Question Answering & Finding Temporal Analogs in News Archives

Adam Jatowt

1 Feb, 2021

adam.jatowt@uibk.ac.at

Today’s Agenda

1. Question answering in news archives
2. Finding and explaining temporal analogs in news archives
3. Related Interactive demos

Big Archival Data

- Massive archives containing past texts are available nowadays, e.g.:
  - Newspaper archives
  - Book archives
  - Scientific publication archives
  - Administrative archives
  - Web archives
  - Social media archives
  - Product review archives
  - Etc.

Born-digital

The New York Times

TIME

Google books

arXiv.org

PubMed

国立公文書館
NATIONAL ARCHIVES OF JAPAN

INTERNET ARCHIVE
WayBackMachine

amazon

Archives are common and span variety of genres

Heritage that is continuously growing and becoming increasingly important to us

Digital Document Archives

- **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 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
- Little usage: very few users utilize document archives, and mainly professionals

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

We want to popularize archives by making them useful and easy to use for everyone

[1] N. Stroeker and R. Vogels, *Survey Report on Digitization in European Cultural Heritage Institutions* 2012

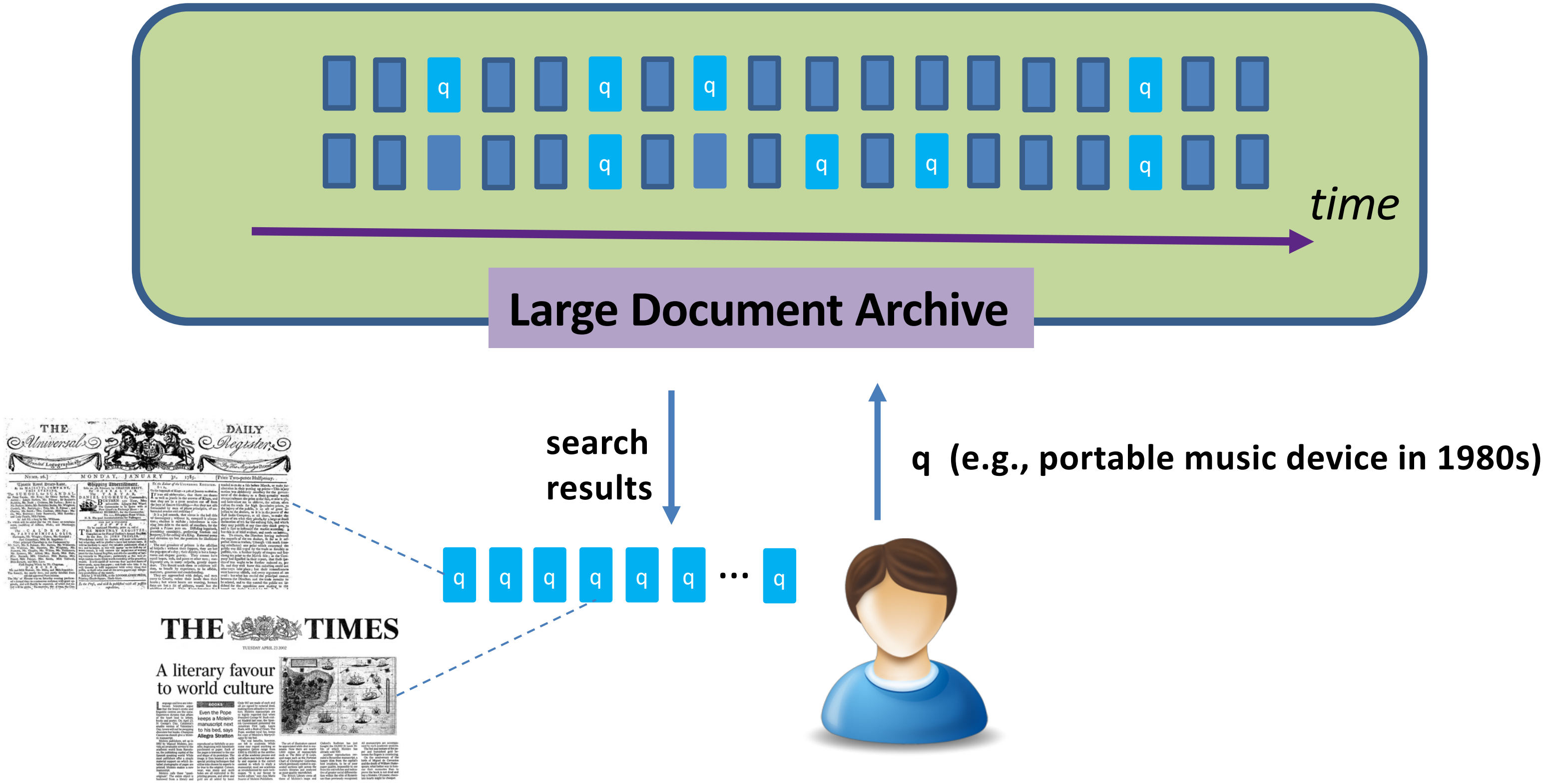
History

"***Those who cannot remember the past are condemned to repeat it*** "
(George Santayana)

- **History** plays important role in our society allowing to understand the past, the present, and even may help to predict the future to some extent
  - Knowledge of history is essential for being prepared for an **active life in contemporary society**

- **Computational approaches to history**: harnessing computational power to support history analysis, writing, usage, studying, etc.
  - Part of larger trend of "**Digital Humanities**"

Current Interfaces to Document Archives



Difficult to make sense of results!

Challenges & Open Questions

- Challenges:
  - Data is large and distributed over time
  - Vocabulary & context in the past changed much
  - Users' knowledge of the past and its context is limited

How can we effectively extract and provide information from document archives (the past) for present users?

How news archives in particular can be made easy to use and accessible to ordinary users?

QA IN NEWS ARCHIVES

Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa and Michael Farber: Improving Question Answering for Event-focused Questions in Temporal Collections of News Articles, Information Retrieval Journal (IRJ) (2021)
Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa and Michael Farber: Answering Event-Related Questions over Long-term News Article Archives, Proceedings of ECIR 2020, pp. 774-789 (2020) {Industrial Impact Paper Honorable Mention}

Question Answering in News Archives

- The idea is to let users ask free natural questions about the past, especially about minor things and events
  - Applications for journalists, professionals researching history and ordinary users
- Automatic Question Answering is a well-established field of Natural Language Processing (NLP)
  - Most systems work either on Wikipedia or recent news
  - Typically an input is a document (e.g., a news article) and a question
  - Few works attempt answering open questions over a large document collections and no works deal specifically with long-term news archives

Questions	Answers	Event Dates
Which party, led by Buthelezi, threatened to boycott the South African elections?	Inkatha Freedom Party	1993.08
What bill was signed by Clinton for firearms purchases?	Brady Bill	1993.11
Which federal prosecutor that led the investigation for the leak of identity of Valerie Plame?	Patrick J. Fitzgerald	2003.11
Riot in Los Angeles occurred because of the acquittal of how many officers in police department?	Four	1992.04
Which American professional pitcher died because his small airplane crashed in New York?	Cory Lidle	2006.10

Examples of questions, their answers and dates

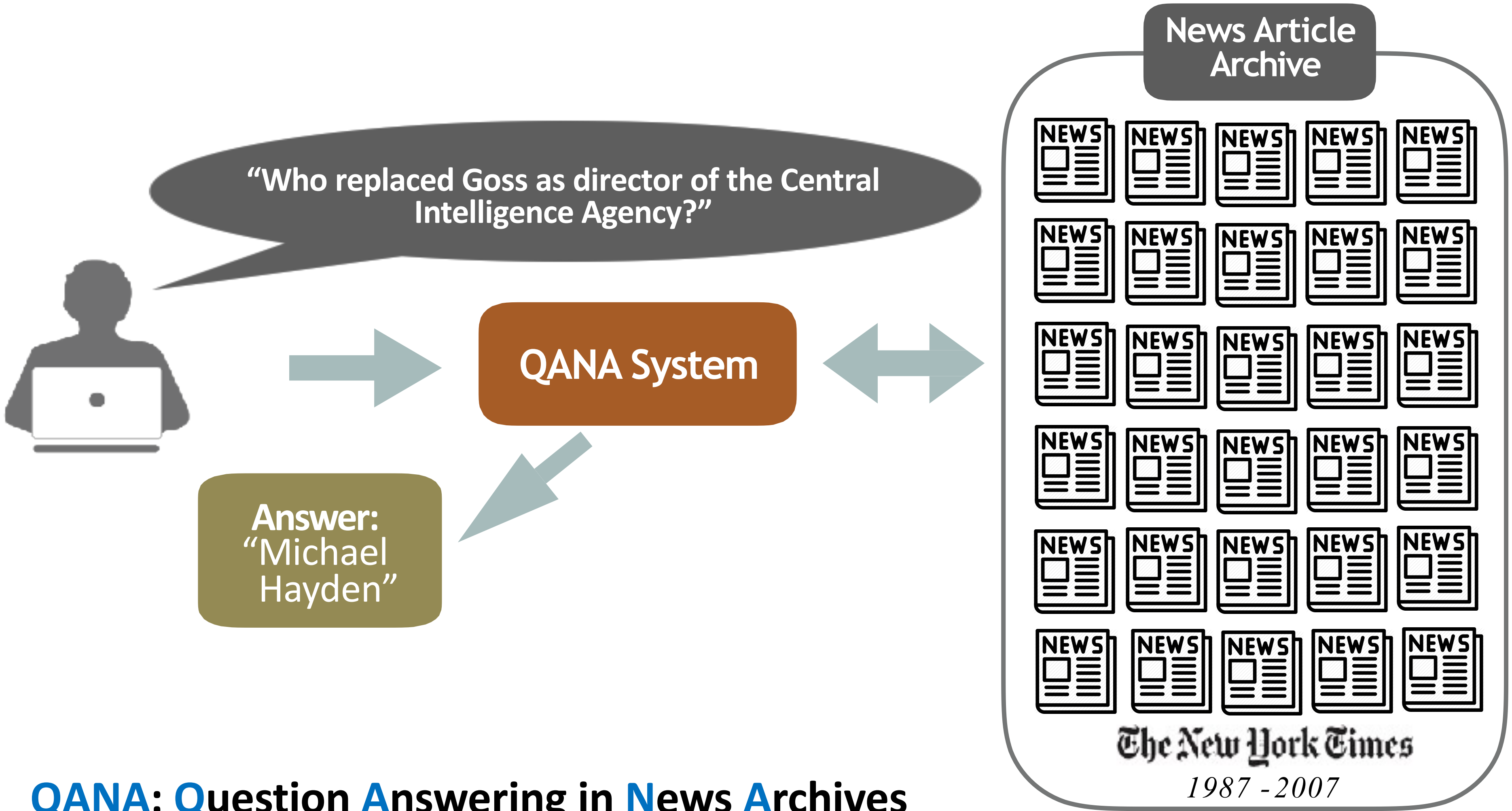
Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa and Michael Farber: Answering Event-Related Questions over Long-term News Article Archives, Proceedings of ECIR 2020, pp. 774-789 (2020) {Industrial Impact Paper Honorable Mention}

Explicit and Implicit Temporal Questions

- Two types of questions:
  - Explicitly time-scoped questions (with time expression)
  - Implicitly time-scoped questions (no time expression)

