Deep Learning Based Extractive Text Summarization

S. Paranika and B. Mayurathan
Department of Computer Science, Faculty of Science, University of Jaffna
paranika517@gmail.com, barathym@univ.jfn.ac.lk

Introduction

Automatic text summarization has become an increasingly important task in natural language processing (NLP). There are two main types of text summarization: extractive summarization and abstractive summarization. Extractive summarization identifies and extracts the most important sentences from a document to create a concise summary, while abstractive summarization generates new sentences that convey the main ideas of the text, even if those exact phrases do not appear in the original document.

Objective

This research focuses on extractive text summarization using Convolutional Neural Networks (CNN) and pre-trained word2vec word embedding techniques to identify and extract the most important sentences from a document, creating a concise summary.

Dataset

This research uses a subset of 50,000 news articles from the CNN/DailyMail Dataset, an English-language dataset containing over 300,000 unique news articles written by journalists at CNN and the Daily Mail.

Evaluation Metrics

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is used to evaluate summarization by comparing generated summaries against reference summaries.

ROUGE-N where N=1 for unigrams, N=2 for bigrams

$$\text{Precision} = \frac{\text{No\_of\_overlapping\_ngrams}}{\text{No\_of\_\_ngrams\_in\_system\_summary}}$$

$$\text{Recall} = \frac{\text{No\_of\_overlapping\_ngrams}}{\text{No\_of\_\_ngrams\_in\_reference\_summary}}$$

