Improving ROUGE for Timeline Summarization

Sebastian Martschat and **Katja Markert**
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
(martschat|markert)@cl.uni-heidelberg.de

Abstract

Current evaluation metrics for timeline summarization either ignore the temporal aspect of the task or require strict date matching. We introduce variants of ROUGE that allow alignment of daily summaries via temporal distance or semantic similarity. We argue for the suitability of these variants in a theoretical analysis and demonstrate it in a battery of task-specific tests.

1 Introduction

There is an abundance of reports on events, crises and disasters. *Timelines* summarize and date these reports in an ordered overview to combat information overload.

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.

Table 1 shows parts of journalist-generated timelines. Approaches for *automatic timeline summarization* (TLS) use such edited timelines as reference timelines to gauge their performance (Chieu and Lee, 2004; Yan et al., 2011b; Tran et al., 2013; Wang et al., 2016). For evaluation, most research uses the standard summarization evaluation metric ROUGE (Lin, 2004) without respecting the specific characteristics of TLS.

In this paper, we identify weaknesses of currently used evaluation metrics for TLS. We devise new variants of ROUGE to overcome these weaknesses and show the suitability of the variants with a theoretical and empirical analysis. A toolkit that implements our metrics is available for download as open source.[1]

2 Task Description and Notation

Given a query (such as *BP oil spill*) TLS needs to (i) extract the most important events for the query and their corresponding dates and (ii) obtain concise daily summaries for each selected date (Allan et al., 2001; Chieu and Lee, 2004; Yan et al., 2011b; Tran et al., 2015; Wang et al., 2016).

Formally, a *timeline* is a sequence $(d_1, s_1), \ldots, (d_k, s_k)$ where the d_i are dates and the s_i are summaries for the dates d_i. Given are a query q and an associated corpus C_q that contains documents relevant to the query. The task of *timeline summarization* is to generate a timeline s_q based on the documents in C_q. The number of dates in the generated timeline as well as the length of the daily summaries are typically controlled by the user. For evaluation we assume access to one or more reference timelines $R_q = \{r_1^q, \ldots, r_{n_q}^q\}$. In our notation we usually drop the query sub-/superscript.

For a timeline t, D_t denotes the set of days in t. For a set of timelines T, we set $D_T = \cup_{t \in T} D_t$.

3 Current Evaluation Metrics

We now describe evaluation metrics for TLS and related tasks.

[1] http://smartschat.de/software

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 285–290,
Valencia, Spain, April 3-7, 2017.

3.1 ROUGE

Most work on TLS adopts the ROUGE toolkit that is used for for standard summarization evaluation (Lin, 2004). ROUGE metrics evaluate a system summary s of one or more texts against a set R of reference summaries (without accounting for dating summaries). The most popular variants of ROUGE are the ROUGE-N metrics which measure the overlap of N-grams in system and reference summaries. Several ROUGE metrics are well correlated with human judgment (Graham, 2015).

For a summary c, let us define the set of c's N-grams as $\texttt{ng}(c)$. $\texttt{cnt}_c(g)$ is the number of occurrences of an N-gram g in c. For two summaries c_1 and c_2, $\texttt{cnt}_{c_1,c_2}(g) = \min\{\texttt{cnt}_{c_1}(g), \texttt{cnt}_{c_2}(g)\}$ is the minimum number of occurrences of g in both c_1 and c_2.

ROUGE-N recall is then defined as[2]

$$\texttt{rec}(R,s) = \frac{\sum_{r\in R}\sum_{g\in\texttt{ng}(r)}\texttt{cnt}_{r,s}(g)}{\sum_{r\in R}\sum_{g\in\texttt{ng}(r)}\texttt{cnt}_r(g)}, \quad (1)$$

while ROUGE-N precision is defined as

$$\texttt{prec}(R,s) = \frac{\sum_{r\in R}\sum_{g\in\texttt{ng}(s)}\texttt{cnt}_{r,s}(g)}{|R|\sum_{g\in\texttt{ng}(s)}\texttt{cnt}_s(g)}. \quad (2)$$

ROUGE-N F_1 is the harmonic mean of recall and precision.

Concatenation-based ROUGE. The simplest and most popular way to apply ROUGE to TLS, which we refer to as *concat*, is to run ROUGE on documents obtained by concatenating the items of the timelines (Takamura et al., 2011; Yan et al., 2011a; Nguyen et al., 2014; Wang et al., 2016). Given a timeline $t = (d_1, s_1), \ldots, (d_k, s_k)$, we concatenate the s_i, which yields a document s'. In s' all date information is lost. We apply this transformation to the reference and the system timelines and use ROUGE on the resulting documents.

