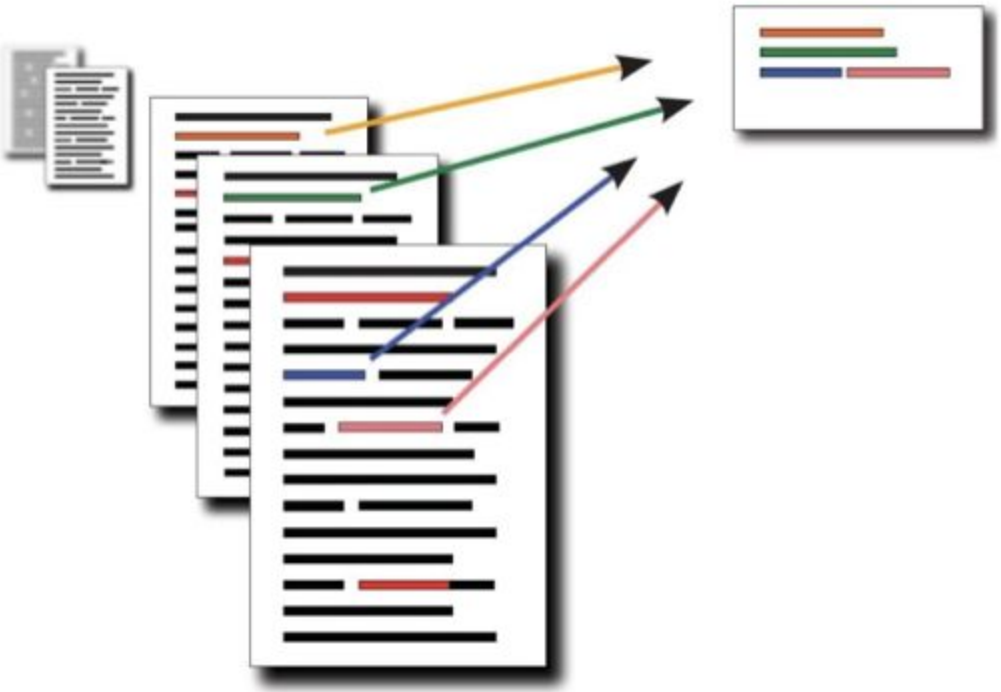
Text Summarization

Long text, big problem?

What about TL;DR?

What is it?

Task: produce an abridged version of a text while retaining the key, relevant information



What is it?

Useful for creating:

- outlines or abstracts of any document, article, etc
- summaries of chat and email
- action items from a meeting
- simplifying text by compressing sentences

Single-doc summarization

Document

Cambodian leader Hun Sen on Friday rejected opposition parties ' demands for talks outside the country , accusing them of trying to `` internationalize " the political crisis .
Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen 's party to form a new government failed .
Opposition leaders Prince Norodom Ranariddh and Sam Rainsy , citing Hun Sen 's threats to arrest opposition figures after two alleged attempts on his life , said they could not negotiate freely in Cambodia and called for talks at Sihanouk 's residence in Beijing .Hun Sen , however , rejected that .``
I would like to make it clear that all meetings related to Cambodian affairs must be conducted in the Kingdom of Cambodia , " Hun Sen told reporters after a Cabinet meeting on Friday .`` No-one should internationalize Cambodian affairs .
It is detrimental to the sovereignty of Cambodia , " he said .Hun Sen 's Cambodian People 's Party won 64 of the 122 parliamentary seats in July 's elections , short of the two-thirds majority needed to form a government on its own .Ranariddh and Sam Rainsy have charged that Hun Sen 's victory in the elections was achieved through widespread fraud .They have demanded a thorough investigation into their election complaints as a precondition for their cooperation in getting the national assembly moving and a new government formed .....