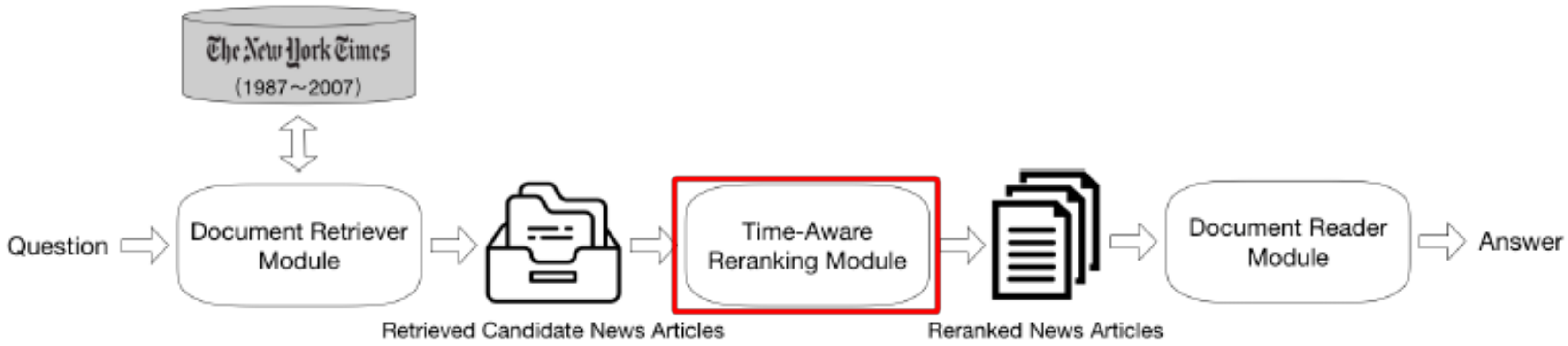
Questions	Time scoped	Answers	Event dates
The USSR flag was lowered and the Russian flag raised over in which building on 25 December 1991?	Explicitly	Kremlin	1991.12
Which country signed an economic accord with Palestinian Liberation Organization in April 1994?	Explicitly	Israel	1994.04
Who famously described his experiences to the media as "a near death experience" during November 2003?	Explicitly	Iain Duncan Smith	2003.11
Democratic U.S. presidential Gary Hart bowed out of the race due to his extra-marital affair with whom?	Implicitly	Donna Rice	1987.05
The dissolution of the Soviet Union occurred after whose resignation?	Implicitly	Mikhail S. Gorbachev	1991.12
Which famous painting by Norwegian Edvard Munch was stolen from the National Gallery in Oslo?	Implicitly	The Scream	2004.08

QANA System



QANA: Question Answering in News Archives

QANA: Question Answering in News Archives (Unsupervised Way)



QANA system exploits temporal information in additional **Time-aware Reranking Module**

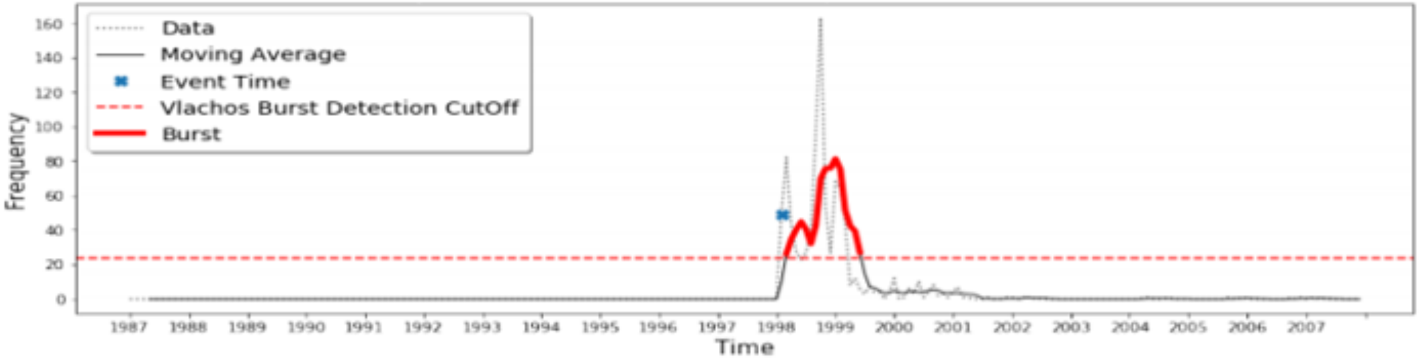
Research Problem

Given typically large number of past documents (~millions), how to select a small set of candidate articles for generating correct answer?

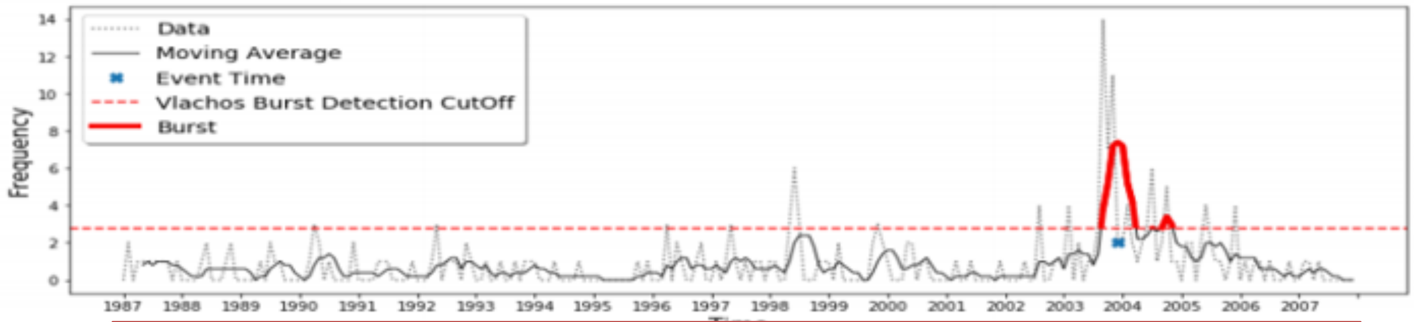
Step 1: Question Time Scope Estimation

The first step is to detect the time scope of the implicit temporal questions by **finding bursts in temporal distribution of search results** returned for question

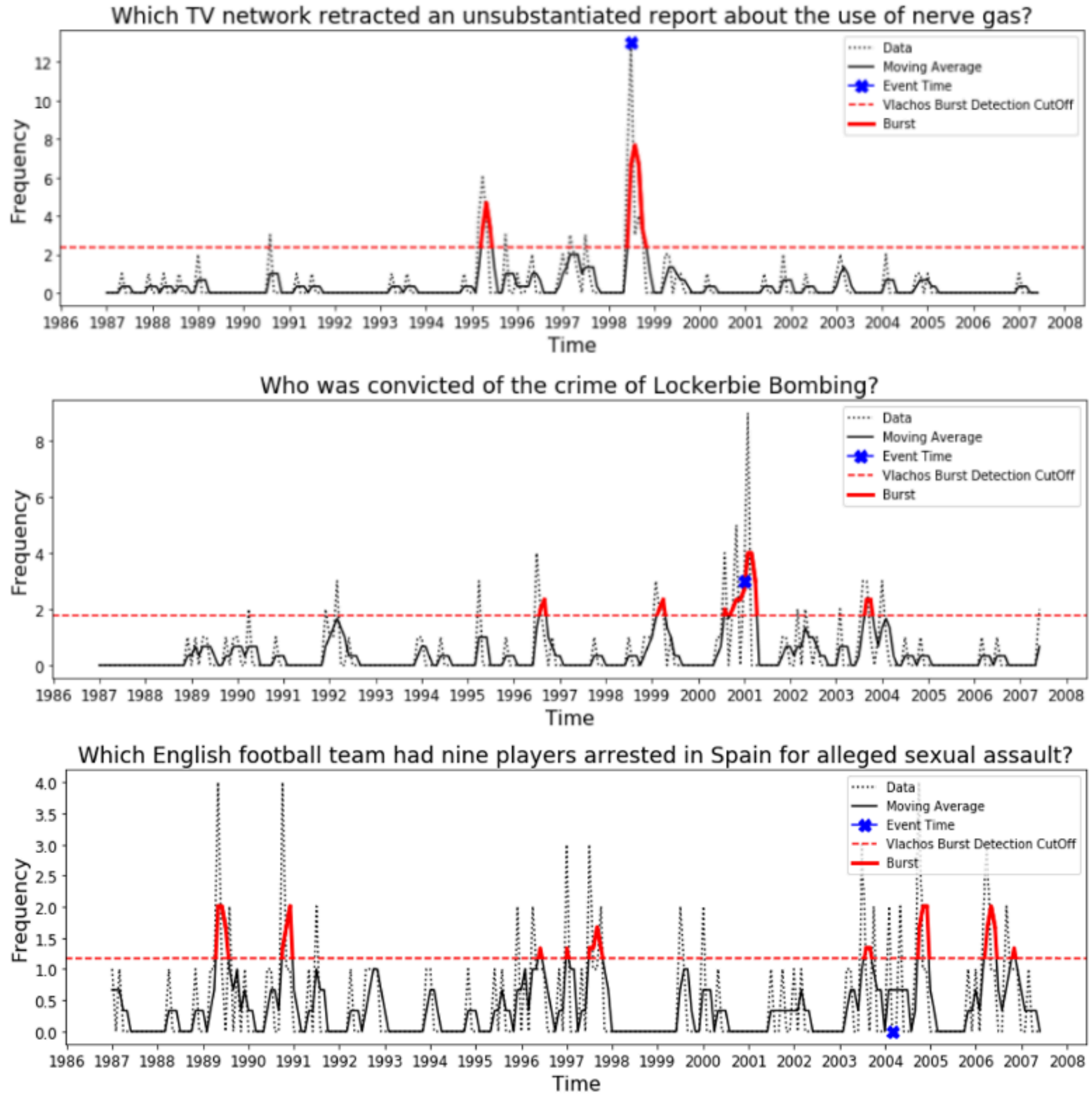
Lewinsky told whom about her relationship with the President Clinton?



Which Hollywood star became governor of California?



Question time scope is represented as a set of time periods


Which TV network retracted an unsubstantiated report about the use of nerve gas?
Data
Moving Average
Event Time
Vlachos Burst Detection CutOff
Burst
Frequency
0
2
4
6
8
10
12
Time
1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Who was convicted of the crime of Lockerbie Bombing?
Data
Moving Average
Event Time
Vlachos Burst Detection CutOff
Burst
Frequency
0
2
4
6
8
Time
1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Which English football team had nine players arrested in Spain for alleged sexual assault?
Data
Moving Average
Event Time
Vlachos Burst Detection CutOff
Burst
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Time
1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008

Step 2: Computing Document Temporal Scores

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

Step 2: Example of Retrospective References

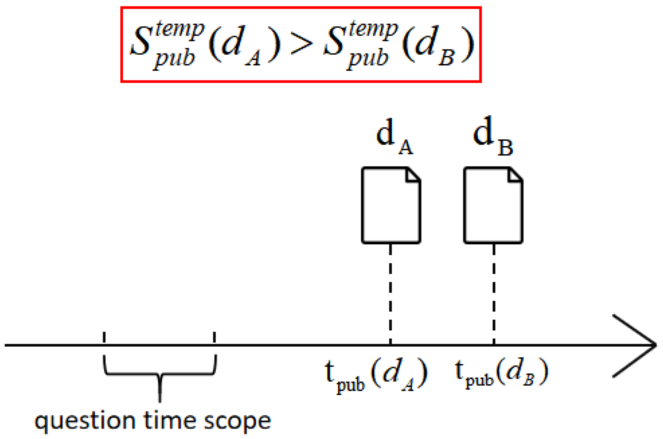
Question: How many people were killed in Concorde crash in 2000?
Answer: **113**
Event Occurred Date: 2000/07/25

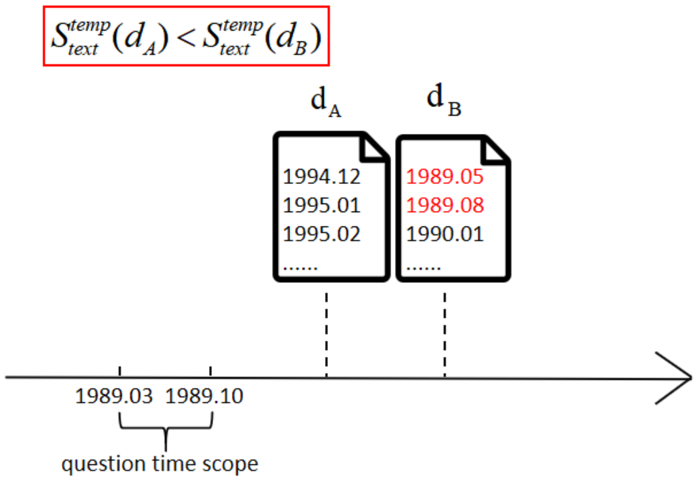
Relevant news article 1:
Title: Brian Trubshaw, 77, Dies; Tested Concorde
Published Time: 2001/03/28
Content:
Brian Trubshaw, a pilot who tested the British-French Concorde supersonic airliner and became its staunchest champion, died on March 24 at his home near Tetbury, ...
...
British Airways and Air France, the only airlines to buy the Concorde, are still struggling to return their fleets to service after grounding them **last year** for safety improvements following an Air France Concorde crash near Paris that killed **113** people.

