on condensing large pieces of text into shorter, more informative summaries. In contrast to extractive summarization, which selects and combines existing sentences or phrases from the source text, abstractive summarization generates summaries in a more human-like way by rephrasing and reformulating the content.

Urdu is a complex and rich language with its own set of linguistic challenges, including a wide vocabulary, diverse sentence structures, and variations in writing style. Urdu script is right-to-left, which requires specific handling in text processing and modeling.

The first step in research is to gather a substantial dataset of Urdu text documents and their corresponding human-generated summaries.

Data preprocessing involves tokenization, stemming, and handling Urdu-specific issues such as script conversion (e.g., from Nastaliq to Naskh). Then to select a suitable deep learning architecture for abstractive summarization. Fine-tune pre-trained models on your Urdu summarization dataset or consider building a model from scratch if resources permit. Urdu abstractive summarization is a challenging and evolving field, but it has the potential to benefit various applications, including news summarization, content recommendation, and information retrieval in the Urdu-speaking world. Continuous experimentation, evaluation, and innovation are key to making meaningful contributions to this research area.

C. *Motivation*

- Urdu summarization in Natural Language Processing (NLP) serves several important purposes and can be motivated by various factors:
- Language Inclusivity: Urdu is one of the world's most widely spoken languages, primarily in Pakistan and parts of India. Summarization in Urdu enables NLP applications to be more inclusive, reaching a broader audience and providing information to Urdu-speaking users.
- Information Accessibility: Summarization in Urdu can make vast amounts of information available in this language more manageable and accessible. It can help users quickly grasp the main points of Urdu text, articles, or documents, thereby saving time and effort.
- Multilingual Communication: In a globalized world, where people from diverse linguistic backgrounds interact online, Urdu summarization can facilitate cross-lingual communication. It allows non-Urdu speakers to understand the essence of Urdu content, promoting cultural exchange and understanding.
- Content Aggregation: For news agencies, blogs, or websites catering to Urdu-speaking audiences, automated summarization can help aggregate and present information efficiently. This is particularly useful for presenting news updates, blog posts, or articles in a concise format.
- Educational Applications: Summarization can support educational efforts in Urdu-speaking regions. It can help students and teachers extract key points from lengthy educational materials, making learning more efficient.
- Business Intelligence: In the business context, organizations operating in Urdu-speaking regions can use summarization to analyze market trends, customer feedback, and competitor activities more effectively. Summaries provide quick insights for decision-making.
- Sentiment Analysis: Summarization can be integrated with sentiment analysis to gauge public sentiment in Urdu-speaking communities, helping businesses and governments understand the mood of the population.
- Legal and Administrative Documents: In legal and administrative settings, summarization can be crucial for quickly extracting essential information from lengthy legal documents, contracts, or government reports, improving efficiency and accuracy.
- Disaster Response and Crisis Management: During emergencies, such as natural disasters or health crises, summarization can assist in processing and disseminating critical information in Urdu to affected populations and responders.
- Search Engines and Recommender Systems: Incorporating Urdu summarization in search engines and recommender systems can enhance the quality of search results and recommendations for Urdu-speaking users.
- Content Curation: Content curation platforms can use summarization to automatically select and display relevant and informative Urdu content to users, improving user engagement and satisfaction.

 DOI : 10.35291/2454-9150.2024.0040

II. RELATED WORK

In the research paper published by M. Asif [1], abstractive and extractive summarization for Urdu language is performed. In extractive summaries, word frequency, Sentence weight, and TF-IDF algorithms are used. Further, a hybrid method is introduced to improve the results of extractive summaries. Bidirectional Encoder Representations from Transformers (BERT) model is used to process the summaries generated by hybrid method for generation of abstractive summary. To evaluate the system-generated summaries, the assistance of the experts of Urdu language is reaped.

Article proposed by Shafiq [2] a deep learning model for the Urdu language by using the Urdu 1 million news dataset and compared its performance with the two widely used methods based on machine learning, such as support vector machine (SVM) and logistic regression (LR). The results show that the suggested deep learning model performs better than the other two approaches. The summaries produced by extractive summaries are processed using the encoder-decoder paradigm to create an abstractive summary.

Mohammed in his paper [3] presented an extractive text summarization methodology for Urdu language documents based on sentence weight algorithm using segmentation, tokenization and stop words as prominent features. ROUGE metric is used for system evaluation by comparing system generated and human generated summaries. System accuracy at Unigram, bigram and trigram level is 67 percent.

