

Automatic Question Answering & Generation in News Article Collections

Adam Jatowt

23 May, 2022; *La Rochelle*

adam.jatowt@uibk.ac.at

Cultural Heritage Informatics & Archival Informatics

- **Cultural Heritage Informatics**

- “emerging field of *interdisciplinary research* and practice concerned with the role of *information* and *computing technologies* to support the *creation, capture, organization, and pluralization of culture*, in whatever form, as *heritage*”

[Kent State University]

- **Archival Informatics**

- “theory and application of *informatics* in and around the realm of *archives* and *record keeping*”

[Wikipedia]



Big Archival Data

- Humanity has generated large amounts of textual data over the last decades:
 - Newspaper archives
 - Book collections
 - Scientific publication archives
 - Administrative document archives
 - Web archives (Internet Archive: 1996-2021)
 - Social media archives
 - Product review collections
 - Product reports, customer complaints
 - etc.



These collections are continuously growing

Temporal aspects (e.g., publication dates) are crucial

Digital Document Archive Examples

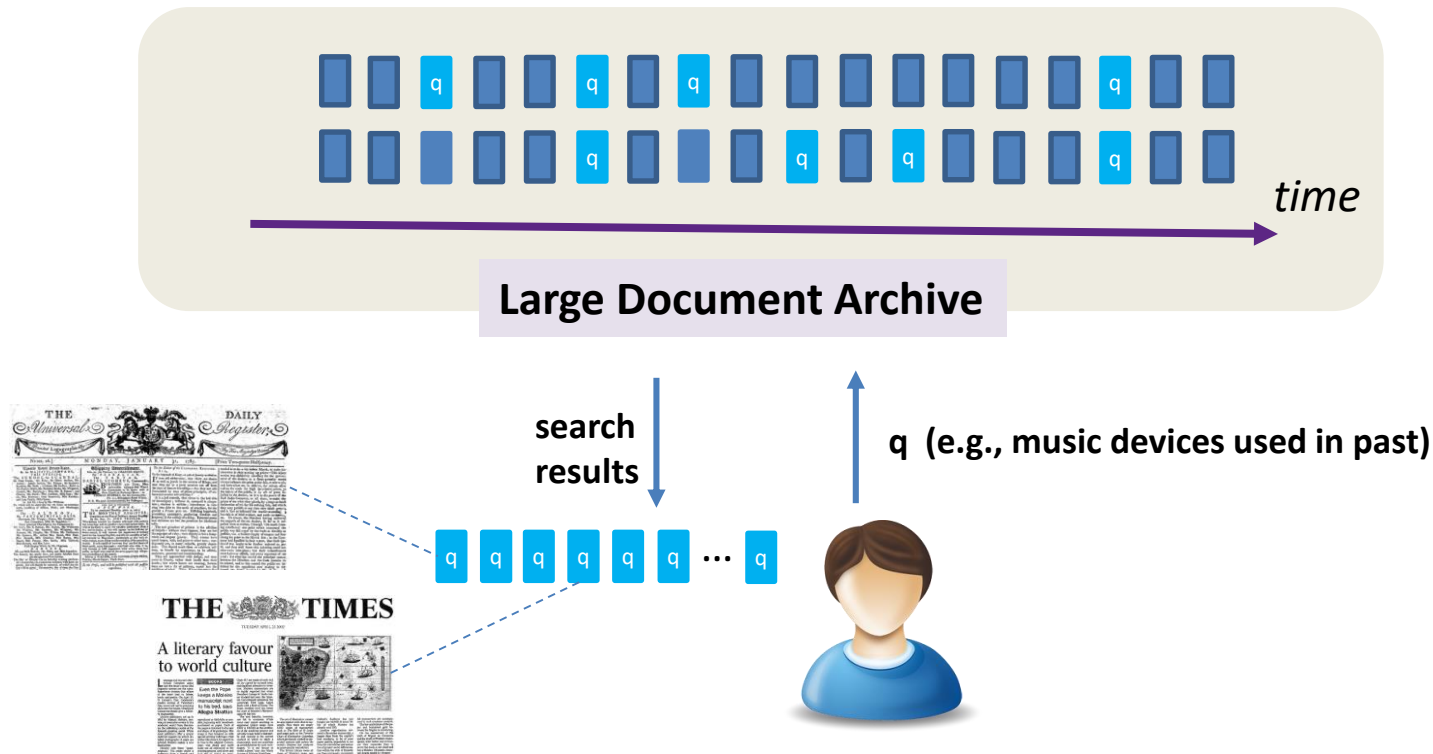
Big archival data, e.g.:

- *Chronicling America* - over 5.2 million individual newspaper pages
 - *The Times Digital Archive* - 3.5 million news articles (1785–2008)
 - *Google Books* - scanned over 5% of books ever published
 - *Internet Archive* - 286 billion web pages since 1996 (>15 petabytes of data)
 - *Amazon* - 142 million product reviews dataset (1994-2014)
 - *etc.*
- Nearly all national libraries and archives have their own digital collections [1]
 - Big Costs: e.g., in 2009 and 2010 the budget of the Japanese National Diet Library for digitization was 137 billion yen

Despite massive data and huge costs the number of users is very small

We could popularize archives more by making them accessible and easy for everyone

Current Realization of Full-text Search in Document Archives



Difficult to make sense of results..

We need more user friendly solutions!

What Kind of Solutions?

- Structuring search results
 - Arranging them chronologically or by the relevance + time?
- Summarizing search results
 - Clustering & comparing by content or time, or their combination?
- Increasing content understanding
 - Adding explanations, or links between results or existing knowledge bases (e.g., Wikipedia)?

Or directly answering what users ask for?

Empowering Users

- Let users ask **free natural questions** against an archival collection, including questions on minor things/events
 - Applications in **education** (e.g., history science) and **entertainment** but also for **journalists**, **lawyers**, **insurance/finance workers**, etc.

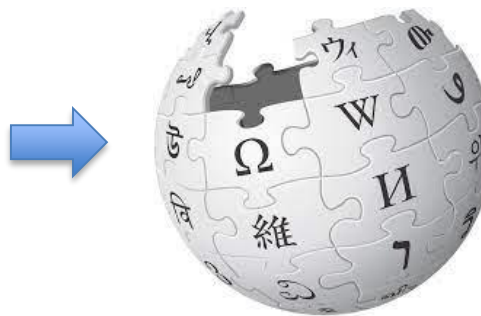
Questions
Which party, led by Buthelezi, threatened to boycott the South African elections?
What bill was signed by Clinton for firearms purchases?
Which federal prosecutor that led the investigation for the leak of identity of Valerie Plame?
Riot in Los Angeles occurred because of the acquittal of how many officers in police department?
Which American professional pitcher died because his small airplane crashed in New York?