Relevant news article 2:
Title: French Report on Concorde Crash Blames Debris and Structural Flaw
Published Time: 2004/12/15
Content:
A metal strip that fell off a Continental Airlines plane was a major element in the crash of an Air France Concorde jet near Paris **in July 2000** that killed **113** people, ...
...
The Concorde crashed into a hotel soon after it took off from Charles de Gaulle airport **on July 25, 2000**, when one of its tires exploded after hitting the titanium strip that had fallen from a Continental DC-10 that had taken off minutes before.

Relevant news article 3:
Title: World Briefing | Europe: France: Ex-Concorde Head In Crash Inquiry
Published Time: 2005/09/28
Content:
Henri Perrier, the former director of the French Concorde program, was questioned for more than 11 hours by a judge in the crash of an Air France Concorde just after takeoff from Paris **in 2000** that killed **113** people, and he was placed under formal investigation -- a step short of formal charges. ...

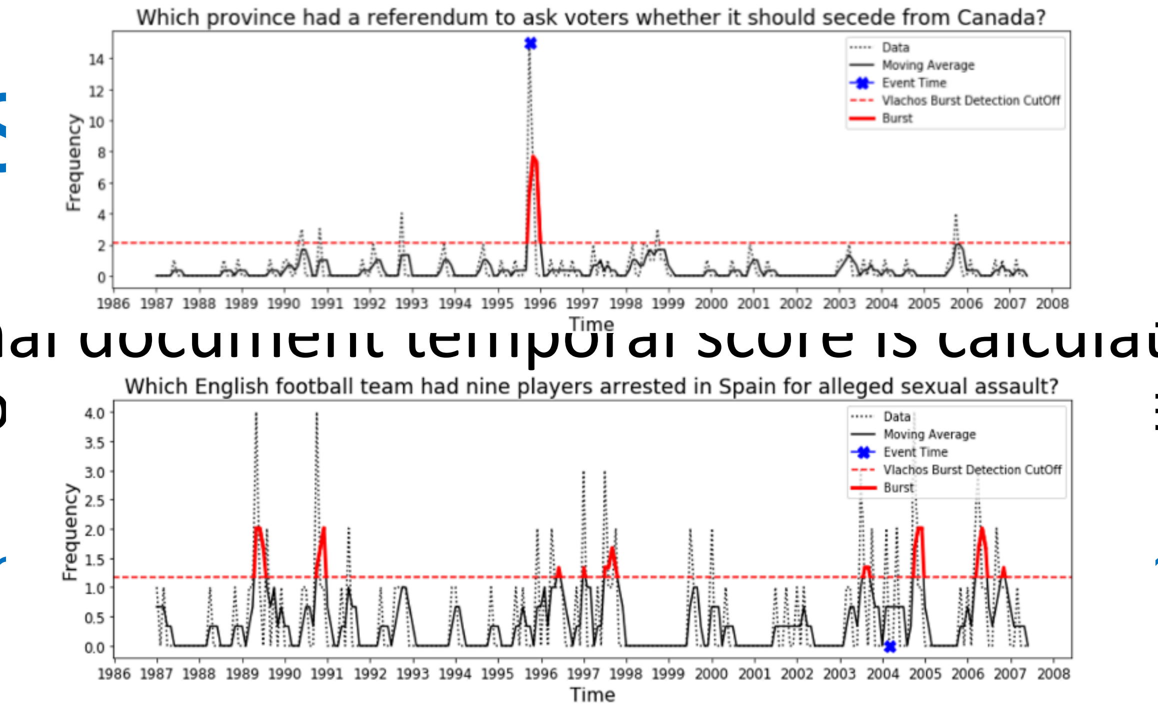
Step 2: Document Temporal Scores Computation





- Use Kernel Density Estimation to reflect the overlap of document time expressions and question time scope
- Question time scope contains multiple periods → aggregate the scores for all periods

S ng



- The fina ted based on two ents
- Finally ar dynamic combir mporal scores

combining timestamp and content temporal scores of documents

$$S^{temp}(d) = \frac{1}{2} \cdot (S_{pub}^{temp'}(d) + S_{text}^{temp'}(d))$$

dynamic linear combination of relevance and temporal score for documents

$$S(d) = (1 - \alpha(Q)) \cdot S^{rel}(d) + \alpha(Q) \cdot S^{temp}(d)$$

$$\alpha(Q) = \begin{cases} 0.0 & when\ burst\_num = 0 \\ ce^{-(1-\frac{1}{burst\_num})} & elsewhere \end{cases}$$

alpha depends on the number of
bursts associated with the question

Step 4: Compute Answers and Select Final Answer

- Take N top-ranked documents from Step 3 and compute answers using DrQA method
- Choose the most common answer as the final answer

Datasets

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


"All the News That's Fit to Print"

The New York Times

Late Edition

VOL. CLXII . . . No. 56,220

NEW YORK, TUESDAY, AUGUST 6, 2013

$2.50

Chicago Sees Pension Crisis Drawing Near

Politics and Downturn Put a City at Risk

ORDER BY QAEDA FOR ATTACK LED TO TERROR ALERT

MESSAGES INTERCEPTED

Bin Laden's Successor Is Said to Have Sent an Edict to Yemen

Amazon's Founder to Buy The Washington Post

For $250 Million — Deal Stuns an Industry

- **Testset:** 500 explicitly and 500 implicitly time-scoped questions

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

https://www.quizwise.com/history-quiz

Results for Explicit Temporal Questions

Model	Top 1		Top 5		Top 10		Top 15	
	EM	F1	EM	F1	EM	F1	EM	F1
DrQA-NYT Chen et al. (2017)	13.20	17.60	18.00	23.73	21.20	26.51	21.00	26.85
QA-No-Re-ranking Seo et al. (2016)	13.60	19.86	18.20	24.97	23.80	31.92	26.20	34.45
QANA-TempPub	17.20	23.31	23.60	30.81	27.20	36.60	30.20	38.91
QANA-TempCont	16.80	23.30	24.00	31.68	27.60	36.19	29.60	38.51
QANA	**18.60**	**25.32**	**24.40**	**32.09**	**30.02**	**39.01**	**31.20**	**40.50**

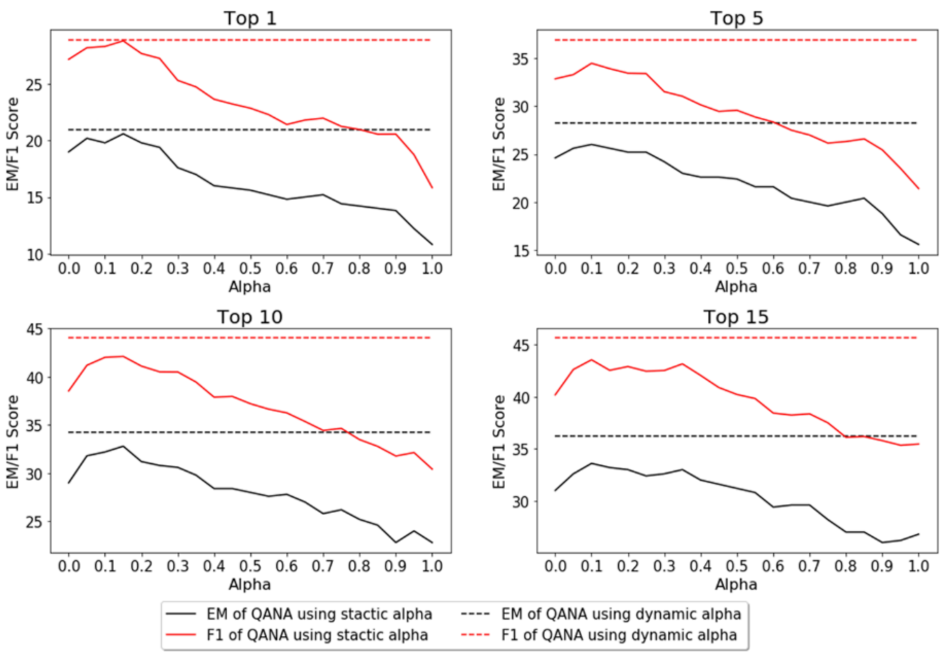
Results for Implicit Temporal Questions

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 (Kanhabua 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**

Additional Experiments

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**

	Top 1		Top 5		Top 10		Top 15	
	EM	F1	EM	F1	EM	F1	EM	F1
Questions with few bursts								
QA-No-Re-ranking (Seo et al. 2016)	20.94	29.81	28.63	37.41	35.89	46.30	39.74	49.49
QANA	22.64	31.54	30.76	40.63	38.03	49.08	41.02	52.17
Questions with many bursts								
QA-No-Re-ranking (Seo et al. 2016)	17.29	24.88	21.05	28.77	22.93	30.90	23.30	31.21
QANA	19.54	26.59	25.93	33.54	30.82	39.56	31.95	39.87



TEMPORAL ANALOG DETECTION & EXPLANATION

Y. Zhang, A. Jatowt, S. Bhowmick and K. Tanaka. Omnia Mutantur, Nihil Interit: *Connecting Past with Present by Finding Corresponding Terms across Time*, ACL 2015, 645-655
Y. Zhang, A. Jatowt, S. Bhowmick and K. Tanaka: *The Past is Not a Foreign Country: Detecting Semantically Similar Terms across Time*, IEEE TKDE, 2793-2807 (2016)
Y. Zhang, A. Jatowt, and K. Tanaka : *Towards Understanding Word Embeddings: Automatically Explaining Similarity of Terms*, IEEE BigData 2016, 823-832 (2016)

Terminology Gap: Background

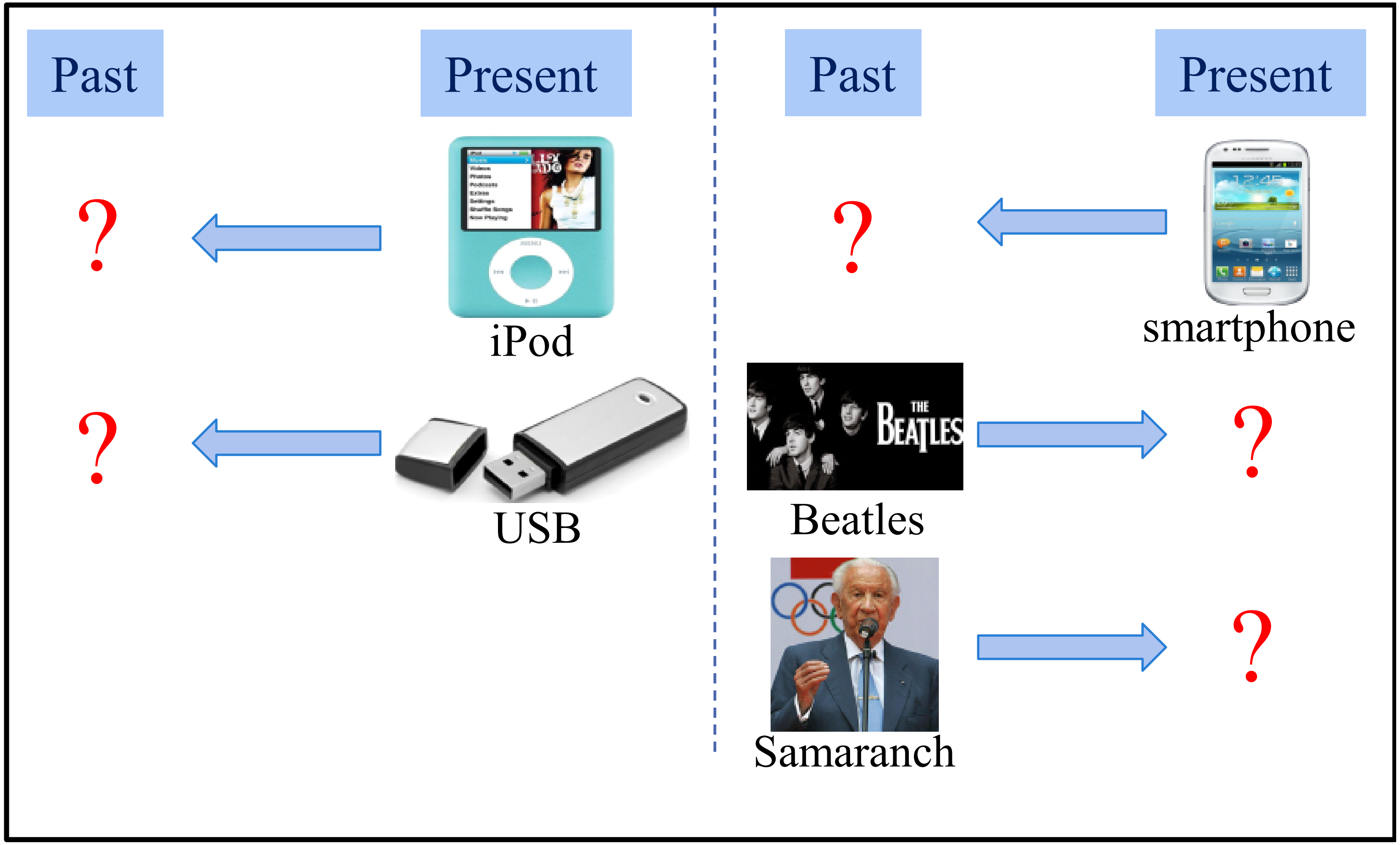
– Many problems for enabling effective search within archives
– We focus on *terminology gap*:
  - Often non-expert users have problems to construct correct queries

