Unsuccessful Approach to Apply Summarization in Information Retrieval

Department: name: Speech and Semantic Lab, Noah's Ark Lab
Author's name: Valentin Malykh
Date: 03/12/2021

HUAWEI

Our Team

Dr. Ilseyar Alimova
Andrey Bout
Dr. Valentin Malykh

Information Retrieval

- Users
- Queries
- Documents

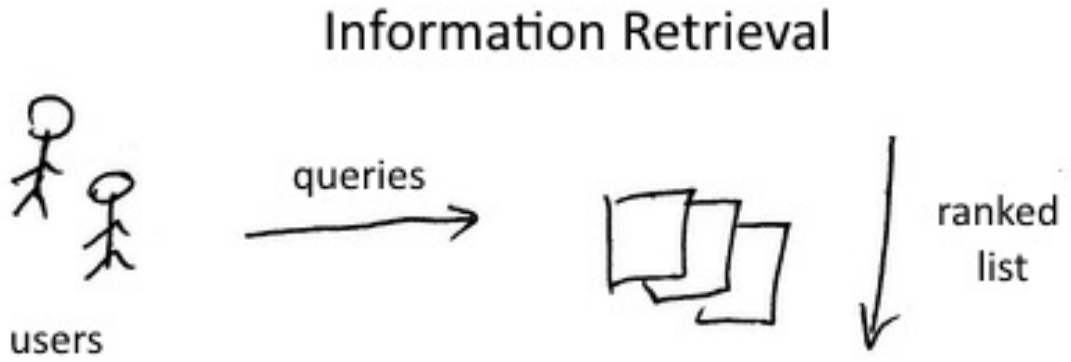
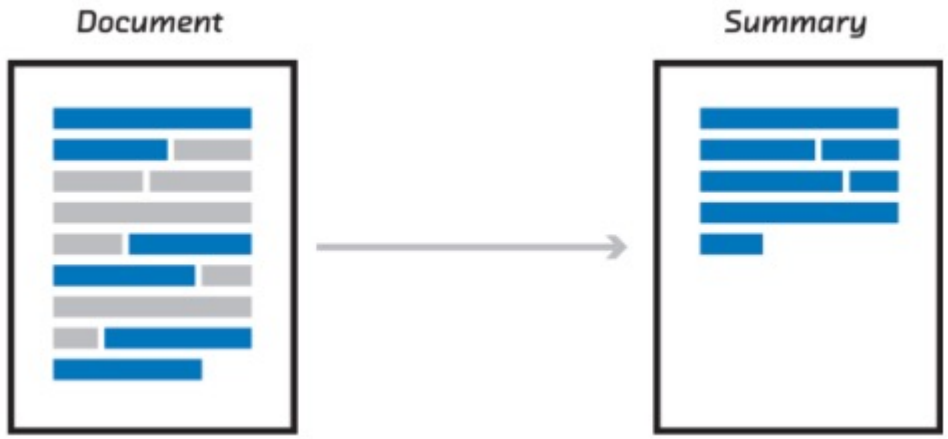


