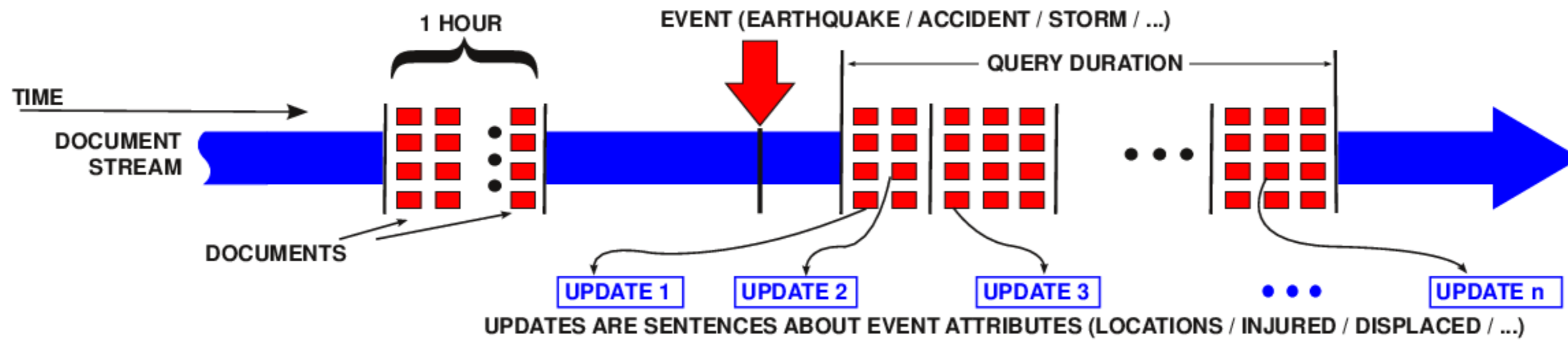


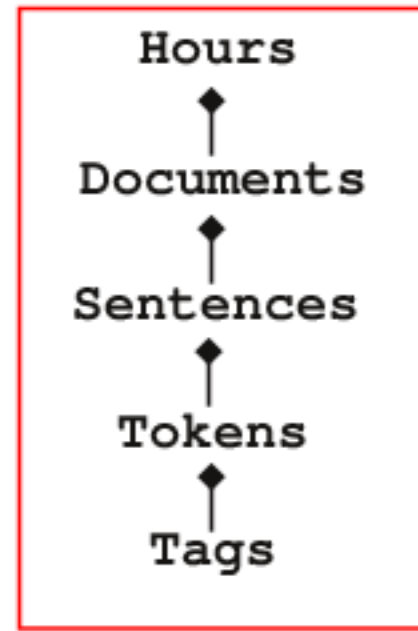
University of Waterloo at the Temporal Summarization Track

Gaurav Baruah¹, Rakesh Guttikonda¹, Adam Roegiest¹ and Olga Vechtomova²

¹ David R. Cheriton School of Computer Science and ² Department of Management Sciences, University of Waterloo



Thrift Format



TREC Style Format

```
<doc>
<docno>...</docno>
<lang>..</lang>
<abs_url>...</abs_url>
<original_url>...</original_url>
<time>...</time>
<tagger>...</tagger>
<sentences>
...
</sentences>
</doc>
```

Query Likelihood with Dirichlet Priors

1. Term-frequency maps for every hour.
2. Document feature vectors for each document in hour.
3. Computed rankings for documents in hours within query duration.

$$score_{LMD} = \sum_{t \in q} q_t \cdot \log \left(1 + \frac{f_{t,d} \cdot l_c}{\mu \cdot l_t} \right) - n \cdot \left(1 + \frac{l_d}{\mu} \right)$$

4. Extracted ranked documents for further processing.

Similarity Metrics

Patterns with grep

Event	Pattern
Accident	"accident calamity casualty disaster hazard mishap pileup setback collision fender-bender smash"
Bombing	"bombing bombe explosion device explosive charge shell projectile rocket missile mine homemade terrorist detonate timer"
Earthquake	"earthquake shock fault quake shakelshaking tremor fault temblor quake quaking"
Shooting	"shooting shots shot fired firing gun fire gun gunning trigger bullet machine fire terrorist firefight automatic rifle shotgun"
Storm	"storm blast blizzard cyclone disturbance downpour gale gust hurricane monsoon snowstorm squall tempest tornado twister windstorm"

