Improving ROUGE for Timeline Summarization

ROUGE Basics and ROUGE Variants for Timeline

Xinyu Liang

January 23, 2024

Heidelberg University

Table of contents

Intro

Timeline

2010-05-06
BP tries to stop the spill by lowering a 98-ton "containment dome" over the leak. The effort eventually fails, as crystallized gases cause the containment dome to become unexpectedly buoyant.

2010-05-26
BP begins "top kill" attempt, shooting mud down the drillpipe in an attempt to clog the leaking well. After several days, the effort is abandoned.

2010-05-27
President Obama announces a six-month moratorium on new deepwater drilling in the gulf.

2010-05-14
Then-BP CEO Tony Hayward tells reporters that the amount of oil spilled is relatively small given the Gulf of Mexico's size.

2010-05-28
Hayward says the "top kill" effort to plug the well is progressing as planned and had a 60 to 70 percent chance of success, the same odds he gave before the maneuver. The next day the company announces that the effort failed.

Table 1: Excerpts from Washington Post (top) and AP (bottom) timelines for the BP oil spill in 2010.

- **Timeline**: combating information overload by reporting in an organised overview
- **Automatic timeline summarization (TLS)** use edited timelines as reference timelines to gauge their performance

ROUGE

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

- **ROUGE-N**: N-gram based co-occurrence statistics.
- **ROUGE-L**: Longest Common Subsequence (LCS) based statistics.
- **ROUGE-W**: Weighted LCS-based statistics that favors consecutive
- **ROUGE-S**: Skip-bigram based co-occurrence statistics.
- **ROUGE-SU**: Skip-bigram plus unigram-based co-occurrence statistics.

$\Rightarrow$ Without respecting the specific characteristics of TLS

Purpose

- Identifying **weaknesses** of currently used evaluation metrics for TLS.
- Devising **new variants** of ROUGE.
  - showing the suitability of the variants with a theoretical and empirical analysis

Task Description and Notation

Task Description

Given a query, TLS needs to :

- Extracting the most important events for the query and their corresponding dates
- Obtaining concise daily summaries for each selected date

Notation

- **Timeline** is a sequence $(d_1, s_1), ..., (d_k, s_k)$
  - d_i are dates
  - s_i are summaries for the dates d_i
- **Timeline summarization**: Generating a timeline s_q based on the documents in C_q.
  - q: query
  - C_q: associated corpus
- **Reference timelines** $R_q = \{r_1^q, ..., r_n^q q\}$
- For a timeline t, D_t denotes **the set of days** in t, For a set of timelines T, $D_T = \cup_{t \in T} D_t$.

Current Evaluation Metrics: ROUGE and Other Metrics

ROUGE-N

ROUGE-N metrics measures the overlap of N-grams in system and reference summaries.

- ROUGE-N recall:

$$\mathsf{rec}(R, s) = \frac{\sum_{r \in R} \sum_{g \in \mathbf{ng}(r)} \mathsf{cnt}_{r,s}(g)}{\sum_{r \in R} \sum_{g \in \mathbf{ng}(r)} \mathsf{cnt}_{r}(g)}, \tag{1}$$

- ROUGE-N precision:

$$\mathsf{prec}(R, s) = \frac{\sum_{r \in R} \sum_{g \in \mathrm{ng}(s)} \mathrm{cnt}_{r,s}(g)}{|R| \sum_{g \in \mathrm{ng}(s)} \mathrm{cnt}_{s}(g)} \tag{2}$$

ROUGE-N Example

A simplified example for ROUGE-2:

- Reference Summary: The cat sat on the mat.
- System Summary: The cat was sitting on the mat.

