



文献アーカイブからのナレッジ抽出と情報アクセス Information Access and Knowledge Extraction from Document Archives

Adam Jatowt

NDL: 11 October, 2019

アジェンダ Talk Schedule

1. はじめに Introduction
2. 異なる時代における類似物(時間的アナログ)の検出
Temporal Analog Detection
 - ー 時間を超えた類似性の説明
Across-time Term Similarity Explanation
3. 時間を超えた比較の要約
Across-time Comparative Summarization
4. 歴史に基づく実体のグループ化と要約
History-based Entity Summarization
5. 面白さに基づくアーカイブからの情報検索
Interestingness-oriented Archival Retrieval
6. 現在との関連性を志向する文献検索に向けて
Towards Present-relevance Oriented Document Search
7. 結び Conclusions

「ビッグアーカイブデータ」 Big Archival Data

- 近年、過去の文献を含む膨大なアーカイブが利用可能に。

例：新聞、図書、学術出版物、行政文書、ウェブサイト、ソーシャルメディア、商品レビュー等のアーカイブ

- 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

ボーン
デジタル



多様なジャンルのアーカイブが身近に存在するようになった。

Archives are common and span variety of genres

アーカイブは絶えず発展し、その重要性を増している。

They are continuously growing and becoming increasingly important to us

デジタル文献アーカイブ Digital Document Archives

- 「ビッグアーカイブデータ」の例 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.
- ほぼ全ての国立図書館がデジタル化資料のアーカイブを持っている。
Nearly all national libraries and archives have own digital collections [1]
- **膨大な費用**: 国立国会図書館の2009・2010年のデジタル化費用は合計137億円。
Big Costs: e.g., in 2009 and 2010 the budget of the Japanese National Diet Library for digitization was 13.7 billion yen
- **僅かな利用**: 主に専門家や研究者によって活用されており、利用者の数が少ない。
Little usage: very few users utilize document archives, and mainly professionals

データ量・費用が膨大にも関わらず利用者の数は少ない。

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

より便利で使いやすいアーカイブにすることで利用を促進する必要がある。

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

歴史学習・理解の重要性

Importance of Studying & Understanding History

「過去から学べない者は、過ちを繰り返す。」(ジョージ・サンタヤーナ)

“Those who cannot remember the past are condemned to repeat it” (George Santayana)

- **歴史**は過去と現在を理解し、未来をある程度予測する上でも、人生において重要な役割を果たしている。

History plays important role in our lives, helps to understand the past, the present and even helps to predict the future to some extent

- Knowledge of history is essential for becoming prepared for an **active life in the contemporary society**

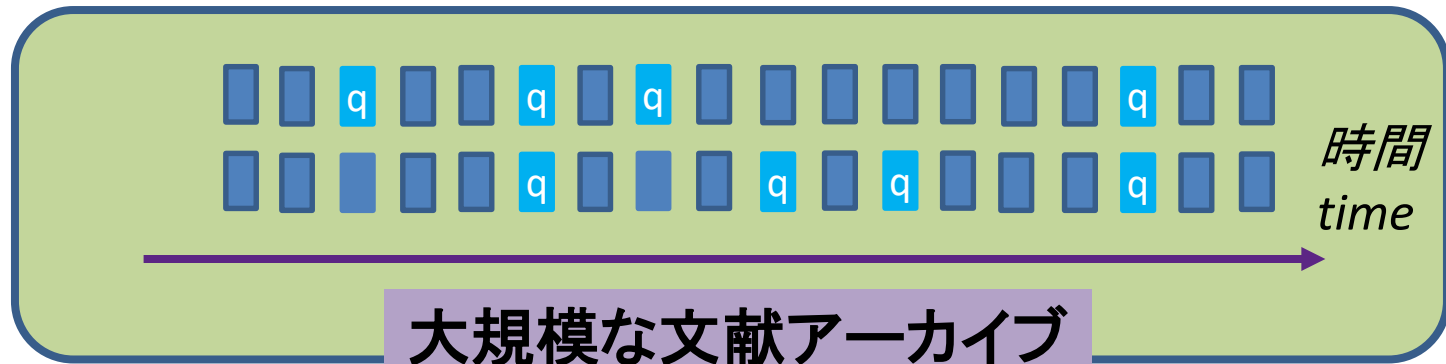
- **歴史への電算的アプローチ**: コンピューターの活用による歴史分析、記述、活用、学習等の支援

Computational approaches to history: Harnessing computational power to support history analysis, writing, usage, studying, etc.

- 「**デジタル人文学**」の潮流の一部 Part of larger trend of “**Digital Humanities**”
 - Related fields: **Computational Social Science** [1], **Web Science**, etc.

文献アーカイブインターフェースの現状

Current Interfaces to Document Archives



大規模な文献アーカイブ
Large Document Archive

検索結果
search results

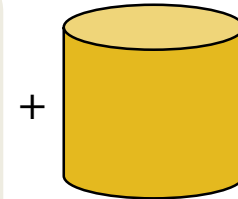
q (例: 1980年代の携帯音楽再生機器
e.g., portable music device in 1980s)

q q q q q q ... q

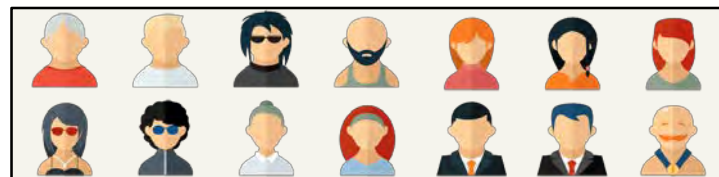


検索結果の理解が困難！
Difficult to make sense of results!

過去を知るための検索エンジン Towards Search Engine for the Past



+ ナレッジベース
Knowledge bases
(Wikipedia, Yago)



課題と未解決の問題

Challenges & Open Questions

- 課題 Challenges:

- データの量は膨大で、長期間に渡っている。
Data is large and distributed over time
- 語彙や文脈は時間を経て大幅に変化している。
Vocabulary & context in the past changed much
- 過去に関する利用者の知識は限られている。
Users' knowledge of the past is limited

どうすれば現代の利用者に対し過去の情報や知識を効果的に渡せるか？
How can we effectively return information or knowledge from the past for present users?

利用されていないデータを活用する方法にはどのようなものがあるか？
What are the possible ways to utilize this untapped source of data?

ニュースアーカイブはどうすれば一般的な利用者にとって
容易かつ面白く利用できるようになるか？

How news archives in particular can be made easy to use and interesting to ordinary users? 8

アジェンダ Talk Schedule

1. はじめに Introduction
2. 異なる時代における類似物(時間的アナログ)の検出
Temporal Analog Detection
 - 時間を超えた類似性の説明
Across-time Term Similarity Explanation
3. 時間を超えた比較の要約
Across-time Comparative Summarization
4. 歴史に基づく実体のグループ化と要約
History-based Entity Summarization
5. 面白さに基づくアーカイブからの情報検索
Interestingness-oriented Archival Retrieval
6. 現在との関連性を志向する文献検索に向けて
Towards Present-relevance Oriented Document Search
7. 結び Conclusions

異なる時代における類似物 (時間的アナログ)の検出 TEMPORAL ANALOG DETECTION

背景：語彙のジェネレーションギャップ

Background: Terminology Gap

- アーカイブ内検索を困難にする多くの要因のうち、**語彙の違い**に着目する。
 - 専門的な知識を持たない利用者は**適切な検索語**を作成するのに苦労することが多い。
- Many different difficulties in enabling search within archives - We focus on *terminology gap*:
 - Often non-expert users have problems to construct **correct queries**



例：「**蓄音機**」を知らない場合

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

検索の目的: 100年前の人々がどのような機器を用いて音楽を聴いていたかに関する資料を見つけたい

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



時間的アナログを探す意義

Motivation for Temporal Analog Detection

- より大きな目標 Larger goal:
 - デジタル歴史学の発展とアーカイブ利用の普及/促進
Fostering Digital History and Popularizing/Facilitating Archival Usage
- 具体的な課題 Concrete task:
 - 時間を超えて類似する実体を発見すること
Finding Analogical Entities across Time
- 適用方法 Applications:
 1. 検索語を提案することでアーカイブ内検索を支援する
Supporting search in archives by query suggestion
 2. 過去と現代における実体の類似性に関する新たな問いに答える
Answering new kind of questions on past-to-present similarity of entities
 3. 特定の事物に関する年表の自動生成
Automatic generation of object timelines
 4. 類似する過去の出来事を見つける
Finding similar events from the past

アナロジーと時間的/空間的アナロジー

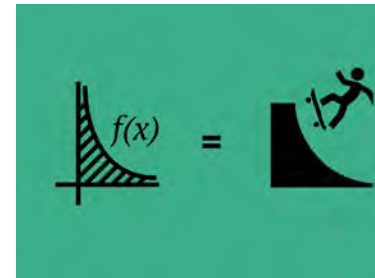
Analogy & Temporal/Spatial Analogy

アナロジー(類推、類比) - 特定の事物に基づく情報を、他の特定の事物へ、それらの間の何らかの類似に基づいて適用する認知過程 [出典: 類推 - Wikipedia]

Analogy (from Greek ἀναλογία, analogia) - *cognitive process of transferring information or meaning from a particular subject (the analogue or source) to another (the target), or a linguistic expression corresponding to such a process* [Wikipedia]

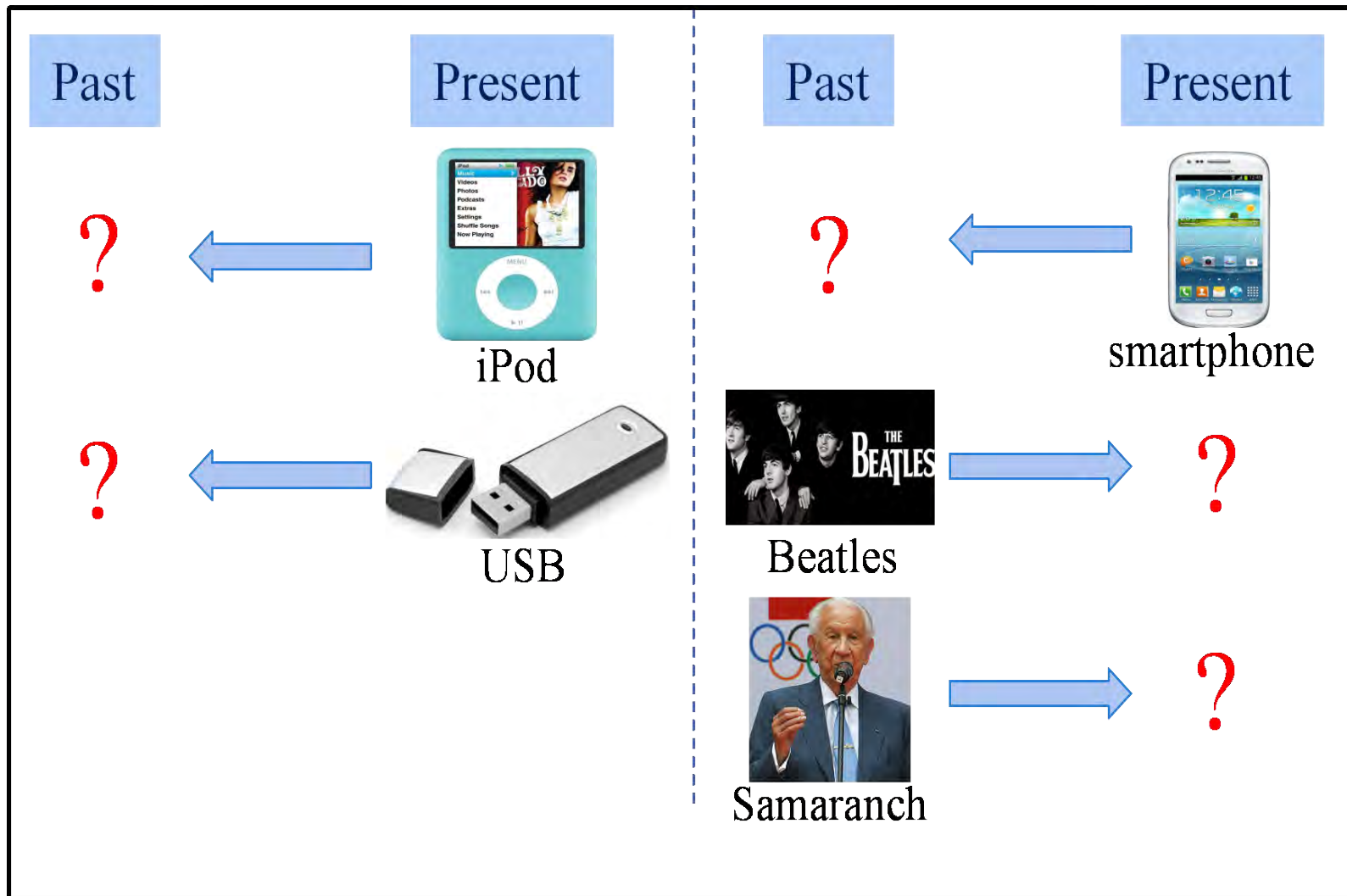
– 例: **時間的アナロジー**: 異なる時代に存在した事物の類似性に基づく比較
E.g., **temporal analogy**: *a comparison of two things based on their being alike where the things existed in different time periods*