E.g., query **“phonograph”** may be unknown

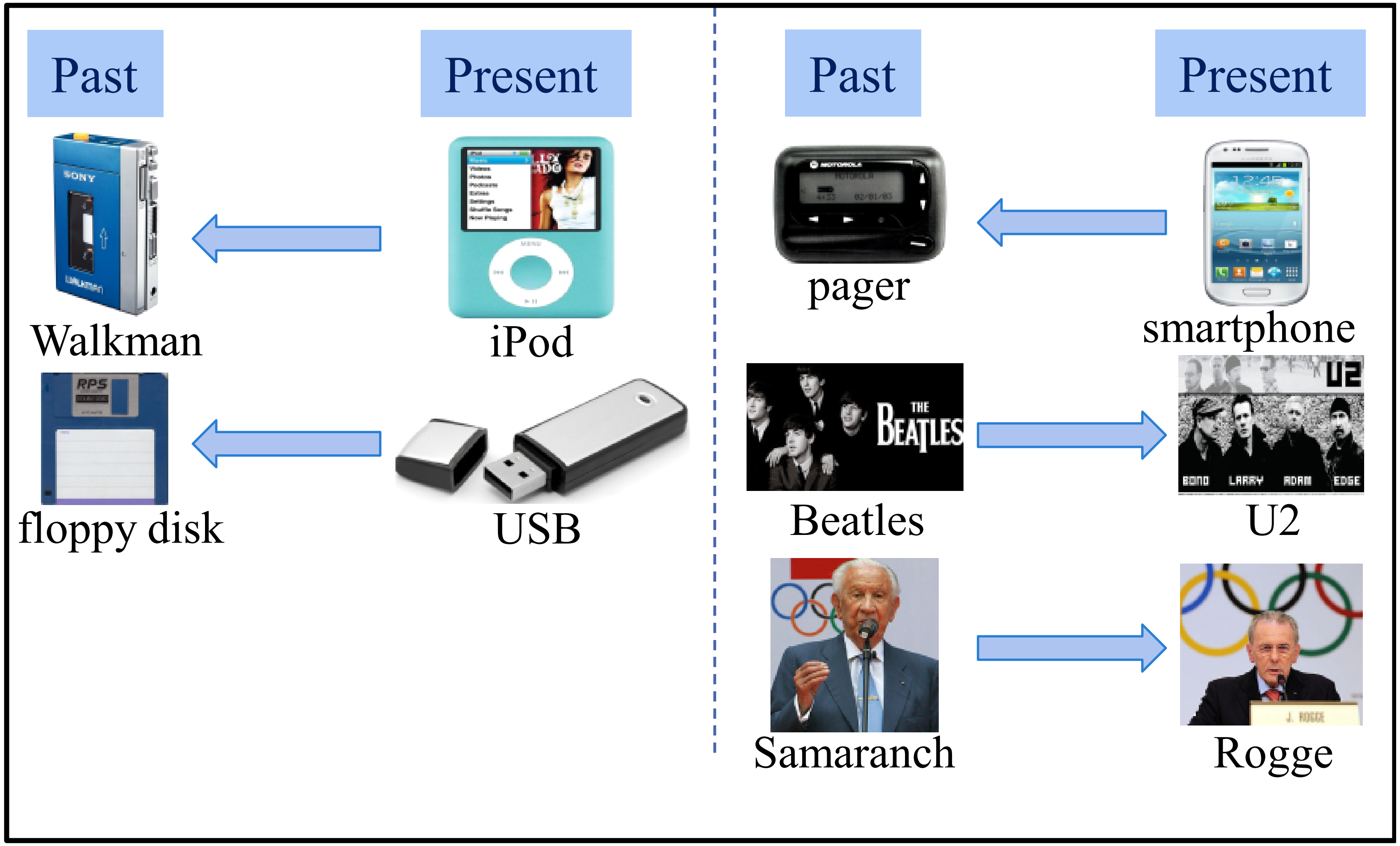
Search intent: Find content on devices people used to listen to music 100 years ago?

Useful not only for search assistance but for historical document understanding, education, etc.

Example Temporal Analogs



Example Temporal Analogs



Types of Temporal Analogs

- Temporal Analogs: entities which are semantically similar, yet which existed in different time periods.

1. **Same entity with different name**

   e.g. Myanmar (after 1989), Burma (before 1989)

2. **Different entities**

   e.g. iPod (2000s), Walkman (1980s)

Panta Rei [Eng: Everything Changes]

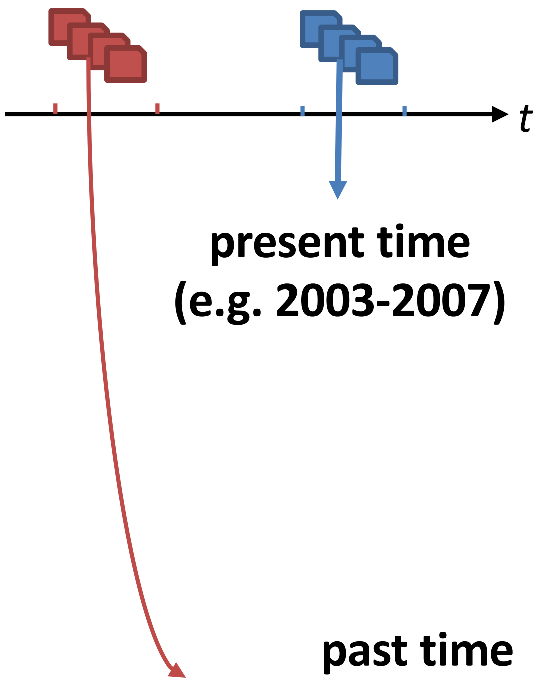
- **Everything changes**: thus contexts surrounding *temporal analogs* are different

Walkman (1980s)	iPod (2010s)
cassette	apple
audio	mp3
video	roqit
tape	player
music	music
sony	geeks
digital	jukebox
stereo	portable
earphone	macintosh
recorder	dlink

** Contexts in the New York Times corpus*

The task is not trivial…

Across-time Similarity: NN-based Term Embedding



present time
(e.g. 2003-2007)

past time
(e.g. 1987-1991)

Distributed Vector Representations (NN) [Mikolov 2013]

$$
\begin{array}{c}
 \\ w_1 \\ w_2 \\ \vdots \\ w_P
\end{array}
\begin{array}{cccc}
D_1 & D_2 & \cdots & D_m \\
\cdots & \cdots & \cdots & \cdots
\end{array}
\begin{bmatrix}
\cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots
\end{bmatrix}
$$

?

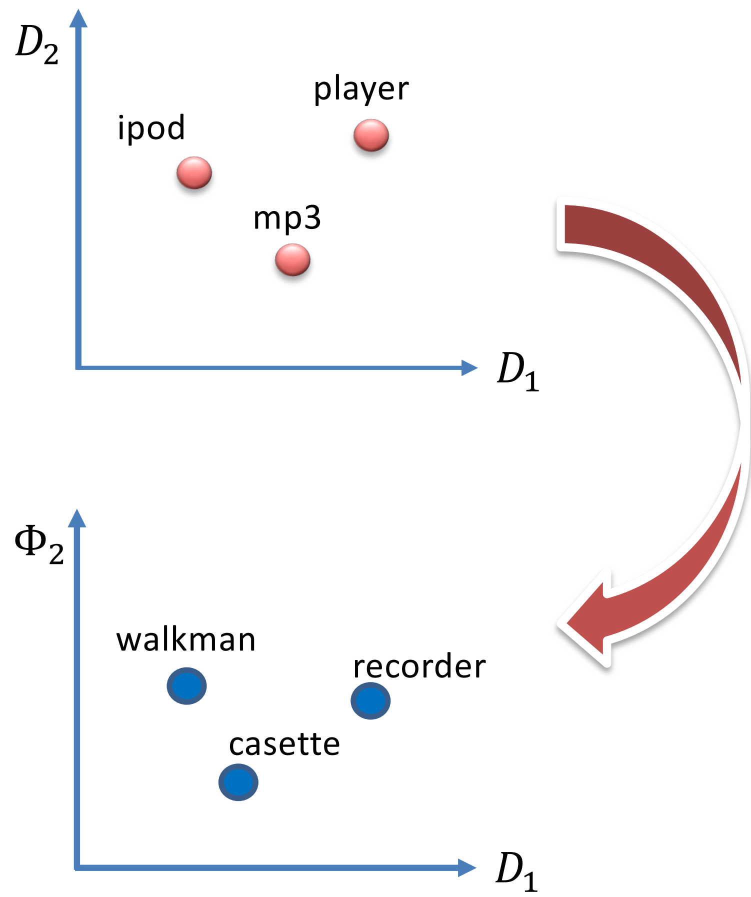
$$
\begin{array}{c}
 \\ \omega_1 \\ \omega_2 \\ \vdots \\ \omega_Q
\end{array}
\begin{array}{cccc}
\Phi_1 & \Phi_2 & \cdots & \Phi_n \\
\cdots & \cdots & \cdots & \cdots
\end{array}
\begin{bmatrix}
\cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots
\end{bmatrix}
$$

D_i and Φ_k are the dimensions of each vector space

T. Mikolov, K. Chen, G. Corrado and J. Dean. *Efficient Estimation of Word Representations in Vector Space*. ICLR Workshop, 2013

Assumption behind Proposed Approach

present time
(e.g. 2003-2007)



past time
(e.g. 1987-1991)

Assumption
The <u>relative</u> positions of terms in each vector space remain stable

Constructing Transformation Matrix

Distributed Vector Representations

K Pairs of corresponding terms (anchors)
$\{(w_i, \omega_i), \ldots, (w_j, \omega_j)\}$

$$\mathbf{M} = \operatorname*{argmin}_{\mathbf{M}} \sum_{i=1}^{u} \left\| \mathbf{M}\mathbf{x}_i^{\mathrm{b}} - \mathbf{x}_i^{\mathrm{t}} \right\|_2^2 + \gamma \left\| \mathbf{M} \right\|_2^2$$

Present time
(e.g. 2003-2007)

$$\begin{array}{c|c} & \begin{matrix} D_1 & D_2 & \cdots & D_m \end{matrix} \\ \begin{matrix} w_1 \\ w_2 \\ \vdots \\ w_P \end{matrix} & \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \end{array}$$

$$M = \begin{array}{c|c} & \begin{matrix} \Phi_1 & \Phi_2 & \cdots & \Phi_n \end{matrix} \\ \begin{matrix} D_1 \\ D_2 \\ \vdots \\ D_m \end{matrix} & \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \end{array}$$

Past time
(e.g. 1987-1991)

$$\begin{array}{c|c} & \begin{matrix} \Phi_1 & \Phi_2 & \cdots & \Phi_n \end{matrix} \\ \begin{matrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_Q \end{matrix} & \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \end{array}$$

We choose ***common, frequent terms***:
the more frequently the word used,
the harder is to change its meaning [Pargel 2007]
e.g. "man", "woman", "water"

M. Pargel et al. *Frequency of word-use predicts rates of lexical evolution throughout Indo-European history*. Nature, 449, 717-720, 2007

Global Term Transformation Approach

"**iPod**" in Present Time

Transformation Matrix

Expected Vector

$$\underbrace{[0.3 \quad \cdots \quad -1.1]}_{\text{Dimensions=300}} \times \boldsymbol{M} = \underbrace{[0.1 \quad \cdots \quad 0.7]}_{\text{Dimensions=200}}$$

$$\begin{bmatrix} 1.1 & -0.3 & \dots & 1.1 \\ 1.3 & -1.1 & \dots & 2.1 \\ 0.4 & 0.8 & \dots & -1.3 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

size=300x200

Past Time

$$\begin{array}{c} \omega_1 \\ \omega_2 \\ \cdots \\ \omega_Q \\ \\ \\ \end{array} \begin{bmatrix} -1.1 & 0.3 & \cdots & 1.2 \\ 0.3 & -1.1 & \cdots & -2 \\ 1.9 & 3.4 & \cdots & 2.3 \\ -0.3 & 0.4 & \cdots & 1.7 \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

cosine similarity

<u>Results:</u>
1. vcr
2. macintosh
3. *walkman*

...

