AnyRef - Document Summarisation and Its Evaluation Using Metrics

Presented By: Lohith Reddy Kalluru, Md Zahid Hasan

IOWA STATE UNIVERSITY

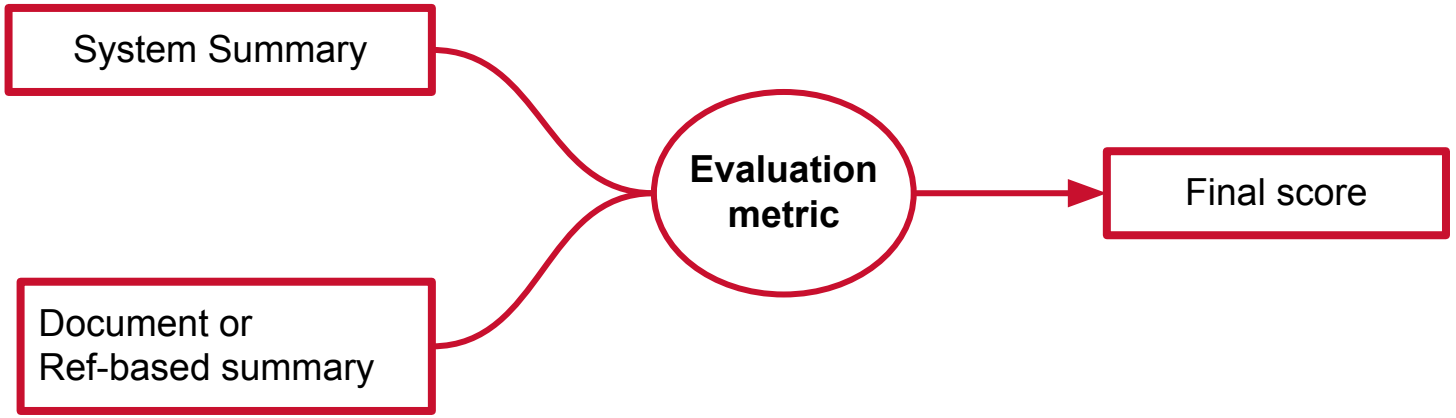
Outline

01	Summarizer	• Models (summarizer) • Datasets (Newsroom)
02	Evaluation Metric	• Pseudo_ref function • EvalBase framework • ROUGE score
03	Results	• Scores • Generalization

Introduction

- The method of extracting these summaries from the original huge text without losing vital information is called as Text Summarization. It is essential for the summary to be a fluent, continuous and depict the significant.
- The approach is to identify the important sections, interpret the context and reproduce in a new way. This ensures that the core information is conveyed through shortest text possible. Note that here, the sentences in summary are generated, not just extracted from original text.

Summarization evaluation/metrics



Dataset

- Currently, we are using newsroom eval dataset for testing purposes.
- There are 60 articles, 7 systems, and 3 ratings by different people for each system for a total of 1,260 rows. Each row is a single rating across the four dimensions from the paper. The columns are:

  1. System - name of the sumarization system/baseline being rated
  2. ArticleID - index of the article in the dataset, can be used to group by article
  3. ArticleText, ArticleTitle, SystemSummary - three inputs the rater sees (HTML encoded for AMT)
  4. CoherenceRating, FluencyRating, InformativenessRating, RelevanceRating - ratings (1-5)

	ArticleID	System	ArticleText	SystemSummary	ArticleTitle	CoherenceRating	FluencyRating	InformativenessRating	RelevanceRating	ReferenceSummary
0	2140	fragments	A worker sets up a polling station the morning...	John Avlon voter turnout in the is a sign of a...	Why has GOP turnout taken a dive?	2	2	2	2	John Avlon says low voter turnout in the prima...
1	2140	fragments	A worker sets up a polling station the morning...	John Avlon voter turnout in the is a sign of a...	Why has GOP turnout taken a dive?	4	5	4	4	John Avlon says low voter turnout in the prima...
2	2140	fragments	A worker sets up a polling station the morning...	John Avlon voter turnout in the is a sign of a...	Why has GOP turnout taken a dive?	2	3	2	3	John Avlon says low voter turnout in the prima...
3	2140	textrank	A worker sets up a polling station the morning...	In New Hampshire , the same dynamic applied --...	Why has GOP turnout taken a dive?	4	5	4	5	John Avlon says low voter turnout in the prima...