– 例: **空間的アナロジー**: 異なる場所に存在した事物の類似性に基づく比較
E.g., **spatial analogy**: *a comparison of two things based on their being alike in where the things exist in different spatial locations*



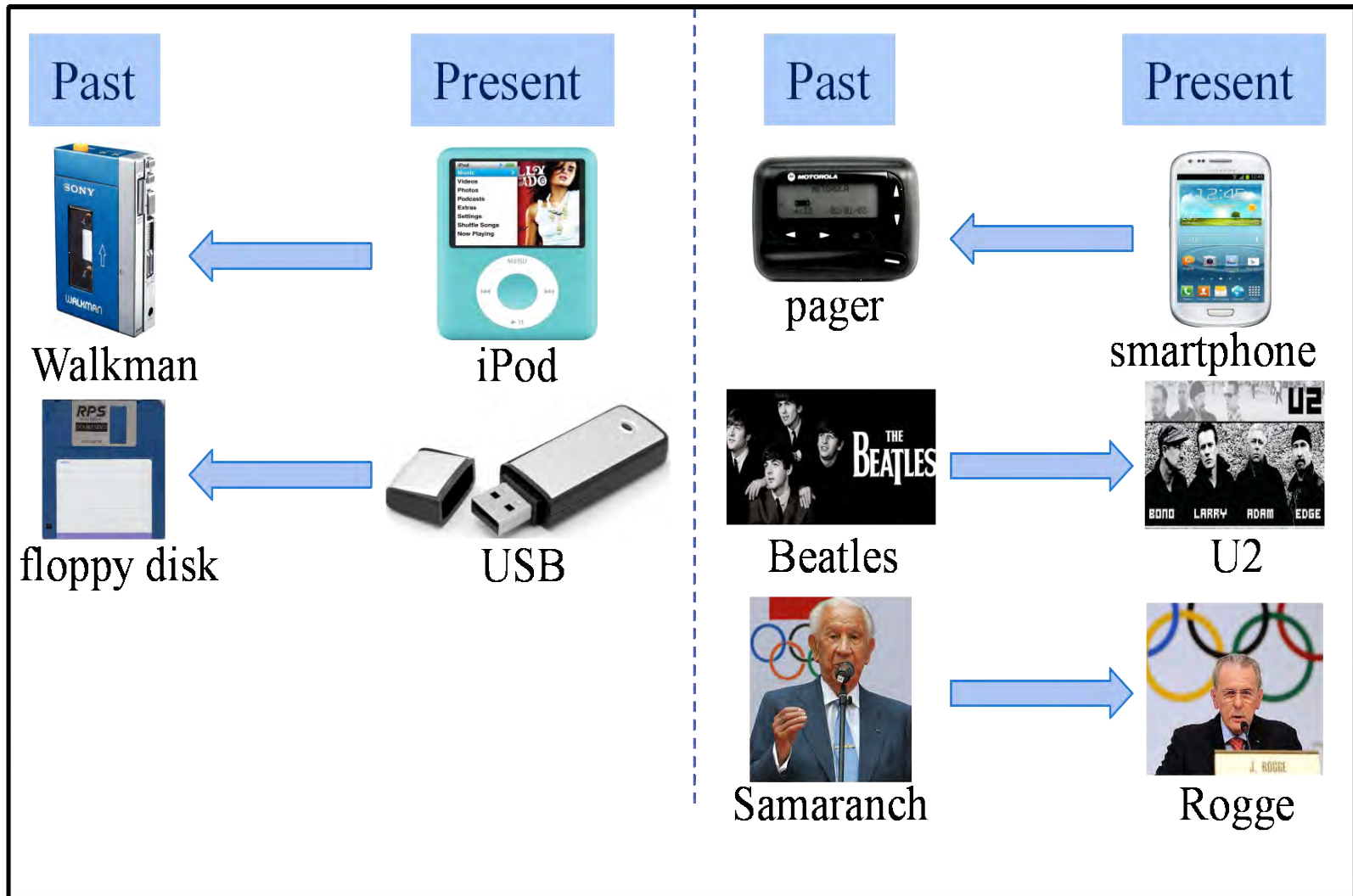
時間的アナログの例

Example Temporal Analogs



時間的アナログの例

Example Temporal Analogs



2種類の時間的アナログ

Two Types of Temporal Analogs

- 時間的アナログ: 意味的に類似しているが、異なる時代に存在した実体。

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

1. 名称が異なる同一の実体

Same entity with different name

例: ミャンマー(1989年以降)、ビルマ(1989年以前)

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

2. 異なる実体 Different entities

例: iPod (2000年代)、Walkman (1980年代)

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

時間的アナログ検出の主要課題

Key Challenges for Temporal Analogy Computation

1. 言葉の変化 Language evolution
2. 辞書ツール(例:古い言葉に関するオンライン辞書や歴史文献に対応可能な品詞タグ付けプログラム)及びナレッジベースの欠如
Lack of lexical tools (e.g., WordNet for the past or POS taggers for historical documents) or dedicated knowledge bases
3. OCRの誤認識 Large number of OCR errors
4. 古い時代に関するデータの少なさ
Sparse data for distant time periods

自然言語処理の伝統的な手法はあまり有効ではないかもしれない
Traditional techniques for natural language processing may not be very effective

「万物は流転する」

Panta Rei [Eng: Everything Changes]

- 「万物は流転する」ため、時間的アナログを取り巻く文脈も変化する。

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

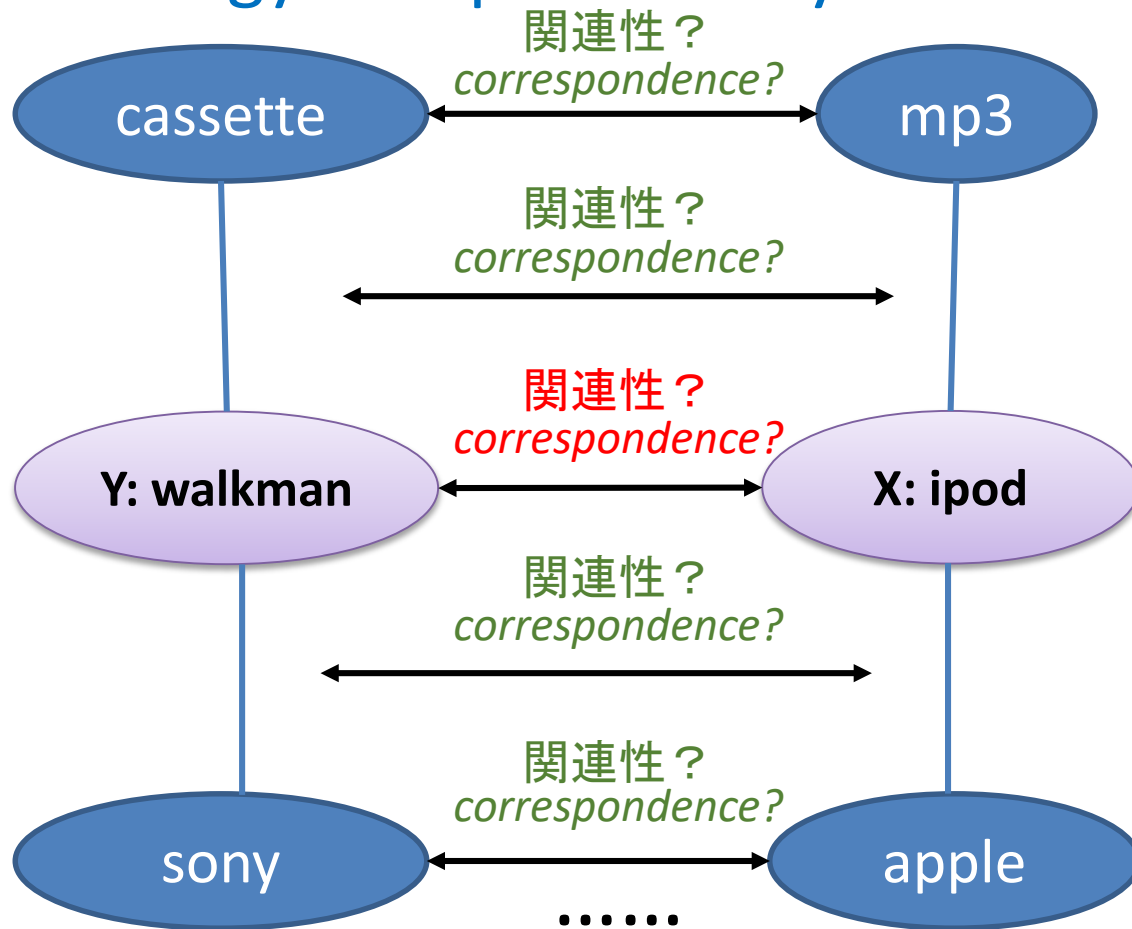
Walkman (1980s)	iPod (2010s)
cassette	apple
audio	mp3
video	roqit
tape	player
music	music
sony	geeks
digital	jukebox
stereo	portable
earphone	macintosh
recorder	dlink

簡単には解決できなさそうだ...
The task is not trivial...

* Contexts in the New York Times corpus

文脈照合による時間的アナログの検出

Temporal Analogy Computation by Context Matching



XとYが似ているかを推定するには、XとYそれぞれが含まれる文脈が類似しているかを分析する必要がある。

To estimate if X and Y are similar, it is necessary to analyze if the contexts of X and context of Y are similar.

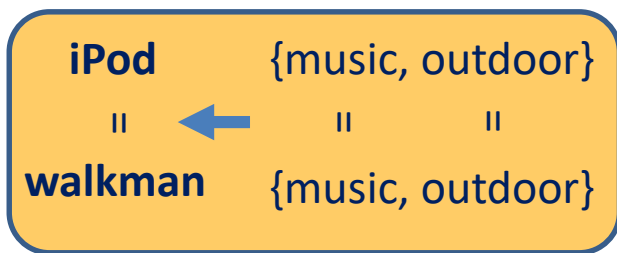
時間的アナログの検出のための文脈比較

Context Comparison for Finding Temporal Analogs

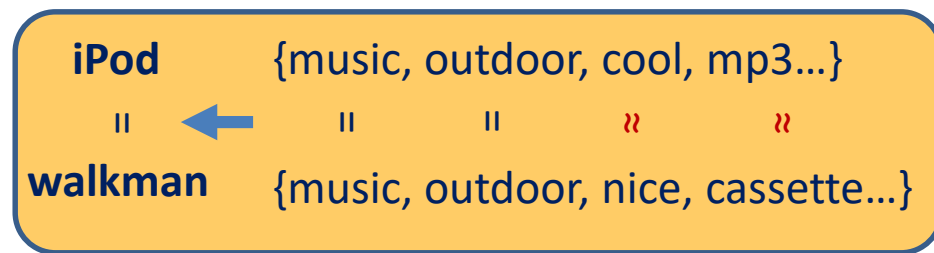
意味的に類似している = 類似の意味を持つ
文脈に含まれている

Semantically similar = surrounded by context with similar meaning

現在
Present



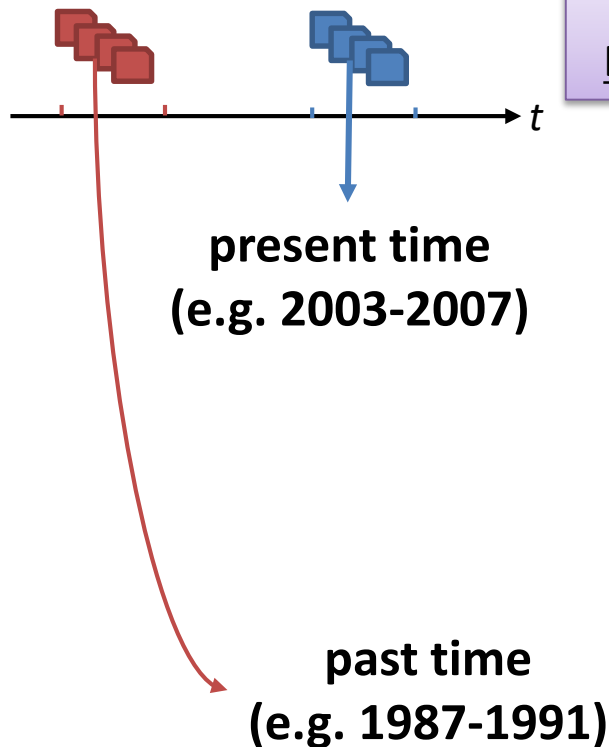
従来のアプローチ(Bag of Words)
Standard Bag of Words Approach