Summary

Cambodian government rejects opposition's call for talks abroad

Output

Extractive summarization

Select from the source text spans that capture the key information

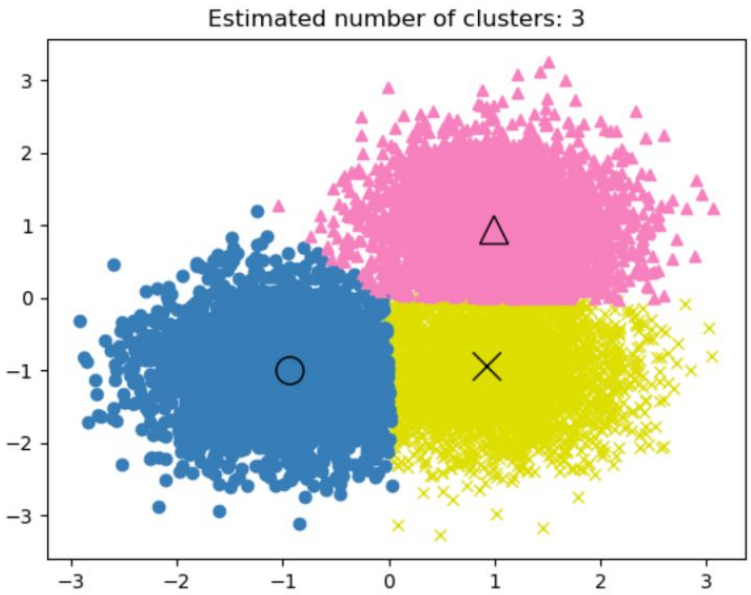
Abstractive summarization

Generate new text that encapsulates the key information from the source text

For abstractive, we generally use transformers and GPT based architecture

Any thoughts about how to do extractive summarization?

Extractive Summarization



Brute force:

Sentence embeddings from BERT or USE

K-means clustering to find common themes

Convert into lower dimensional space

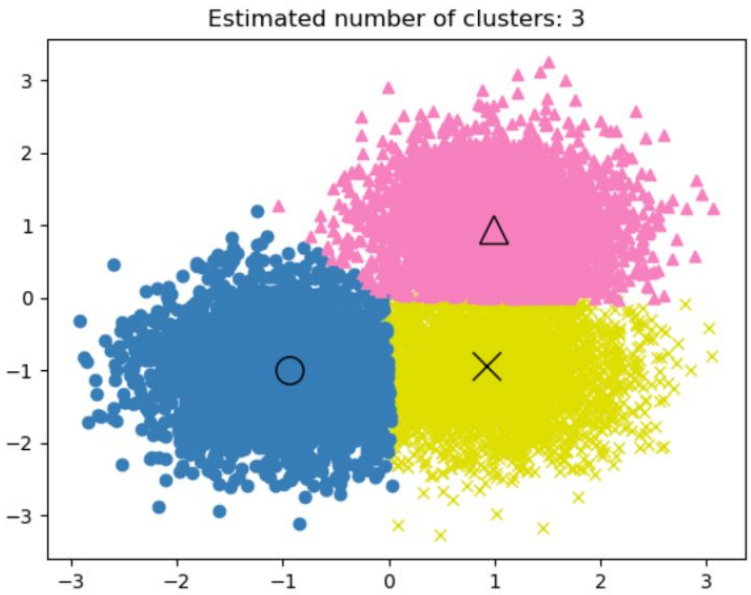
Take the sentence at or the closest to centroid

Sometimes this approach works

But do you see any problems? Let's discuss!

https://www.width.ai/post/bert-for-extractive-text-summarization-on-lectures

Extractive Summarization



Brute force:

Sentence embeddings from BERT or USE

K-means clustering to find major themes

~~Convert into lower dimensional space~~

Take the sentence at or closest to the centroid

Problems still exist!