$$\text{FI Score} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

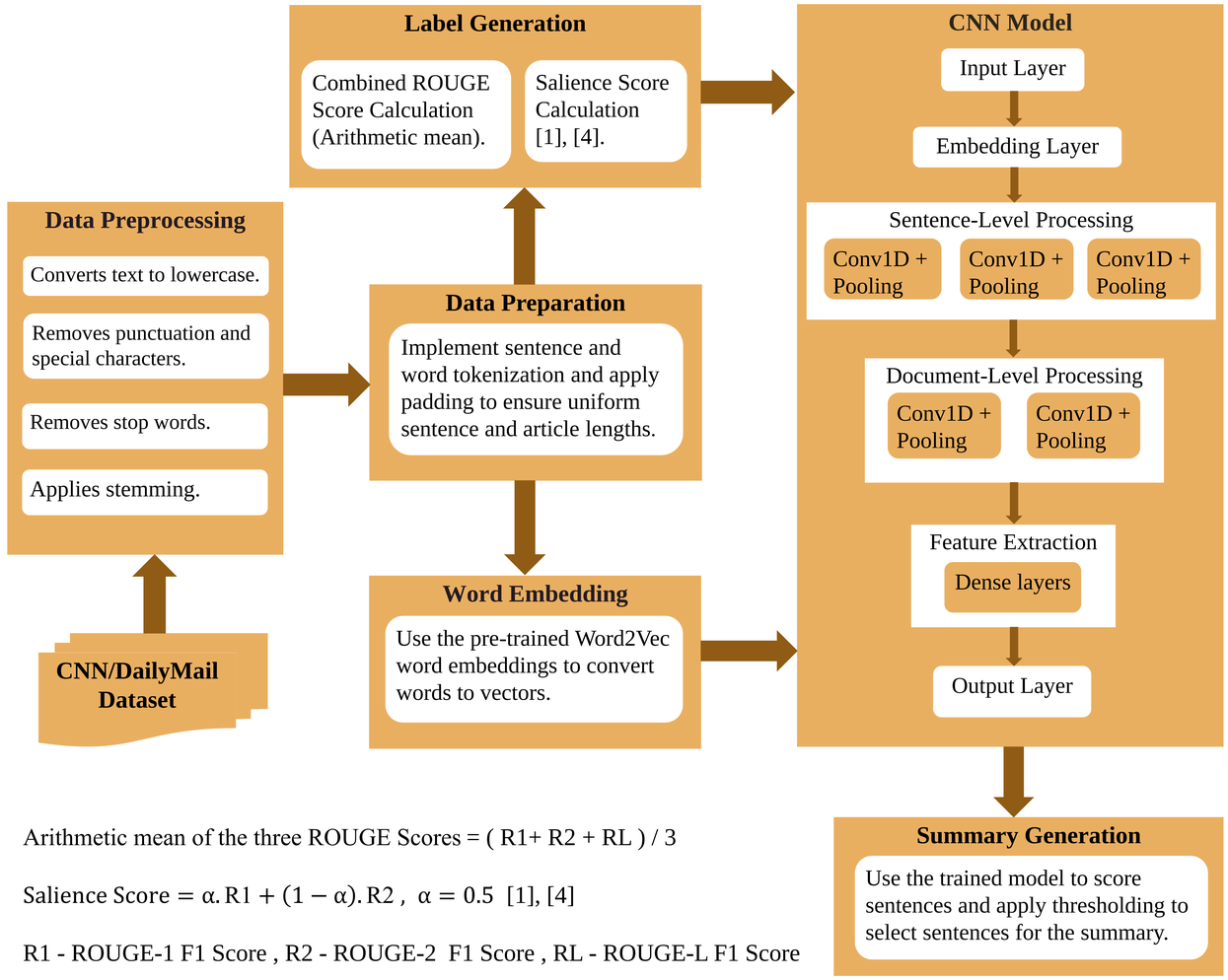
ROUGE-L

$$\text{Precision\_L} = \frac{\text{Length\_of\_longest\_common\_subsequence}}{\text{Length\_of\_system\_summary}}$$

$$\text{Recall\_L} = \frac{\text{Length\_of\_longest\_common\_subsequence}}{\text{Length\_of\_reference\_summary}}$$

$$\text{FI Score} = \frac{2 * (\text{Precision\_L} * \text{Recall\_L})}{(\text{Precision\_L} + \text{Recall\_L})}$$

Methodology



Arithmetic mean of the three ROUGE Scores = (R1+ R2 + RL) / 3

Salience Score $= \alpha.\,R1 + (1-\alpha).\,R2$, $\alpha = 0.5$ [1], [4]

R1 - ROUGE-1 F1 Score , R2 - ROUGE-2 F1 Score , RL - ROUGE-L F1 Score

Test Results

The table presents the results obtained from this research.

ROUGE Metrics	Evaluation Metrics	Combined ROUGE Score Calculation	Salience Score Calculation
ROUGE-1 (Unigram Overlap)	Precision	0.6948	0.6903
	Recall	0.8660	0.8504
	F1 Score	0.7401	0.7283
ROUGE-2 (Bigram Overlap)	Precision	0.6488	0.6451
	Recall	0.8254	0.8075
	F1 Score	0.6910	0.6793
ROUGE-L (LCS)	Precision	0.6896	0.6850
	Recall	0.8597	0.8441
	F1 Score	0.7347	0.7229

The table shows the performance of other approaches.

Ref	Dataset	Methods	ROUGE-1 F1 Score	ROUGE-2 F1 Score	ROUGE-L F1 Score
[1]	DUC 2004	CNN Model	0.3846	0.0823	-
[3]	CNN/ Daily Mail	Random Forest	0.82845	0.2884	0.79694
		Naïve Bayes (Bernoulli)	0.28015	0.02206	0.02855
		SVM (Radial Basis Function)	0.11009	0.00003	0.11753
[4]	DUC 2007	CNN Model	0.3968	0.1026	-

Conclusion

- In this research, an extractive text summarization approach was implemented using a CNN model.
- This approach combines word embeddings, sentence-level processing, and document-level analysis to score the importance of sentences within a document..
- Two methods were employed for label generation: one combining ROUGE scores, and the other using salience score calculation.
- The combined ROUGE score method produced better results.
- The proposed model performs well compared to existing state-of-the-art methods [3].

References

[1] Zhang, Yong, Joo Er Meng, and Mahardhika Pratama. "Extractive document summarization based on convolutional neural networks." IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, pp. 918-922, 2016.
[2] Alquliti, Wajdi Homaid, and Norjihan Binti Abdul Ghani. "Convolutional neural network based for automatic text summarization." International Journal of Advanced Computer Science and Applications 10, no. 4 (2019).
[3] Jian, Hew Zi. "Text Summarization for News Articles by Machine Learning Techniques." Applied Mathematics and Computational Intelligence (AMCI) 11, no. 1 (2022): 174-196.
[4] Charitha, Sangaraju, Nagaratna B. Chittaragi, and Shashidhar G. Koolagudi. "Extractive document summarization using a supervised learning approach."2018 IEEE Distributed Computing, pp. 1-6, 2018.
[5] Shirwandkar, Nikhil S., and Samidha Kulkarni. "Extractive text summarization using deep learning."2018 fourth international conference on computing communication control and automation (ICCUBEA), pp. 1-5. IEEE, 2018.
[6] Alwis, Varuni. "Intelligent E-news summarization.“ 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 189-195. IEEE, 2018.