今回提唱するアプローチ
Our Approach

時間を超えた類似性: ニューラルネットワークを用いた単語埋め込み

Across-time Similarity: NN-based Term Embedding



分散ベクトル表現 [ミコロフ, 2013年]
Distributed Vector Representations (NN) [Mikolov 2013]

$$\begin{matrix} & D_1 & D_2 & \cdots & D_m \\ w_1 & \begin{bmatrix} \cdots \\ \cdots \\ \cdots \\ \cdots \end{bmatrix} \\ w_2 & \begin{bmatrix} \cdots \\ \cdots \\ \cdots \\ \cdots \end{bmatrix} \\ \vdots & & & & \\ w_P & \begin{bmatrix} \cdots \\ \cdots \\ \cdots \\ \cdots \end{bmatrix} \end{matrix}$$
$$\begin{matrix} & \Phi_1 & \Phi_2 & \cdots & \Phi_n \\ \omega_1 & \begin{bmatrix} \cdots \\ \cdots \\ \cdots \\ \cdots \end{bmatrix} \\ \omega_2 & \begin{bmatrix} \cdots \\ \cdots \\ \cdots \\ \cdots \end{bmatrix} \\ \vdots & & & & \\ \omega_Q & \begin{bmatrix} \cdots \\ \cdots \\ \cdots \\ \cdots \end{bmatrix} \end{matrix}$$

D_i と Φ_k が各ベクトル空間の次元
 D_i and Φ_k are the dimensions of each vector space

単語のベクトル表現: Bag of Words vs ニューラルネットワーク

Word Representation: BOW vs. NNs

- Bag of Wordsを用いたベクトル空間

Vector Space using Bag of Words (BOW)

Size of Vocabulary $\approx 400,000$

$$\vec{\text{ipod}} = \begin{matrix} & \overbrace{\text{a} \quad \text{an} \quad \dots \quad \text{music} \quad \dots \quad \text{road} \quad \dots \quad \text{usa} \quad \dots \quad \text{zebra} \quad \dots \quad \text{zoo}} \\ \left[\begin{array}{cccccccccccc} 50 & 150 & \dots & 200 & \dots & 5 & \dots & 100 & \dots & 0 & \dots & 0 \end{array} \right]$$

$$\vec{\text{car}} = \begin{matrix} & \text{a} \quad \text{an} \quad \dots \quad \text{music} \quad \dots \quad \text{road} \quad \dots \quad \text{usa} \quad \dots \quad \text{zebra} \quad \dots \quad \text{zoo} \\ \left[\begin{array}{cccccccccccc} 60 & 180 & \dots & 50 & \dots & 200 & \dots & 350 & \dots & 0 & \dots & 0 \end{array} \right]$$

- ニューラルネットワーク(Skip-gramモデル)を用いたベクトル空間

Vector Space using NNs (Skip-gram Model)

Size of Neurons of Hidden layer, usually 200 - 800

$$\vec{\text{ipod}} = \begin{matrix} & \overbrace{\chi_0 \quad \chi_1 \quad \dots \quad \chi_i \quad \dots \quad \chi_{200}} \\ \left[\begin{array}{cccccc} 0.1 & -0.2 & \dots & 0.3 & \dots & 0.8 \end{array} \right]$$

$$\vec{\text{car}} = \begin{matrix} & \chi_0 \quad \chi_1 \quad \dots \quad \chi_i \quad \dots \quad \chi_{200} \\ \left[\begin{array}{cccccc} 0.4 & 0.1 & \dots & -0.8 & \dots & 1.1 \end{array} \right]$$

- Captures semantic meaning
- Semantically similar words are located close in vector spaces

時間を超えた類似性 Across-Time Similarity

- 時間を超えた **マッピング** *Mapping* Across Time
 - 異なる時代における文脈を変換する
Transform background context across time

Present Vector Space (D=300)

<i>iPod</i>	0.3	1.2	...	-1.1
<i>music</i>	1.1	0.4	...	0.8
<i>mp3</i>	1.3	-1.2	...	-0.3
ω_i
ω_m

Past Vector Space (D=200)

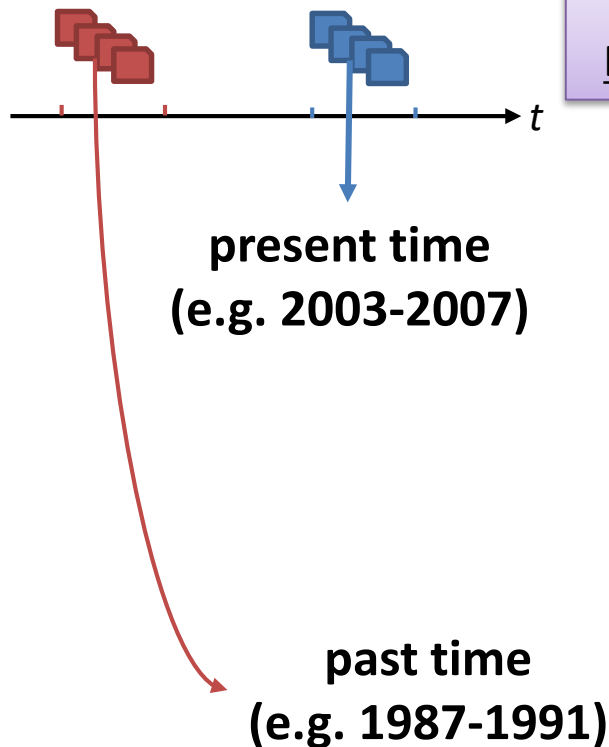
<i>walkman</i>	-1.1	0.3	...	1.2
<i>Music</i>	0.3	1	...	2
<i>Cassette</i>	0.4	-0.3	...	1.3
w_i
w_n

直接比較することはできない！
Can not be directly compared!

現在のベクトル空間から過去のベクトル空間へのマッピングを構築する必要がある
We have to **build the mapping** from the **present** vector space to **past** vector space

時間を超えた類似性: ニューラルネットワークを用いた単語埋め込み

Across-time Similarity: NN-based Term Embedding



分散ベクトル表現 [ミコロフ, 2013年]
Distributed Vector Representations (NN) [Mikolov 2013]

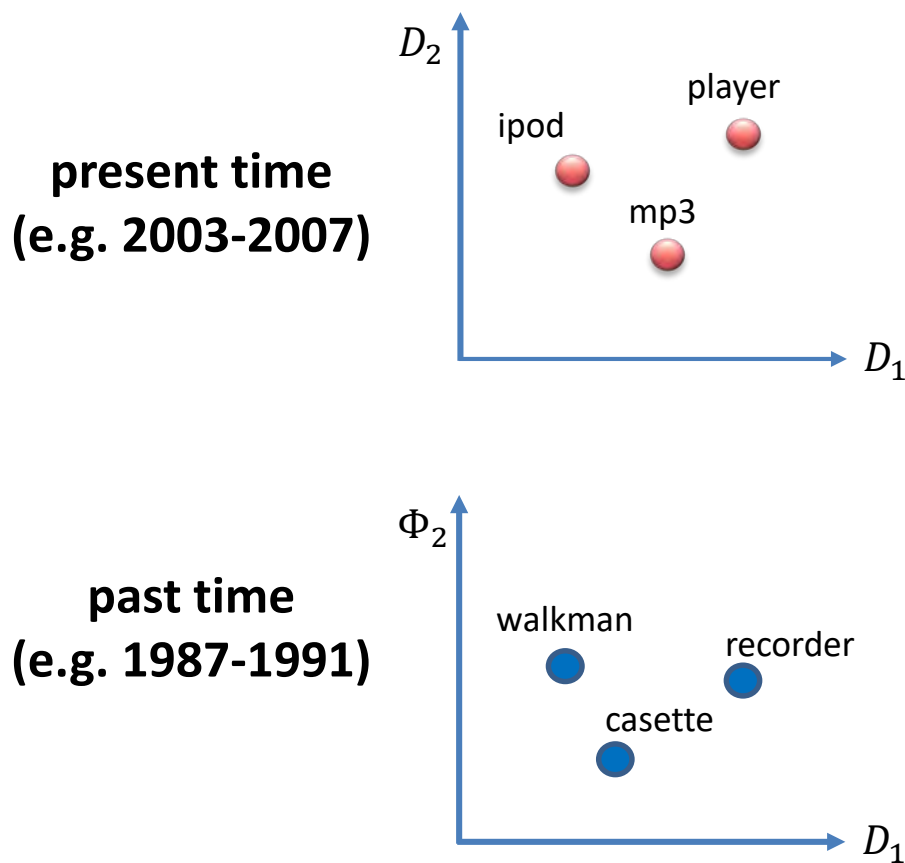
$$\begin{array}{c} w_1 \\ w_2 \\ \vdots \\ w_P \end{array} \begin{bmatrix} D_1 & D_2 & \cdots & D_m \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$
$$\begin{array}{c} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_Q \end{array} \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_n \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$



D_i と Φ_k が各ベクトル空間の次元
 D_i and Φ_k are the dimensions of each vector space

分析にあたっての仮説

Assumption behind Proposed Approach



仮説

それぞれのベクトル空間
における単語の相対的な
位置は一定である

Assumption

The relative positions of
terms in each vector space
remain stable

変換マトリクスの構築

Constructing Transformation Matrix

分散ベクトル表現
Distributed Vector
Representations

対応する単語のK個のペア
K Pairs of corresponding terms (anchors)
 $\{(w_i, \omega_i), \dots, (w_j, \omega_j)\}$



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

$$\mathbf{M} = \begin{matrix} & \Phi_1 & \Phi_2 & \dots & \Phi_n \\ \begin{matrix} D_1 \\ D_2 \\ \vdots \\ D_m \end{matrix} & \begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \end{matrix}$$



Present time
(e.g. 2003-2007)

$$\begin{matrix} w_1 \\ w_2 \\ \vdots \\ w_P \end{matrix} \begin{matrix} D_1 & D_2 & \dots & D_m \\ \begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \end{matrix}$$

Past time
(e.g. 1987-1991)

$$\begin{matrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_Q \end{matrix} \begin{matrix} \Phi_1 & \Phi_2 & \dots & \Phi_n \\ \begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \end{matrix}$$

一般的かつ頻出する単語を選択: 単語が頻繁に使用される程、意味が変化しづらい [パーゲル, 2007]