Result: ranked list of temporal analogs

Problems with Global Term Transformation

Not the best answers..

VCR was found to be a counterpart of iPod due to allowing to record/playback
Macintosh was found to be a counterpart of iPod as being produced by Apple

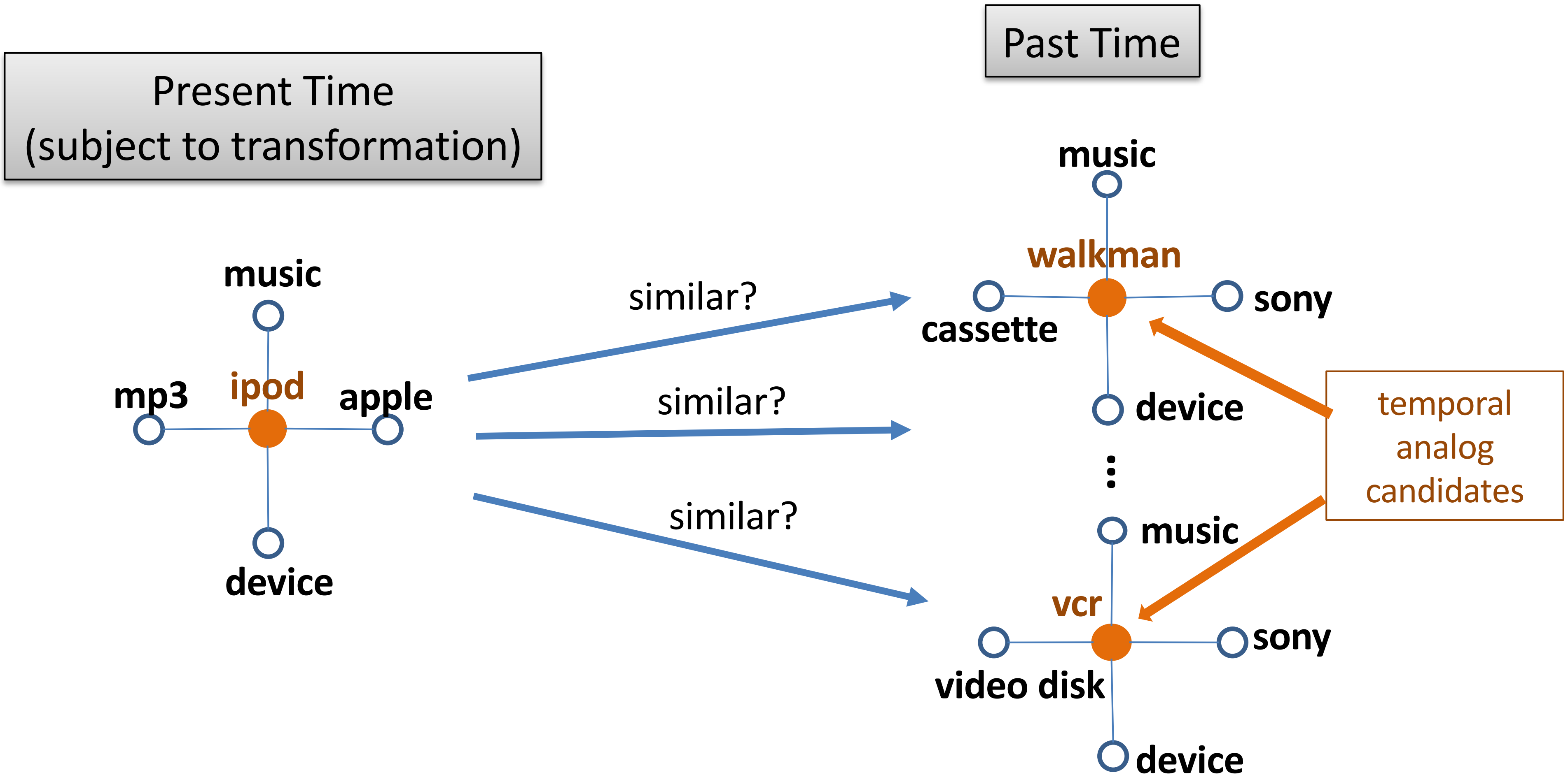
Transformation Matrix → **Global Correspondence**

Relations between query and its local context are neglected

↓

Local Correspondence

Transformation Using Local Graph by Using Reference Points



Desired Characteristics of Reference Points

- Reference Points - terms in query’s context which help to build effective across-time connection
- Desired criteria:
  a) have **high relation** with the query
  b) be sufficiently **general**
  c) **independent** from each other

Reference Point Detection

- Three methods for finding reference points using:
  1. Term co-occurrence **(LT-Cooc)**
     - Uses terms with high frequency and high relatedness as captured by Chi-square test

     e.g. ***iPod****: music, Apple, computer, digital, iTunes*
  2. Lexico-Syntactic Patterns **(LT-Lex)**
     - Uses term hypernyms [Ohshima, 2010]

     e.g. ***iPod****: music, music device, music player*
  3. Semantic Clustering **(LT-Cluster)**
     - Bisecting k-means is first used to obtain clusters of words with similar meanings
     - Chooses typical term from each semantic cluster

     e.g. ***iPod****: music, digital, iTunes, company, store*

H. Ohshima, K. Tanaka. *High-speed Detection of Ontological Knowledge and Bi-directional Lexico-Syntactic Patterns from the Web*. Journal of Software, 2010, 5(2): 195-205.

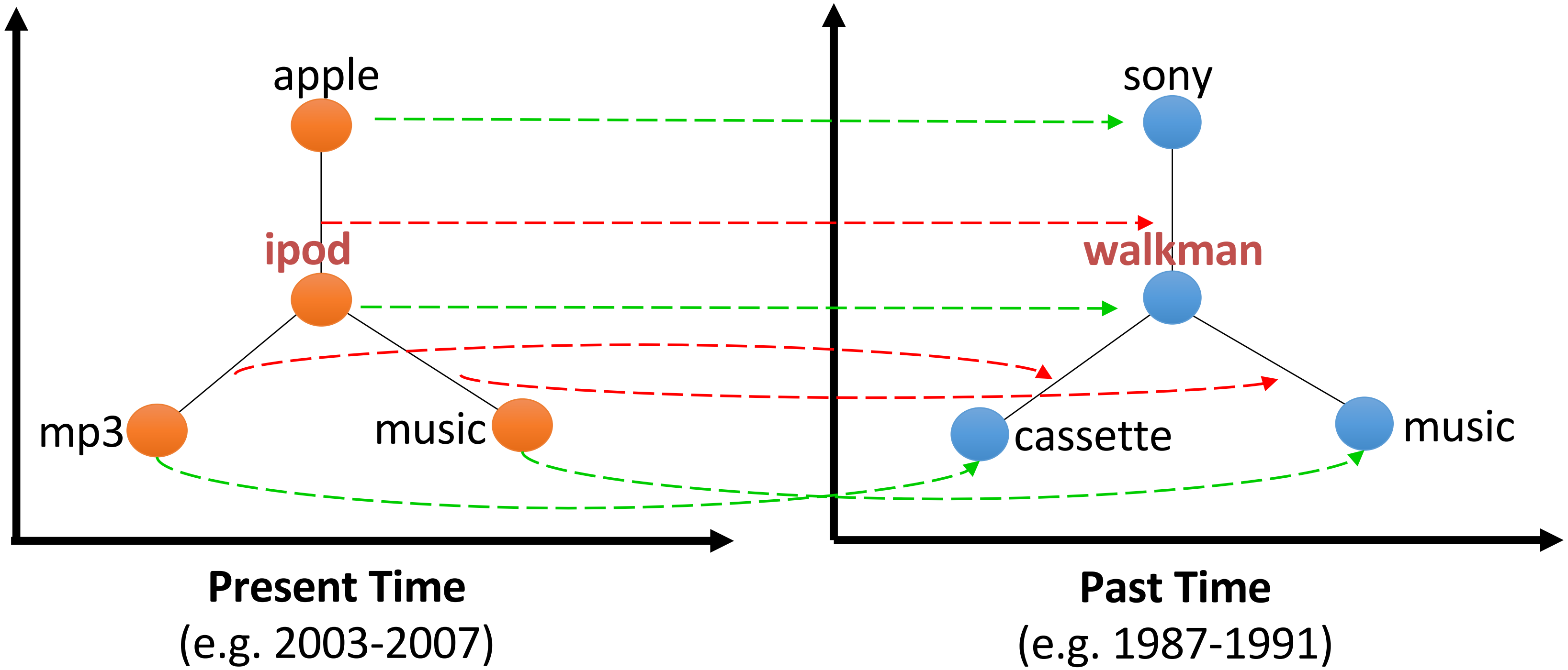
Local Graph Similarity Measurement

Our approach:

Measure by concept similarity and relational similarity

absolute position

difference between vector representations



Experiments: Dataset and Settings

- **Dataset**: New York Times Annotated Corpus (1987-2007)
  – 1.8 million articles in total, **0.45 million articles** in the present and past time period, on average. Vocabulary size: **300K**
- **Test sets** (persons, locations, objects):
  – **95 pairs** of <query, temporal counterpart> for ***[2002-2007]*** *to* ***[1987-1991]***

- **Training *Transformation Matrix***
  – *Feature dimension for Skip-gram model: 200*
  – *Number of Common Frequent Terms (CFTs): top frequent common words (5%)*


"All the News That's Fit to Print"
The New York Times
Late Edition
VOL. CLXII . . . No. 56,220
NEW YORK, TUESDAY, AUGUST 6, 2013
$2.50

Chicago Sees Pension Crisis Drawing Near

Politics and Downturn Put a City at Risk

ORDER BY QAEDA FOR ATTACK LED TO TERROR ALERT

MESSAGES INTERCEPTED

Bin Laden's Successor Is Said to Have Sent an Edict to Yemen

Amazon's Founder to Buy The Washington Post

For $250 Million — Deal Stuns an Industry

Experiments: Test Set

- Manually created a test set with 52 queries and 95 pairs of (query, temporal analog)

ID	q [2002,2007]	t [1987,1991]
1	Putin	Yeltsin
2	Chirac	Mitterrand
3	iPod	Walkman
4	Facebook	Usenet
5	Linux	Unix
6	spam	junk mail, autodialers, junk fax
7	spreadsheet	database, word processor
8	email	messages, letters, mail, fax
9	superman	superman, batman
10	Pixar	Tristar, Disney
11	Euro	Mark, Lira, Franc
12	Myanmar	Burma
13	Koizumi	Kaifu
14	Rogge	Samaranch
15	Serbia, Croatia, Macedonia, Montenegro, Kosovo, Slovenia, Bosnia	Yugoslavia
16	fridge	fridge, freezer, refrigerator, ice_cubes
17	NATO	NATO
18	Google	IBM, Microsoft, Matsushita, Panasonic
19	Boeing	Boeing, Airbus, Mcdonnell Douglas
20	Flash drive, USB, CDROM, DVD	floppy disc
..	…	…

Table 1. Examples of test sets where term *q* is input and term *t* is the expected temporal analog (*t* can be multiple)

Type of queries:

1. Persons
2. Locations
3. Objects

Experiments: [2002,2007] and [1987,1991] on NYT News Corpus

1. Searching **from present to past** (95 query-answer pairs)

	Method	MRR	P@1	P@5	P@10	P@20
baselines	BOW	4.1e-5	0	0	0	0
	LSI+Com	0.206	15.8	27.3	29.5	38.6
	LSI+Tran	0.112	7.9	13.6	21.6	22.7
	HMM	0.161	13.2	20.9	20.9	24.2
methods	Global_Tran	0.298	16.8	44.2	56.8	**73.7**
	Local_Tran (Cooc)	0.283	18.8	35.3	50.6	62.4
	Local_Tran (Cluster)	0.285	14.7	42.1	55.1	65.2
	Local_Tran (Lex)	**0.369**	**24.2**	**49.5**	**63.2**	71.6

2. Searching **from past to present** (95 query-answer pairs)

	Method	MRR	P@1	P@5	P@10	P@20
baselines	BOW	3.4e-5	0	0	0	0
	LSI+Com	0.181	13.2	19.7	28.9	35.5
	LSI+Tran	0.109	5.3	17.1	21.1	23.7
methods	GT	0.226	15.2	27.3	33.3	45.5
	GT+LT (Cooc)	0.231	14.7	**30.7**	**36**	46.7
	GT+LT (Cluster)	0.228	13.6	28.8	31.8	47
	GT+LT (Lex)	**0.235**	**16.7**	28.8	31.8	**48.5**

Example Results: Finding Past Analogs for Present Queries

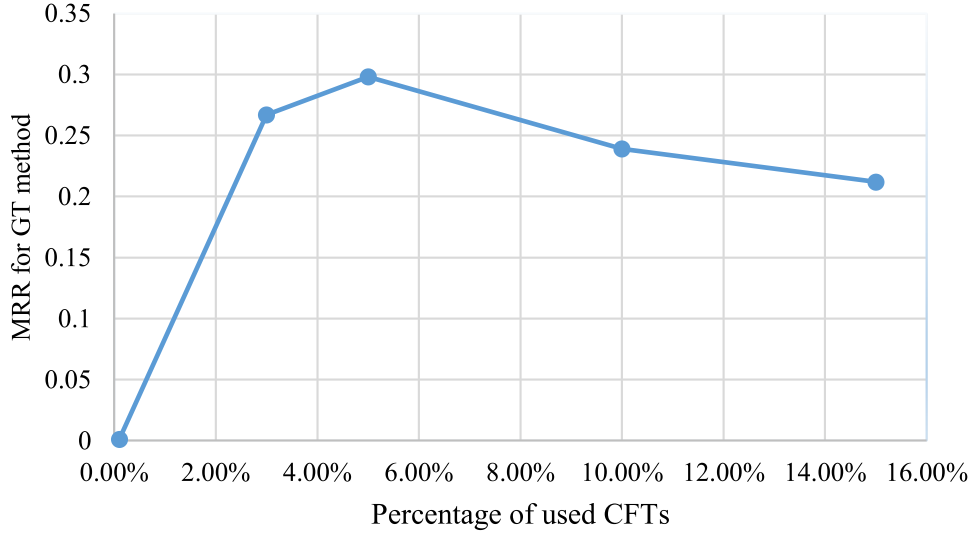
**Lexico-Syntactic Pattern* used to detect reference points

	queries	correct answers	baselines		methods	
	[2002,2007]	[1987,1991]	BOW (baseline)	LSI+Com (baseline)	Global_Tran	Local_Tran (Lex)
1	Putin	Yeltsin	1000+	51	24	2
2	Chirac	Mitterrand	1000+	6	7	2
3	iPod	Walkman	1000+	6	3	1
4	Facebook	Usenet	1000+	1000+	1	1
5	Linux	Unix	1000+	5	20	1
6	spam	junk mail	1000+	1000+	5	1
7	spreadsheet	database	1000+	395	3	1
9	email	messages	1000+	1	2	7
10	email	letters	1000+	1000+	1	1
11	email	mail	1000+	119	7	6
12	email	fax	1000+	1000+	3	4
14	superman	batman	1000+	46	5	2
15	Pixar	Tristar	1000+	110	1	1
16	Pixar	Disney	1000+	1	3	2
17	Euro	Mark	1000+	1000+	2	1
19	Euro	Franc	1000+	1000+	7	3
20	Myanmar	Burma	1000+	3	64	46
21	Koizumi	Kaifu	1000+	66	2	1
22	NATO	NATO	1000+	1	304	141
24	fridge	freezer	1000+	7	1	1
25	fridge	refrigerator	1000+	4	2	2
27	Serbia	Yugoslavia	1000+	12	1	1
28	Kosovo	Yugoslavia	1000+	27	14	10
30	mp3	compact disk	1000+	44	58	19
…	…	…	…	…	…	…

Rank of correct answers

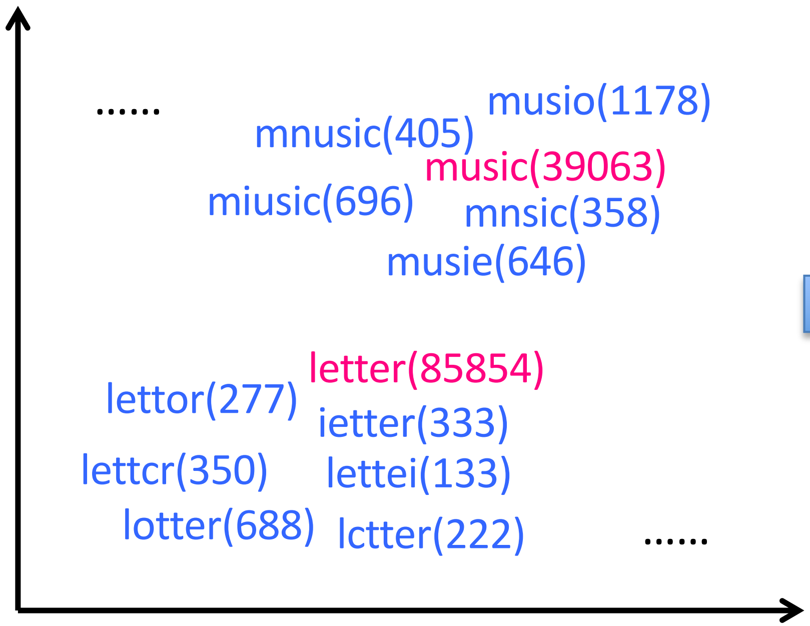
Evaluation: Effect of the number of Common Frequent Terms (CFT)

- 0.1%, 3%, **5%**, 10%, 15%



Solution to Alleviate OCR Errors

- OCR problem (Optical Character Problem)
  - Build **dictionary** to map wrong spellings to correct ones
    - **Input:** vector representation of all the words
    - **Output:** dictionary {wrong spelling: correct spelling}

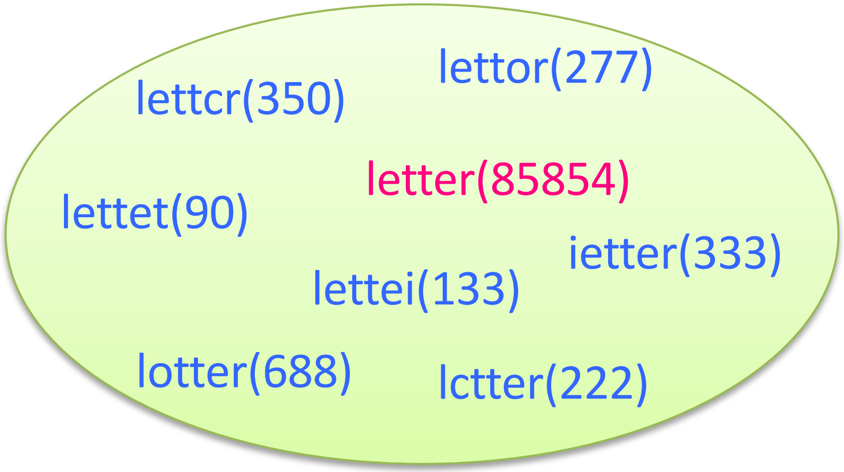


Vector Space of [1906, 1915]

Original Spelling	Correct Form
mnusic	music
miusic	music
musie	music
……	……
lettcr	letter
lettor	letter
lotter	letter
……	……

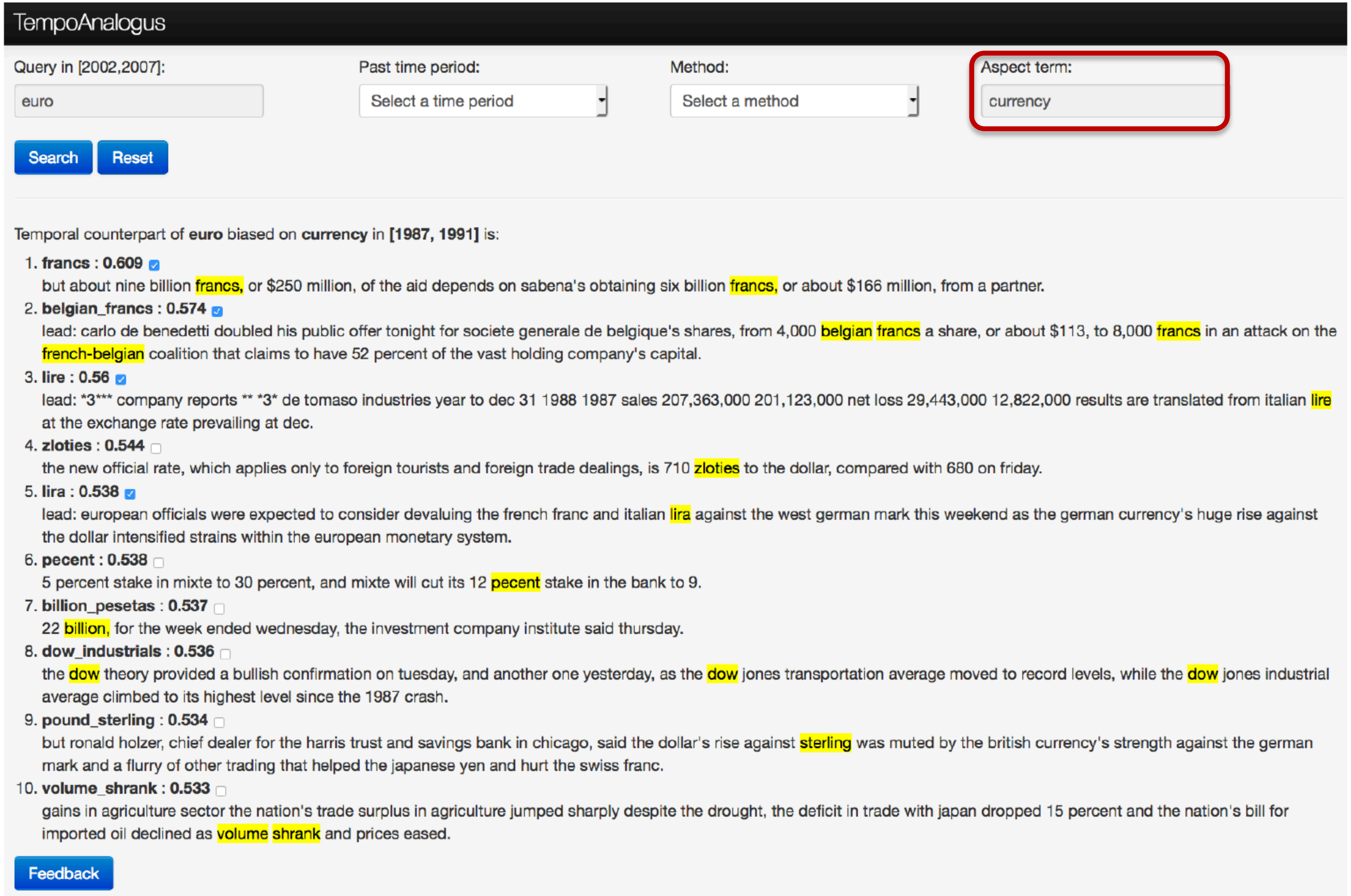
Solution to Alleviate OCR Errors

- Assumptions for Alleviating OCR Problem:
  (1) Wrongly spelled term has similar context with its correctly spelled term;
  (2) The correct term is more dominant (or frequent) compared to its wrongly spelled ones;
  (3) Wrongly spelled term has one edit-distance from its correct term.