Examples of archival questions & their answers

Question Answering Field

- Automatic **Question Answering (QA)** is an established field of Natural Language Processing (NLP)
 - An input is: **question** and **document collection**
- Most systems work on **synchronic document collections** like Wikipedia
 - No work deals with longitudinal temporal news collections like archives

**What is the population
of La Rochelle?**
[Question]



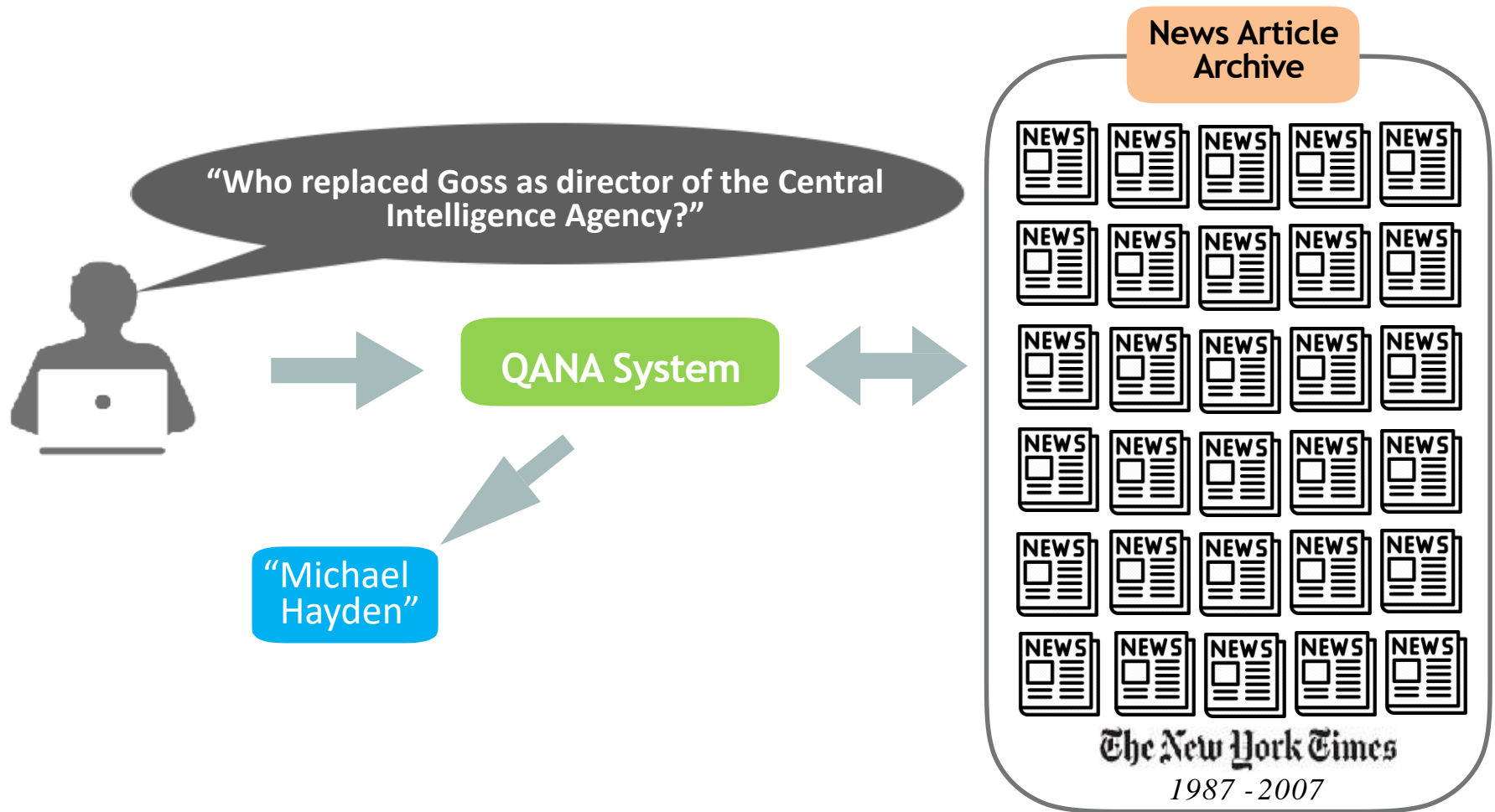
77,205
[Answer]

Agenda for Remaining Part of the Talk

1. Background
2. Unsupervised QA approach
3. Event time estimation to support QA search
4. Large-scale open-domain QA dataset

UNSUPERVISED APPROACH

QANA: Question Answering in News Archives

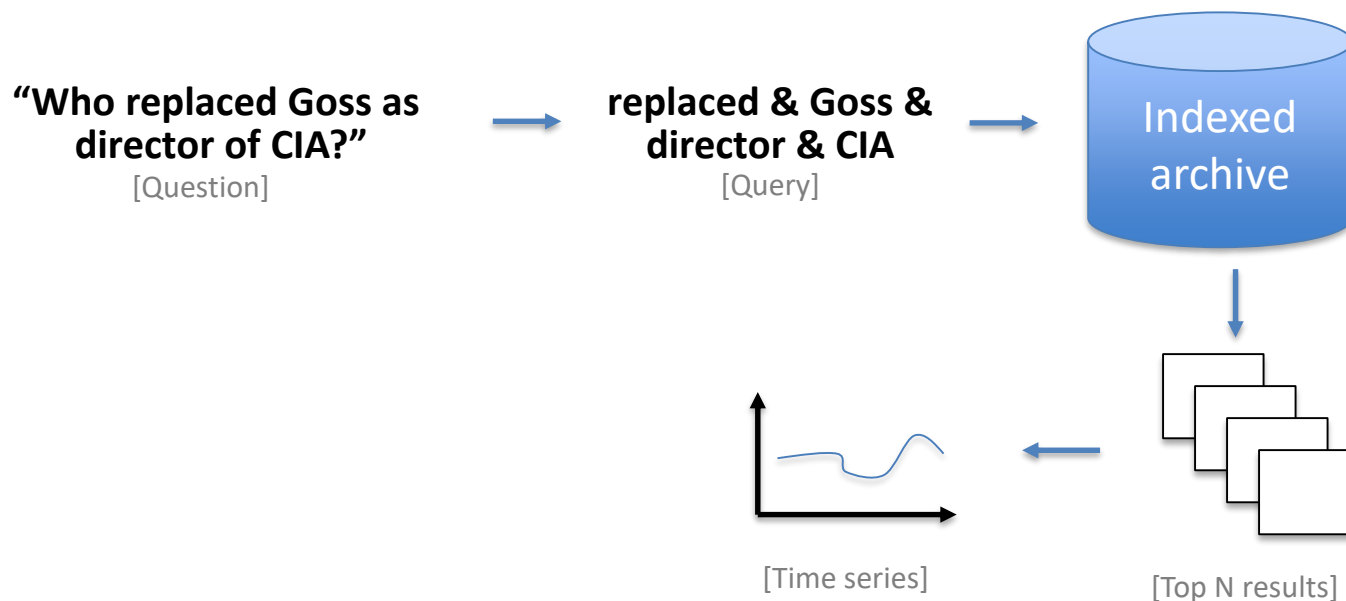


Our Research Focus: IR Angle

Given millions of documents, how to select a **small subset** of them for extracting correct answer to the user's question?