Generated 2-grams:

- Reference 2-grams: [”The cat”, ”cat sat”, ”sat on”, ”on the”, ”the mat”]
- Generated 2-grams: [”The cat”, ”cat was”, ”was sitting”, ”sitting on”, ”on the”, ”the mat”]

Counting how many of these 2-grams are in common:

- Common 2-grams: [”The cat”, ”on the”, ”the mat”]

Calculating the ROUGE-2 score for reference:

- ROUGE-2 for Reference: $rec(R, s) = 3/5$

Concatenation-based ROUGE: concat

Running ROUGE on documents obtained by concatenating the items of the timelines

- **Timeline** $t = (d_1, s_1), ..., (d_k, s_k)$
- **Concatenating** the $s_i \Rightarrow$ document s'
- Using ROUGE on the resulting documents

Shortcoming:

- how to set this constant is inconclusive
- different datings of the same event below the threshold difference would again not receive any penalty

Date-agreement ROUGE: agreement

Main Idea: Evaluating the quality of the summary for each day individually

- Recall:

$$\text{rec}(d, R, s) = \frac{\sum_{r \in R(d)} \sum_{g \in \mathbf{ng}(r)} \text{cnt}_{r,s(d)}(g)}{\sum_{r \in R(d)} \sum_{g \in \mathbf{ng}(r)} \text{cnt}_r(g)}. \tag{3}$$

- By micro-averaging:

$$\text{rec}(R, s) = \frac{\sum_{d \in D_R} \sum_{r \in R(d)} \sum_{g \in \mathbf{ng}(r)} \text{cnt}_{r,s(d)}(g)}{\sum_{d \in D_R} \sum_{r \in R(d)} \sum_{g \in \mathbf{ng}(r)} \text{cnt}_r(g)} \tag{4}$$

Shortcoming: Requiring the dates in the reference timeline and the generated timeline match exactly

Alignment-based ROUGE

Design standards

Requirement:

- Temporal and semantic similarity of the daily summaries
- Without requiring an exact match between days

The main idea: daily summaries that are close in time and that describe the same event or very similar events should be compared for evaluation

Formal Definition

R: a set of reference timelines
s: a system timeline.

Mapping:

$$f : D_R \rightarrow D_s \tag{5}$$

Penalize date differences:

$$t_{d_r,d_s} = \frac{1}{|d_r - d_s| + 1} \tag{6}$$

Alignment-based ROUGE recall:

$$\text{rec}(R, s) = \frac{\sum_{d \in D_R} t_{d,f(d)} \sum_{r \in R(d)} \sum_{g \in \textbf{ng}(r)} \text{cnt}_{r,s(f(d))}(g)}{\sum_{d \in D_R} \sum_{r \in R(d)} \sum_{g \in \textbf{ng}(r)} \text{cnt}_r(g)} \tag{7}$$

Computing Alignments

Cost $c_{dr,ds}$ of assigning *dr* to *ds*.

Goal: Finding a mapping $f : D_R \rightarrow D_s \Rightarrow$ minimizes the sum of the costs:

$$f^* = \arg\min_{f} \sum_{d_r \in D_R} c_{d_r, f(d_r)} \tag{8}$$

Instantiations

- **Date Alignment**: the cost only depends on date distance, ignoring semantic similarity.

$$c_{d_r,d_s} = 1 - \frac{1}{|d_r - d_s| + 1} \tag{9}$$

- **Date-content Alignment**: Includes semantic similarity in the costs, $R1(d_r, d_s)$ is the ROUGE-1 F1 score that compares the reference summaries for date d_r with the system summary for date d_s.

$$c_{d_r,d_s} = \left(1 - \frac{1}{|d_r - d_s| + 1)}\right) \cdot (1 - \mathrm{R1}(d_r, d_s)) \tag{10}$$

- **Many-to-one Date-content Alignment**: drop the injectivity requirement from Date-content Alignment: If $|D_R| > |D_s|$, some $d_r \in D_R$ will be unaligned. For these dates we set the n-gram counts to 0 in the numerator of Equation.

Example

2010-05-06 BP tries to stop the spill by lowering a 98-ton "containment dome" over the leak. The effort eventually fails, as crystallized gases cause the containment dome to become unexpectedly buoyant.
2010-05-26 BP begins "top kill" attempt, shooting mud down the drillpipe in an attempt to clog the leaking well. After several days, the effort is abandoned.
2010-05-27 President Obama announces a six-month moratorium on new deepwater drilling in the gulf.