Image from MLWiki.org

HUAWEI

Summarization

- Document
- Summary



HUAWEI

Our Setup

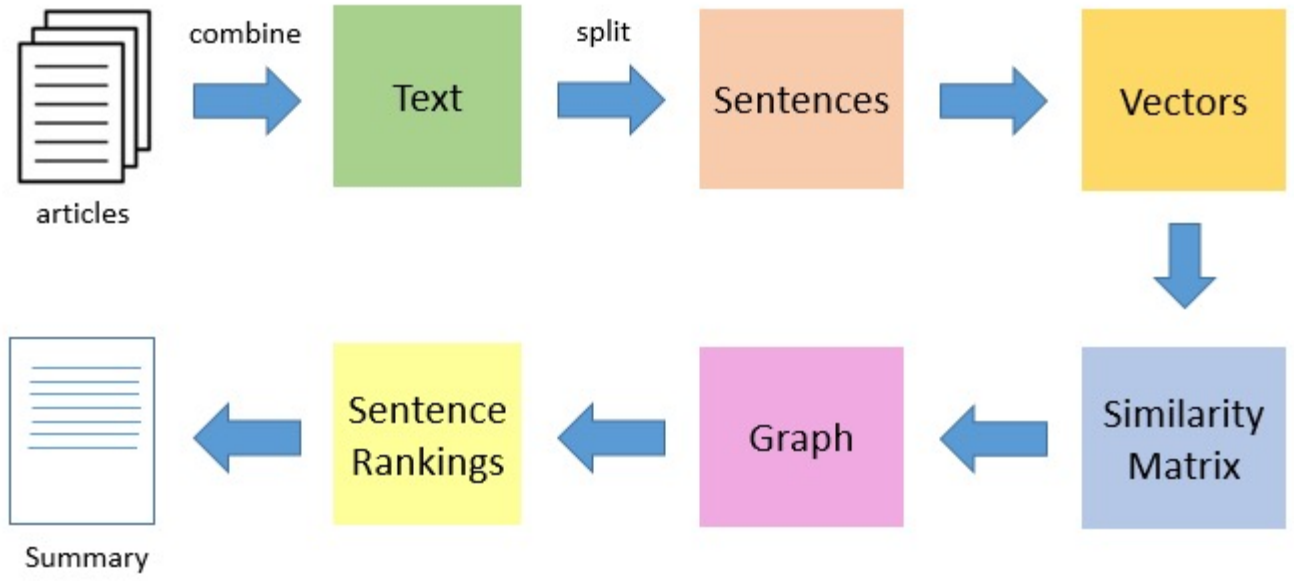
- Corpora
- MS MARCO
- CNN / DailyMail

- Summarization
- TextRank
- mBART

- Ranking
- ColBERT

HUAWEI

Summarization Models



HUAWEI

Summarization Models

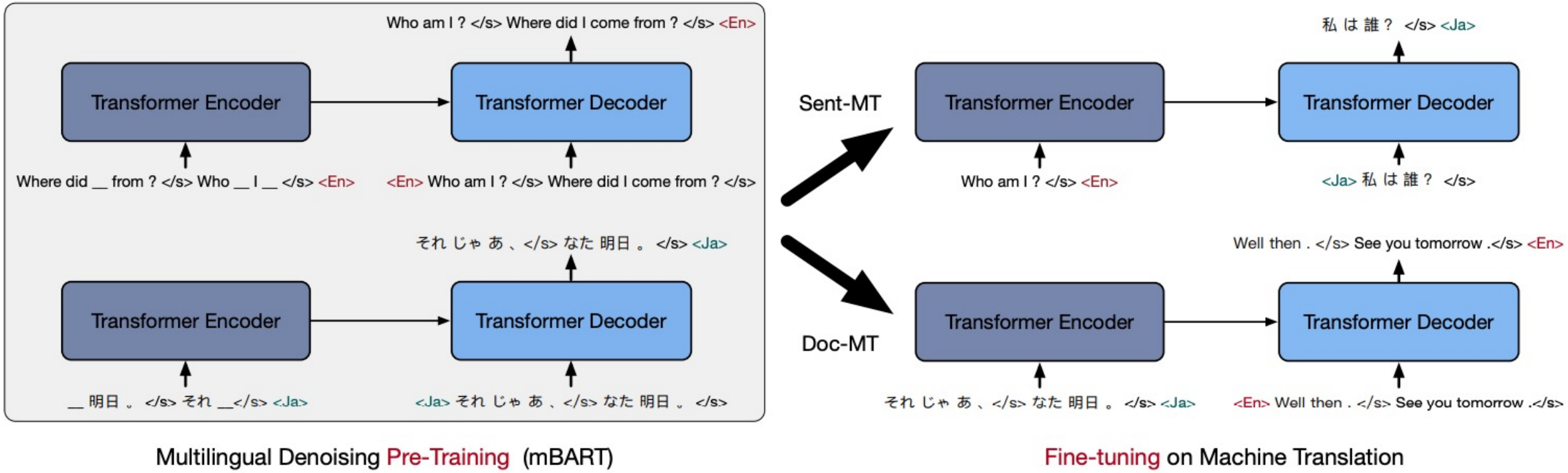


Figure 1: Framework for our Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained model is used for all tasks.