Retrieving Initial Documents

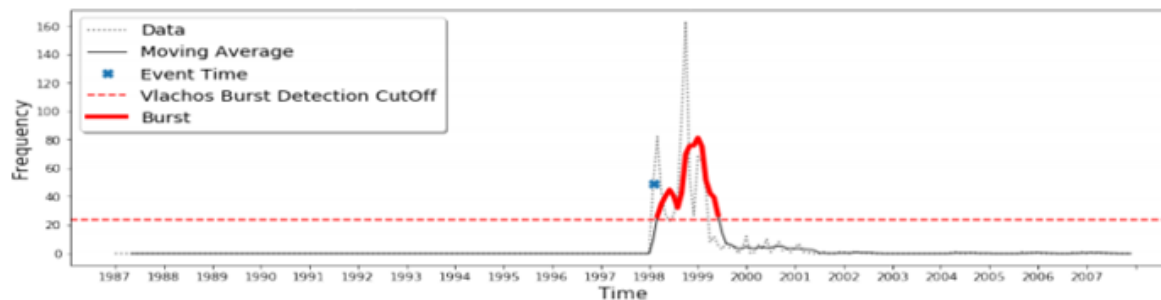
- Construct search query from an input question
 - Keyword selection approach: modified Yake! [1]
- Use BM25 to collect top N candidate documents



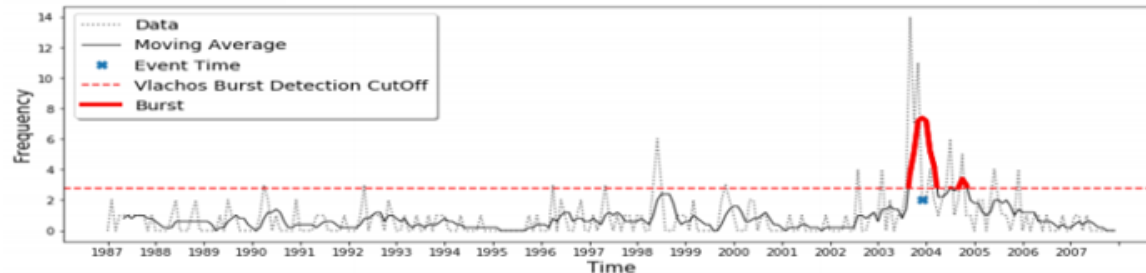
Question's Time Scope Estimation

Detect **question time scope** by automatically **finding bursts** in distribution of results

Lewinsky told whom about her relationship with the President Clinton?

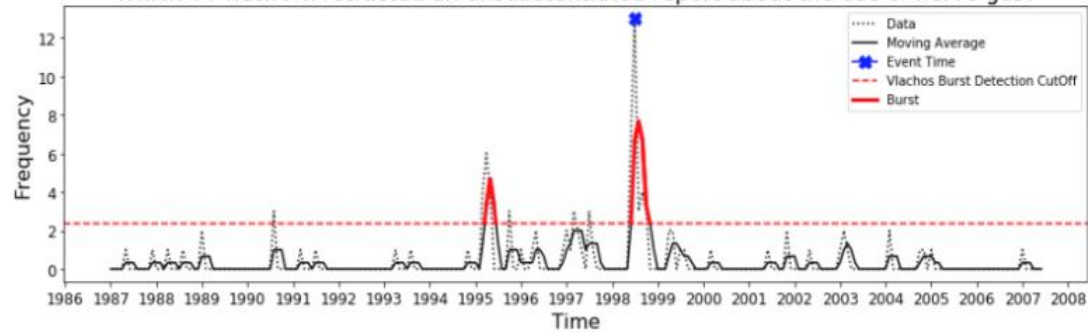


Which Hollywood star became governor of California?

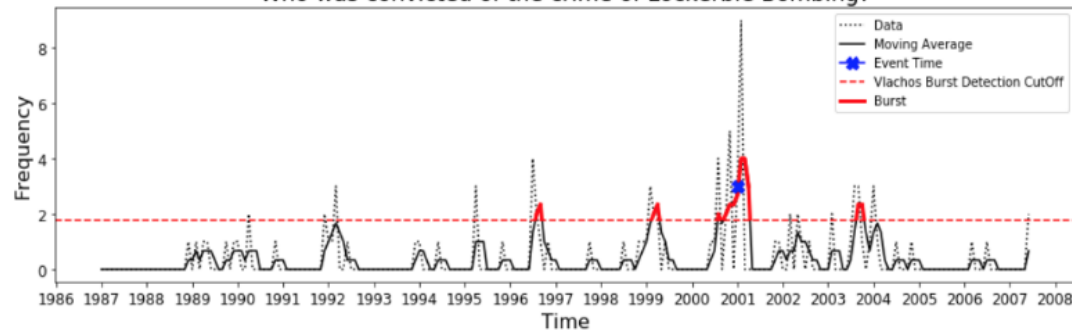


Question's time scope is represented as a set of time periods

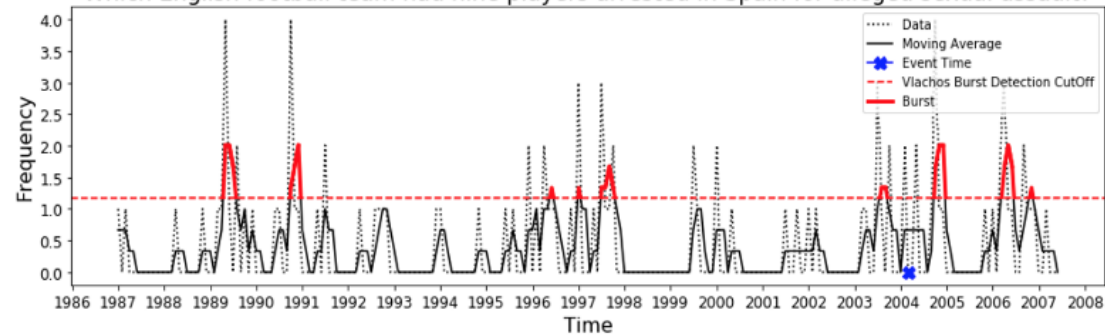
Which TV network retracted an unsubstantiated report about the use of nerve gas?



Who was convicted of the crime of Lockerbie Bombing?

