Yale

Summ_eval library for summarization evaluation

Alexander R. Fabbri,[1] Wojciech Kryściński,[2] and Dragomir Radev [1]

[1]Yale University and [2]Salesforce Research

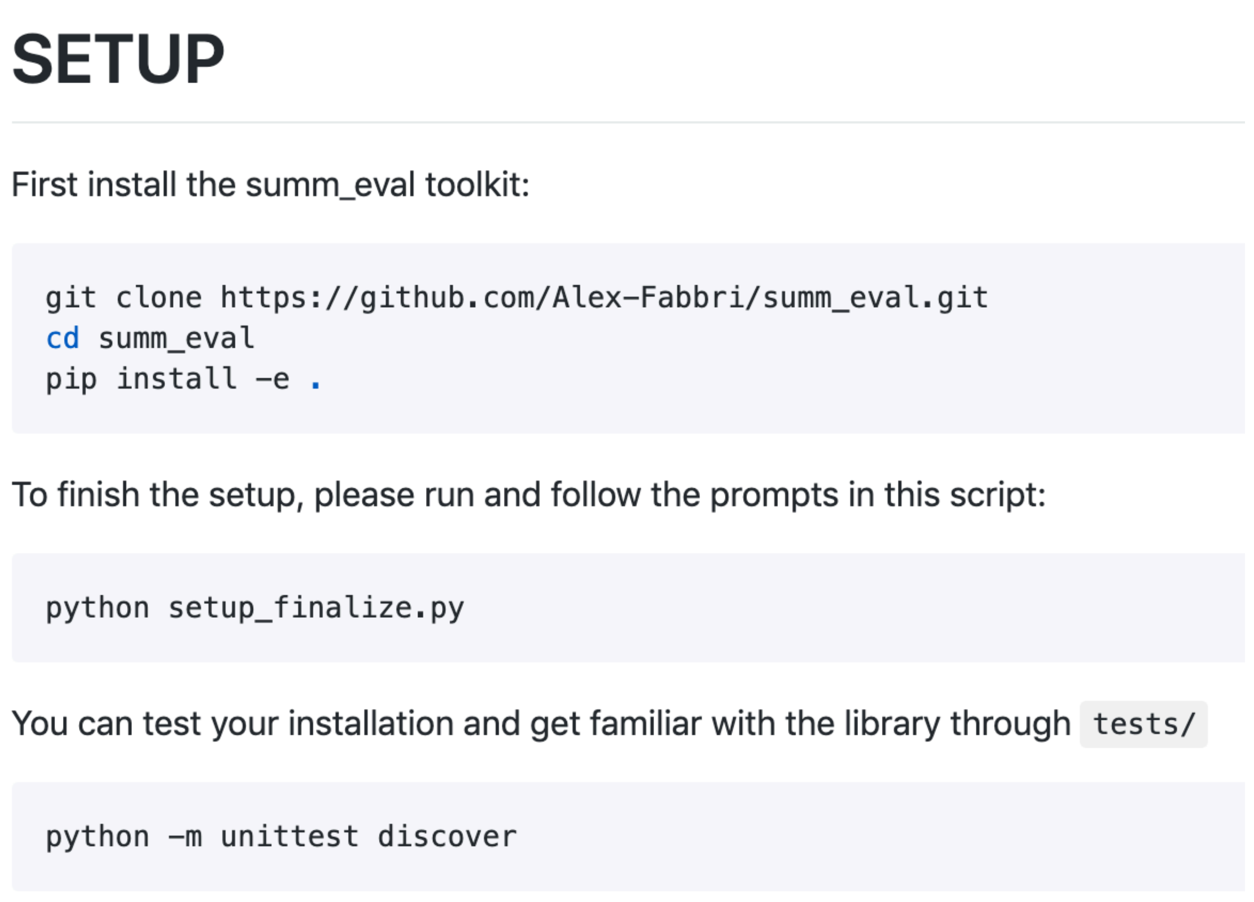
LILY Lab

Introduction

Automatic text summarization is often evaluated using the ROUGE metric (Lin 2004), an automatic metric which calculates n-gram overlap between a reference summary and a machine-generated summary. Despite many variations of ROUGE as well as several follow-up metrics being proposed, ROUGE remains the default evaluation metric. However, ROUGE has been shown to poorly correlate with human judgment outside of its original setting (Nenkova 2006). Part of the reason for ROUGE's prominence is the presence of an easy-to-use package for calculating this score. To make the comparison of text summarization algorithms across additional metrics easier as well as facilitate the growth of well-correlated evaluation metrics, we introduce summ_eval, a toolkit for summarization evaluation consisting of 13 evaluation metrics. Additionally we have begun work collecting human judgments for over 14 recently introduced models.