- **Example Results**
  - Without Error Correction:
    - car [2004,2009] → [1906,1015] vehicle, tricycle, mnotor, rmotor, car, eycles
  - With Error Correction:
    - car [2004,2009] → [1906,1915] vehicle, tricycle, motor, car, cycles

Aspect-based Retrieval +Demo



Y. Zhang, A. Jatowt, S. Bhowmick and Y. Matsumoto: *ATAR: Aspect-based Temporal Analog Retrieval System for Document Archives*, WSDM 2019

From Detection to Explanation

- What is an analog of ***q*** in past?
  - e.g., *What* is counterpart of ***iPod*** in 1980s?

- Why ***t*** is an analog of ***q*** in past?
  - e.g., *Why* is ***iPod*** similar to ***Walkman*** in 1980s?

Y. Zhang, A. Jatowt, and K. Tanaka : *Towards Understanding Word Embeddings: Automatically Explaining Similarity of Terms*, IEEE BigData 2016, 823-832 (2016)

Across-time Similarity Explanation: Problem Statement

Input:

iPod : ? ≒ **Walkman : ?**

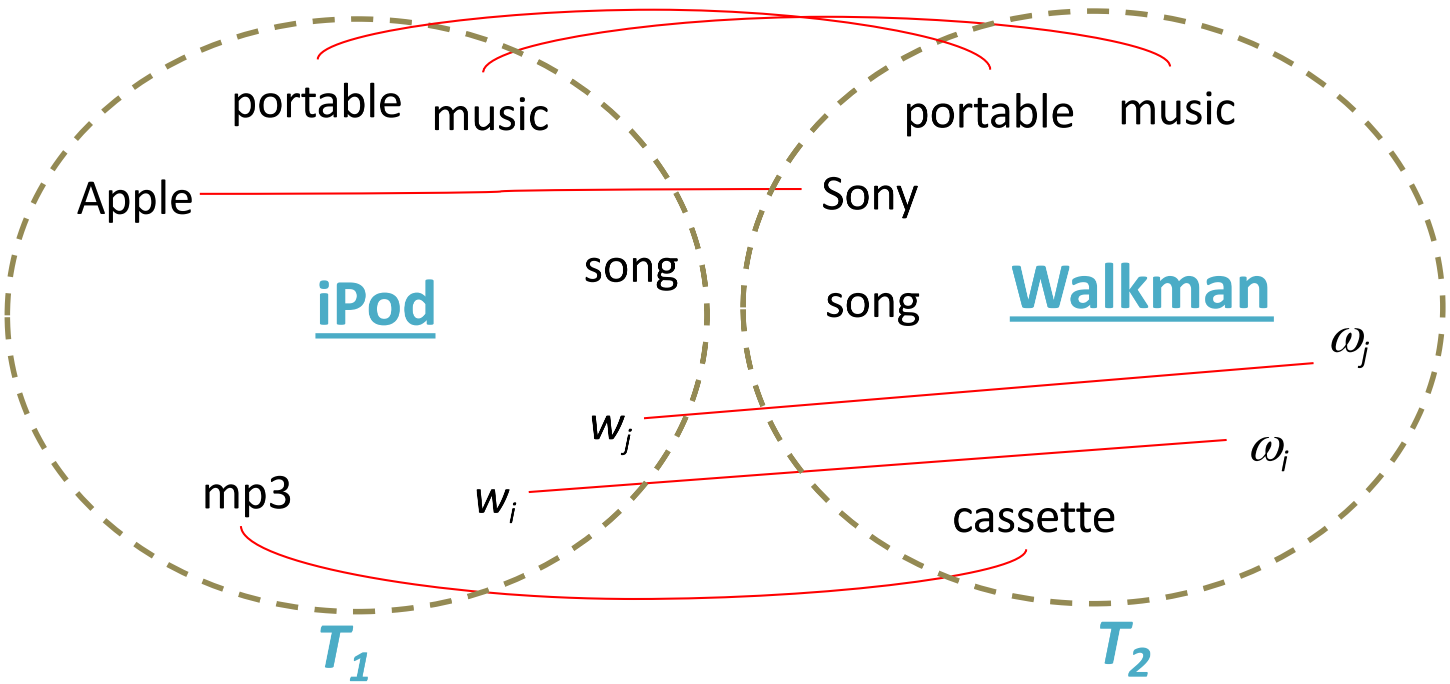
Based on several criteria

Output:

iPod : music	≒	Walkman : music	usage
iPod : portable	≒	Walkman : portable	characteristic
iPod : MP3	≒	Walkman : cassette	storage media
iPod : Apple	≒	Walkman : Sony	company

Providing evidence to support understanding of similarity between two entities across time

Conceptual View of Problem



Output

<music, music>

<portable, portable>

<Apple, Sony>

<mp3, cassette>

…

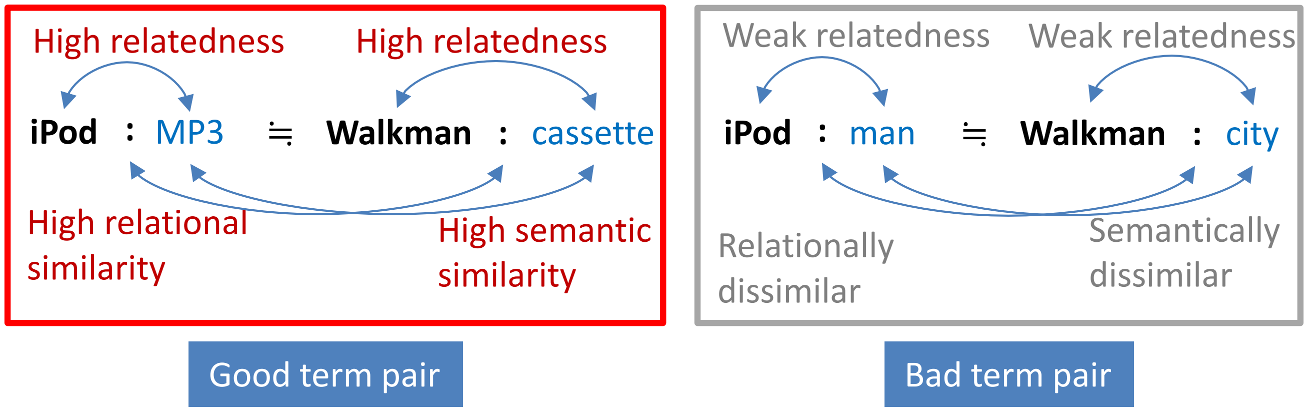
<w_i, ω_i>

Context terms of a given entity are derived from frequently co-occurring terms

Task: find good word pairs denoting commonalities or aligned differences

Explaining Across-time Similarity

1. **Relatedness**
   – Terms in a pair should be related to their entities
2. **Semantic similarity**
   – Terms should be similar to each other
3. **Relational similarity**
   – Terms should have similar relation to their entities



Local Computing of Word Pair Quality

- Aggregating relevance, semantic similarity, relational similarity

$$quality(<w_u,\omega_v>) = \underbrace{rel(<w_u,\omega_v>)}_{\text{Relevance of each word pair to the entities}}{}^{\alpha} \cdot (\underbrace{SimIntraPair(<w_u,\omega_v>)}_{\text{Semantic similarity between word pair in each word pair}} \cdot \underbrace{\mathrm{SimRela}Pair(<w_u,\omega_v>)}_{\text{Relational similarity between word pair and the entities}})^{1-\alpha}$$

+ **Global method** – a Random Walk on a graph with nodes being pairs of terms (details in [Zhang et al. 2016])

Y. Zhang, A. Jatowt, and K. Tanaka : *Towards Understanding Word Embeddings: Automatically Explaining Similarity of Terms*, IEEE BigData 2016, 823-832 (2016)

Results

	Methods	Precision	Recall	F_1-score
baselines	Overlap	0.63	0.48	0.55
	BOW	0.23	0.17	0.20
	Com	0.46	0.34	0.39
methods	Local	0.66	0.50	0.57
	Global	**0.72***†	**0.54***†	**0.61***†

[2002, 2007]: "Bustamante, a democrat, is the leading candidate to replace him if the recall succeeds, holding a narrow margin over his closest competitor, *Arnold Schwarzenegger*, a republican."

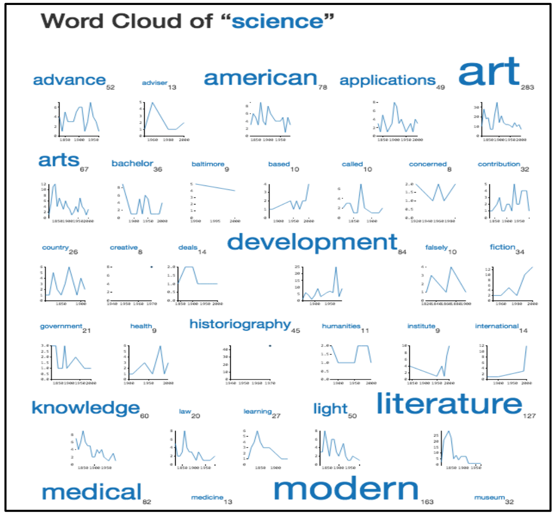
[1987, 1991]: "In theatrical-release films, the big roles, and the gigantic salaries, are dominated by fellows with names like Newman, Redford, Stallone, *Schwarzenegger* and Costner."

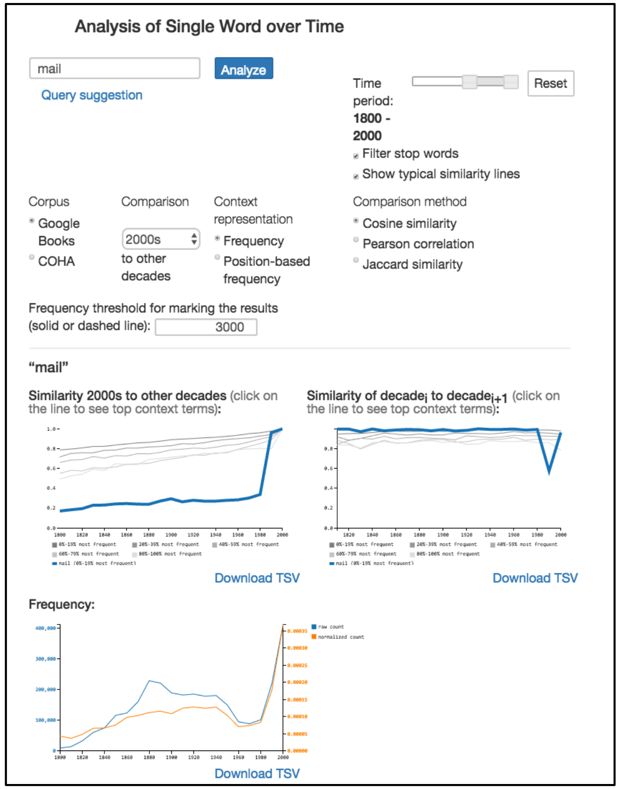
	baselines			methods	
Correct pairs	**Overlap**	**BOW**	**Com**	**Local**	**Global**
	iPod vs. *Walkman*				
Apple - Sony (company)		✓		✓	✓
MP3 - cassette (media)				✓	✓
portable - portable (characteristic)	✓			✓	✓
music - music (usage)	✓				✓
	Arnold Schwarzenegger vs. *Arnold Schwarzenegger*				
Bustamante - Stallone (competitor)				✓	✓
Californians - moviegoers (supporter)			✓	✓	✓
Hollywoord - Hollywood (industry)	✓			✓	✓
Terminator - Terminator (movie)	✓		✓	✓	✓
	Sepp Blatter vs. *Joao Havenlange*				
Klinsmann - Osim (coach)				✓	✓
Zidane - Vautrot (controversy)					✓
FIFA - FIFA (organization)	✓	✓	✓	✓	✓
soccer - soccer (field)	✓	✓	✓	✓	✓
	Germany vs. *East Germany*				
Schröder - Kohl (president)				✓	✓
Europe - Soviet (union)			✓		
Berlin - Berlin (capital)	✓		✓	✓	✓
Germans - Germans (citizen)	✓		✓	✓	✓