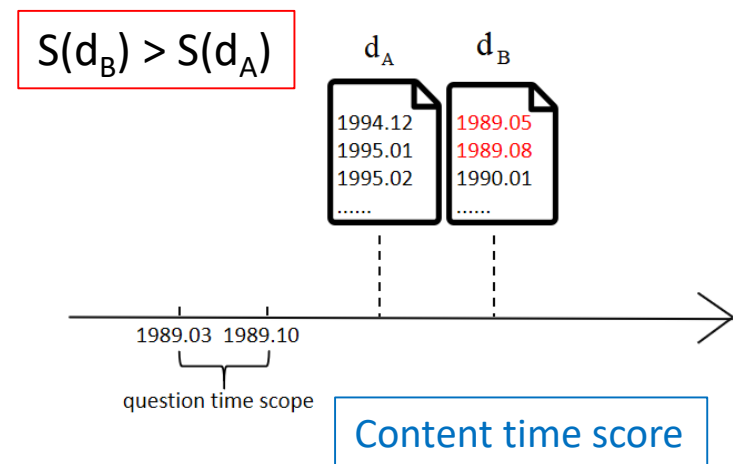
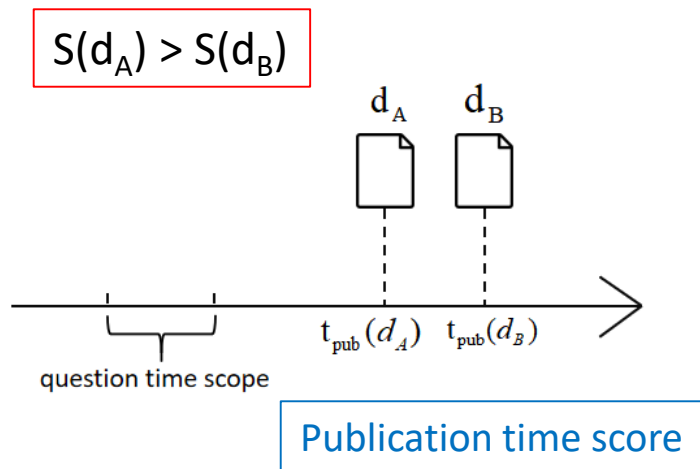


Which English football team had nine players arrested in Spain for alleged sexual assault?



Document Temporal Scores Estimation

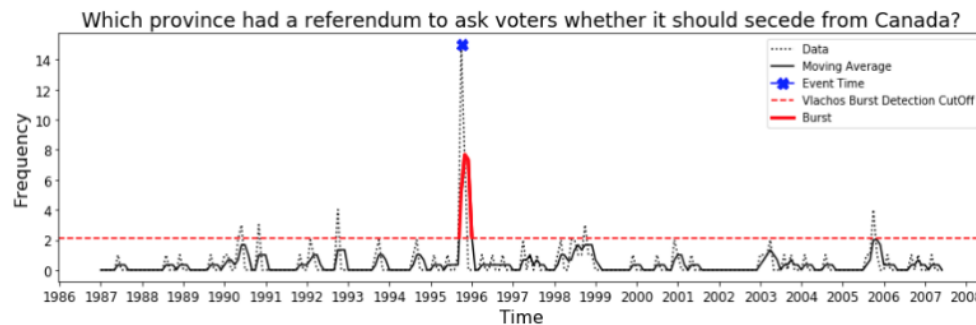
1. Take the estimated **time scope** of a question
2. Score documents wrt. **degree to which they refer to this time scope**



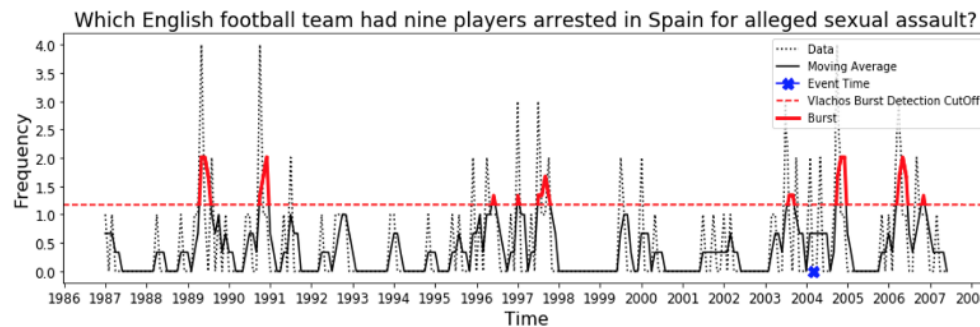
- + Use Kernel Density Estimation to compute the overlap of content time expressions and question time scope
- + If the question time scope contains multiple periods, aggregate the scores for all periods

Document Reranking

1. Combine the **two scores** of documents
2. Rerank documents by a linear **dynamic combination** of their **relevance scores** and **temporal scores**



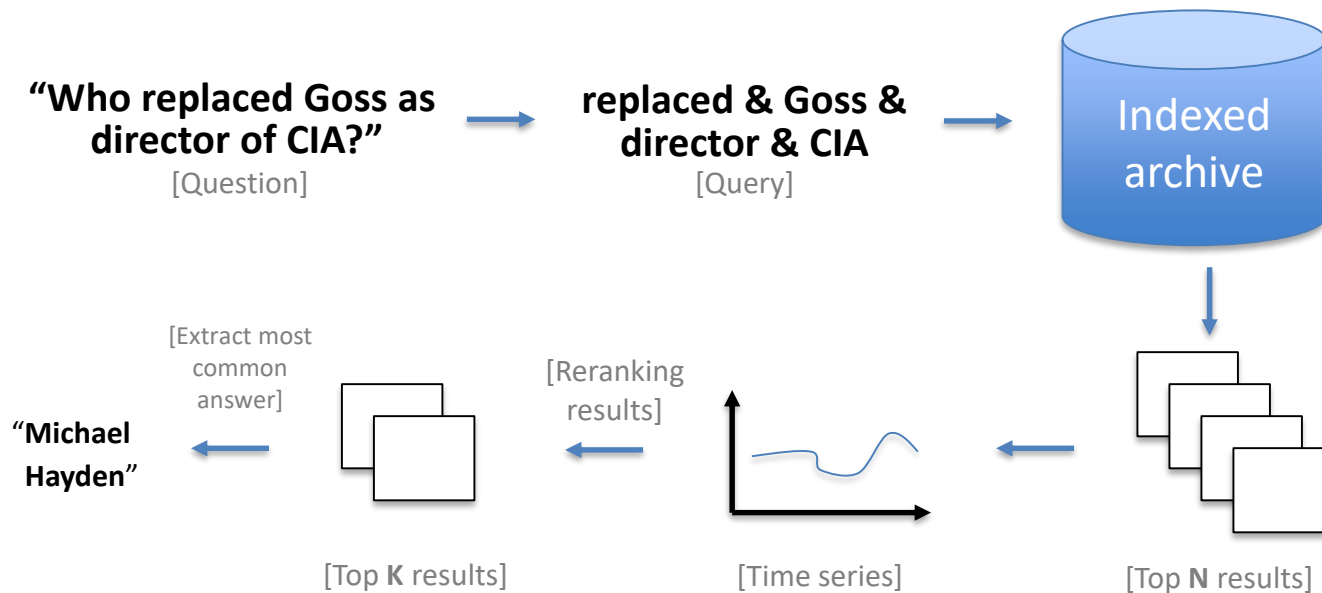
Document **temporal** scores are used more