Choice of Metrics

We choose metrics across a wide range of approaches. We include an interface to ROUGE as well as an embedding-based Rouge metric ROUGE-we (Ng and Abrecht, 2015). Additionally, we add metrics such as Mover Score (Zhao et al., 2019) BERT Score (Zhang et al., 2019) and Sentence Mover's Similarity (Clark et al., 2019) which all make use of some form of Word Mover's Distance (Kusner et al., 2015) and recent advances in representational power. We also include a Question Answering-based metric called SummaQA (Scialom et al., 2019) and a learned regression-based metric S3 (Peyard et al., 2017). We include a metric which calculates dataset statistics as well as the syntactic complexity of a dataset. Finally, we include standard metrics which have been applied in machine translation, such as METEOR (Banerjee and Lavie, 2005), BLEU (Papineni et al., 2002), CHRF (Popovic 2017) and Cider (Vedantam et al., 2015)



SETUP

First install the summ_eval toolkit:

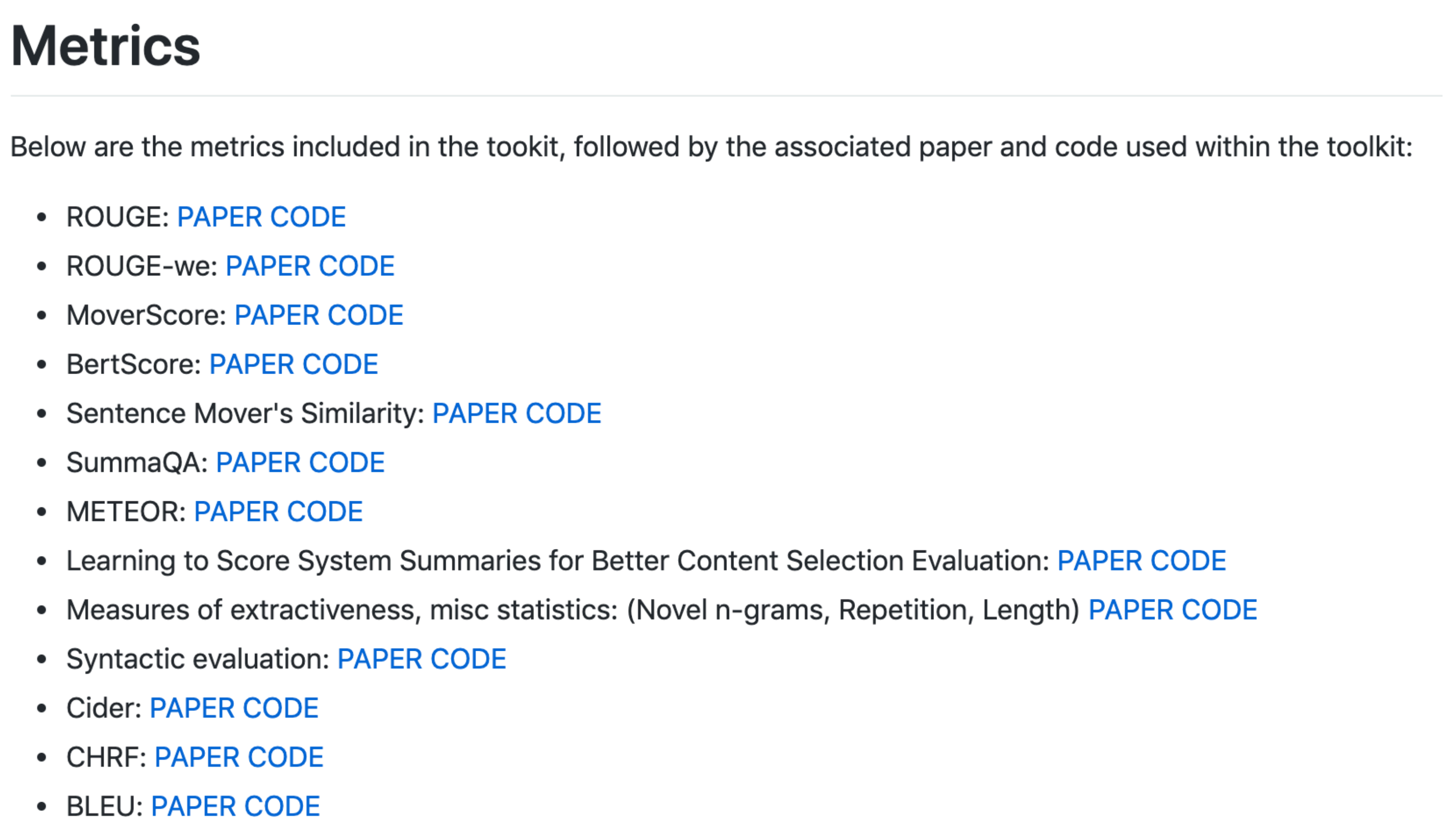
```
git clone https://github.com/Alex-Fabbri/summ_eval.git
cd summ_eval
pip install -e .
```