2010-05-14 Then-BP CEO Tony Hayward tells reporters that the amount of oil spilled is relatively small given the Gulf of Mexico's size.
2010-05-28 Hayward says the "top kill" effort to plug the well is progressing as planned and had a 60 to 70 percent chance of success, the same odds he gave before the maneuver. The next day the company announces that the effort failed.

Table 1: Excerpts from Washington Post (top) and AP (bottom) timelines for the BP oil spill in 2010.

Discussions

- **Complexity:** greedy algorithm: for every date in D_R we choose the date in D_s such that the cost is minimal
- **Generalizing agreement:** Agreement also fits in this framework: set t_{d_r,d_s} =1, c_{d_r,d_s}=0 iff $d_r = d_s$, and t_{d_r,d_s} =0, $c_{d_r,d_s} = \infty$

Tests for Metrics

Test Definitions

Comparing a modified version to the original timeline should decrease precision and/or recall, depending on the operation.

Testing operations:

- Remove
- Add
- Merge
- Shift k days

Evaluation

Test	Metric	ΔP	ΔR	ΔF_1
Remove	concat	0.000	-0.051	-0.026
	agreement	0.000	-0.051	-0.026
	align	0.000	-0.051	-0.026
	align+	0.000	-0.051	-0.026
	align+ m:1	0.000	-0.045	-0.023
Add	concat	-0.032	0.000	-0.016
	agreement	-0.032	0.000	-0.016
	align	-0.032	0.000	-0.016
	align+	-0.032	0.000	-0.016
	align+ m:1	-0.030	0.000	-0.015
Merge	concat	0.000	0.000	0.000
	agreement	-0.045	-0.045	-0.045
	align	-0.045	-0.045	-0.045
	align+	-0.045	-0.045	-0.045
	align+ m:1	-0.045	-0.023	-0.034
Shift 1 day	concat	0.000	0.000	0.000
	agreement	-0.887	-0.887	-0.887
	align	-0.679	-0.679	-0.679
	align+	-0.500	-0.500	-0.500
	align+ m:1	-0.500	-0.622	-0.569
Shift 5 days	concat	0.000	0.000	0.000
	agreement	-0.927	-0.927	-0.927
	align	-0.878	-0.878	-0.878
	align+	-0.833	-0.833	-0.833
	align+ m:1	-0.833	-0.817	-0.825

Table 2: Tests on *timeline17*. Numbers are difference to 1 according to ROUGE-1-based metrics.

- Timeline17 data set: 17 timelines across nine topics and associated corpora.
- Compare each modified timeline to the corresponding original timeline.
- Evaluate using variants based on ROUGE-1 and ROUGE-2
- ROUGE-2 yielded similar results

Discussion

- The frequently used **concat** is not a suitable metric for TLS.
- **Agreement** has the expected behavior for all tests, but, due to the required exact date matching, faces a very high drop for even minor date shifting and does not differentiate well between shifting one day and shifting five days.
- **Alignment-based metrics** pass all tests and the drops caused by shifts are lower and differentiation is better than agreement
- semantic similarity (**align+**) further decreases drops in date shifting.
- Except for the Shift 1 day test, many-to-one-alignments (**align+ m:1**) yield the most lenient results

Conclusions and Future Work

Conclusions

- Identified weaknesses of metrics encountered in the literature
- Devised a family of alignment-based ROUGE variants tailored to TLS

Future Work

- The correlation of TLS metrics with human judgment.
- investigate more content and date similarity measures for computing and weighting optimal alignments.

References

References

Sebastian Martschat and Katja Markert. 2017. Improving ROUGE for Timeline Summarization. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 285–290, Valencia, Spain. Association for Computational Linguistics..

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Question

1. Why do we use F1 scores in the Date-Content Alignment section, why not others?
2. Why do we require injectivity?