Document **relevance** scores are used more

Extracting Answers and Select Final Answer

- Take K ($K \ll N$) top-ranked reranked documents and find answers using [DrQA method](#)
- Aggregation step: choose the [most common answer](#) as the final answer



Dataset and Testset

- **Dataset:** New York Times Annotated Corpus (1987-2007)
 - 1.8 million articles in total



- **Testset:** 1,000 selected questions (for 1987-2007)

Resources	Number of explicitly time-scoped questions	Number of explicitly time-scoped questions
History quizzes from funtrivia ^a	235	204
History quizzes from quizwise ^b	67	75
Wikipedia pages	140	143
Questions from datasets Rajpurkar et al. (2016), Jia et al. (2018)	58	78

<http://www.funtrivia.com/quizzes/history/index.html>

<http://www.quizwise.com/history-quiz>

Results

Model	Top 1		Top 5		Top 10		Top 15	
	EM	F1	EM	F1	EM	F1	EM	F1
DrQA-NYT (Chen et al. 2017)	19.40	25.65	25.40	32.14	26.20	34.13	27.00	35.86
QA-NLM-U (Kanhbua and Nørvåg 2010)	20.40	28.34	25.00	33.50	30.40	38.58	31.40	39.95
QA-No-Re-ranking (Seo et al. 2016)	19.00	27.19	24.60	32.81	29.00	38.52	31.00	40.17
QANA-TempPub	20.40	28.27	26.20	34.27	32.80	42.88	35.60	45.06
QANA-TempCont	20.00	28.03	26.00	33.76	32.20	42.17	33.80	43.71
QANA	21.00	28.90	28.20	36.85	34.20	44.01	36.20	45.63

Comparison with
Wikipedia-based
system

Model	Top 1		Top 5		Top 10		Top 15	
	EM	F1	EM	F1	EM	F1	EM	F1
DrQA-Wiki (Chen et al. 2017)	21.20	25.76	22.00	26.30	23.00	26.97	24.40	28.70
DrQA-NYT (Chen et al. 2017)	19.40	25.65	25.40	32.14	26.20	34.13	27.00	35.86
QANA	21.00	28.90	28.20	36.85	34.20	44.01	36.20	45.63

IMPROVING TIME SCOPE ESTIMATION

Detecting Question Time Scope

- Burst detection relies on redundancy
 - Questions on minor events may produce no bursts
 - Recurring events may have multiple or noisy bursts

We need more effective approach for
detecting question time scope

Task Definition

Given:

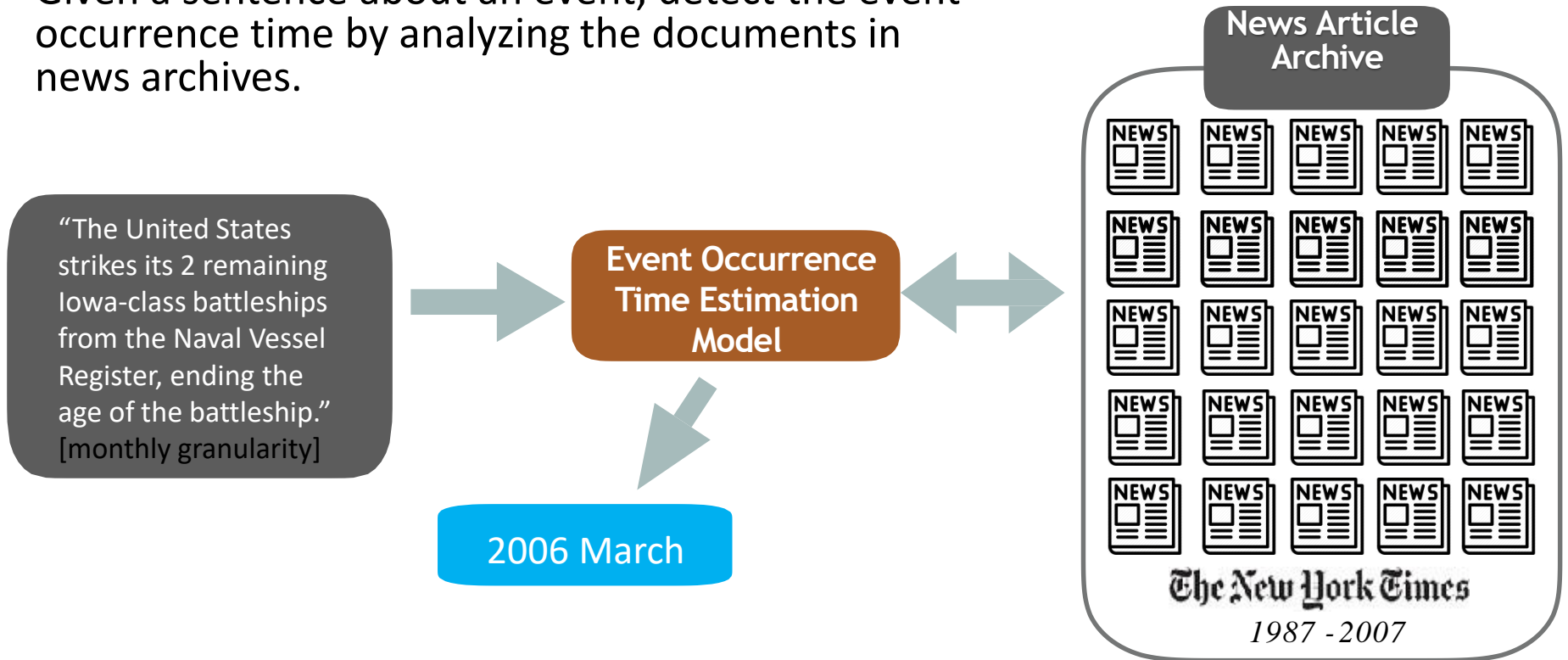
- a description of an event *e*
- a required temporal granularity *g*
- an underlying archival news collection *D*

Find *e*'s occurrence date under *g* using *D* as data source.

- E.g.:
 - *e*: "A bombing of a Superferry by Abu Sayyaf in the Philippines killed 116"
 - *g*: month granularity
 - *D*: New York Times news archive [1987-2007]
 - **Answer**: "2004-02"

Model

Given a sentence about an event, detect the event occurrence time by analyzing the documents in news archives.