To finish the setup, please run and follow the prompts in this script:

```
python setup_finalize.py
```

You can test your installation and get familiar with the library through `tests/`

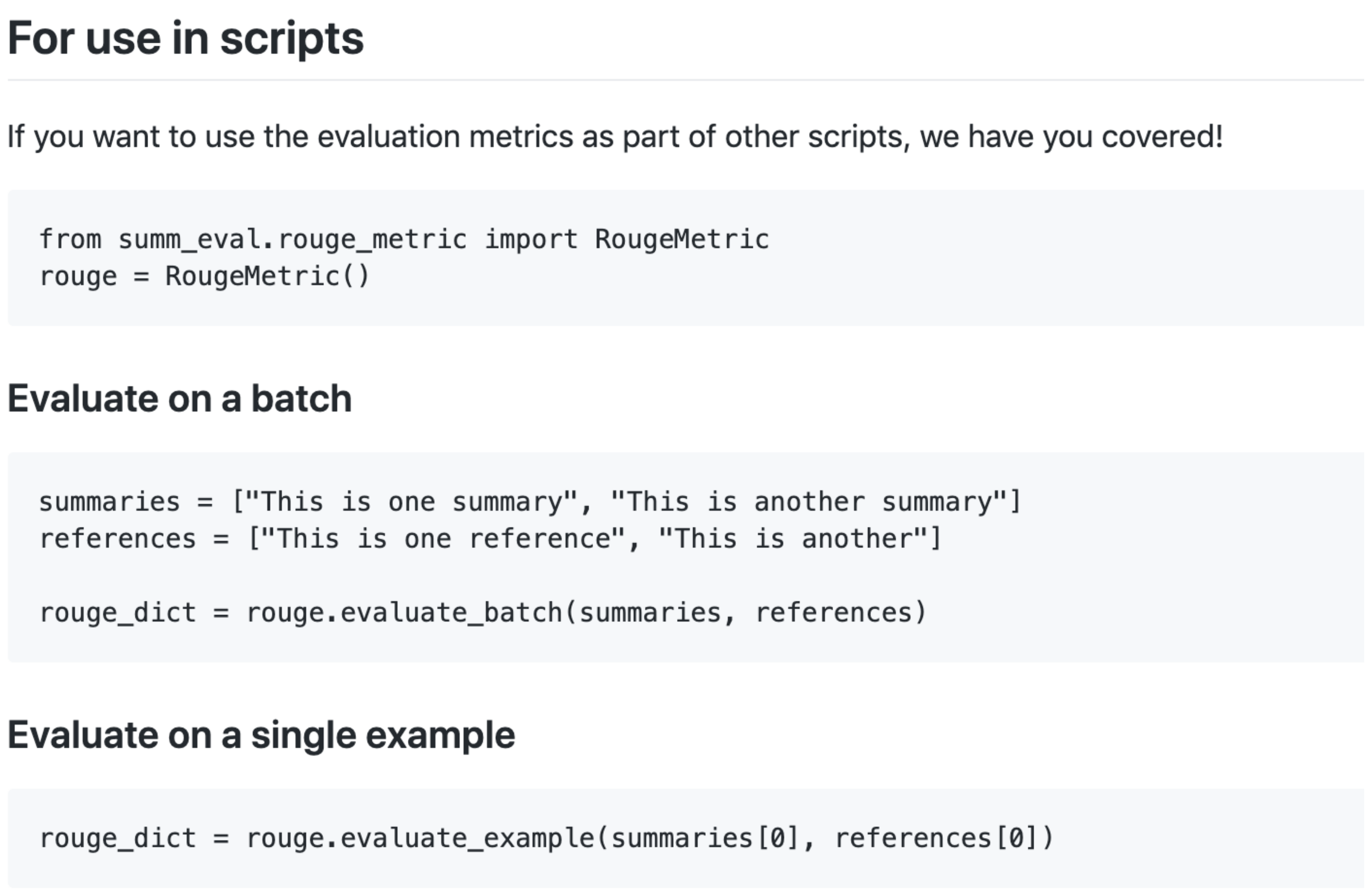
```
python -m unittest discover
```



Metrics

Below are the metrics included in the tookit, followed by the associated paper and code used within the toolkit:

- ROUGE: PAPER CODE
- ROUGE-we: PAPER CODE
- MoverScore: PAPER CODE
- BertScore: PAPER CODE
- Sentence Mover's Similarity: PAPER CODE
- SummaQA: PAPER CODE
- METEOR: PAPER CODE
- Learning to Score System Summaries for Better Content Selection Evaluation: PAPER CODE
- Measures of extractiveness, misc statistics: (Novel n-grams, Repetition, Length) PAPER CODE
- Syntactic evaluation: PAPER CODE
- Cider: PAPER CODE
- CHRF: PAPER CODE
- BLEU: PAPER CODE



For use in scripts

If you want to use the evaluation metrics as part of other scripts, we have you covered!

```
from summ_eval.rouge_metric import RougeMetric
rouge = RougeMetric()
```

Evaluate on a batch

```
summaries = ["This is one summary", "This is another summary"]
references = ["This is one reference", "This is another"]

rouge_dict = rouge.evaluate_batch(summaries, references)
```

Evaluate on a single example

```
rouge_dict = rouge.evaluate_example(summaries[0], references[0])
```

Design of Toolkit and Human Judgments

Each metric consists of two functions, evaluate_example and evaluate_batch for dealing with single summary-reference pairs as well as corpus-level calculations. This allows the user to incorporate the toolkit as part of a large script or even within training as a reward function. We use gin configuration files to allow the user to quickly change metric parameters and re-run experiments. We also provide a command-line tool for calculating all metrics across a user-provided dataset. Additionally, there is an option to aggregate scores over the entire corpus or to calculate the score for each example in the corpus individually.

To better understand the pros and cons of current evaluation metrics, we ran each metric over 17 model outputs from recent summarization models over the CNN-Daily Mail summarization corpus (Hermann et al., 2015). We calculate correlations with human judgments over the 4 dimensions of Coherence (collective quality of all sentences), consistency (factual alignment between the summary and the source), fluency (quality of individual sentences) and relevance (selection of important content from the source). Human judgments were taken from the paper "Neural Text Summarization: A Critical Evaluation." However, correlations with human judgments were found to be extremely poor, despite positive inter-annotator agreement

Conclusion and Future Work

We have introduced a new toolkit for summarization evaluation and have studied correlation among metrics and human judgments. We are currently collecting high-quality human judgments from expert annotators to bridge a gap in the study of which summarization metric should be the gold standard.