Centroid may not be the best choice.
Corset could be a solution

https://www.width.ai/post/bert-for-extractive-text-summarization-on-lectures

Extractive Summarization

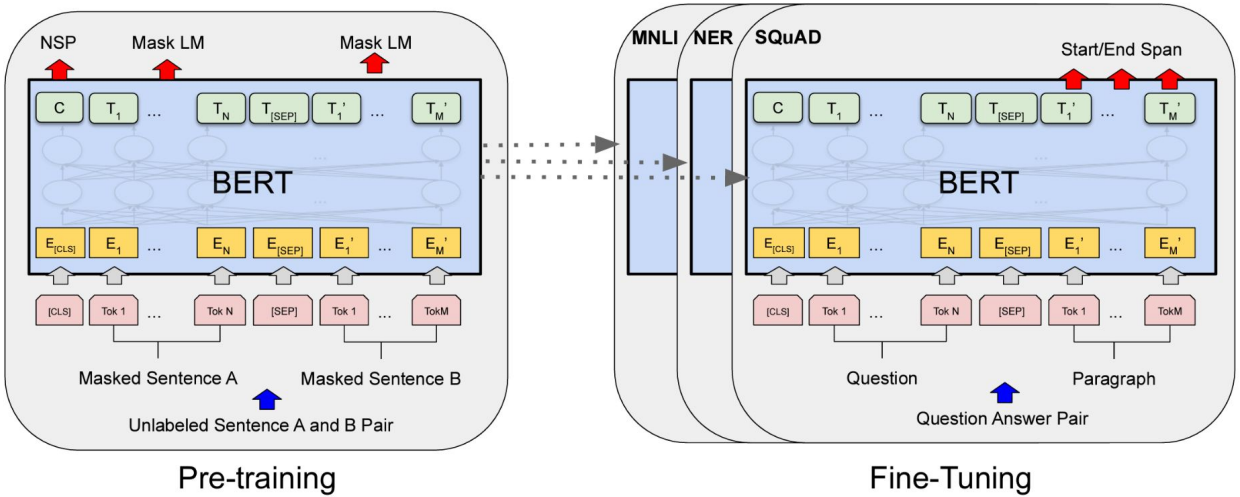


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

Let's discuss! How can you fine-tune BERT for summarization?