Example Event Descriptions

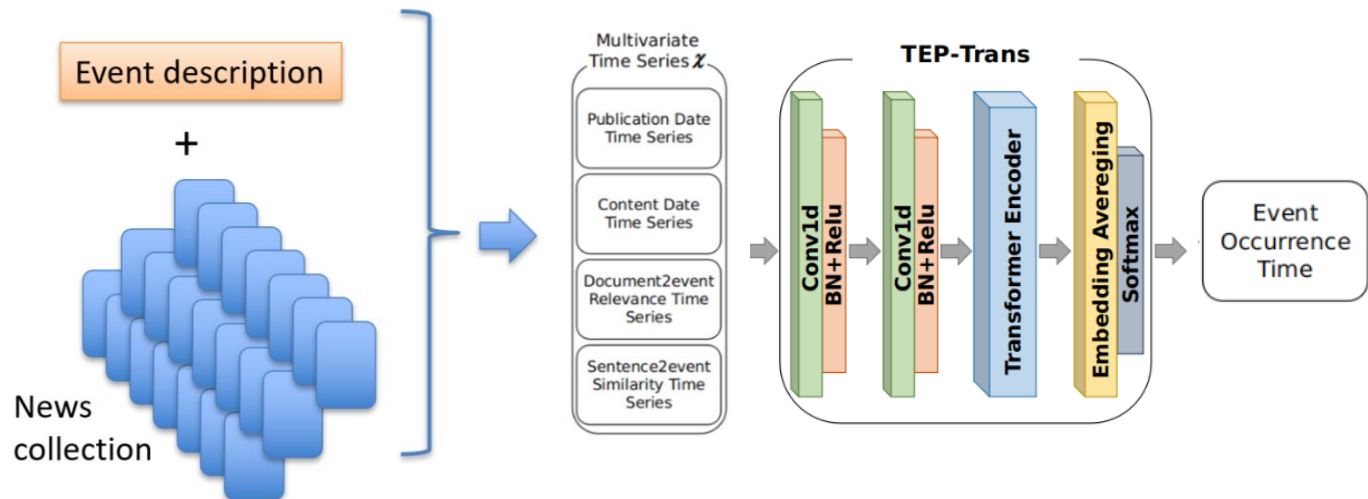
Only year info
in Wikipedia

Description
An official news agency in the Soviet Union reports the landing of a UFO in Voronezh.
Antonov-26 plane crashes at Gyumri, Armenia, 36 killed.
FBI agent Earl Pitts pleads guilty to selling secrets to Russia.
President of Pakistan Pervez Musharaf narrowly escaped an assassination attempt.
Former White House aide I. Lewis Libby, Jr. is found guilty on four of five counts of perjury and obstruction of justice trial.
George Bell is 1st Blue Jay ever to win the AL MVP.
Toru Takemitsu's "Archipelago" premieres in Aldeburgh England.
Will Clark, National League's Most Valuable Player signs a \$15 million four-year contract with San Francisco Giants.

Not in
Wikipedia

Our Approach: Generating Multivariate Time Series

- Use multiple temporal and frequency-based signals from the collection after keyword extraction and BM25 search
- **Publication date time series**
 - Aggregated top- N relevant documents over time
- **Content date time series**
 - Aggregated time references mapped to timeline
- **Document2event relevance time series**
 - Similarity score of relevant documents to the event sentence
- **Sentence2event similarity time series**
 - Similarity score of sentences with time expressions to the event sentence



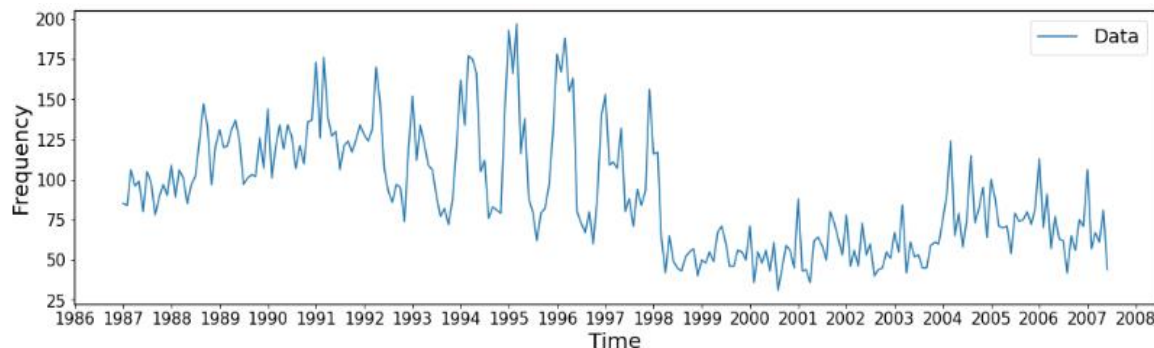
Dataset and Training/Test set

- **Dataset:** New York Times Annotated Corpus (1987-2007)

- 1.8 million articles in total
- Indexed by ElasticSearch



- **Event sentence set:** 22k event descriptions from Wikipedia year pages and onthisday.com website, (80% train, 10% dev, 10% test)



Results

- Random guess accuracy for day granularity: 0.01%
- Improved to: 16.42%

Model	Day		Week		Month		Year	
	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE
RG	0.01	2482.12	0.08	355.25	0.40	81.57	4.77	6.91
DPD	0.04	2690.47	0.17	252.34	0.93	56.71	7.90	5.51
BD [44]	1.42	1418.26	14.01	215.80	18.75	49.70	27.09	4.37
NLM [21]	1.38	1300.34	15.53	194.16	21.87	45.85	33.52	3.80
AA [12]	6.02	1508.73	16.96	216.02	21.65	48.39	32.54	3.99
MSSD	9.50	1268.47	17.05	181.22	22.32	44.32	34.82	3.67
CNN-LSTM [24]	1.38	1382.38	7.49	174.26	23.30	37.04	37.54	3.21
HEO-LSTM [14]	-	-	-	-	-	-	15.58	4.81
TEP-CNN	8.39	1518.93	19.41	194.86	25.35	44.17	34.01	3.87
TEP-Trans	16.42	1235.67	23.66	166.64	30.89	36.19	40.93	3.01

Main results

Features	Day		Week		Month		Year	
	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE
X_{temp}^{pub}	7.76	1563.20	13.25	216.92	17.63	48.04	30.26	3.80
X_{temp}^{cont}	6.60	1623.37	12.32	213.85	16.96	48.60	29.10	3.85
X_{text}^{doc}	8.52	1358.91	16.42	197.48	21.29	44.78	33.48	3.59
X_{text}^{sent}	9.86	1480.49	16.24	194.46	20.66	43.75	31.91	3.62
$X_{temp}^{pub}, X_{temp}^{cont}$	7.41	1578.50	15.31	211.88	19.28	46.16	30.53	3.73
$X_{text}^{doc}, X_{text}^{sent}$	13.34	1301.54	18.39	183.91	24.06	41.34	34.46	3.43
$X_{temp}^{pub}, X_{text}^{doc}$	11.02	1217.29	18.92	174.43	25.93	40.14	38.12	3.27
$X_{temp}^{cont}, X_{text}^{sent}$	12.18	1435.37	18.43	182.97	23.70	41.30	33.12	3.58
X	16.42	1235.67	23.66	166.64	30.89	36.19	40.93	3.01