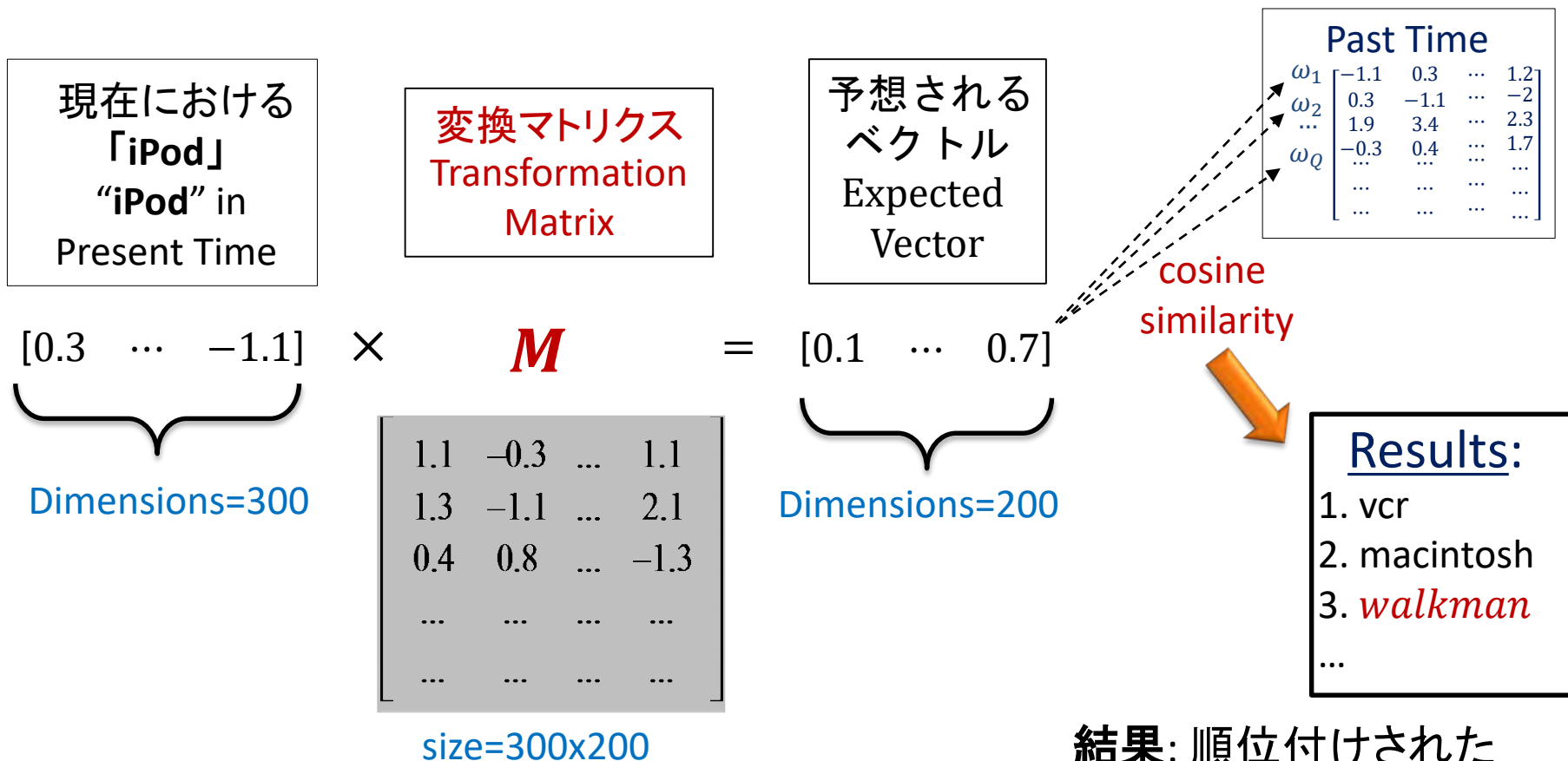
We choose **common, frequent terms**: the more frequently the word used, the harder is to change its meaning [Pargel 2007]

e.g. "man", "woman", "water"

26

単語変換の大局的なアプローチ

Global Term Transformation Approach



結果: 順位付けされた
時間的アナログのリスト
Result: ranked list of
temporal analogs

大局的な単語変換の問題点

Problems with Global Term Transformation

最良の答えではない。。
Not the best answers..

記録/再生機能を持つビデオカセットレコーダーがiPodの類似物として発見された。
Appleにより生産されたMacintoshがiPodの類似物として発見された。
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

変換マトリクス
Transformation Matrix



大局的な一致
Global Correspondence

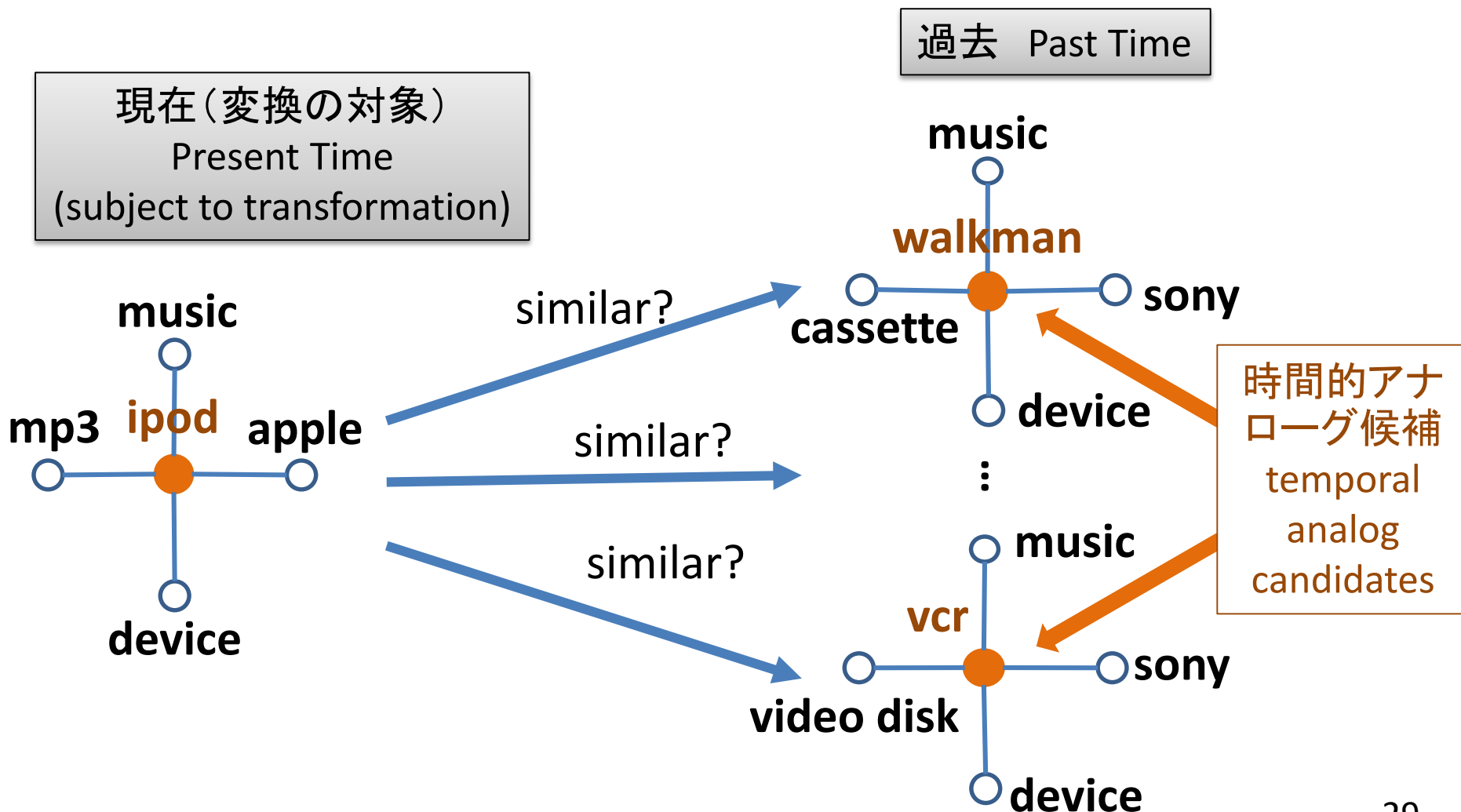
検索語とその周囲の文脈の関係が反映されていない
Relations between query and its local context are neglected



局所的な一致 Local Correspondence

局所的なグラフと参照ポイントを用いた変換方法

Transformation Using Local Graph by Using Reference Points



実験: データセットと設定 Experiments: Dataset and Settings

The New York Times

- データセット 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**
- テストセット (人物、場所、事物) Test sets (persons, locations, objects):
 - 95 pairs of <query, temporal counterpart> for **[2002-2007]** to **[1987-1991]**
- 変換マトリクスを学習させる Training Transformation Matrix
 - Feature dimension for Skip-gram model: 200
 - Number of Common Frequent Terms (CFTs): top frequent common words (5%)



実験: テストセット Experiments: Test Set

- 52のクエリと95の(クエリ, 時間的アナログ)のペアによって構成されたテストセットを手動で作成。

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
..

表1: テストセットの例 (qが入力語、tが予測される時間的アナログ。tは複数存在する場合もある。)

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

クエリの種類:

1. 人物
2. 場所
3. 事物

Type of queries:

1. Persons
2. Locations
3. Objects

実験結果例：現在のクエリに対する過去のアナログ検出

Example Results: Finding Past Analogs for Present Queries

クエリ queries		正解 correct answers	ベースライン baselines		方法 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
...

*Lexico-Syntactic Pattern used to detect reference points

正解の
順位付け
Rank of
correct
answers

OCRの誤認識を軽減する方法

Solution to Alleviate OCR Errors

光学的文字読み取り(OCR)の問題(Optical Character Problem)

- 辞書を構築して誤った綴りを正しいものにマッピングする

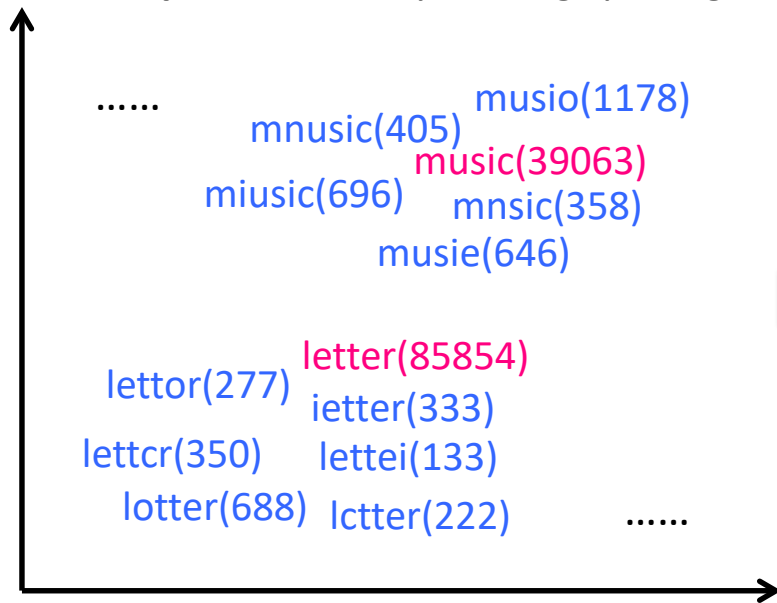
Build **dictionary** to map wrong spellings to correct ones

- 入力: 全ての単語のベクトル表現

Input: vector representation of all the words

- 出力: 辞書{誤った綴り, 正しい綴り}

Output: dictionary {wrong spelling: correct spelling}



Vector Space of [1906, 1915]

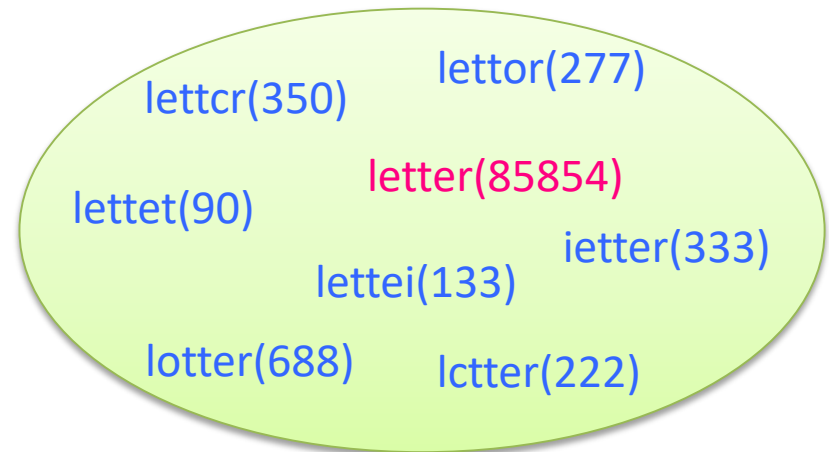
Original Spelling	Correct Form
mnusic	music
miusic	music
musie	music
.....
lettcr	letter
lettcr	letter
lotter	letter
.....

OCRの誤認識を軽減する方法

Solution to Alleviate OCR Errors

- OCR問題を軽減するための前提 Assumptions for Alleviating OCR Problem:

- (1) 誤って綴られた単語は正しい綴りと類似する文脈の中にある。
Wrongly spelled term has similar context with its correctly spelled term;
- (2) 正しい綴りは誤った綴りより頻出する。
The correct term is more dominant (or frequent) compared to its wrongly spelled ones;
- (3) 誤った綴りは一箇所修正すれば正しい綴りになる。
Wrongly spelled term has one edit-distance from its correct term.