This method discards any temporal information. As a result, different datings of the same event are not penalized. Most work does not address this issue at all. An exception is Takamura et al. (2011), who ignore word matches when the matched word only appears in a summary where the time difference exceeds a pre-specified constant. However, it is left open how to set this constant and different datings of the same event below the threshold difference would again not receive any penalty.

[2]We rely on the representation of ROUGE-N presented in Lin and Bilmes (2011).

Date-agreement ROUGE. A more principled method of accounting for temporal information is to evaluate the quality of the summary for each day individually (Tran et al., 2013; Wang et al., 2015). We refer to this method as *agreement*. For a date d, a set of reference timelines R and a system timeline s, we set $R(d)$ to the set of summaries for d in R.[3] $R(d)$ can be empty if the date is not included in any timeline. $s(d)$ is the (possibly empty) summary of d in s. We define recall for a date d as

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

$\texttt{rec}(d,R,s)$ can be extended to the set of dates D_R, typically by micro-averaging, that is

$$\texttt{rec}(R,s) = \frac{\sum_{d\in D_R}\sum_{r\in R(d)}\sum_{g\in\texttt{ng}(r)}\texttt{cnt}_{r,s(d)}(g)}{\sum_{d\in D_R}\sum_{r\in R(d)}\sum_{g\in\texttt{ng}(r)}\texttt{cnt}_r(g)}. \quad (4)$$

The handling of precision is analogous: instead of the formula for ROUGE recall we use the formula for ROUGE precision and average with respect to D_s instead of D_R.

While this metric accounts for temporal information, it requires that dates in reference and generated timelines match exactly. Otherwise, a score of 0 is assigned. For example, in the BP oil spill example in Table 1, the first timeline would get a score of 0 when comparing it with the second timeline, even though both timelines report on the existence and later failure of the "top kill" effort, although on different dates. This effect can be particularly problematic for longer-lasting events.

3.2 Other Metrics

Some work evaluates TLS manually (Chieu and Lee, 2004; Tran et al., 2015). However, such evaluation is costly.

A related task to TLS is the TREC *update summarization* task (Aslam et al., 2015). In contrast to TLS, this task requires *online* summarization by presenting the input as a stream of documents. The metric employed relies on manually matching sentences of reference and system timelines. Kedzie et al. (2015) modify TREC metrics for a fully

[3]For convenience, we slightly overload notation. In the definition of standard ROUGE R and s were summaries, now they are timelines which contain summaries.

automatic setting, but still need a manually optimized threshold for establishing semantic matching. Moreover, the matching is binary: two summaries either match or do not match. The metric does not incorporate information about the degree of similarity between two summaries.

Lastly, in the DUC 2007 and TAC 2008–2011 evaluation campaigns a different type of update summarization was evaluated: the objective was to create and then update a multi-document summary with new information (see, e.g., Owczarzak and Dang (2011)). This task differs fundamentally from TLS and TREC-style update summarization, since no individual summaries for dates have to be created. Evaluation metrics specifically designed for the task employ a combination of ROUGE scores to simultaneously reward similarity to human-generated summaries and penalize redundancy with respect to the original machine-generated summary (Conroy et al., 2011).

4 Alignment-based ROUGE

From the analysis in the previous section we see that a metric for TLS should take temporal and semantic similarity of daily summaries into account, while not requiring an exact match between days.

We now propose variants of ROUGE that fulfill this desideratum. The main idea is that daily summaries that are close in time and that describe the same event or very similar events should be compared for evaluation. For example, the daily summaries that report on the "top kill" effort in the example in Table 1 should be compared. To do so, we first *align* dates in system and reference timelines.[4] ROUGE scores are then computed for the summaries of the aligned dates.

4.1 Formal Definition

Let R be a set of reference timelines and let s be a system timeline. The proposed alignment-based ROUGE recall relies on a mapping

$$f\colon\ D_R \to D_s \tag{5}$$

that assigns each date $d_r \in D_R$ in some reference timeline a date $d_s \in D_s$ in the system timeline. For evaluation, the summaries for the aligned dates are compared.[5]

In order to penalize date differences when comparing summaries, each date pair $(d_r, d_s) \in D_R \times D_s$ is associated with a *weighting factor* t_{d_r,d_s}. In this paper, we only consider the weighting factor

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

where $d_r - d_s$ is the difference between d_r and d_s in number of days. Given some alignment f, alignment-based ROUGE recall $\texttt{rec}(R, s, f)$ is then defined as

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

4.2 Computing Alignments

For computing alignments, we associate to every date pair $(d_r, d_s) \in D_R \times D_s$ another value, which is the *cost* c_{d_r,d_s} of assigning d_r to d_s. We will study costs that depend on date distance and/or semantic similarity of the corresponding summaries. The goal is to find a mapping $f^*\colon\ D_R \to D_s$ that minimizes the sum of the costs, i.e.