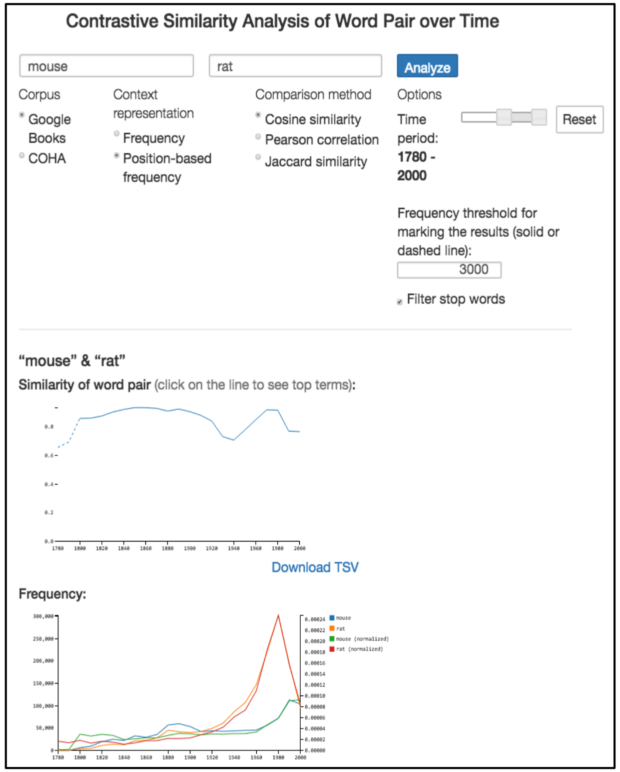
RELATED INTERACTIVE SYSTEMS

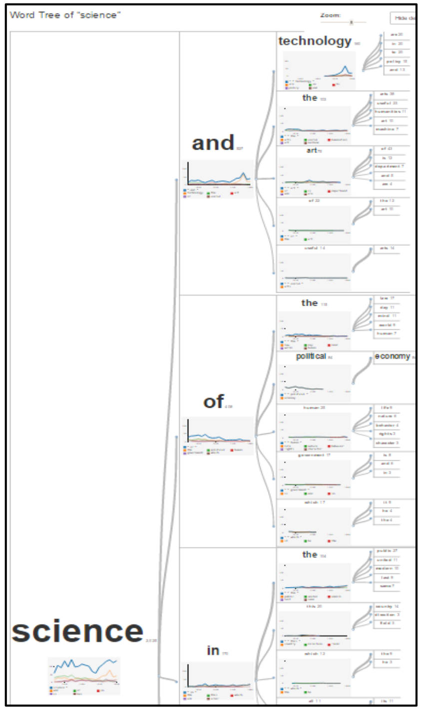
Word Semantic Evolution Analysis

http://tinyurl.com/WordEvolutionStudy









Adam Jatowt, Ricardo Campos: Interactive System for Reasoning about Document Age. CIKM 2017: pp., 2471-2474

Adam Jatowt et al.: Every Word has its History: Interactive Exploration and Visualization of Word Sense Evolution. CIKM 2018: 1899-1902

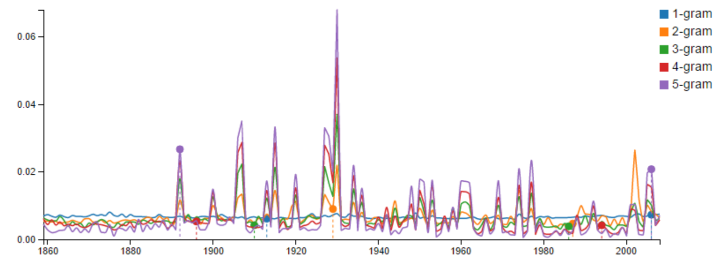
Framework for Analysing Archival Documents

Input Text:

To Sherlock Holmes she is always THE woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex. It was not that he felt any emotion akin to love for Irene Adler. All emotions, and that one particularly, were abhorrent to his cold, precise but admirably balanced mind. He was, I take it, the most perfect reasoning and observing machine that the world has seen, but as a lover he would have placed himself in a false position. He never spoke of the softer passions, save with a gibe and a sneer. They were admirable things for the observer--excellent for drawing the veil from men's motives and actions. But for the trained reasoner to admit such intrusions into his own delicate and finely adjusted temperament was to introduce a distracting factor which might throw a doubt upon all his mental results. Grit in a sensitive instrument, or a crack in one of his own high-power lenses, would not be more disturbing than a strong emotion in a nature such as his. And yet there was but one woman to him, and that woman was the late Irene Adler, of dubious and questionable memory.

Estimated document age

Result:



Evidence for age estimation

Which ngrams contributed the most to the spikes on the plot:

at 1930

#	ngram	contribution (frequency × weight ÷ sumOfWeights)	cumulative percentage	frequency	weight	count in text
1	any emotion akin	0.000595	3.37 %	0.111920	1.000000	1
2	and questionable memory	0.000536	6.42 %	0.100856	1.000000	1
3	and predominates the	0.000521	9.37 %	0.097908	1.000000	1
4	questionable memory .	0.000515	12.29 %	0.096764	1.000000	1
5	for Irene Adler	0.000511	15.18 %	0.095978	1.000000	1

at 1907

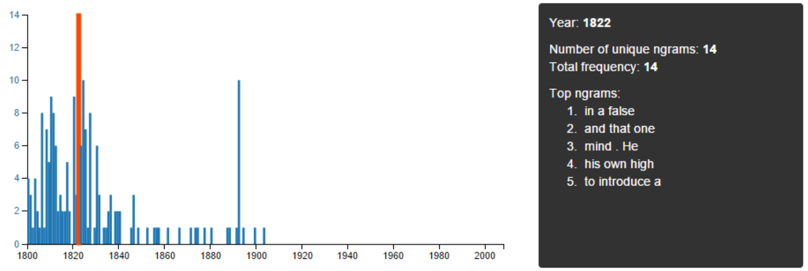
#	ngram	contribution (frequency × weight ÷ sumOfWeights)	cumulative percentage	frequency	weight	count in text
1	observing machine that	0.000543	4.65 %	0.102142	1.000000	1
2	perfect reasoning and	0.000509	9.01 %	0.095771	1.000000	1
3	and observing machine	0.000473	13.06 %	0.088899	1.000000	1
4	most perfect reasoning	0.000417	16.62 %	0.078332	1.000000	1
5	reasoning and observing	0.000251	18.77 %	0.047189	1.000000	1

at 1915

Dates of first appearance of text ngrams over time

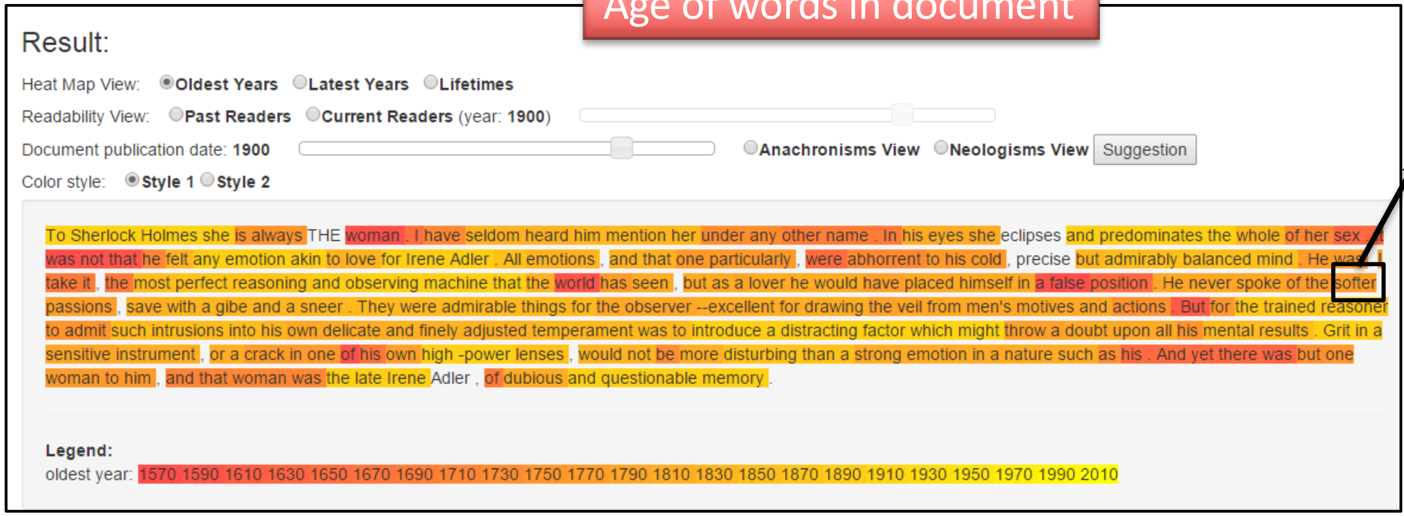
Cut-off Year View:

Oldest Years Latest Years

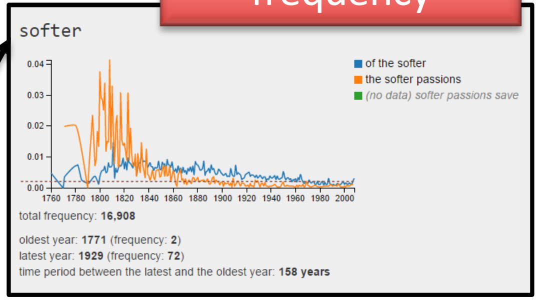


Framework for Analysing Archival Documents

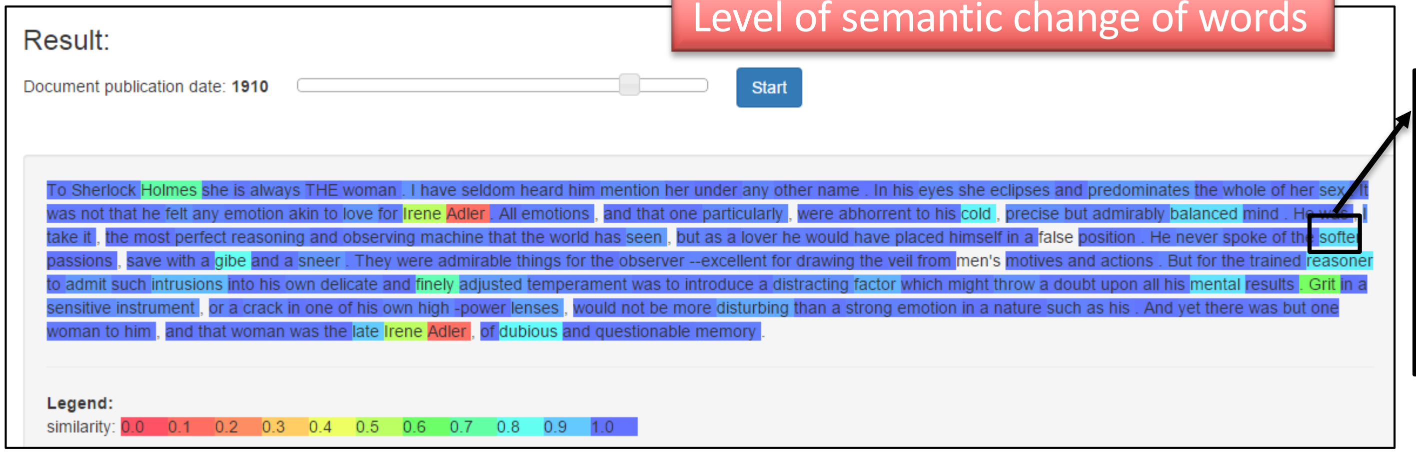
Age of words in document



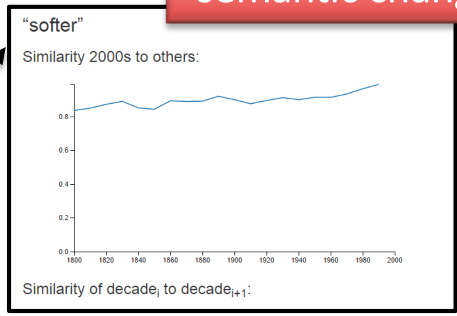
Word across-time frequency



Level of semantic change of words



Word across-time semantic change



Conclusions

Novel Ways of Information Access & Knowledge Extraction from Long-term News Archives

1. Open question answering in archival collections
2. Research task of *across-time analogy detection* & *explanation*
   – Approaches using vector space transformation: **global** and **local**
3. Examples of related interactive systems for archival documents and term evolution analysis

Thank you!