An Analytical Study of Text Summarization Techniques

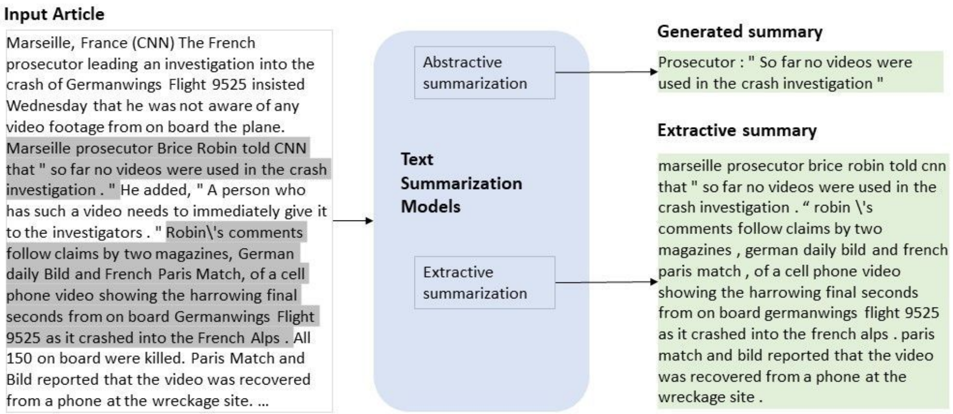
By Bavrabi Ghosh, Aritra Ghosh, Subhojit Ghosh, Anupam Mondal

INSTITUTE OF ENGINEERING & MANAGEMENT,KOLKATA

IEMTRONICS
International Conference

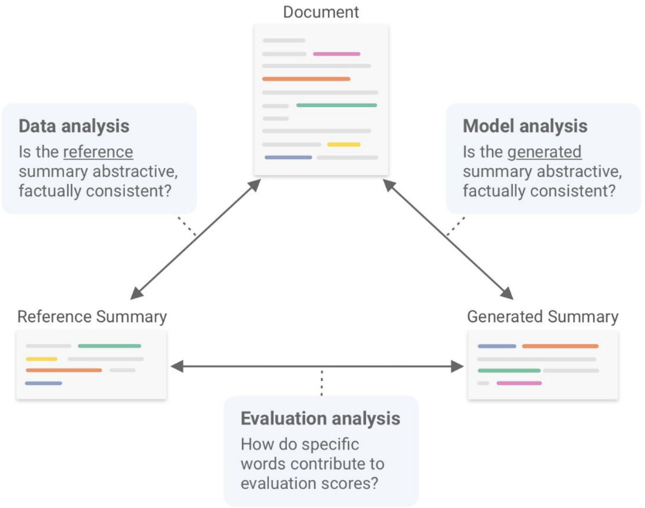
Introduction

Text summarization is crucial for condensing large information into concise overviews, enabling efficient comprehension across various domains like journalism, research, business, and social media. However, existing summarization models often fail to capture the nuances and details required by users. To address these limitations and ease research efforts, we aimed to create a comprehensive document on summarization techniques and applications.



