Big Data Analytics and Text Mining (Module 1)

Last update: 02 October 2024

Academic Year 2024 – 2025
Alma Mater Studiorum · University of Bologna

Contents

1 Automatic text summarization

Extractive summarization Select fragments of text.

Extractive summarization

Abstractive summarization Rephrase the content of the text.

Abstractive summarization

Hybrid summarization Apply an extractive method followed by an abstractive one.

Hybrid summarization

Generic vs query-focused

- **Generic** Summary of the whole document.
- **Query-focused** Summary that replies to given questions

Technical vs lay

- **Technical** Summary using scientific language.
- **Lay** Summary using common language.

Narrative vs bullet point

- **Narrative** Standard textual summary.
- **Bullet point** Set of key phrases.

Single document vs multi document

- **Single document** Summary covering a single document.
- **Multi document** Summary covering multiple documents.

Short document vs long document

- **Short document** Summary of a document with a few tokens.
- **Long document** Summary of a document with many tokens.

1.1 Metrics

Summarization metrics can evaluate different levels:

Syntactic Check word overlapping (e.g., ROUGE).

Semantic Check semantic coverage (e.g., BERTScore).

Factuality Check factuality to the source (e.g., BARTScore).

Fluency Check for redundancies (e.g., unique N-gram ratio).

Efficiency Measure trade-off between performance and costs (e.g., CARBURACY).

1.1.1 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

ROUGE

ROUGE N-gram oriented metric that compares the generated summary and the ground truth.

ROUGE-1 Overlap of 1-grams.

ROUGE-2 Overlap of 2-grams.

ROUGE-L Length of the common longest subsequence.

Precision

$$\texttt{ROUGE}_{\texttt{precision}} = \frac{|\text{overlaps}|}{|\text{generated summary}|}$$

Recall

$$\texttt{ROUGE}_{\texttt{recall}} = \frac{|\text{overlaps}|}{|\text{ground truth}|}$$

1.1.2 Limitations

- ROUGE only evaluates on a syntactic level.
- ROUGE-2 and ROUGE-L are sensitive to the position of words.

1.2 State-of-the-art generative summarizers

1.2.1 BART

BART

- Encoder-decoder Transformer with an input size of 1024 tokens.
- It is suited for short document summarization.
- It is pre-trained using a denoising sequence-to-sequence approach.

1.2.2 Longformer encoder-decoder

Longformer encoder-decoder

- Encoder-decoder Transformer with an input size of 16k tokens.
- It is suited for long document summarization.
- It uses a linear encoder self-attention based on global and local attention that reduces the quadratic complexity of the standard attention mechanism.

1.2.3 PRIMERA

PRIMERA

- Encoder-decoder Transformer based on Longformer with an input size of 4K tokens.
- It is suited for long document summarization.
- It has an ad-hoc pre-training for multi document summarization.