- 結果の例 Example Results

- 修正なし Without Error Correction:

- car [2004,2009] → [1906,1015] vehicle, tricycle, mnotor, rmotor, car, eycles

- 修正あり With Error Correction:

- car [2004,2009] → [1906,1915] vehicle, tricycle, motor, car, cycles

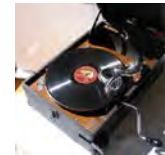
アスペクトベースの時間的アナログ抽出

Aspect-Based Temporal Analog Retrieval

クエリの形式: 実体, 観点, 時期

Query Pattern: *Entity, Viewpoint, Target Time*

- What similar to iPod 20 years ago, 80 or 100 years ago?
 - iPod, *listening to music*, 20 years ago → Discman, minidisc, Walkman
 - iPod, *listening to music*, 80 years ago → Disc-based phonograph
 - iPod, *listening to music*, 100 years ago → Cylinder-based phonograph
- What was the computing/writing device in the past analogous to PC?
 - PC, *computing*, past → calculator, abacus
 - PC, *writing*, past → typewriter
- Which president before Kennedy was also assassinated?
 - Kennedy, *assassinated president*, past → Lincoln, Garfield, McKinley, Spencer Perceval
- What was the old currency in China?
 - 人民币(renminbi), *currency*, past → 银元(Yinyuan), 交子(Jiaozi)
- What was a similar vehicle recall in the past to the one of Toyota in 2007?
 - Toyota recall (2007), *floor mat problem*, past → Audi recall (1982)



アーカイブから類似の実体を抽出する

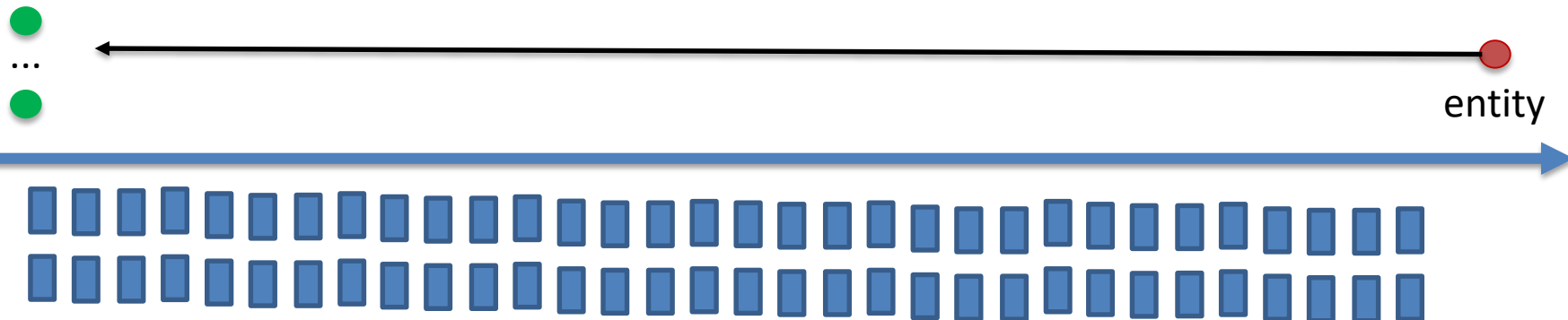
Similar Entity Retrieval from Archives

過去の類似する実体
(アスペクトを考慮に入れる)

Similar entities in past
considering aspect



過去の類似する実体
Similar entities in past



アスペクトベースの抽出システム

System for Aspect-based Retrieval

TempoAnalogus

Query in [2002,2007]:
euro

Past time period:
Select a time period

Method:
Select a method

Aspect term:
currency

Search

Reset

Temporal counterpart of **euro** biased on **currency** in [1987, 1991] is:

- francs : 0.609** ☒
but about nine billion **francs**, or \$250 million, of the aid depends on sabena's obtaining six billion **francs**, or about \$166 million, from a partner.
- belgian_francs : 0.574** ☒
lead: carlo de benedetti doubled his public offer tonight for societe generale de belgique's shares, from 4,000 **belgian francs** a share, or about \$113, to 8,000 **francs** in an attack on the **french-belgian** coalition that claims to have 52 percent of the vast holding company's capital.
- lire : 0.56** ☒
lead: "3*** company reports ** "3* de tomaso industries year to dec 31 1988 1987 sales 207,363,000 201,123,000 net loss 29,443,000 12,822,000 results are translated from italian **lire** at the exchange rate prevailing at dec.
- zloties : 0.544** ☐
the new official rate, which applies only to foreign tourists and foreign trade dealings, is 710 **zloties** to the dollar, compared with 680 on friday.
- lira : 0.538** ☒
lead: european officials were expected to consider devaluing the french franc and italian **lira** against the west german mark this weekend as the german currency's huge rise against the dollar intensified strains within the european monetary system.
- percent : 0.538** ☐
5 percent stake in mixte to 30 percent, and mixte will cut its 12 **percent** stake in the bank to 9.
- billion_pesetas : 0.537** ☐
22 **billion**, for the week ended wednesday, the investment company institute said thursday.
- dow_industrials : 0.536** ☐
the **dow** theory provided a bullish confirmation on tuesday, and another one yesterday, as the **dow** jones transportation average moved to record levels, while the **dow** jones industrial average climbed to its highest level since the 1987 crash.
- pound_sterling : 0.534** ☐
but ronald holzer, chief dealer for the harris trust and savings bank in chicago, said the dollar's rise against **sterling** was muted by the british currency's strength against the german mark and a flurry of other trading that helped the japanese yen and hurt the swiss franc.
- volume_shrank : 0.533** ☐
gains in agriculture sector the nation's trade surplus in agriculture jumped sharply despite the drought, the deficit in trade with japan dropped 15 percent and the nation's bill for imported oil declined as **volume shrank** and prices eased.

Feedback

時間的アナログの検出: 時間を超えた類似性の説明

TEMPORAL ANALOG DETECTION: ACROSS-TIME SIMILARITY EXPLANATION

アジェンダ Talk Schedule

1. はじめに Introduction
2. 異なる時代における類似物(時間的アナログ)の検出
Temporal Analog Detection
 - 時間を超えた類似性の説明
Across-time Term Similarity Explanation
3. 時間を超えた比較の要約
Across-time Comparative Summarization
4. 歴史に基づく実体のグループ化と要約
History-based Entity Summarization
5. 面白さに基づくアーカイブからの情報検索
Interestingness-oriented Archival Retrieval
6. 現在との関連性を志向する文献検索に向けて
Towards Present-relevance Oriented Document Search
7. 結び Conclusions

検出から説明へ

From Detection to Explanation

- q の過去におけるアナログは何か?
 - 例: 1980年代におけるiPodの類似物は何か?

What is an analog of q in past?

 - e.g., What is counterpart of **iPod** in 1980s?
- なぜ t は q の過去におけるアナログなのか?
 - 例: なぜ1980年代におけるWalkmanはiPodに似ているのか?

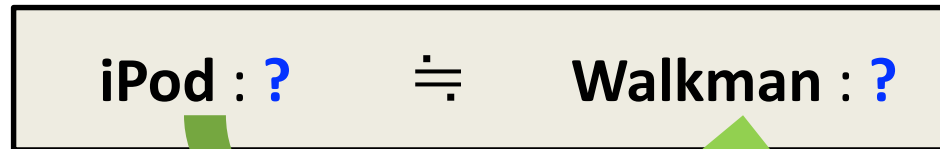
Why t is an analog of q in past?

 - e.g., Why is **iPod** similar to **Walkman** in 1980s?

時間を超えた類似性の説明: 問題の所在

Across-time Similarity Explanation: Problem Statement

Input:



複数の基準に基づく
Based on several criteria

Output:

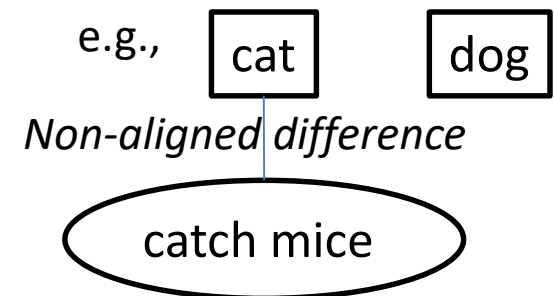
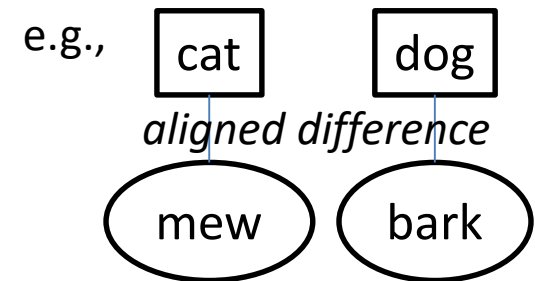
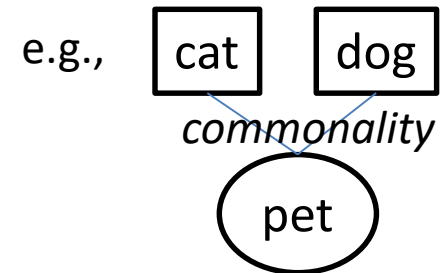


2つの実体が時間を超えて類似していることの理解を促すエビデンスの提供
Providing evidence to support understanding of similarity between two entities across time

構造整列仮説 (гентナー&マークマン, 1997)

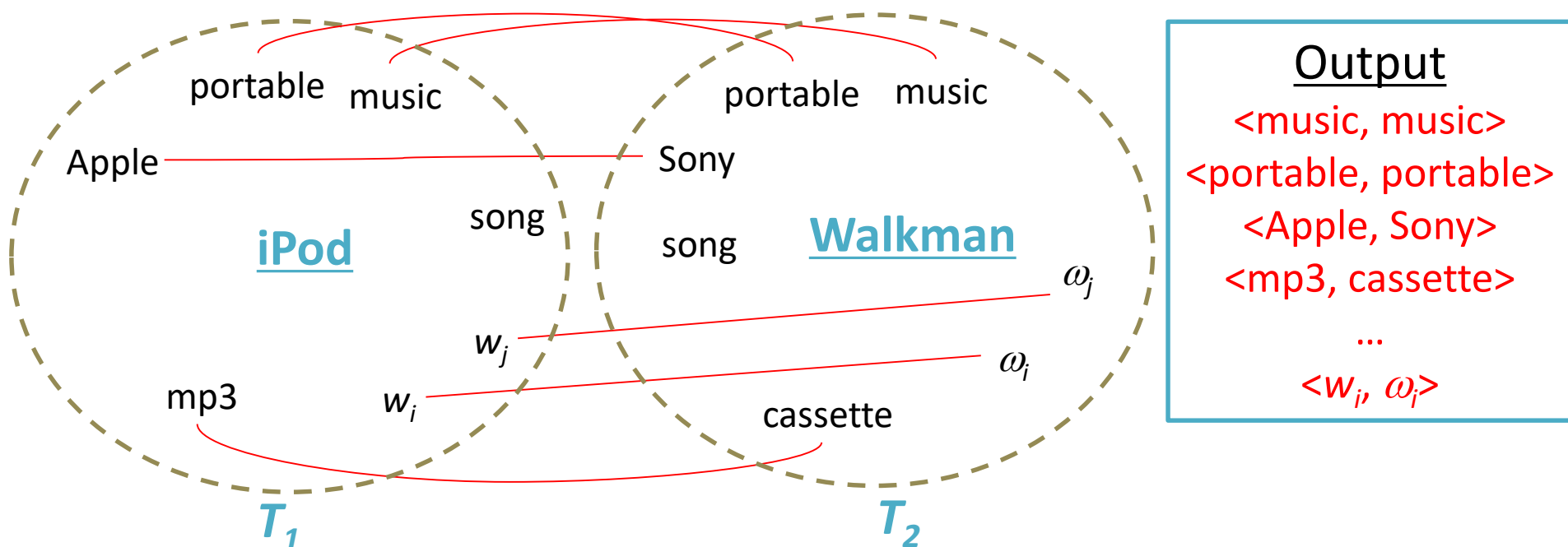
Theory of Structural Alignment Model (Gentner & Markman, 1997)

- **共通性:** 二つの実体の**同一の**属性
Commonality: **identical** attributes of two entities
 - E.g., music (iPod) = music (Walkman)
- **整列した差異:** 二つの実体と**同じ関係を持つ**が値が異なる属性
Aligned Difference: attributes which have **same relation** to the two entities but have different values
 - E.g., mp3 (iPod) ≈ cassette (Walkman) [**storage media**]
- **整列していない差異:** 一方の実体のみに当てはまり、もう一方に**対応する要素が存在しない**属性
Non-aligned Difference: the element of one entity that has **no corresponding** element in the other entity
 - E.g., display panel (iPod) [**no corresponding item in Walkman**]