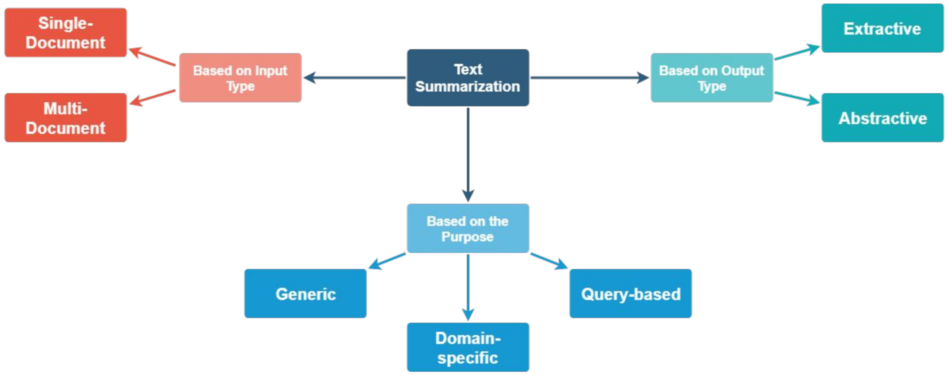
Motive

We explored various text summarization approaches and techniques to gain insights into existing methods' complexities and identify factors behind suboptimal performance. An analysis of research papers revealed extractive and abstractive summarization as prevalent methods. Extractive summarization utilizes ranking algorithms to extract salient sentences, while abstractive summarization attempts to generate new sentences capturing the meaning. We delved into technical details of early systems like the cueing, title, location methods and the Trainable Document Summarizer.



Background

We explored various text summarization approaches and techniques to gain insights into existing methods' complexities and identify factors behind suboptimal performance. An analysis of research papers revealed extractive and abstractive summarization as prevalent methods. Extractive summarization utilizes ranking algorithms to extract salient sentences, while abstractive summarization attempts to generate new sentences capturing the meaning. We delved into technical details of early systems like the cueing, title, location methods and the Trainable Document Summarizer.



Summarization Methods

Extractive Method

- Inverse Document Frequency method
- Cluster Method
- Graph Based Approach
- Latent Semantic Analysis Method

Abstractive Summarization

- Seq2Seq (Sequence-to-Sequence) Models
- Pointer-Generator Networks
- BERTSUM

Deep Learning Summarization

- Bidirectional and Auto-Regressive Transformers
- Generative Pre-trained Transformer
- Text-To-Text Transfer Transformer
- Pegasus

Overview of Important Features

Text representation models in NLP transform words into numerical forms for pattern detection.

- **N-grams** - Group words into N components, offering reasonable vector sizes but computational challenges.
- **Bag of Words** - It disregards word order but captures multiplicity.
- **TF-IDF** - Determines term importance relative to corpus but assumes independent term counts.
- **Word Embedding** - maps words to vectors, placing similar words close together, capturing semantic and syntactic relationships through algorithms like FastText, GloVe, and Word2Vec.

Overview of Findings

<table>
<tr><th colspan="3">Algorithms</th><th>Limitations</th></tr>
<tr><td rowspan="6">Extractive</td><td rowspan="3">Unsupervised</td><td>Fuzzy logic</td><td>Post-processing should remove redundancies to improve the quality of summarization.</td></tr>
<tr><td>Concept-based</td><td>The summary should use similarity measures to reduce redundancy, which can affect quality.</td></tr>
<tr><td>Latent-Semantic</td><td>LSA-generated summaries take a long time.</td></tr>
<tr><td rowspan="3">Supervised</td><td>Machine Learning</td><td>To make good summaries, it has to be trained and improved on a large set of data.</td></tr>
<tr><td>Neural Network</td><td>Both the training phase and the application phase are quite slow with neural networks. Training data also requires human interruption.</td></tr>
<tr><td>Conditional Random Fields</td><td>Linguistic features are not taken into account in the use of CRF. It also needs an external domain specific corpus.</td></tr>
<tr><td rowspan="7">Abstractive</td><td rowspan="4">Structural</td><td>Trees</td><td>The text ignores context and important phrases in the text, resulting in a failure to recognize the relationships between sentences. Another issue is that it consistently emphasises syntax rather than meaning.</td></tr>
<tr><td>Template based</td><td>As the templates are pre-defined using this technique, the summaries lack variation.</td></tr>
<tr><td>Based on Rules</td><td>It takes a long time to create regulations. It is also difficult to manually write the rules.</td></tr>
<tr><td>Ontology method</td><td>The process of creating a suitable ontology is time-consuming and limited to a single domain.</td></tr>
<tr><td rowspan="3">Semantic</td><td>Multi-model semantic</td><td>The framework must be automatically analysed because humans now manually evaluate it.</td></tr>
<tr><td>Information item</td><td>Generating grammatical and meaningful sentences from the material is difficult. The linguistic quality of summaries is low due to incorrect parses.</td></tr>
<tr><td>Based on Semantic Graph</td><td>Limited to single document abstractive summarization.</td></tr>
</table>