Ranking Model

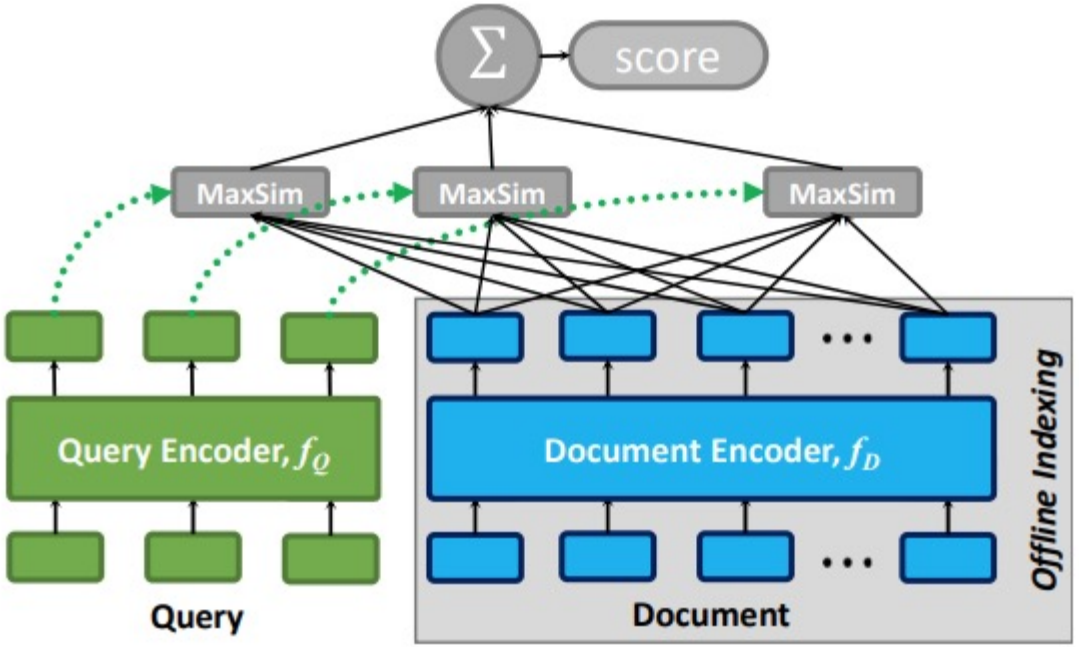
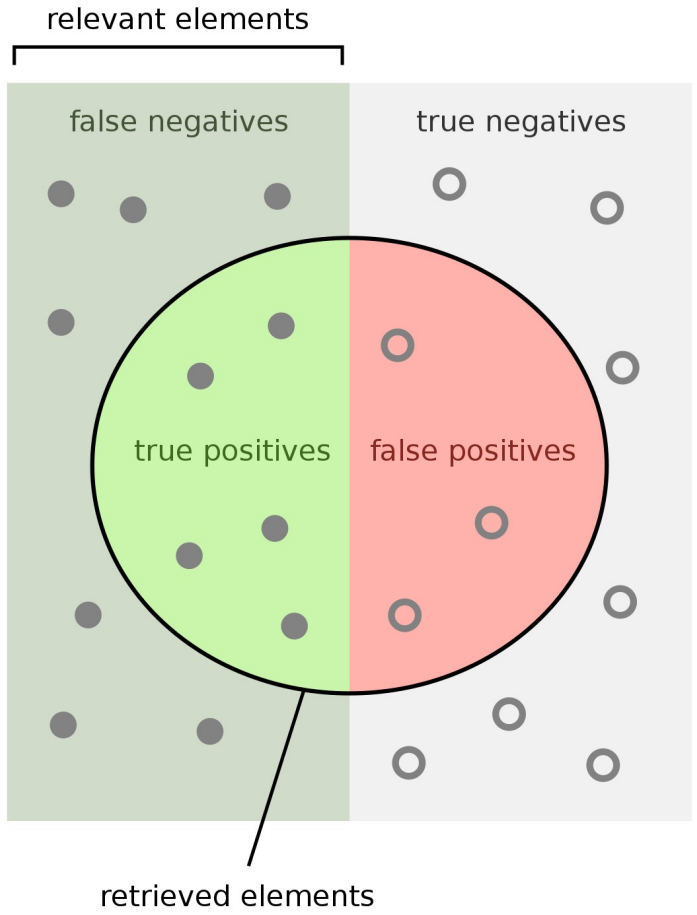


Figure 3: The general architecture of ColBERT given a query q and a document d.

HUAWEI

Metrics

Recall@k



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{retrieved elements}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

HUAWEI

Preliminary Results

Summarizer Model	Index size (Gb)	Avg. document length	Recall@100	Recall@10
Full documents	59	1124	0.866	0.603
First 512	10	512	0.854	0.605
TextRank	10	165	0.776	0.507
mBART (fine-tuned on CNN)	10	194	0.824	0.575

HUAWEI

Key Takeaways

- Even the best abstract summarization could be not enough
- There is more than main idea in the text

HUAWEI

Thank you

Valentin Malykh
Speech and Semantics Laboratory
Huawei Noah’s Ark lab

valentin.malykh@huawei.com