問題の概念図

Conceptual View of Problem



特定の実体に文脈上関連する単語が、頻繁に共起する語の中から導き出される。
Context terms of a given entity are derived from frequently co-occurring terms

タスク: 共通性や整列した差異を示す単語のペアを見つける。
Task: find good word pairs denoting commonalities or aligned differences

時間を超えた類似性を説明する

Explaining Across-time Similarity

1. 関係性 Relatedness

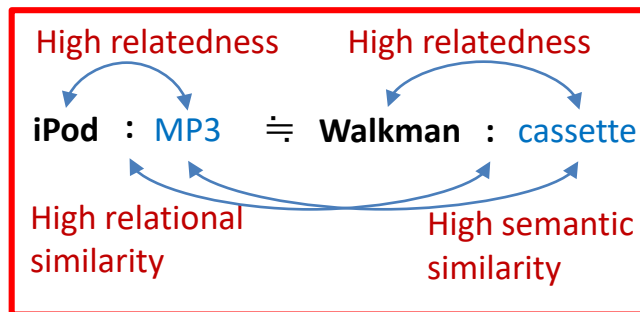
- ペアとなる単語はそれぞれ対象の実体と関係している。
- Terms in a pair should be related to their entities

2. 意味的類似性 Semantic similarity

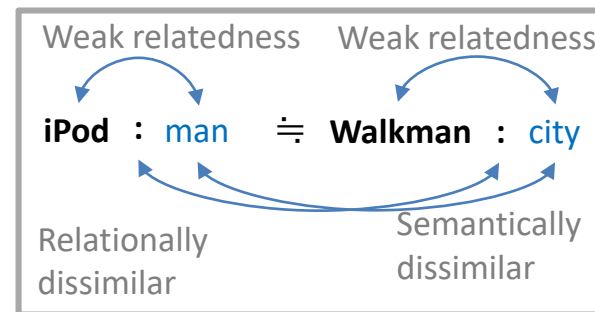
- ペアとなる単語自体が 互いに類似している。
- Terms should be similar to each other

3. 関係的類似性 Relational similarity

- ペアとなる単語とそれらが関係する実体との関係性が類似している。
- Terms should have similar relation to their entities



良い単語の組み合わせ
Good term pair



悪い単語の組み合わせ
Bad term pair

実験結果 Experimental Results

Methods	Precision	Recall	F ₁ -score
Overlap	0.63	0.48	0.55
BOW	0.23	0.17	0.20
Com	0.46	0.34	0.39
Local	0.66	0.50	0.57
Global	0.72*†	0.54*†	0.61*†

baselines

methods

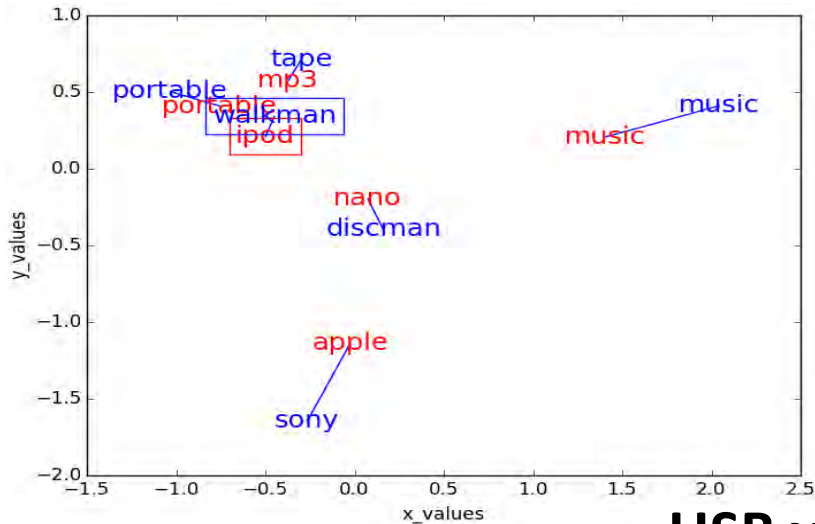
[2002, 2007]: "Bustamante, a democrat, is the leading candidate to replace him if the recall succeeds, holding a narrow margin over his closest competitor, *Arnold Schwarzenegger*, a republican."

[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."

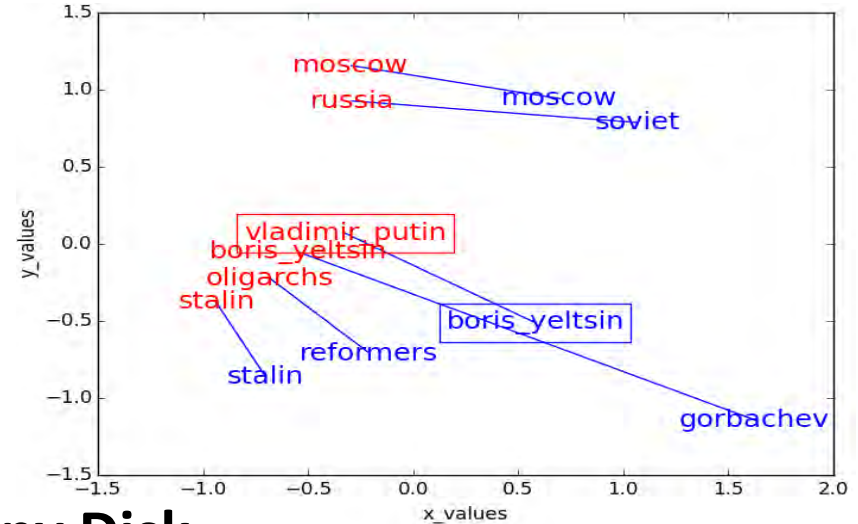
Correct pairs	baselines			methods	
	Overlap	BOW	Com	Local	Global
<i>iPod vs. Walkman</i>					
Apple - Sony (company)		✓		✓	✓
MP3 - cassette (media)				✓	✓
portable - portable (characteristic)	✓			✓	✓
music - music (usage)	✓				✓
<i>Arnold Schwarzenegger vs. Arnold Schwarzenegger</i>					
Bustamante -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

主成分分析に基づく視覚化 PCA based Visualization

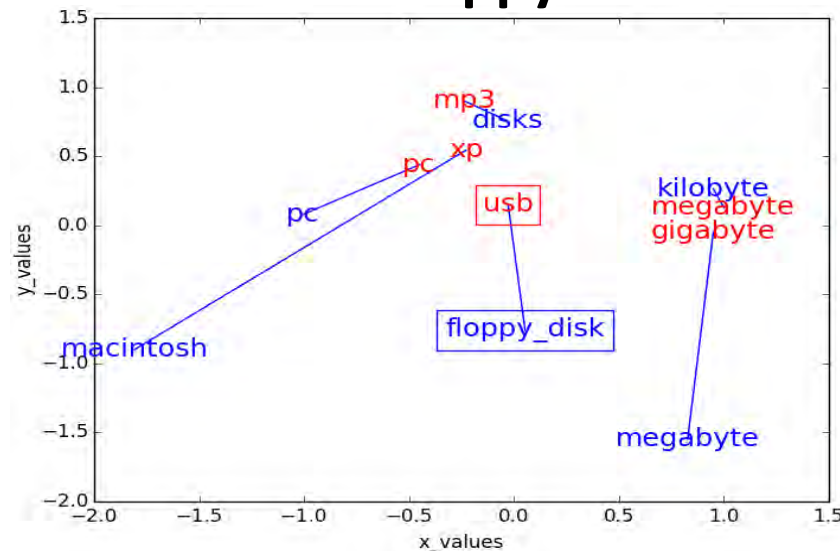
iPod vs. Walkman



Vladimir Putin vs. Boris Yeltsin



USB vs. Floppy Disk



時間を超えた比較の要約

ACROSS-TIME COMPARATIVE SUMMARIZATION

アジェンダ Talk Schedule

1. はじめに Introduction
2. 異なる時代における類似物(時間的アナログ)の検出
Temporal Analog Detection
 - 時間を超えた類似性の説明
Across-time Term Similarity Explanation
3. 時間を超えた比較の要約
Across-time Comparative Summarization
4. 歴史に基づく実体のグループ化と要約
History-based Entity Summarization
5. 面白さに基づくアーカイブからの情報検索
Interestingness-oriented Archival Retrieval
6. 現在との関連性を志向する文献検索に向けて
Towards Present-relevance Oriented Document Search
7. 結び Conclusions

背景 Background

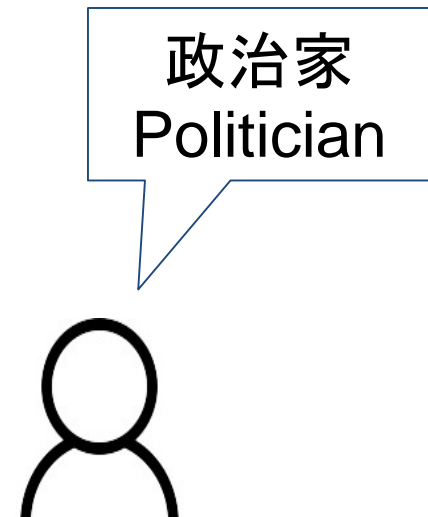
- ニュースは質の良い情報の源泉として最も重要なものの一つである。
- News is one of the most important channels for acquiring high-quality information
- 利用者の調べたい内容によっては、異なる二つの時期のニュースの比較を行いたいと望む場合もある。
- Sometimes, users wish to compare two collections of news, which can be from distant times, biased on query