Challenges

1. Evaluation
2. Important Sentence Selection
3. Anaphora Problem
4. Predefined Template
5. Long Sentences and Jargon
6. Interpretability
7. Cataphora Problem

Future Scope

1. Multilingual Summarization
2. Customizable Summaries
3. Real-Time Summarization
4. Video Summarization

References

1. Aker, A., Cohn, T., Gaizauskas, R.: Multi-document summarization using a* search and discriminative learning. In: Pro- ceedings of the 2010 conference on empirical methods in natural language processing. pp. 482–491 (2010)
2. AL-Banna,A.A.,AL-Mashhadany,A.K.:Naturallanguageprocessingforautomatictextsummarization[datasets]-survey. Wasit Journal of Computer and Mathematics Science 1(4), 156–170 (2022)
3. Bala, A., Mitra, R., Mondal, A.: Recommendation system to predict best academic program. In: 2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech). pp. 1–6. IEEE (2023)
4. Dey,M.,Mondal,A.,Das,D.:Ntcir-12mobileclick:Sense-basedrankingandsummarizationofenglishqueries.In:Ntcir (2016)
5. Edmundson,H.P.:Newmethodsinautomaticextracting.JournaloftheACM(JACM)16(2),264–285(1969)
6. Ferreira, R., de Souza Cabral, L., Lins, R.D., e Silva, G.P., Freitas, F., Cavalcanti, G.D., Lima, R., Simske, S.J., Favaro, L.: Assessing sentence scoring techniques for extractive text summarization. Expert systems with applications 40(14), 5755–5764 (2013)
7. Garć ıa-Herna ́ndez, R.A., Ledeneva, Y.: Word sequence models for single text summarization. In: 2009 Second International Conferences on Advances in Computer-Human Interactions. pp. 44–48. IEEE (2009)
8. Kupiec,J.,Pedersen,J.,Chen,F.:Atrainabledocumentsummarizer.In:Proceedingsofthe18thannualinternationalACM SIGIR conference on Research and development in information retrieval. pp. 68–73 (1995)
9. Mahata,S.K.,Mondal,A.,Dey,M.,Sarkar,D.:Sentimentanalysisusingmachinetranslation.In:ApplicationsofMachine intelligence in Engineering, pp. 371–377. CRC Press (2022)
10. Mondal, A., Cambria, E., Dey, M.: An annotation system of a medical corpus using sentiment-based models for summarization applications. In: Computational Intelligence Applications for Text and Sentiment Data Analysis, pp. 163–178. Elsevier (2023)
11. Mondal, A., Dey, M., Das, D., Nagpal, S., Garda, K.: Chatbot: An automated conversation system for the educational domain. In: 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). pp. 1–5. IEEE (2018)
12. Mondal,A.,Dey,M.,Mahata,S.K.,Sarkar,D.:Anautomaticsummarizationsystemtounderstandtheimpactofcovid-19 on education. In: Applications of Machine intelligence in Engineering, pp. 379–386. CRC Press (2022)
13. Mridha, M.F., Lima, A.A., Nur, K., Das, S.C., Hasan, M., Kabir, M.M.: A survey of automatic text summarization: Progress, process and challenges. IEEE Access 9, 156043–156070 (2021)
14. Sinha,S.,Mandal,S.,Mondal,A.:Questionansweringsystem-basedchatbotforhealthcare.In:ProceedingsoftheGlobal AI Congress 2019. pp. 71–80. Springer (2020)
15. Syed,A.A.,Gaol,F.L.,Matsuo,T.:Asurveyofthestate-of-the-artmodelsinneuralabstractivetextsummarization.IEEE Access 9, 13248–13265 (2021)
16. Tas,O.,Kiyani,F.:Asurveyautomatictextsummarization.PressAcademiaProcedia5(1),205–213(2007)
17. Zhang, Y., Zincir-Heywood, N., Milios, E.: Narrative text classification for automatic key phrase extraction in web document corpora. In: Proceedings of the 7th annual ACM international workshop on Web information and data management. pp. 51–58 (2005)

Thank you