
Dataset description

This is the Timeline17 dataset used for experiments of the paper .

(1) G. B. Tran, T.A. Tran, N.K. Tran, M. Alrifai and N. Kanhabua. 2013. Leverage Learning to rank in an optimization framework for timeline summarization. In TAIA workshop, SIGIR 2013 .

(2) G. B. Tran, M. Alrifai and D. Q. Nguyen. 2013. Predicting Relevant News Events for Timeline Summaries In 22nd World Wi de Web (WWW), Brazil, May 2013 .

Please cite the paper (1) if you find it helpful for your experiment.

Briefly, the dataset consists of 17 manual-created timelines and their associated news articles. They belong to 9 news topics namely: BP Oil Spill, Michael Jackson Death (~Dr. Murray Trial), Haiti Earthquake, H1N1 (Influenza), Financial Crisis, Syrian Crisis, Libyan War, Iraq War, Egyptian Protest.
Each timeline and its news articles belongs to one source news agency, such as BBC, Guardian, CNN, Foxnews, NBCNews, etc. The contents of these news are in plain text file format and noise filtered.

The statistics about our dataset

<Topics > , < Source> , <#Docs> <#Ground Truth>, <#Dates> <Average Sentence per date of manually created timeline>, <#Since>

BP Oil , Washington Post ,	296 , 1 , 12 , 1.6 , 2010
BP Oil , Reuters ,	298 , 1 , 16, 1.9 , 2010
BP Oil , BBC ,	293 , 1 , 48 , 2.0 , 2010
BP Oil , Foxnews ,	286 , 1 , 13 , 4.0 , 2010
BP Oil , Guardian ,	288 , 1, 102 , 3.0 , 2010
Michael Jackson death ,BBC ,	142 , 1 , 38 , 2.1 , 2009
Haiti Earthquake , BBC ,	296 , 1 , 11 , 7.8 , 2010
H1N1 , Reuters ,	207 , 1 , 15 , 1.5 , 2009
H1N1 , BBC ,	122 , 1 , 7 , 4.6 , 2009
H1N1 , Guardian ,	76 , 1 , 12 , 2.8 , 2009
Financial Crisis , WP ,	298 , 1, 65 , 6.9 , 2008
Syrian Crisis , Reuters ,	346 , 3 , 76 , 1.6 , 2011
Syrian Crisis , BBC ,	308 , 1 , 13 , 2.4 , 2011
Libyan War , Reuters ,	379 , 1 , 28 , 1.3 , 2011
Libyan War , CNN ,	398 , 1 , 38 , 2.1 , 2011
Iraq War , Guardian ,	344 , 1 , 155 , 2.6 , 2005
Egyptian Protest , CNN ,	273 , 1 , 20 , 2.8 , 2011

Is it helpful?

Firstly, there has been no available dataset published for Timeline Summarization (TS/timeline) at the moment, therefore, we construct this dataset for our evaluation. We believe it will not only benefit TS research community but also the traditional multi-document summarization research.

Secondly, collecting news articles from Internet and extracting their main content is very time costly, especially from html format. It often requires lots of efforts to obtain a (nearly) clean dataset. Thus, we believe our dataset can save time and effort for others.

Format

Folder tree:

News4TS/
Topic_SourceAgency/
InputDocs/
Date1/
article1 article2, …
Date2/
article1 article2, …
Date3/
article1 article2, …
….

timelines/
manually created manually created created timeline 1
Manually created timeline 2 (if there are more than one timeline)
……..

-Each timeline is in a folder namely {Topics_sourceAgency), for example: SyrianCrisis_bbc, LibyaWar_reuters
-News articles are organized by published date. Each Date is represented by a sub-folder, such as: 2011-07-01, 2011-07-14
-Inside each Date folder, there are news articles published in this date.
-Each article is in the plain text format; its name ends with ".htm.txt"; each line is a sentence.
-Each manual-created timeline is in one plain text file, in format:

<date1>
<summary of date1, each sentence is in a line>

<date1>
<summary of date1, each sentence is in a line>

<*date*> is converted into the format YYYY-MM-DD

for example,

2011-04-22
The Syrian uprising , then a month old , experienced its bloodiest day so far on 22 April when 72 protesters were killed by security forces firing on crowds .
Many of the dead were in the southern village of Ezra , near Deraa and in a suburb of Damascus .

2011-10-03
Opposition groups formed the Syrian National Council and pledge to overthrow President Bashar al-Assad .

How did we collect this dataset?

-1- Manually created timelines:

We looked for available timelines published by famous news agencies such as CNN, BBC, NBCnews, etc. that discuss famous topics happening recently ``BP Oil Spill'', ``Influenza H1N1''. We only took English timeline where the timestamps are mostly explicit dates, such as '07 July 2011' and ignored the timeline where the timestamps are at the year or month or week level, such as ``Middle 2005'' or ``July 2006''. Finally, we get 17 different timelines in different topics.

-2- News articles for timelines:
We used Google to retrieve news articles from the same agencies of the timelines, for example, BBC news articles for BBC-published timeline, using topic queries (e.g, BP Oil Spill, Influenza, etc.). We filtered out news articles that are NOT published during timelines timespan by using the time filter option from Google. We retrieved top 400 returned results. The reason is the returned results are often repeated after top 40 pages (which is corresponding to 40 returned pages from Google).
At the end, we obtained total 4650 news articles after duplication removal. We used BoilerPipe {freely available at http://www.l3s.de/~kohlschuetter/boilerplate/} (Kohlschutter,2010) to extract the content of the news articles. To the best of our knowledge, it is one of the best tools for doing this task in an automatic manner.

We noticed that the extracted content contains a lot of errors such as advertisements, links to other pages, etc. We hence manually created rules for removing
non-content text, however, there are still some errors left.
Sentence that contain some following patterns should be removed:

("stm USER_NAME = repman DOCUMENT_NAME");
("co. uk navigation blq_lang_ss = Skip to bbc .");
("co. uk FIDDLER_VERSION = 5.0.0 HTTP_USER_AGENT =");
("co. uk HTTP_X_FORWARDED_SERVER = news .");
("While you will be able to view the content of this page in your current browser , you will not be able to get the full visual experience .");
("Please consider upgrading your browser software or enabling style sheets -LRB- CSS -RRB- if you are able to do so .");
("co. uk search blq_lang_ak = Access keys help blq_lang_ak_u");
("blq_lang_css = This page is best viewed in an up-to-date web browser with style sheets -LRB- CSS -RRB- enabled .");
("co. uk\\/bbc\\/bbc \\ \\/ s ?");
("Sign up for free e-mail news alerts");
("Letters for publication should be sent to :");
("co. uk at");
("If you see a comment that you believe is irrelevant or inappropriate , you can flag it to our editors by using the report abuse links .");
("-LRB- Additional reporting by ");
("-LRB- Writing by ");
("Explore the three plans to stop the oil flow .");
("See the step-by-step procedure involved in an oil burn .");
("This Story : Read + ");
("Posted by : ");
("on.cnn.com");
("We ve seen this movie");
("THIS COPY MAY NOT BE IN ITS FINAL FORM AND MAY BE UPDATED .");
("stm storyID = ");
("stm SCRIPT_NAME = ");
("SCRIPT_URL = ");

("Please consider upgrading your browser software");
("co. uk search blq_lang_ak");
("It was last modified at");
("It was first published at");
("This article was amended");
("Copyright ");
(">> reporter :");
("This material may not be published , broadcast , rewritten or redistributed .");
("This image contains graphic content that some viewers may find disturbing .");