Paper Review

Abstractive Text Summarization for Sanskrit Prose: A Study of Methods and Approaches

Authors: Shagun Sinha;Girish Nath Jha| Published in: Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 60–65 |
Published at:Language Resources and Evaluation Conference (LREC 2020)|
Publisher: European Language Resources Association (ELRA)|
Issue Date: 11–16 May 2020|

Name: TANISHA TABASSUM
ID: 20101577
GROUP NO: 26
SECTION: 1
SUBMISSION: 2

What this paper is about?

The paper explores abstractive text summarization for Sanskrit Prose (SATS), focusing on semantics and graph-based techniques. It argues for the effectiveness of ATS in capturing the essence of the text and proposes a supervised-learning approach. The paper discusses limitations, suggests cross-lingual summarization, and emphasizes the importance of domain-specific knowledge and external resources. Overall, it examines ATS for Sanskrit Prose, presents experimental results, and discusses future directions.

Literature Survey

- Explores the extractive and abstractive approaches in TS for Indian Languages.
- Discusses structure-based and semantics-based approaches, as well as machine learning and corpora-based methods.
- Highlights the significance of ontology, graph-based methods, and corpus annotation in Indian Language summarization.

Methodology

Proposed approach for abstractive text summarization is based on supervised learning, leveraging neural network models.

1. Encoder-Decoder Architecture:
   - Utilize a sequence-to-sequence (Seq2Seq) model with an encoder-decoder architecture.
   - The encoder processes the input text and encodes it into a fixed-length vector representation.
   - The decoder generates the summary by decoding the encoded representation.
2. Attention Mechanism:
   - Incorporate an attention mechanism to enhance the decoder's ability to focus on relevant parts of the input during decoding.
   - Attention helps capture the most important information for summary generation.

Methodology

3. Training Process:

- Train the model using a large dataset of paired source texts and human-generated summaries.
- Employ a variant of the backpropagation algorithm, such as Adam optimization, to update the model parameters.
- Use techniques like teacher forcing, where the model is fed with ground-truth summaries during training.

Dataset Processing

Annotated dataset and preprocessing steps used for training the abstractive text summarization model

Annotated Dataset:

- Acquire a diverse dataset of news articles or documents with associated human-generated summaries.
- Ensure the dataset covers a wide range of topics and exhibits variations in writing styles and summary lengths.
- Perform a thorough annotation process, involving multiple annotators to maintain quality and consistency.

.

Dataset Processing

Preprocessing Steps:

- **Clean and tokenize the source texts and summaries.**
- **Remove irrelevant information, such as advertisements or metadata.**
- **Apply text normalization techniques, including lowercasing, stemming, or lemmatization.**
- **Address specific challenges, such as handling out-of-vocabulary (OOV) words or rare entities.**

Dataset Processing

Data Split:

- Split the dataset into training, validation, and testing subsets to evaluate the model's performance accurately.
- Use the majority of the data for training, a smaller portion for validation, and a separate unseen subset for final testing.

Conclusion

In conclusion, this paper presents a preliminary attempt to develop a Sanskrit abstractive text summarizer for current prose, particularly news articles. By surveying existing approaches to abstractive summarization in Indian languages, the authors have identified the need for a semantic approach due to the inherent complexities and rich morphology of Sanskrit. A graph-based method is suggested as a suitable starting point, and the possibility of emulating a supervised method used for news articles and headlines is explored. While the language of the output summary is beyond the scope of this paper, it is worth considering generating summaries in Sanskrit, Hindi, or English through existing Machine Translation. This work lays the foundation for future advancements in Sanskrit abstractive text summarization, including the potential expansion to other prose styles and the exploration of annotation possibilities.