Results when using only some time series

Error Examples & Question Answering Results

Table 8: Examples of event descriptions that are wrongly estimated by TEP-Trans, based on month granularity

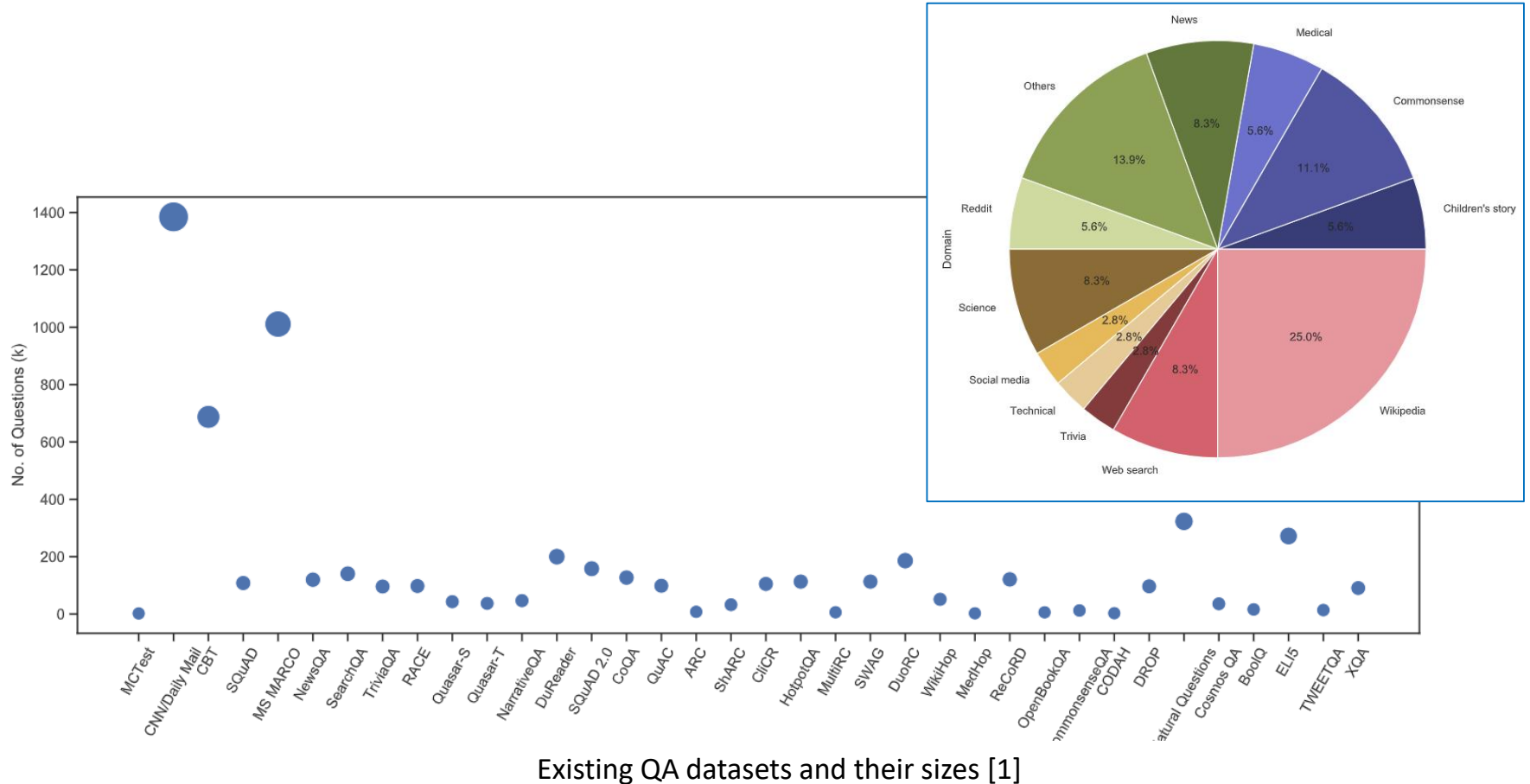
No.	Description	Occurrence Time	Estimated Time
1	William Anthony Odom, North Carolina 15-year-old, accidentally hangs himself staging a gallows scene at a Halloween party.	1990-10	1996-10
2	The flu outbreak in Britain puts pressure on NHS.	2000-01	2005-11
3	Turin, Italy, is awarded the 2006 Winter Olympics.	1999-06	2006-02

Table 9: Performance of different models in QA task

Model	Top 1		Top 5		Top 10		Top 15	
	EM	F1	EM	F1	EM	F1	EM	F1
QANA [45]	21.00	28.90	28.20	36.85	34.20	44.01	36.20	45.63
QANA + TEP-Trans	23.00	30.89	29.60	38.17	35.40	45.49	38.00	48.35

LARGE SCALE OPEN QA DATASET FOR TEMPORAL NEWS COLLECTIONS

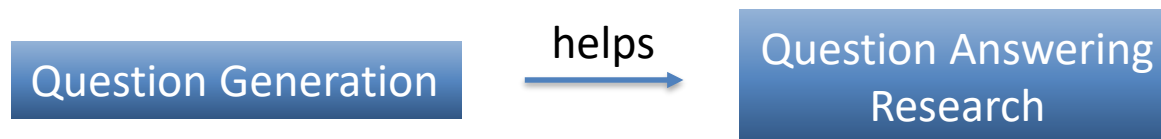
Large Scale QA Datasets



[1] Zhu, Fengbin, et al. *Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering*. arXiv preprint arXiv:2101.00774 (2021)

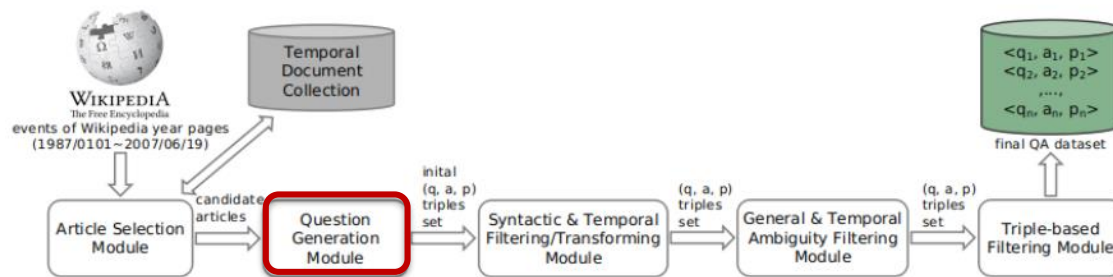
Building Large Scale QA Dataset for Archival News Collection

- **Main idea:** instead of manually preparing questions use **Question Generation** + **aggressive filtering**
- **Question Generation** (QG): the task of automatically generating questions from natural language text
 - “Adam makes a presentation today at DAS” -> “Who makes a presentation today at DAS?”



