Extractive Text Summarization

Shreya Ballijepalli, Janani Vijayarajan and Vishnu Bharadwaj Suresh

Purdue University

September 28, 2022

1 Introduction

The problem statement we are trying to solve is extractive text summarization, which involves generating a concise and meaningful summary from a document by identifying relevant sentences. We are looking for approaches to improve the diversity and reduce the redundancy in the generated summary, and analyzing the effectiveness of architectures such as Word2Vec, BERT and ELMo in this task.

2 Research Question

We have analyzed the impact of different word representations on the extractive summarization task based on a centroid approach, which is based on the geometric meaning of the centroid vector of a document and identifies sentences which have a high similarity score to the centroid to be part of the summary. A concise and diverse summary can only be generated if the sentence and centroid representations effectively capture context and semantics of the words. The main research topic we are focusing on is whether models such as Word2Vec, BERT and ELMo succeed in generating such a diverse and concise summary.

3 Modeling

3.1 Centroid-based Approach

Given a document corpus $[D_1, D_2, ..D_n]$ with a vocabulary size of $|V|$, we represent each word w in the vocabulary using different word representations. Each sentence is represented as a summation over all word representations.

$$Embedding(S) = \sum_{w \in S} Embedding(w)$$

The centroid embedding is generated for each document D based on the words having the $tfidf$ weight score above a certain threshold K. The centroid embedding is represented as follows:

$$C = \sum_{w \in D, tfidf(w) > K} Embedding(w)$$

Since the centroid embedding consists of all top words in the document, it represents a pseudo-document which condenses the meaning of the document. The sentences with the highest cosine similarity are selected and included in the summary. The intuition behind this is that the sentence with the highest cosine similarity to the centroid embedding will give a good representation of the document.

3.2 Word Representations

3.2.1 Word2Vec Skip-gram model

The Word2Vec algorithm uses a neural network model to learn word associations from a large corpus of text. It is modeled in a way such that the cosine similarity indicates the level of semantic similarity. We chose this representation to evaluate if its effectiveness in capturing semantic relatedness is reflected in the similarity between the sentence and centroid representations.

3.2.2 ELMo

ELMo word vectors are computed on top a bi-directional language model. The generated word representation is based on the entire sentence and is effective is capturing polysemy and out of vocabulary words. We chose this representation that takes the sentence's context into the representation to analyze its effectiveness to map diverse sentences in the summary.

3.2.3 BERT

BERT is a transformer language model with a variable number of encoder layers and self-attention heads. The generated word representations has a increased capacity for understanding context and ambiguity in language. We want to analyze the ability of contextual representation and disambiguation in the effectiveness of summary generation.

3.3 Tradeoffs

We fine-tuned a pre-trained version of the BERT model on the dataset for the summarization task. This could be trained for lesser than 10 epochs given the time taken in training and the computational power required. Since this didn't give reasonable results in terms of the ROUGE score, we used a pre-trained version of BERT.

4 Experimental setup

As part of the evaluation, we used the CNN/DailyMail dataset which is a popular summarization dataset consisting of news articles and the highlights representing the gold summary. We performed both qualitative and quantitative evaluation to compare the three approaches.

1. **Qualitative Evaluation**

   (a) Semantic Relevance
   We visualize the sentence embeddings with the highest similarity score to the centroid. This is done to analyze if the embeddings of the words part of the sentence overlap with the word embeddings in the centroid representing that words with similar context/semantic representation are closer to one another.

   (b) Context and Diversity
   We manually analyze the generated summaries for all three approaches to understand if they represent the main context of the document and analyze the diversity and validate the lack of redundancy in the summary.

2. **Quantitative Evaluation**

    (a) ROUGE scores
    ROUGE is a standard metric in text summarization that calculates the similarity between a candidate document and a collection of reference documents on the basis of the n-gram overlap. We want to analyze which approach gives the highest ROUGE score with the gold standard summaries.

    (b) Sentence Similarity Metric
    We plot the sentence similarity scores with the centroid for all three approaches. This will evaluate which approach produces the sentences with the highest semantic similarity.

5 Experimental results

ROUGE Scores
We evaluated the ROUGE recall scores on all the three approaches in 1.

	Rouge-1	Rouge-2	Rouge-L
Word2Vec	28.4	10.5	18
ELMo	56.7	20.5	35.8
BERT	**56.8**	**20.8**	**36.3**

Table 1: ROUGE scores % on CNN/DailyMail dataset

The BERT embeddings have the highest ROUGE-1, ROUGE-2 and ROUGE-L scores comparable to ElMo representing these two approaches capture semantics and context better than Word2Vec. 1.