Extractive Summarization

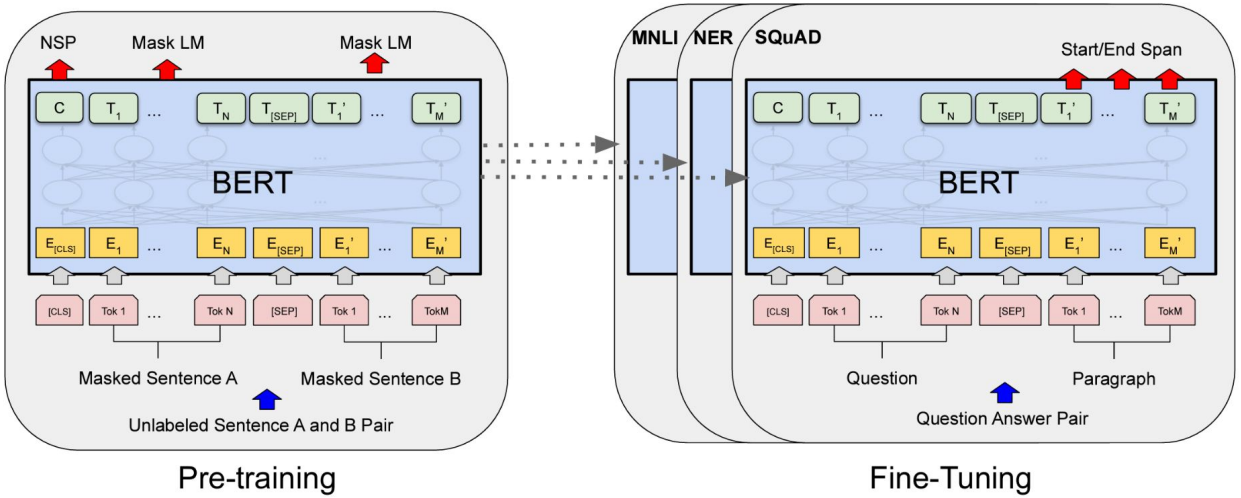
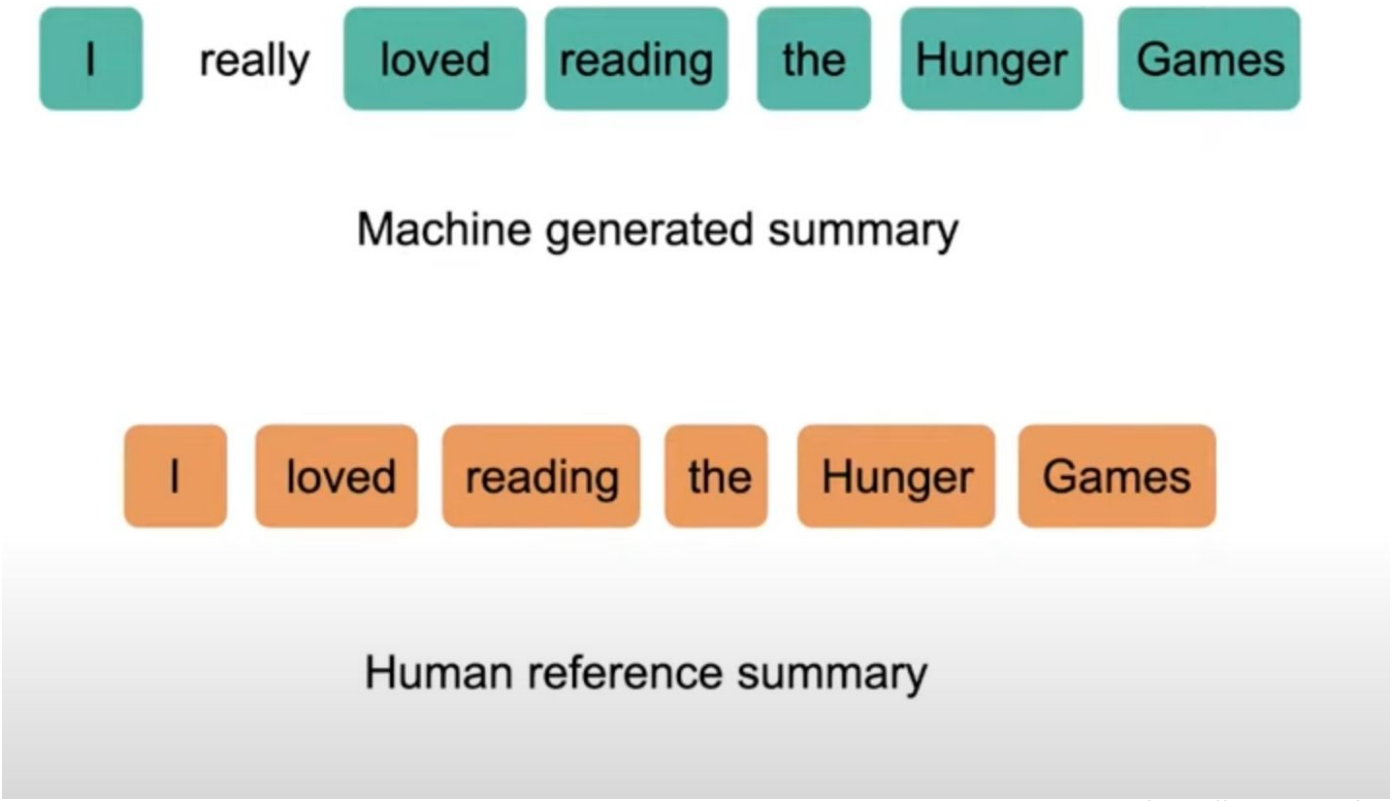


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

In transfer learning, what are we exactly transferring?