News of 1990s



News of 2010s



動機 Motivation

クリントン大統領は
1998年に中国を訪
問した。

President Clinton
visited China in
1998



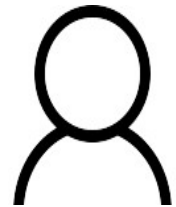
News of 1990s

トランプ大統領は
2017年に中国を訪
問した。

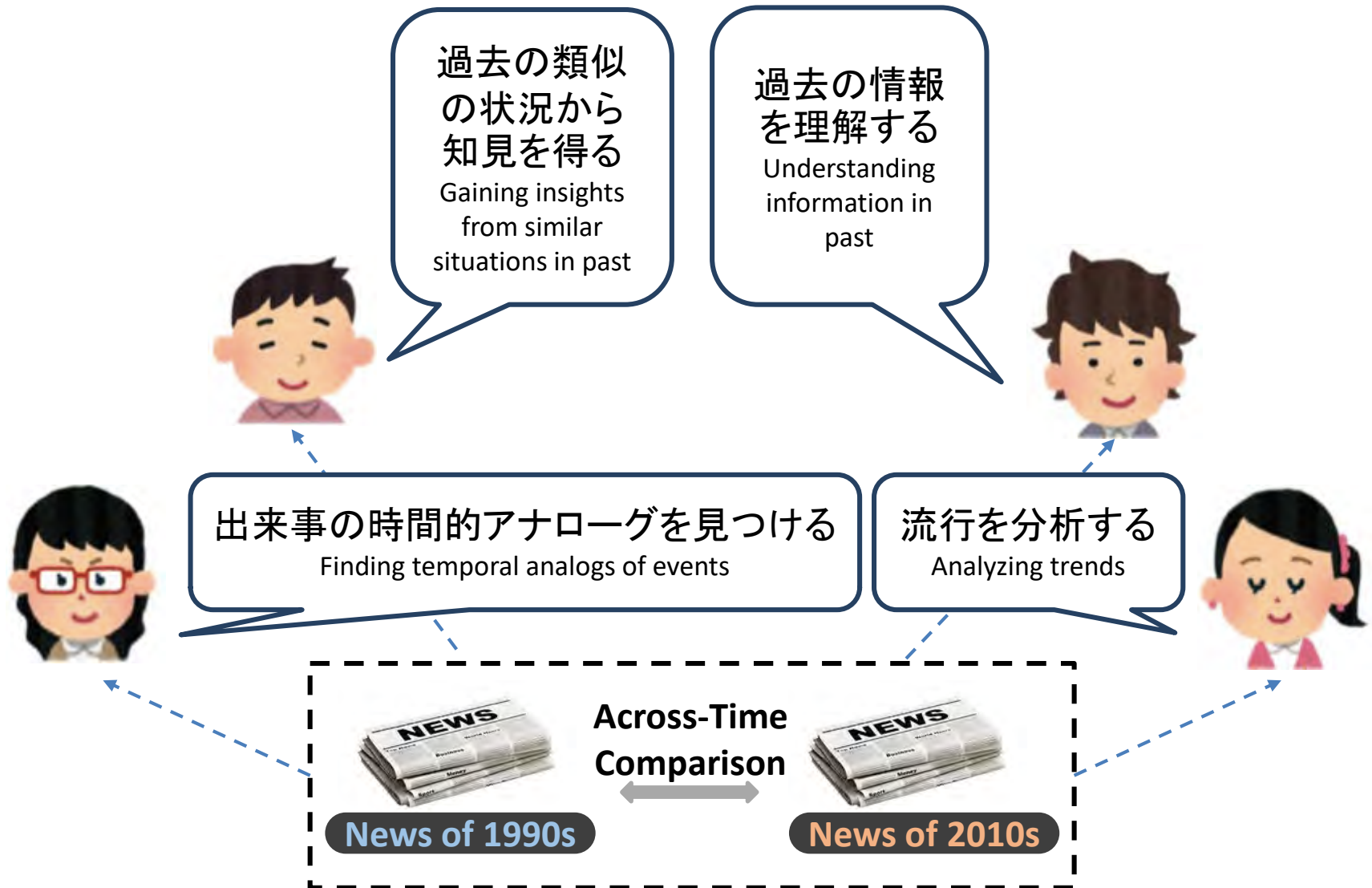
President Trump
visited China in 2017



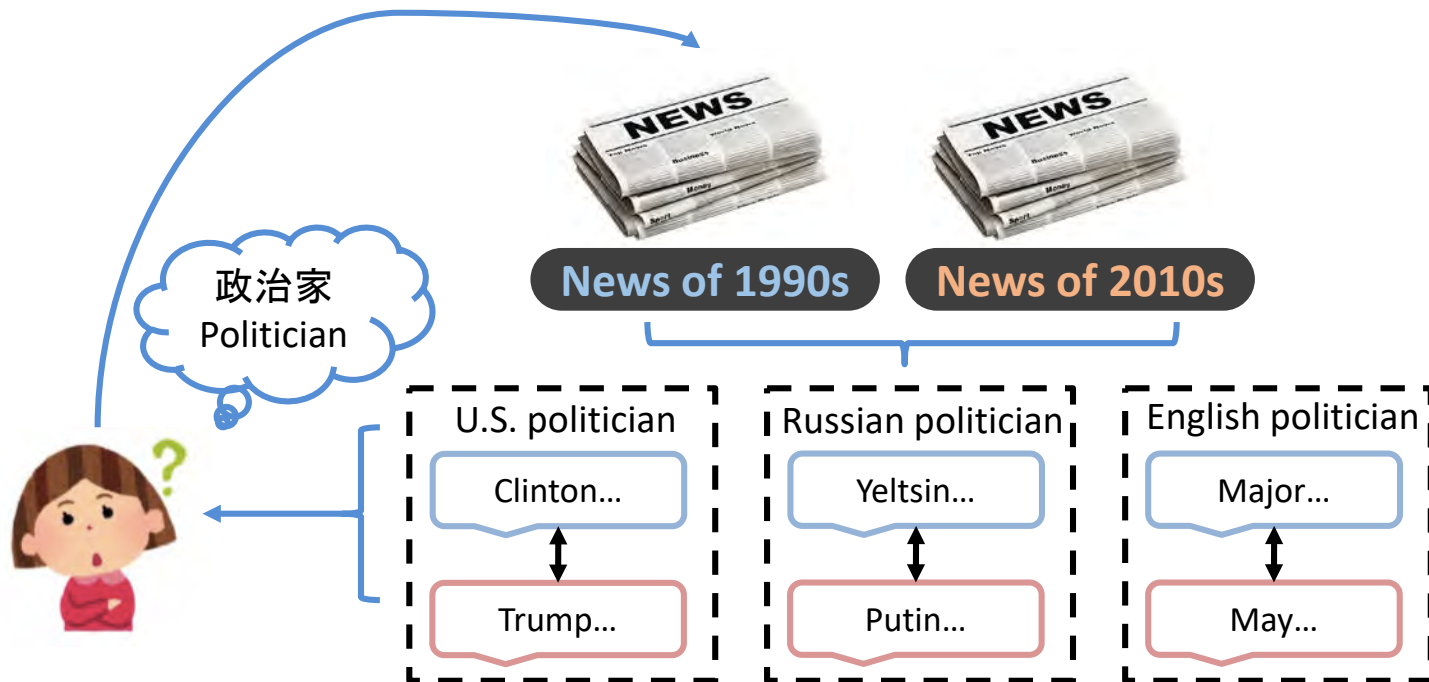
News of 2010s



動機 Motivation



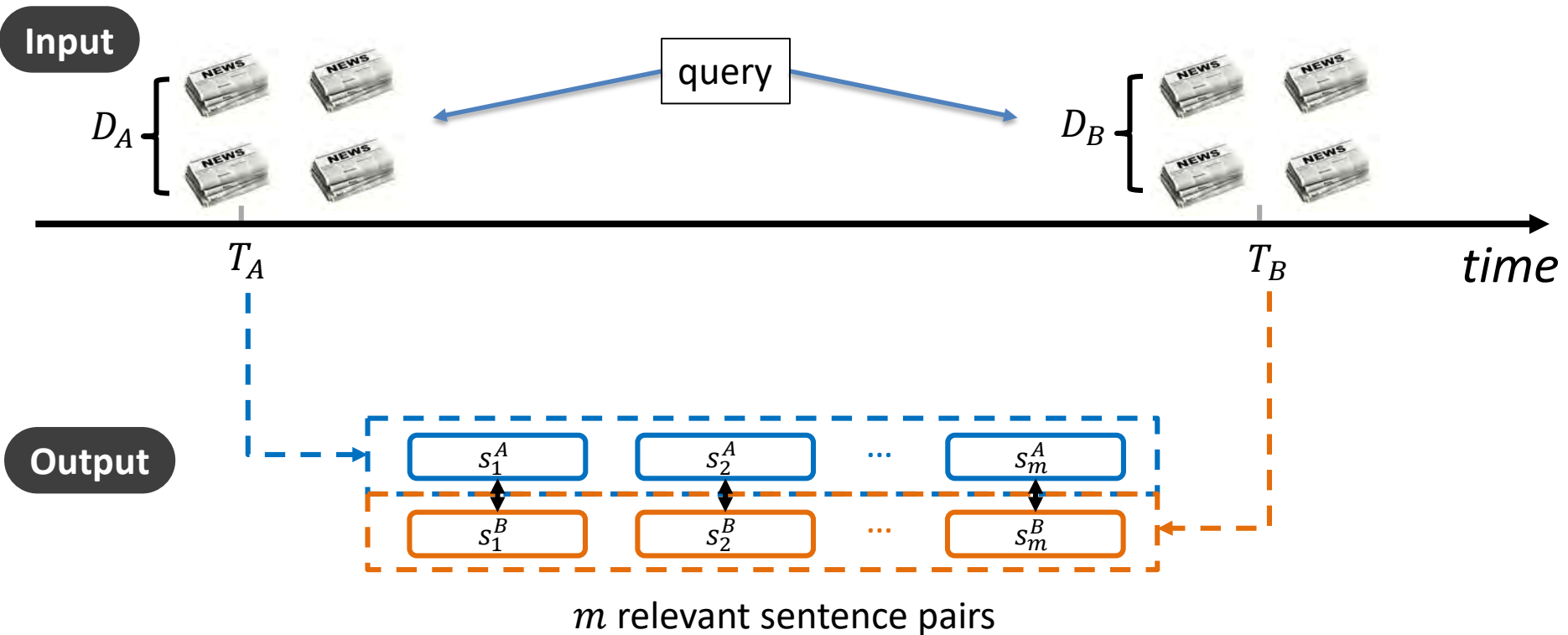
動機 Motivation



特定の検索クエリに基づいて、二つの異なる時期のニュースを比較する。

Compare two collections of news from distant times biased to a given specific query.

問題の定式化 Problem Formulation



関連性
Relevance

類似性
Correspondence

重要性
Saliency

多様性
Diversity

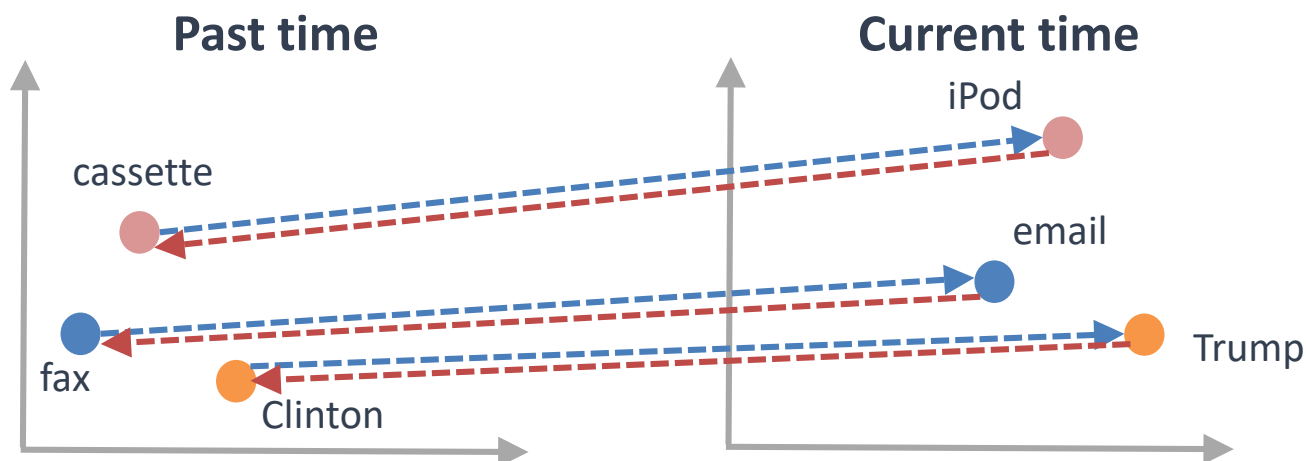
課題 Challenges

- 重要度の高い文の割合は非常に低い。
The ratio of salient candidate sentences over trivial candidate sentences is very low
- 情報源となるニュースアーカイブが多様であるために、複数の潜在的サブグループに及ぶ可能性がある。
The input news archives can be very diverse as well as may cover multiple latent subgroups
- 文章の対応関係の測定は困難である(ニュース全体の文脈は異なっているかもしれない)。
To measure sentence correspondence is a difficult task
 - The entire context of input news collections may be fairly different
- 上述の全ての要素を考慮するのは困難な問題である。
Considering all the listed above factors is a challenging problem

方法：高品質な要約を作るために上述の四要素を全て考慮に入れた
結合整数線形計画法 (J-ILP)

Method: **Joint integer linear programming framework (J-ILP)** considering all the four factors for high quality summary

直交変換 Orthogonal Transformation



ベクトル空間表現
Vector Space Representation

L 個のペアの学習用データが、過去と現在の両方の文書集合の中で学習され $[(a_1, b_1), (a_2, b_2), \dots (a_n, b_n)]$ 、以下の式を満たす変換行列 W が得られる。

Given L pairs of training data trained in both document collections $[(a_1, b_1), (a_2, b_2), \dots (a_n, b_n)]$, transformation W should be learned as follows:

$$W = \operatorname{argmin}_W \sum_{i=1}^L \|a_i - Wb_i\|_F^2, s.t. W^T W = I$$

WMDによる異なる時期の文の類似性計算

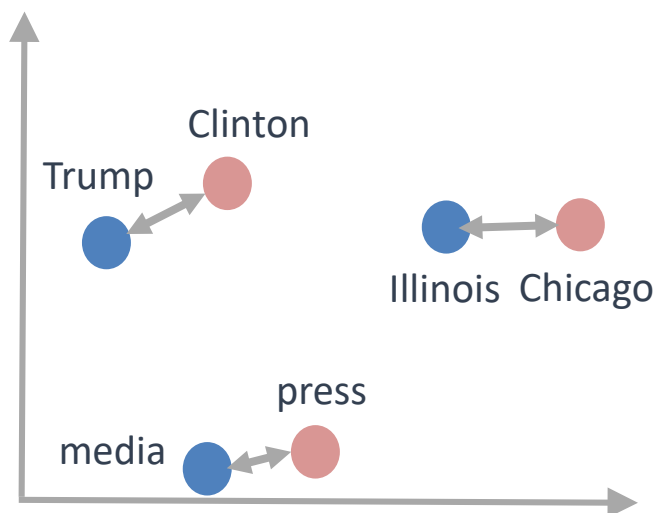
Across-time Sentence Similarity Computation using WMD