Sentence Similarity Scores
We plotted the cosine similarity scores between each sentence to the centroid embedding in one of the summaries and plotted it in Figure 1.

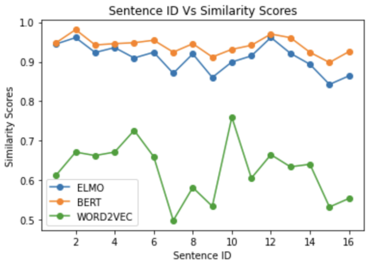

Figure 1: Comparison of sentence similarity scores with the centroid

The similarity scores show that BERT performs a bit higher than ELMo which is higher than Word2Vec embeddings.

Visualization of sentence and centroid embeddings
We plot the word embeddings part of the highest scored sentence using TSNE with the embeddings in the centroid vector in Figure 2. We can see that the overlap between centroid

and sentence embeddings is higher in BERT and ELMo whereas it is scattered in Word2Vec representing that both BERT and ElMo can place words with higher semantic similarity closer.

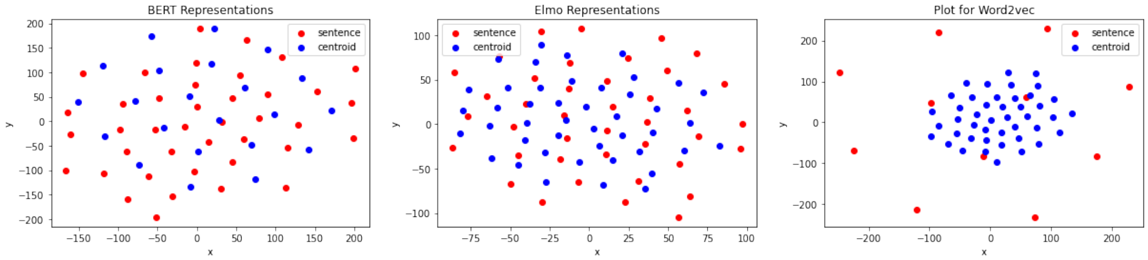

Figure 2: Visualisation of centroid and sentence embeddings

Diversity Analysis in the Summary

One of the documents from the dataset was randomly selected and the summaries were generated for all three approaches.

About the document: The document deals with the aftermath of the 2008 Beijing Olympics on China and starts off on a positive note about all the major changes that continued on after the end of the Olympic Games, then transitions into a negative tone, critiquing some things that have not changed and highlighting some of the promises made before the Games that were not upheld.

The summaries generated for this document for all the three approaches are represented in Table 2.

Word2Vec	“The legacy is bad and good,” opined James McGregor. He credits it for speeding up the modernization of the city’s infrastructure, from roads to telecoms, to subway lines. “The world had a chance to see a different China from the one that is making the headline news,” he said.
BERT	Beijing residents have become more aware of environment issues, but many “green” projects remain unfinished. In recent months, foreign reporters have encountered a number of obstacles, especially in sensitive areas. For good or for ill, Beijing is changing fast.
ELMo	Inbound tourism remains robust, thanks to the massive media exposure China got from the 17-day event. But four years after the Games, Beijing’s pollution indexes are still hitting record highs. China still routinely blocks internet access and locks up whistle-blowing journalists.

Table 2: Parts of summaries generated by different approaches

The summaries provided by BERT and ELMo are very similar to each other, capturing the fact that although some positive points about China’s development were mentioned, the overall tone of the document is slightly negative, with more sentences in the summary relating to the criticisms. The summary by Word2Vec, on the other hand, captures very little about the dual tone of the paper, and has very little pertaining to the criticisms.

6 Discussion

The centroid-based extractive summarization task was implemented for different word embedding models, with summaries being obtained for the CNN/DailyMail summarization dataset and qualitative analysis was performed using the results. It was found that the BERT embeddings generated the best representation of the summary, followed closely by ELMo and Word2Vec.

The only major change from the project proposal to the final implementation of the project was the replacement of the LSTM-based embedding model by ELMo embeddings.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[2] Demian Gholipour Ghalandari. Revisiting the centroid-based method: A strong baseline for multi-document summarization, 2017.

[3] Salima Lamsiyah, Abdelkader El Mahdaouy, Bernard Espinasse, and Saïd El Alaoui Ouatik. An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. *Expert Systems with Applications*, 167:114152, 2021.

[4] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[6] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[7] Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21, Valencia, Spain, April 2017. Association for Computational Linguistics.