[CLS] token before each sentence. BERTSUM is one such approach

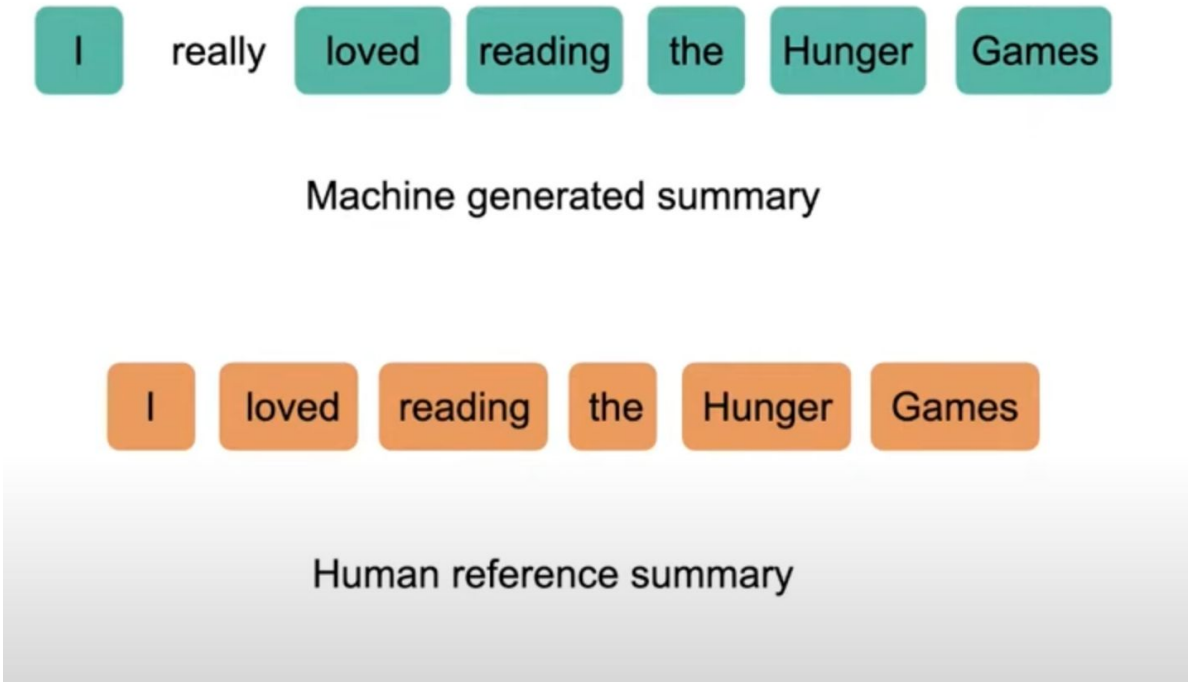
Evaluation–Rogue Score



6 common words

https://www.youtube.com/@HuggingFace

Evaluation–Rogue Score



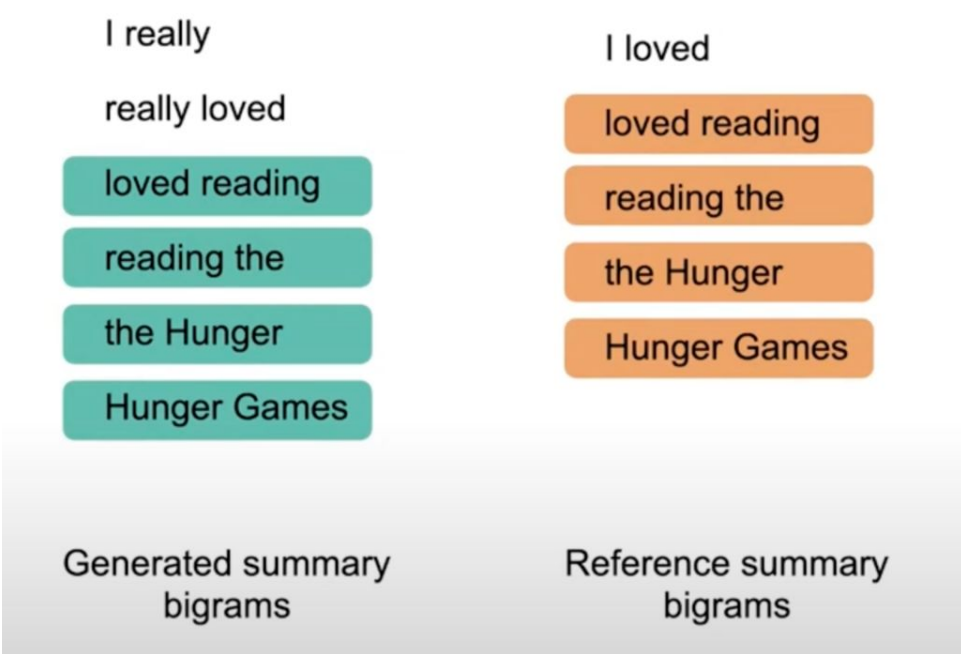
$$\text{ROUGE-1 recall} = \frac{\text{Num word matches}}{\text{Num words in reference}} = \frac{6}{6}$$