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

4.3 Instantiations

We consider three instantiations of the alignment problem presented above. They vary in the cost function and with respect to constraints on the alignment.

Date Alignment. For the first instantiation, which we call *date alignment* or *align*, the cost only depends on date distance, ignoring semantic similarity. We set

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

We require that the alignment is injective.[6]

In Table 1, for example, the daily summaries for 2010-05-27 and 2010-05-28 would be aligned.

[4] We are inspired by Luo (2005) who devises an alignment-based metric for coreference resolution.

[5] We only discuss how recall is computed. For computing precision we instead consider alignments $f\colon\ D_s \to D_R$ and apply the corresponding formulas for precision as discussed in Section 3.

[6] 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 7.

Date-content Alignment. The second instantiation, *date-content alignment* or *align+*, also includes semantic similarity in the costs. An approximation of semantic similarity is represented by the ROUGE-1 F_1 score between two daily summaries. We set

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

where $\mathrm{R1}(d_r, d_s)$ is the ROUGE-1 F_1 score that compares the reference summaries for date d_r with the system summary for date d_s. Here, too, we require that the alignment is injective.

The two daily summaries referring to the "top kill" effort in Table 1 would be aligned when this metric is employed.

Many-to-one Date-content Alignment. For our last metric (*many-to-one date-content alignment* or *align+ m:1*) we drop the injectivity requirement from *align+*.

4.4 Discussion

Complexity. If we require that f^* is injective, as in *align* and *align+*, we face a linear assignment problem, for which polynomial-time algorithms exist (Kuhn, 1955). The optimal assignment for *align+ m:1* can be computed by a simple greedy algorithm: for every date in D_R we choose the date in D_s such that the cost is minimal.

Generalizing *agreement*. Note that *agreement*, which relies on exact date match, also fits in our framework: we require f^* to be injective and set $t_{d_r,d_s} = 1$, $c_{d_r,d_s} = 0$ iff $d_r = d_s$, and $t_{d_r,d_s} = 0$, $c_{d_r,d_s} = \infty$ otherwise for all $(d_r, d_s) \in D_R \times D_s$.

5 Tests for Metrics

An evaluation metric should behave as expected when task-specific operations are performed on output (Moosavi and Strube, 2016). For example, in TLS, removing a date (and its summary) from a reference timeline should decrease recall when comparing the timeline to itself. A metric cannot be suitable if it does not pass such tests.

We now devise and evaluate tests for the metrics discussed in this paper. Eventually, metrics that pass the tests should be checked for correlation with human judgment. We defer such an experiment to future work.

5.1 Test Definitions

We derive tests that examine whether well-defined basic *operations* on reference timelines affect the metrics as expected. An example is the date removal operation described above. Other basic operations are date addition, merging and shifting. In order to have a controlled environment we apply all operations to copies of reference timelines. Comparing a reference timeline to itself gives precision, recall and F_1 score of 1. Comparing a modified version to the original timeline should decrease precision and/or recall, depending on the operation. We apply the following operations:

- **Remove**: remove a random date and its summary. Precision should stay 1, recall should decrease.
- **Add**: for the first date not in the reference timeline, add a summary consisting of the first sentence of the first article of that day from the associated corpus. Precision should decrease, recall should stay 1.
- **Merge**: merge summaries of the closest pair of dates, breaking ties by temporal order. Precision and recall should decrease slightly.
- **Shift k days**: shift each day by k days to the future. Precision and recall should decrease. The drop should increase as k increases.

5.2 Evaluation

We run the proposed tests[7] on the publicly available *timeline17* data set (Tran et al., 2013), which contains 17 timelines across nine topics and associated corpora. We apply each operation to each timeline. We then compare each modified timeline to the corresponding original timeline.

We evaluate using variants based on ROUGE-1 and ROUGE-2, which are the most popular ROUGE-N metrics for evaluating TLS. Table 2 shows averaged results over all timelines for ROUGE-1 (ROUGE-2 yielded similar results).