Nawaz [4] presented in his paper, in an automatic extractive summary generation, the sentences with the highest weights are given importance to be included in the summary. The sentence weight is computed by the sum of the weights of the words in the sentence. There are two famous approaches to compute the weight of the words in the English language: local weights (LW) approach and global weights (GW)

approach. The sensitivity of the weights depends on the contents of the text, the one word may have different weights in a different article, known as LW based approach. Whereas, in the case of GW, the weights of the words are computed from the independent dataset, which implies the weights of all words remain the same in different articles. In the proposed framework, LW and GW based approaches are modeled for the Urdu language.

The extractive summaries are generated by LW and GW based approaches and evaluated with ground-truth summaries that are obtained by the experts. The VSM is used as a baseline framework for sentence weighting. Experiments show that LW based approaches are better for extractive summary generation. The F-score of the sentence weight method and the weighted term-frequency method are 80% and 76%, respectively. The VSM achieved only 62% accuracy on the same dataset. Both, the datasets with ground-truth, and the code are made publicly available for the researchers.

The paper by Humayun [5] reports the construction of a benchmark corpus for Urdu summaries (abstracts) to facilitate the development and evaluation of single document summarization systems for Urdu language. In Urdu, space does not always mark word boundary. Therefore, we created two versions of the same corpus. In the first version, words are separated by space. In contrast, proper word boundaries are manually tagged in the second version. In this paper, the author has applied, part-of-speech tagging, morphological analysis, lemmatization, and stemming for the articles and their summaries in both versions. To apply these annotations, some NLP tools for Urdu are re-implemented. This paper provided Urdu Summary Corpus, all these annotations and the needed software tools (as open-source) for researchers to run experiments and to evaluate their work including but not limited to single-document summarization task.

Paper published by Basit [6] focuses on studying the existing systems and proposing an approach for Urdu documents providing a better semantic similarity score. This approach proposes an extended and improved formula for generating an initial matrix of LSA representing the documents. This results in better and more accurate SS

scores. To evaluate a system called TripleS4Urdu and show better results.

III. RESEARCH GAPS

While automatic summarization has made significant progress in languages like English, there are unique challenges and opportunities in summarizing Urdu text due to its distinct linguistic characteristics. Here are some research gaps in Urdu summarization:

Lack of Large Annotated Corpora: One of the major challenges in Urdu summarization is the limited availability of large, high-quality annotated corpora. Building comprehensive datasets for training and evaluating summarization models is essential to make progress in the field.

Abstractive Summarization: While extractive summarization techniques have been explored to some extent, abstractive summarization, which generates summaries by paraphrasing and restructuring the content, is still an underexplored area in Urdu summarization.

Handling Code-Switching: Urdu is often mixed with English and other languages in digital content, which makes it challenging to develop summarization models that can effectively handle code-switching and language variations.

Evaluation Metrics: Developing appropriate evaluation metrics for Urdu summarization is crucial. Existing metrics may not capture the nuances of the Urdu language and the specific requirements of Urdu summarization.

 DOI : 10.35291/2454-9150.2024.0040

Sentiment and Tone Preservation: Summaries should preserve the sentiment and tone of the original text. Research is needed to develop models that can generate summaries while maintaining the emotional context of the content.

Extracting Key Information from Noisy Text: Urdu text on social media and user-generated content often contains noise, slang, and abbreviations. Research is needed to develop methods to extract key information from noisy text while generating coherent summaries.

Author	Title	Year and Publication	Technique used/ Methodology	Dataset	Results	Limitations
M. Asif, S. A. Raza, J. Iqbal, N. Perwaiz, T. Faiz and S. Khan	Bidirectional Encoder Approach for Abstractive Text Summarization of Urdu Language	2022, ieee.org	BERT Model	50 News Articles	Abstractive summary was created by retaining the meaning of the text.	Small Data set. No Quantitative result.
Nida Shafiq, Isma Hamid, Muhammad Asif, Qamar Nawaz, Hanan Aljuaid and Hamid Ali	Abstractive text summarization of low-resourced languages using deep learning	2023, peerj.com	Encoder-Decoder Model, LSTM	Urdu 1 million news dataset (914 articles)	Suggested model performs better than the other two models (SVM and LR)	Data Set is small.
Ali Nawaz, Maheen Bakhtyar, Junaid Baber, Ihsan Ullah, Waheed Noor, Abdul Basit	Extractive Text Summarization Models for Urdu Language	2020, Elsevier	Vector Space Model (VSM), LW (Local Weight) and GW (Global Weight) based approach	UTD – 600 News Articles	LW based approach outperforms the GW based approach	All approaches show limited performance on abstractive summary ground-truth
EUTS: EXTRACTIVE URDU TEXT SUMMARIZER	Muhammad, Aslam	2018, IEEE	Sentence Weight Algorithm	Humayoun Dataset – 50 articles	system accuracy is 67 % at the 50 documents as the comparison with human generated summary.	Accuracy can be improved.