$$\text{ROUGE-1 precision} = \frac{\text{Num word matches}}{\text{Num words in summary}} = \frac{6}{7}$$

$$\text{ROUGE-1 F1-score} = 2\left(\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\right)$$

https://www.youtube.com/@HuggingFace

Evaluation–Rogue Score

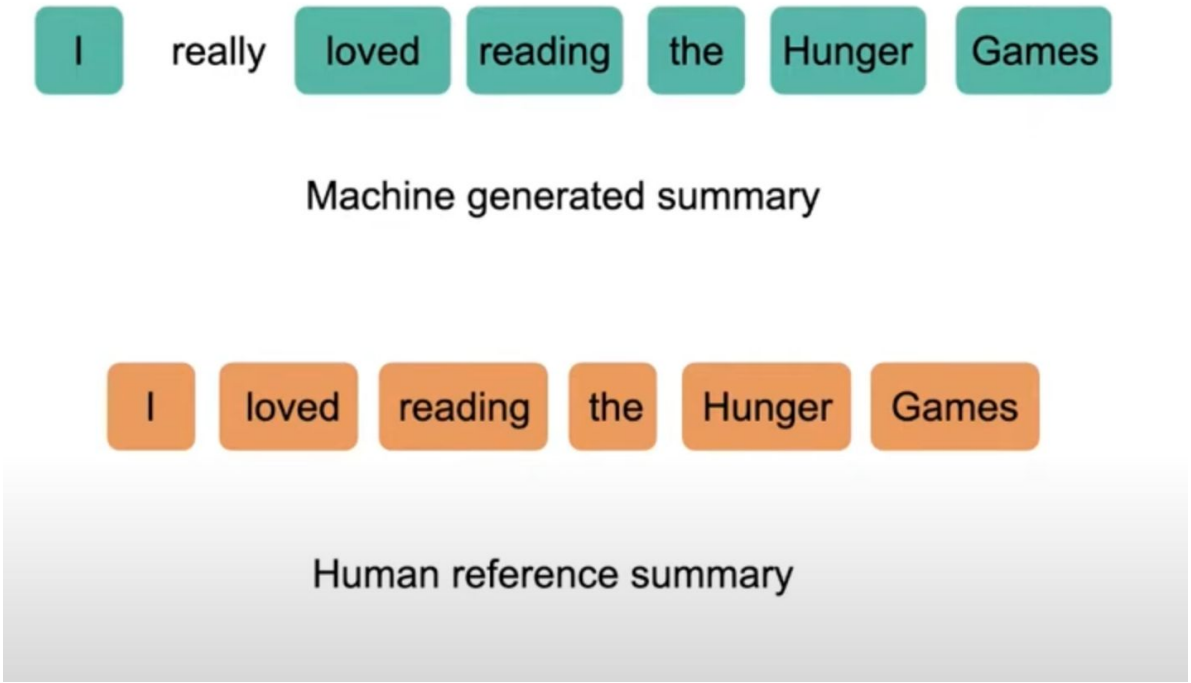


$$\text{ROUGE-2 recall} = \frac{\text{Num bigram matches}}{\text{Num bigrams in reference}} = \frac{4}{5}$$

$$\text{ROUGE-2 precision} = \frac{\text{Num bigram matches}}{\text{Num bigram in summary}} = \frac{4}{6}$$

https://www.youtube.com/@HuggingFace

Evaluation–Rogue Score



$$\text{ROUGE-L recall} = \frac{\text{LCS(gen, ref)}}{\text{Num words in reference}} = \frac{6}{6}$$

$$\text{ROUGE-L precision} = \frac{\text{LCS(gen, ref)}}{\text{Num words in summary}} = \frac{6}{7}$$

Longest Common Subsequence!

https://www.youtube.com/@HuggingFace

Acknowledgements

Dr. Chris Tanner at Harvard

Hugging Face YouTube Lectures