Live Blog Corpus for Summarization

Avinesh P.V.S., Maxime Peyrard, Christian M. Meyer

UBIQUITOUS KNOWLEDGE PROCESSING

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES
ADAPTIVE PREPARATION OF INFORMATION FROM HETEROGENEOUS SOURCES

Overview

Motivation

- Live blogs is popular.
  (BBC , Guardian, Der Spiegel)
- Short snippets of heterogeneous information.

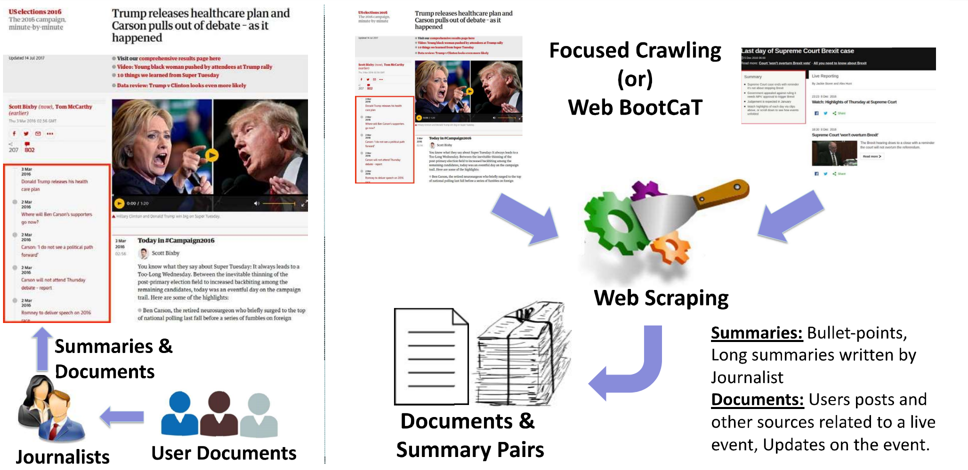
Contributions:

- Live blog corpus construction.
- Benchmark baseline system scores.

Applications:

A journalistic aid to assist live blog compilation.

Live Blog Summarization



US elections 2016
The 2016 campaign, minute-by-minute

Trump releases healthcare plan and Carson pulls out of debate – as it happened

Updated 14 Jul 2017

- Visit our comprehensive results page here
- Video: Young black woman pushed by attendees at Trump rally
- 10 things we learned from Super Tuesday
- Data review: Trump v Clinton looks even more likely

Scott Bixby (now), Tom McCarthy (earlier)
Thu 3 Mar 2016 02.56 GMT

207 802

3 Mar 2016
Donald Trump releases his health care plan

2 Mar 2016
Where will Ben Carson's supporters go now?

2 Mar 2016
Carson: 'I do not see a political path forward'

2 Mar 2016
Carson will not attend Thursday debate – report

2 Mar 2016
Romney to deliver speech on 2016 race

Hillary Clinton and Donald Trump win big on Super Tuesday.

3 Mar 2016 02:56
Today in #Campaign2016
Scott Bixby

You know what they say about Super Tuesday: It always leads to a Too-Long Wednesday. Between the inevitable thinning of the post-primary election field to increased backbiting among the remaining candidates, today was an eventful day on the campaign trail. Here are some of the highlights:

- Ben Carson, the retired neurosurgeon who briefly surged to the top of national polling last fall before a series of fumbles on foreign

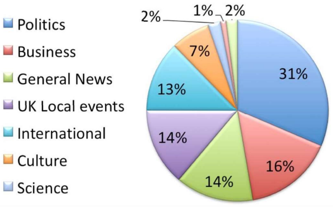
Summaries: Bullet-points, Long summaries written by Journalist
Documents: Users posts and other sources related to a live event, Updates on the event.

Corpus Analysis and Experiments

Corpus Statistics

Statistic	BBC	Guardian
# topics	974	1,681
# documents	92,537	94,462
# documents/ topic	95.01	56.19
# words/ document	61.75	107.53
# words/ summary	59.48	42.23

Number of Topics per Domain



Textual Heterogeneity

	BBC	Guardian	DUC'04	TAC'08A
TH_{js}	.5917	.5689	.3019	.3188

System scores & extractive upper bounds

Systems	BBC (L)			Guardian (L)			BBC (2*L)			Guardian (2*L)		
	R1	R2	SU4	R1	R2	SU4	R1	R2	SU4	R1	R2	SU4
TF*IDF	.227	.067	.064	.153	.021	.027	.367	.115	.147	.248	.037	.065
LexRank	.276	.080	.079	.188	.029	.038	.421	.138	.176	.297	.051	.089
LSA	.212	.046	.052	.135	.013	.021	.341	.084	.123	.220	.034	.051
KL	.267	.086	.080	.178	.026	.035	.397	.132	.165	.272	.041	.076
ICSI	.302	.104	.091	.210	.046	.046	.461	.176	.201	.322	.071	.101
UB-1	**.514**	.273	.218	**.422**	.177	.145	**.754**	.388	.435	**.640**	.256	.304
UB-2	.494	**.312**	.210	.389	**.230**	.137	.709	**.453**	.419	.584	**.334**	.277

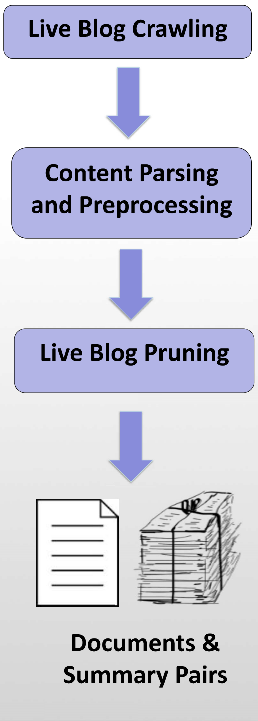
- State-of-the-art ICSI system is .2 ROUGE-1 and .3 ROUGE-2 lower than upper bounds (UB-1 and UB-2)

Challenges of Live blog summarization:

- Systems miss to identify topic shits.
  E.g. Trumps Health care plan, Carson missing debate etc.
- Large compression ratio (Input size vs Summary size)

Corpus Construction

- **Live Blog Crawling**
  - Focused Crawling
  - Web BootCaT
- **Content Parsing & Preprocessing**
  - Boilerplate removal
  - Summary and document extraction
  - Metadata extraction
    (e.g. URL, genre, date, author etc.)
- **Live Blog Pruning**
  - Multi-event blog
    (e.g., Latest updates from Essex)
  - Sports, games and live chats
  - Summary size < 3 sentences



Live Blog Crawling

Content Parsing and Preprocessing

Live Blog Pruning

Documents & Summary Pairs

Topics for BBC and Guardian

Dataset	Crawling	Preprocessing	Pruning
BBC	9,931	7,307	974
Guardian	16,246	6,405	1,681

Summary & Conclusions

- Live blog summarization corpus useful in a direct application for journalists and news readers.
- A pipeline to collect live blog corpus from BBC and Guardian.
- Benchmark summarization results far from the upper bound.
- Need solutions to handle multiple topic shifts and large input size.

Try it out, get in touch

Code and data: https://github.com/AIPHES/lrec2018-live-blog-corpus
Questions or comments: avinesh@aiphes.tu-darmstadt.de

RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

Heidelberger Institut für Theoretische Studien HITS

TECHNISCHE UNIVERSITÄT DARMSTADT

Acknowledgements: This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogenous Sources" (AIPHES) under grant No. GRK 1994/1.