Improving ROUGE for Timeline Summarization

Sebastian Martschat and **Katja Markert**
Department of Computational Linguistics, Heidelberg University

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Motivation

- **Timeline Summarization**: **date** and **summarize** events

Reference

2010-05-26
BP begins "top kill" attempt. After several days, the effort is abandoned.

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

Predicted

2010-05-14
BP CEO Hayward states that the amount of oil spilled is relatively small.

2010-05-28
Hayward says the "top kill" effort to plug the well is progressing as planned.

- previous work: evaluate with ROUGE without considering **specific temporal characteristics**
- we propose **alignment-based ROUGE variants**
- formal and empirical **test-driven analysis** of metrics

Software

- available as a **Python library**:

```
pip install tilse
```

- documentation and code:

`https://github.com/smartschat/tilse`

Previous Work

- mainly **concat**: ignore dates, treat timeline as one big document

$$\text{ROUGE}_{\texttt{recall}}(r, p) = \frac{\sum_{g \in \texttt{ng}(r)} \texttt{cnt}_{r,p}(g)}{\sum_{g \in \texttt{ng}(r)} \texttt{cnt}_{r}(g)}$$

- cannot penalize wrong datings of events!
- also in previous work: compare summaries where **dates exactly match** – assigns zero credit in the example!

Alignment-based ROUGE

Date costs

2010-05-26
BP begins "top kill" attempt. After several days, the effort is abandoned.

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

→

2010-05-14
BP CEO Hayward states that the amount of oil spilled is relatively small.

2010-05-28
Hayward says the "top kill" effort to plug the well is progressing as planned.

Date-content costs

2010-05-26
BP begins "top kill" attempt. After several days, the effort is abandoned.

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

2010-05-14
BP CEO Hayward states that the amount of oil spilled is relatively small.

2010-05-28
Hayward says the "top kill" effort to plug the well is progressing as planned.

Key idea: align dates in reference and predicted timelines

Framework

- compute **cost-optimal alignment** f^***:** $D_r \to D_s$ between dates in reference and predicted timelines:

$$f^* = \arg\min_{f} \sum_{d \in D_r} c_{d,f(d)}$$

- compute ROUGE by **comparing aligned dates**, weighting scores with date difference:

$$\text{align-ROUGE}_{\texttt{recall}}(r, p) = \frac{\sum_{d \in D_r} t_{d,f(d)} \sum_{g \in \texttt{ng}(r(d))} \texttt{cnt}_{r(d),p(f(d))}(g)}{\sum_{d \in D_r} \sum_{g \in \texttt{ng}(r(d))} \texttt{cnt}_{r(d)}(g)}$$

- explore **different variants** of alignments: **parameterizable** by cost function, weighting function and restrictions on alignment f

Instantiations

- **date costs** or **date-content costs**

$$c_{d,f(d)} = \left(1 - \frac{1}{|d - f(d)| + 1}\right) \cdot (1 - \text{R1}(d, f(d)))$$

- **weighting**: $t_{d,f(d)} = \frac{1}{|d - f(d)| + 1}$
- f **injective** or **many-to-one**

Metric Tests

- check whether metrics **behave as expected** when adding, removing, merging and date shifting daily summaries

	add	remove	merge	shift 1 day	shift 5 days
concat	✓	✓	✗	✗	✗
match	✓	✓	✓	✗	✗
align	✓	✓	✓	✓	✓

Conclusion

- previously used metrics for timeline summarization are **not suitable**
- proposed alignment-based metrics are **theoretically and empirically satisfying**

Future Work

- devise **more sophisticated cost functions**
- perform a **human judgment correlation study**

http://cl.uni-heidelberg.de | http://www.smartschat.de {martschat|markert}@cl.uni-heidelberg.de