Question Generation Module

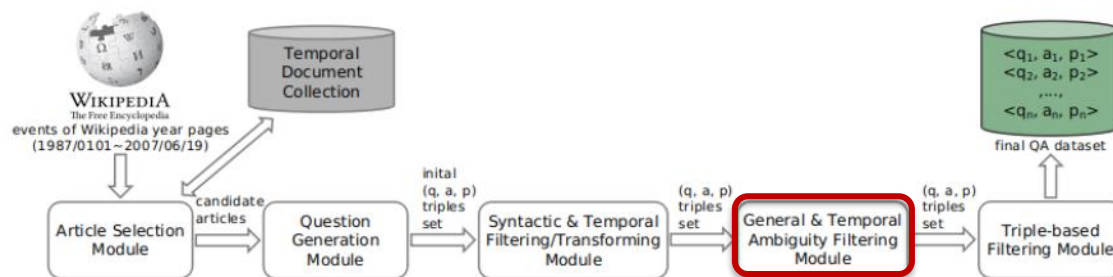
- Question Generation Model: T5 Model fine-tuned on SQuAD 1.1
 - Each **entity** in the paragraph is labeled as an answer
 - The model uses (**answer**, **paragraph**) to generate the question



General & Temporal Ambiguity Filtering Module

Two steps:

- Filtering by **Content Specificity**
- Filtering by **Temporally Ambiguity**



Filtering General Answer Sentences

- General Sentence
 - *“Despite recent declines in yields, investors continue to pour cash into money funds”*
- Specific Sentence
 - *“While American PC sales have averaged roughly 25 % annual growth since 1984 and West European sales a whopping 40 % , Japanese sales were flat for most of that time .”*
- Filtering setup:
 - Dataset [1]: 4,342 manually annotated sentences
 - Classifier: Roberta-base
 - Accuracy: 84.49%

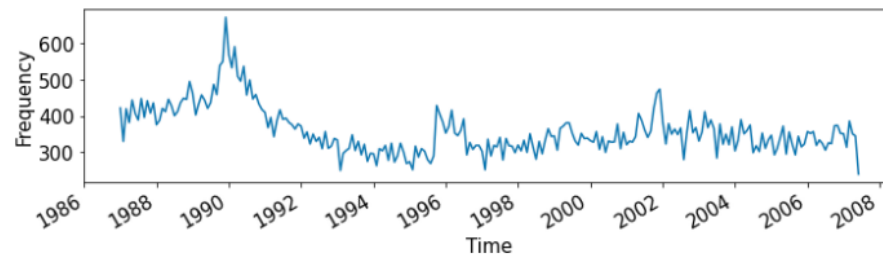
Filtering by Temporally Ambiguity

- Temporally ambiguous question:
 - A question with several correct answers from different time points
- Examples:
 - (OK) *Who sent 20,000 American troops to Bosnia?* [Clinton]
 - (OK) *How old was Evelyn Sabin when she died?* [90]
 - (NG) *Who was the Senator of West Virginia?* [John D. Rockefeller]
 - (NG) *How many points does Ashley McElhiney have?* [5]
 - (NG) *What did the Syrians want to end?* [the civil war]
 - (NG) *What country is the current U.S. policy?* [Iran]
- Filtering setup:
 - Dataset: 5,500 manually annotated questions obtained from the previous filtering steps
 - Classifier: BERT-base
 - Accuracy: 81.82%

ArchivalQA

Data Statistics:

- **532k** question-answer pairs derived from 313k paragraphs of 88k news articles from New York Times dataset (1987-2007)



Distribution of articles used in ArchivalQA

id	question	answer	org_answer	answer_start	para_id	trans_que	trans_ans	source
train_0	Who claimed responsibility for the bombing of Bab Ezzouar?	Al Qaeda	Al Qaeda	184	1839755_20	0	0	wiki
train_4	When did Tenneco announce it was planning to sell its oil and gas operations?	May 26, 1988	today	103	148748_0	0	1	rand
val_45	What threat prompted Mr. Paik's family to flee to Hong Kong?	the Korean War	the Korean War	327	1736040_7	0	0	wiki
test_84	Along with the French Open, what other tournament did Haarhuis win in 1998?	Wimbledon	Wimbledon	527	1043631_15	1	0	rand

ArchivalQA Data Examples


ArchivalQA Sub-datasets

Four sub-datasets are created based on the question difficulty levels and the containment of temporal expressions:

- ArchivalQAEasy (100k)
- ArchivalQAHard (100k)
- ArchivalQATime (75k)
- ArchivalQANoTime (75k)

Table 10: Performance of different models over different Sub-Datasets

Model	ArchivalQAEasy		ArchivalQAHard		ArchivalQATime		ArchivalQANoTime		
	EM	F1	EM	F1	EM	F1	EM	F1	
DrQA-NYT [7]	42.10	51.97	22.81	31.24	31.32	42.17	39.59	47.18	
DrQA-NYT-TempRes [7]	48.41	57.26	27.37	34.02	33.19	44.01	46.39	54.91	
BERTserini-NYT [66]	59.15	69.16	25.00	33.73	50.65	63.24	55.36	68.37	
BERTserini-NYT-TempRes [66]	61.80	71.56	29.88	38.44	51.12	65.67	58.27	70.19	
DPR-NYT [22]	46.24	59.63	39.99	48.03	42.29	53.73	45.28	57.92	} Dense Retriever
DPR-NYT-TempRes [22]	52.10	64.51	41.65	48.96	42.91	54.27	51.13	62.75	



Conclusions

1. Novel research task: open question answering on archival document collections
2. Unsupervised approach
3. Question time scope estimation using multivariate time series analysis
4. Building a large-scale question-answer dataset for temporal news collections following question generation approach

Our Prior & Future Work in Archival Informatics



- In the past we used news archives for:
 - Causal relation detection [WWW'16, TOIS'16]
 - Temporal analogy estimation [ACL'15, TKDE'16, CIKM'17]
 - Comparative timeline summarization [WSDM'19, ECIR'19]
 - Estimating contemporary relevance of content [JCDL'21]
 - Finding interesting/surprising content [ECIR'21]
 - Automatic question answering [ECIR'20, IRJ'21, SIGIR'21, SIGIR'22]
 - Multi-hop question answering & generation [Mavi 2022]
 - Comparative question answering & generation
 - Comprehension-oriented question answering & generation

**How can we effectively return data
from the past for present users?**