We find that the frequently used *concat* is not a suitable metric for TLS. It is insensitive to merging and date shifting as it does not respect temporal information. *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.

[7] We show results for the date-shifting test with $k \in \{1, 5\}$. Other values of k yield the expected behavior.

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.

The alignment-based metrics show the most desirable behavior according to our criteria: they pass all tests and the drops caused by shifts are lower and differentiation is better than for *agreement*. For the other tests, these metrics behave similarly to *agreement*. Including 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 of all alignment-based metrics.

6 Conclusions and Future Work

Current evaluation metrics for TLS are not suitable. In a formal and empirical analysis we identified weaknesses of metrics encountered in the literature. We devised a family of alignment-based ROUGE variants tailored to TLS. We found that these metrics exhibit the desired behavior when applying a battery of task-specific tests.

In future work we will study the correlation of TLS metrics with human judgment. In order to optimize correlation, we will also investigate more content and date similarity measures for computing and weighting optimal alignments.

Acknowledgments

We thank the anonymous reviewers and our colleague Esther van den Berg for feedback on earlier drafts of this paper. We are grateful to Lu Wang and William Yang Wang for providing us with more details on the evaluation setup of the work presented in their respective papers.

References

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of new topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* New Orleans, Louis., 9–12 September 2001, pages 49–56.

Javed A. Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Richard McCreadie, Virgil Pavlu, and Testuya Sakai. 2015. TREC 2015 temporal summarization track overview. In *Proceedings of the Twenty-Fourth Text REtrieval Conference,* Gaithersburg, Md., 17–20 November 2015.

Hai Leong Chieu and Yoong Keok Lee. 2004. Query based event extraction along a timeline. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* New York, N.Y., 25–29 July 2004, pages 425–432.

John M. Conroy, Judith D. Schlesinger, and O'Leary Dianne P. 2011. Nouveau-ROUGE: a novelty metric for update summarization. *Computational Linguistics*, 37(1):1–8.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, 17–21 September 2015, pages 128–137.

Chris Kedzie, Kathleen McKeown, and Fernando Diaz. 2015. Predicting salient updates for disaster summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Beijing, China, 26–31 July 2015, pages 1608–1617.

Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics*, 2:83–97.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Portland, Oreg., 19–24 June 2011, pages 510–520.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of*

the Text Summarization Branches Out Workshop at ACL '04, Barcelona, Spain, 25–26 July 2004, pages 74–81.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing,* Vancouver, B.C., Canada, 6–8 October 2005, pages 25–32.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Berlin, Germany, 7–12 August 2016, pages 632–642.

Kiem-Hieu Nguyen, Xavier Tannier, and Véronique Moriceau. 2014. Ranking multidocument event descriptions for building thematic timelines. In *Proceedings of the 25th International Conference on Computational Linguistics,* Dublin, Ireland, 23–29 August 2014, pages 1208–1217.

Karolina Owczarzak and Hoa Dang. 2011. Overview of the TAC 2011 summarization track: guided task and AESOP task. In *Proceedings of the Text Analysis Conference,* National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.

Hiroya Takamura, Hikaru Yokono, and Manabu Okumura. 2011. Summarizing a document stream. In *Proceedings of the 33rd European Conference on Information Retrieval,* Dublin, Ireland, 18–21 April 2011, pages 177–188.

Giang Tran, Mohammad Alrifai, and Dat Quoc Nguyen. 2013. Predicting relevant news events for timeline summaries. In *Proceedings of the 22nd World Wide Web Conference,* Rio de Janeiro, Brasil, 13–17 May, 2013, pages 91–92.

Giang Tran, Mohammad Alrifai, and Eelco Herder. 2015. Timeline summarization from relevant headlines. In *Proceedings of the 37th European Conference on Information Retrieval,* Vienna, Austria, 29 March – 2 April 2015, pages 245–256.

Lu Wang, Claire Cardie, and Galen Marchetti. 2015. Socially-informed timeline generation for complex events. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Denver, Col., 31 May – 5 June 2015, pages 1055–1065.

William Yang Wang, Yashar Mehdad, Dragomir R. Radev, and Amanda Stent. 2016. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* San Diego, Cal., 12 – 17 June 2016, pages 58–68.

Rui Yan, Liang Kong, Congrui Huang, Xiajun Wan, Xiaoming Li, and Yan Zhang. 2011a. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* Edinburgh, Scotland, U.K., 27–29 July 2011, pages 433–443.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaming Li, and Yan Zhang. 2011b. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Beijing, China, 25–29 July 2011, pages 745–754.