pseudo_func

- Input
  - documents
  - system_summaries
  - summarizer_name
  - ref_based_metric_name
- Output
  - scores

```
def pseudo_func(predictions,references, model_name, ref_based_metric_name):
  device = "cuda" if torch.cuda.is_available() else "cpu"
  tokenizer = AutoTokenizer.from_pretrained(model_name)

  # https://stackoverflow.com/questions/70544129/transformers-asking-to-pad-but-the-tokenizer-does-not-have-a-padding-token
  if tokenizer.pad_token is None:
    tokenizer.add_special_tokens({'pad_token': '[PAD]'})

  if "google" in model_name.lower():
    model = PegasusForConditionalGeneration.from_pretrained(model_name).to(device)
  else :
    model = AutoModelForCausalLM.from_pretrained(model_name).to(device)
```

summarizer_name: Hugging face model Hub

ref_metric: [ROUGE, BERTscore]

EvalBase/env.py

Takes documents and extracts

- ArticleText
- SystemSummary
- ReferenceSummary
- hum_eval_path
- refs_path
- hum_with_ref_path

```
datasets = {
    "newsroom": {
        "human_metrics": ["InformativenessRating", "RelevanceRating", "CoherenceRating", "FluencyRating"],
        "docID_column": "ArticleID",
        "document_column": "ArticleText",
        "system_summary_column": "SystemSummary",
        "reference_summary_column": "ReferenceSummary",
        "approaches": ["trad", "new"],
        "human_eval_only_path": "dataloader/newsroom-human-eval.csv",  # you need to get this file. See ReadMe.
        "refs_path": "dataloader/test.jsonl",  # you need to get this file. See ReadMe.
        "human_eval_w_refs_path": "dataloader/newsroom_human_eval_with_refs.csv"
    },
```

https://github.com/SigmaWe/EvalBase/blob/main/env.py

Test framework (EvalBase)

- We plugged in the functions through env.py which are partial functions with certain arguments sent to the eval_utils.py when called from newsroom.py.
- Datasets were read automatically without the need for any change.
- It can evaluate through multiple metrics and aggregate the results in json and text format.

```
"google-pegasus-xsum":functools.partial(pseudo_func.pseudo_func,model_name = "google/pegasus-xsum", ref_based_metric_name= "rouge"),
"bert-base-cased":functools.partial(pseudo_func.pseudo_func,model_name = "bert-base-cased", ref_based_metric_name= "rouge"),
"distilgpt2":functools.partial(pseudo_func.pseudo_func,model_name = "distilgpt2", ref_based_metric_name= "rouge"),
"openai-gpt":functools.partial(pseudo_func.pseudo_func,model_name = "openai-gpt", ref_based_metric_name= "rouge"),
"google-pegasus-cnn_dailymail":functools.partial(pseudo_func.pseudo_func,model_name = "google/pegasus-cnn_dailymail",ref_based_metric_name= "rouge"),
"gpt2":functools.partial(pseudo_func.pseudo_func,model_name = "gpt2", ref_based_metric_name= "rouge")
```

Initial Results (Dec 8)

- Model= Pegasus and Eval_metric= ROUGE:

Rogue-1	Rogue-2	RougeL	RougeLsum
0.04743833017077	0.03802281368821	0.04174573055028	0.04174573055028
0.04743833017077	0.03802281368821	0.04174573055028	0.04174573055028
0.08379888268156	0.08208955223880	0.08379888268156	0.08379888268156

Initial Results (Dec 8)

- Model= bert-base-uncased, Eval_metric= ROUGE for **first 15 newsroom documents**:

	Rogue-1	Rogue-2	RougeL	RougeLsum
BERT/ROUGE	0.34360992309416	0.3031325629862	0.33172716963627	0.32889254151974

- Model= bert-base-uncased, Eval_metric=BERTScore for **first 15 newsroom documents**:

	avg_precison	avg_recall	avg_F1-score
BERT/ROUGE	0.83231127262115	0.89182925224304	0.8610429709

Sample of final Results (Dec 17)

corr_metric	aspect	approach	model	score_name	
pearsonr	InformativenessRating	new	google-pegasus-xsum	rouge1	0.369
				rouge2	0.336
				rougeL	0.295
				rougeLsum	0.295
			bert-base-cased	rouge1	0.764
				rouge2	0.716
				rougeL	0.740
				rougeLsum	0.740
			distilgpt2	rouge1	0.776
				rouge2	0.766
				rougeL	0.773
				rougeLsum	0.773
			openai-gpt	rouge1	0.766
				rouge2	0.719
				rougeL	0.741
				rougeLsum	0.741
			pegasus-cnn_dailymail	rouge1	0.638
				rouge2	0.536
				rougeL	0.572
				rougeLsum	0.572
			gpt2	rouge1	0.776
				rouge2	0.766
				rougeL	0.773
				rougeLsum	0.773

Full result folder in the Git repository: https://github.com/SigmaWe/AnyRef_team_1/tree/main/results

Future directions

- Extend the function to accept other arguments related to summariser model or metric → generalize
- Other downstream tasks and comparison

THANK YOU