2-Norm Similarity

$$\|x - q\| \geq \gamma \|q\|$$

where $\gamma = 0.75$

Cosine Similarity

$$\cos \theta = \frac{x \cdot q}{\|x\| \|q\|} \geq 0.5$$

Passage Scoring and Term Expansion with PRF

Seed Queries

Generic All Attributes: <event-type> <attribute-*>
earthquake injuries locations deaths displaced financial impact
 Specific All Attributes: <query-string> <attribute-*>
iran earthquake injuries locations deaths displaced financial impact

Expansion Terms (hour-wise)

Hour	Top 10 expansion terms
Hour 2012-08-11-18:	top 10 expansion terms
GAA	earthquake quake magnitude injured hundreds killed seismic iran earthquakes northwestern
SAA	injured earthquake magnitude haris varzaqan ahar quake killed hundreds iran
RRF	earthquake injured magnitude quake killed hundreds haris ahar varzaqan iran
Hour 2012-08-11-21:	top 10 expansion terms
GAA	earthquake magnitude seismogram anss nsmpr mss recenteqsw crustal shakemap seismologist
SAA	earthquake iran villages magnitude least northwestern earthquakes tv injured news
RRF	earthquake seismogram anss nsmpr mss magnitude recenteqsw crustal shakemap seismologist

Ranking Sentences

$$score_{cover} = \sum_{t \in q} \left(\log \left(\frac{l_c}{l_t} \times rrf(t) \right) \right) - m \cdot \log(l)$$

$$rrf(t) = \sum_i \frac{1}{k + r_i(t)}$$

BM25 and Term Expansion with DSTE

Distributional Similarity with Seed Words

seed word	Top 10 Expansion Terms
quake	earthquake tremor disaster magnitude aftershock temblor toll damage province death
damaged	destroyed killed left hit injured struck wounded leveled brought died
cities	counties areas towns regions provinces parts villages people states residents
assistance	aid help food relief money work medicine team sympathy water
disaster	earthquake quake emergency relief tremor crisis aftershock catastrophe development region

Ranking Sentences

$$score_{sentence} =$$

$$\sum_{t \in q} q_t \times \frac{f_{t,d}(k_1 + 1)}{k_1((1 - b) + b(l_d/l_{avg})) + f_{t,d}} \times w_t$$

Confidence Rating for Updates

$$confidence = score_{sentence} \times score_{LMD}$$

Performance

Method	RunID	Expected Latency Gain	Expected Comprehensiveness
Similarity Metric	CosineEgrep	0.0104	0.0179
	CosineEgrepIgnoreCase	0.0146	0.0130
	NormEgrep	0.0011	0.0611
	NormEgrepIgnoreCase	0.0010	0.0534
Passage Retrieval and PRF	UWMDSqlc2t25	0.0173	0.5375
	UWMDSqlc4t50	0.0176	0.5304
BM25 and Distributional Similarity	rg1	0.0205	0.5705
	rg2	0.0218	0.5624
	rg3	0.0275	0.5165
	rg4	0.0261	0.5063
Average Scores	All Track Submissions	0.0599	0.2996

Conclusions

1. Query Term Expansion and Pseudo-Relevance Feedback help improve recall (EC) in general.
2. Cosine Similarity is better for ELG, 2-Norm Similarity is better for EC.
3. Future Work: Improve ELG for PRF/Term Expansion methods.

References

- BUETTCHER, S., CLARKE, C. L. A., AND CORMACK, G. V. *Information retrieval: Implementing and Evaluating Search Engines*. MIT Press, Cambridge, Mass., 2010.
- CORMACK, G. V., CLARKE, C. L. A., AND BUETTCHER, S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2009), SIGIR '09, ACM, pp. 758–759.
- SPERLING, M., JIN, R., RAYVYCH, I., LI, J., AND YI, J. Similar document detection and electronic discovery: So many documents, so little time. In *DESI V Workshop at ICAI, 2013* (June 2013).
- VECHTOMOVA, O. A semi-supervised approach to extracting multiword entity names from user reviews. In *Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search* (New York, NY, USA, 2012), JIWES '12, ACM, pp. 2:1–2:6.