Table 1: Related Work Summary

Name of the Dataset	**Contributors**	**Size**	**Link**
Urdu Summary Corpus	Muhammad Humayoun, Mohammed Adeel	50 Articles	https://github.com/humsha/USCorpus
Urdu News Dataset	Khalid Hussain, Nimra Mughal, Irfan Ali, Saif Hassan, Sher Muhammad Daudpota	1M Articles	https://data.mendeley.com/datasets/834vsxnb99/3
Urdu 1 Million news dataset	Nida Shafeeq, Isma Hamid	914 Articles	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10280265/bin/peerj-cs-09-1176-s002.xlsx
Urdu Abstractive Summaries	Ali Nawaz	50 Articles	https://github.com/AliNawazUoB
BBC -Urdu News Articles	M. Irfan	8.46K Articles	https://huggingface.co/datasets/mirfan899/usummary

Table 2: Urdu Summary Corpus

Addressing these research gaps in Urdu summarization will contribute to the development of more effective and accurate automatic summarization systems for Urdu language content, improving information access and dissemination in Urdu-speaking regions and communities.

IV. PROPOSED METHODOLOGY

The task of abstractive text summarization in the Urdu language presents a challenging natural language processing problem. Despite the growing volume of Urdu content on the internet, there is a lack of effective automated summarization tools tailored to this specific language. The problem at hand is to develop a robust and accurate abstractive text summarization system for Urdu that can generate concise and coherent summaries of lengthy texts while preserving the original meaning and context.

This problem encompasses several key challenges:

Language Complexity: Urdu is a morphologically rich and highly inflected language with a diverse vocabulary. This complexity poses challenges for accurate semantic understanding and paraphrasing during summarization.

Lack of Resources: Compared to languages like English, there is a scarcity of linguistic resources such as annotated corpora, pre-trained models, and linguistic tools for Urdu,

 DOI : 10.35291/2454-9150.2024.0040

making it difficult to leverage the advantages of deep learning approaches.

Coherence and Meaning Preservation: The summarization system should not only condense the text but also ensure that the summary maintains the coherence of ideas and preserves the essential meaning of the original content.

Evaluation Metrics: Developing appropriate evaluation metrics for measuring the quality of abstractive summaries in Urdu is crucial, as existing metrics may not be directly applicable or accurate for this language.

Scalability: The system should be designed to handle large volumes of text efficiently, making it suitable for real-world applications such as news aggregation, content curation, and document summarization.

Addressing these challenges is essential to advance the field of NLP for Urdu and enable the creation of valuable tools for both native speakers and researchers interested in the language. Developing an effective abstractive text summarization system for Urdu will contribute to improved information access and understanding in the Urdu-speaking community and beyond.

The Aim of my research is to create titles for the news articles using Abstractive Text Summarization using Large language model. And compare the same with human generated titles

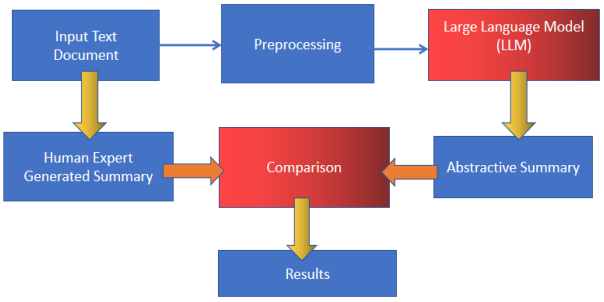


Fig 1: Proposed framework

V. CONCLUSION

In this paper we have provided a survey of papers that focuses on Text summarization of Urdu Language using various techniques for both Extractive and Abstractive text summarization. We have also listed the available data sets for summarization. We have proposed a methodology that uses Large Language Models, could be LLAMA, MTO, BLOOM etc to summarize urdu text. And the summary generated by the model can be compared by human generated summary to check the correctness of the result.

References

[1] Asif, Muhammad, et al. "Bidirectional Encoder Approach for Abstractive Text Summarization of Urdu Language." 2022 International Conference on Business Analytics for Technology and Security (ICBATS). IEEE, 2022.

[2] Shafiq, Nida, et al. "Abstractive text summarization of low-resourced languages using deep learning." PeerJ Computer Science 9 (2023): e1176.

[3] Nawaz, Ali, et al. "Extractive text summarization models for Urdu language." Information Processing & Management 57.6 (2020): 102383.

[4] Nawaz, Ali, et al. "Extractive text summarization models for Urdu language." Information Processing & Management 57.6 (2020): 102383.

[5] Humayoun, Muhammad, et al. "Urdu summary corpus." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016.

[6] Basit, Rida Hijab, et al. "Semantic similarity analysis of urdu documents." Pattern Recognition: 9th Mexican Conference, MCPR 2017, Huatulco, Mexico, June 21-24, 2017, Proceedings 9. Springer International Publishing, 2017.

[7] El-Kassas, Wafaa S., et al. "Automatic text summarization: A comprehensive survey." Expert systems with applications 165 (2021): 113679. Elsvier.

[8] Burney, Aqil, et al. "Urdu text summarizer using sentence weight algorithm for word processors." International Journal of Computer Applications 46.19 (2012): 38-43. Acdemia.edu.

[9] Tahir, Bilal, and Muhammad Amir Mehmood. "Corpulyzer: A novel framework for building low resource language corpora." IEEE Access 9 (2021): 8546-8563.

[10] Abdul-Mageed, Muhammad, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. "ARBERT & MARBERT: deep bidirectional transformers for Arabic." arXiv preprint arXiv:2101.01785 (2020).

[11] Al-Maleh, Molham, and Said Desouki. "Arabic text summarization using deep learning approach." Journal of Big Data 7 (2020): 1-17. Springer.

[12] Anwar, Waqas, Xuan Wang, and Xiao-long Wang. "A survey of automatic Urdu language processing." 2006 international conference on machine learning and cybernetics. IEEE, 2006.

[13] Himaja, Prathi, et al. "A Survey On Text Summarization in Urdu Language using Machine Learning Techniques." 2023 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2023.

[14] Naseer, Asma, et al. "Analysis of Corpus Development for Urdu Language." 2021 International Conference on Innovative Computing (ICIC). IEEE, 2021.

[15] Iqbal, Muntaha, Bilal Tahir, and Muhammad Amir Mehmood. "CURE: Collection for urdu information retrieval evaluation and ranking." 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2). IEEE, 2021.

 DOI : 10.35291/2454-9150.2024.0040

[16] Nadeem, Maaz Ali, et al. "Sequence-driven Neural Network models for NER Tagging in Roman Urdu." 2022 International Conference on Frontiers of Information Technology (FIT). IEEE, 2022.
[17] Nazir, Shahzad, et al. "Toward the development of large-scale word embedding for low-resourced language." IEEE Access 10 (2022): 54091-54097.
[18] Shaikh, M. Kashif, HH Ali Khowaja, and M. Ahmed Khan. "Urdu text translation with natural language processing." Student Conference On Engineering, Sciences and Technology. IEEE, 2004.
[19] Farooq, Aman, Safiyah Batool, and Zain Noreen. "Comparing Different Techniques of Urdu Text Summarization." 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC). IEEE, 2021.
[20] Hamza, Syed Ali, Bilal Tahir, and Muhammad Amir Mehmood. "Domain identification of urdu news text." 2019 22nd International Multitopic Conference (INMIC). IEEE, 2019.
[21] Farooq, Aman, et al. "Urdu News Classification: An Empirical Study Using Machine Learning Techniques." 2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC). IEEE, 2022.
[22] El-Kassas, Wafaa S., et al. "Automatic text summarization: A comprehensive survey." Expert systems with applications 165 (2021): 113679. Science Direct.
[23] Humayoun, Muhammad, and Naheed Akhtar. "CORPURES: Benchmark corpus for urdu extractive summaries and experiments using supervised learning." Intelligent Systems with Applications 16 (2022): 200129. Science Direct.
[24] Shafi, Jawad, et al. "UNLT: Urdu natural language toolkit." Natural Language Engineering 29.4 (2023): 942-977. Cambridge.
[25] Arora, Gaurav. "inltk: Natural language toolkit for indic languages." arXiv preprint arXiv:2009.12534 (2020).
[26] Chandrasekaran, Dhivya, and Vijay Mago. "Evolution of semantic similarity—a survey." ACM Computing Surveys (CSUR) 54.2 (2021): 1-37. Arxiv.org.
[27] Vekariya, Darshana V., and Nivid R. Limbasiya. "A novel approach for semantic similarity measurement for high quality answer selection in question answering using deep learning methods." 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2020.
[28] Elsaid, Asmaa, et al. "A comprehensive review of arabic text summarization." *IEEE Access* 10 (2022): 38012-38030.
[29] Abu Nada, Abdullah M., et al. "Arabic text summarization using arabert model using extractive text summarization approach." (2020).

 DOI : 10.35291/2454-9150.2024.0040