Sentence s
(Past time)

Clinton
speaks
to
the
media
in
Illinois

Sentence s'
(now)

Trump
greet
s the
press
in
Chicago



ベクトル空間表現
Vector Space Representation

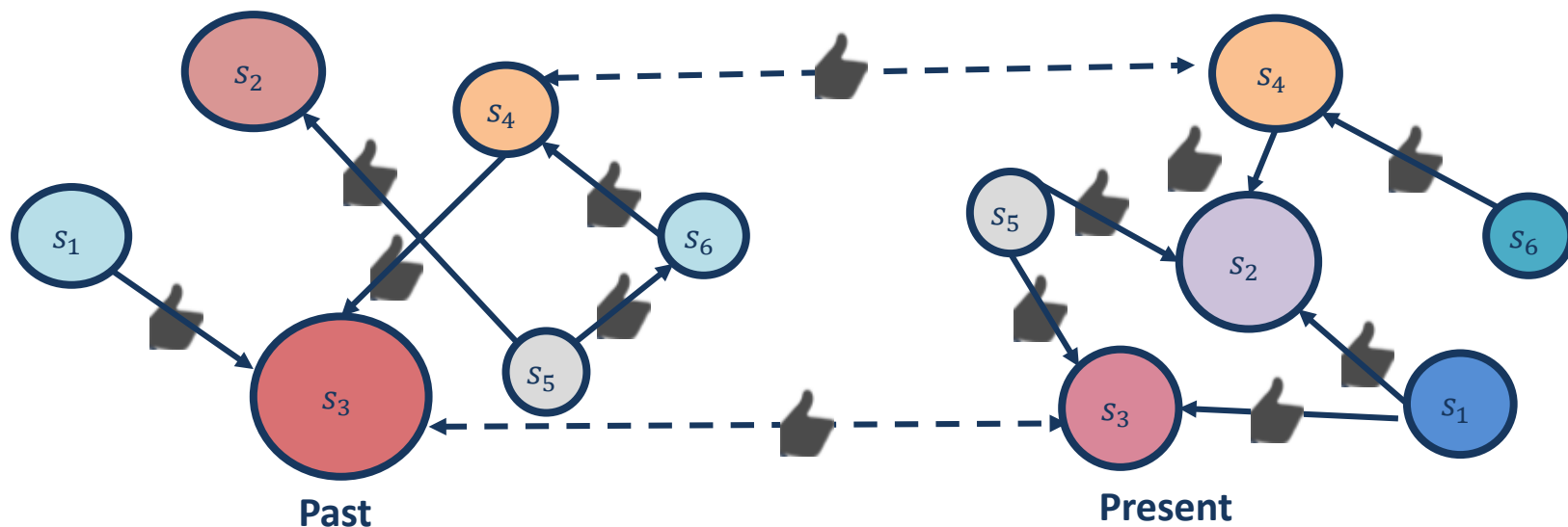
文書間距離
Word Movers Distance

結合整数線形計画法

Joint Integer Linear Programming

多様で典型的な模範データを検出し、同時に対応する実体のペアを順位付けする、簡潔な結合整数線形計画法(J-ILP)フレームワークを提案します。

We propose a concise *joint integer linear programming (J-ILP)* framework which detects diverse and representative exemplars and concurrently ranks correspondent entity pairs from the detected exemplars.

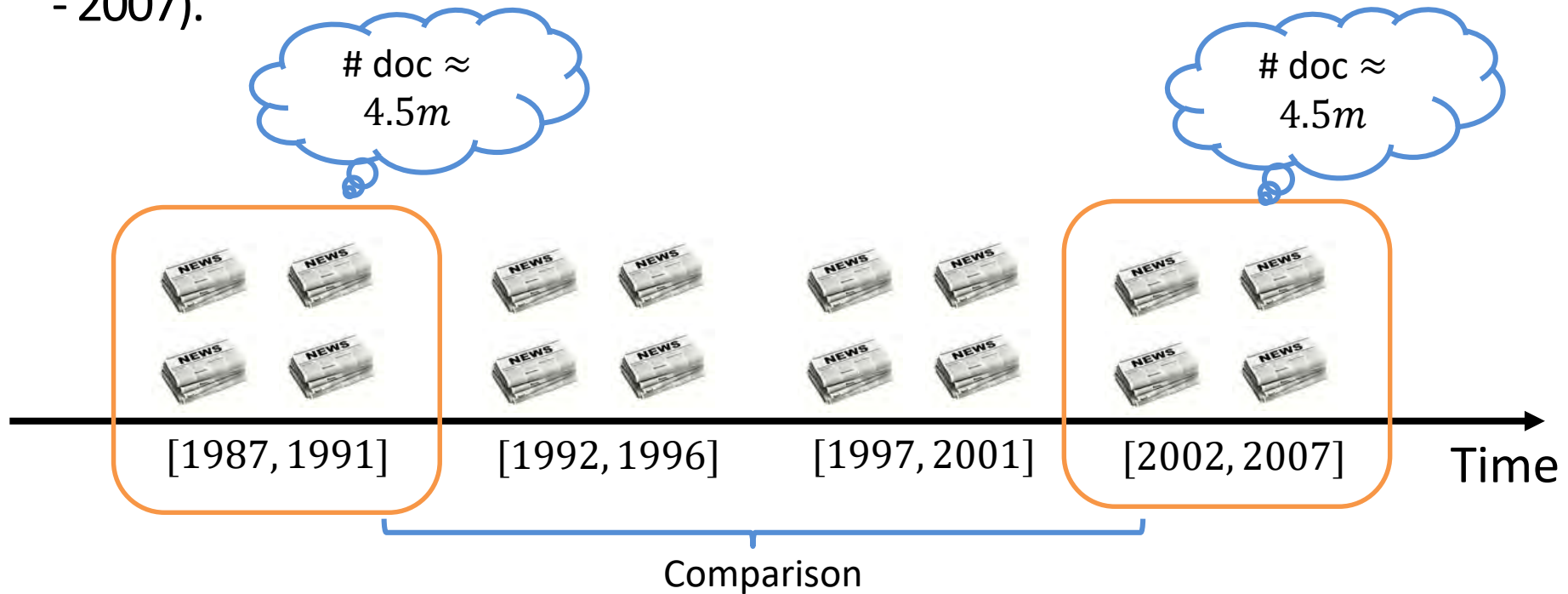


他の実体に支持される実体を模範と呼ぶ。

We call an entity as exemplar if it is voted by other entities

実験: セットアップ Experiments: Setup

- 実験には、New York Times Annotated Corpus (1987～2007年)を使用する。
- For the experiments we use the New York Times Annotated Corpus (1987 - 2007).



実験: 結果 Experiments: Results

ROUGE (要約システムの自動評価法)を用いたパフォーマンスの比較 Performance Comparison using ROUGE

Type	System	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
Proposed Methods	J-ILP (OT)	0.485	0.218	0.447	0.195	0.148
	J-ILP (LT)	0.478	0.234	0.423	0.162	0.128
	J-ILP (Non-Tran)	0.171	0.061	0.098	0.040	0.017
Multi-Document Summarization Methods	LexRank	0.260	0.138	0.186	0.089	0.064
	LSA	0.313	0.144	0.242	0.096	0.073
	KLSUM	0.235	0.048	0.169	0.085	0.021
Comparative Summarization Methods	DSS	0.179	0.075	0.151	0.074	0.057
	MRRW	0.346	0.180	0.283	0.103	0.077
	LPCS	0.386	0.190	0.312	0.122	0.082

適合率(Precision)を用いたパフォーマンスの比較 Performance Comparison using Precision

System	<i>Precision_S</i>	<i>Precision_P</i>
J-ILP (OT)	0.875	0.333
J-ILP (LT)	0.875	0.217
J-ILP (Non-Tran)	0.266	0.002
LexRank	0.350	0.100
LSA	0.375	0.083
KLSUM	0.208	0.042
DSS	0.283	0.053
MRRW	0.325	0.067
LPCS	0.467	0.133

例 Example

- (4.1) Cecil Fielder, who led the major leagues in 1990 with 51 homers after he played a season in Japan, was among 26 players named yesterday to go to Japan for an eight-game tour next month.
- (4.2) Japanese soccer officials announced yesterday that its men's national team would not travel to the United States for two exhibition games because of the war in Iraq.
-
- (5.1) Pertamina, Indonesia's state-owned oil company, and Japanese buyers have agreed in principle to a new one-year contract for sales of crude oil.
- (5.2) Japan is looking to the Russian Far East, ensuring that Sakhalin Island will become a major supplier of oil and gas to Japan within a decade.
-
- (6.1) American and Japanese negotiators met today in the opening round of talks intended to follow through on agreements reached last summer to remove "structural impediments" to trade.
- (6.2) With President Vicente Fox of Mexico here to sign a free trade pact with Japan, talks broke down Thursday over Japan's dogged defense of its pork and orange juice producers.
-
- (7.1) Japan's Fair Trade Commission said today that its international committee was considering applying anti-monopoly regulations to all foreign companies whose business practices affect Japan.
- (7.2) The Fair Trade Commission in Japan ruled on Tuesday that the Intel Corporation violated the country's antimonopoly law in the way it sold semiconductors and ordered the company to stop some of its sales practices.
-
- (8.1) Japan recently announced the end to five decades of commercial whaling.
- (8.2) Japan's latest effort to resume commercial whaling was strongly rebuffed in two votes at the biennial meeting of 160 countries adhering to the Convention on Trade in Endangered Species.
-
- (9.1) Foreign car sales in Japan rose 59 percent from last year's levels to a record 9,597 in July, a spokesman for the Japan Automobile Importers Association said.
- (9.2) Sales at Japan's largest industrial electronics companies rebounded in the October through December quarter on strong demand for optical disk drives, cellphones and the semiconductors used in digital cameras and other hot-selling gadgets.
-
- (10.1) Japan said today that its trade figures with the United States had improved strikingly in the last six months, and it predicted that the trend would continue for the rest of the year.
- (10.2) Sharply increased trade with China has lifted the Japanese economy out of a lost decade of feeble growth and recurring recession, while cheap imports from China have driven costs down significantly for Japan's long-suffering consumers.
-

Table 6: Example of generated summary by J-ILP with orthogonal transformation (Query: *Japan*). Each row contains a pair of two across-time comparative sentences. The first sentence in each pair is extracted from documents published in [1987, 1991], while the second one is taken from documents published in [2002, 2007].

歴史に基づく実体の
グループ化と要約

**HISTORY-FOCUSED ENTITY
GROUPING AND SUMMARIZING**

アジェンダ Talk Schedule

1. はじめに Introduction
2. 異なる時代における類似物(時間的アナログ)の検出
Temporal Analog Detection
 - 時間を超えた類似性の説明
Across-time Term Similarity Explanation
3. 時間を超えた比較の要約
Across-time Comparative Summarization
4. 歴史に基づく実体のグループ化と要約
History-based Entity Summarization
5. 面白さに基づくアーカイブからの情報検索
Interestingness-oriented Archival Retrieval
6. 現在との関連性を志向する文献検索に向けて
Towards Present-relevance Oriented Document Search
7. 結び Conclusions

文献の時系列 Timeline Documents

- 実体に関する多くの文献は時系列順に整理された出来事についての記述を含む。(例:「京都」のWikipediaページの「歴史」の項目)
Many documents about entities contain descriptions of chronologically arranged events (e.g., history section of Wikipedia page on La Rochelle)
- そうした文献に含まれる文の一つ一つには、その文に書かれた出来事がいつ起きたかを明かすタイムスタンプが付けられていると考えることができる。
We can assume that each sentence in such a document can be annotated with a timestamp revealing when an event expressed in the sentence took part

History [edit]

See also: *Timeline of La Rochelle*

Antiquity [edit]



The area of La Rochelle was occupied in antiquity by the Gallic tribe of the *Santonnes*, who gave their name to the nearby region of *Saintonge* and the city of *Saintes*.^[*citation needed*]

The Romans subsequently occupied the area, where they developed salt production along the coast as well as wine production, which was then re-exported throughout the Empire.^[*citation needed*] *Bordeaux* has been found at Saint-Estève and at *Les Moulins*, as well as at *Le Grand-Miroir*.

From the 12th to the 15th century Bordeaux was ruled by the Plantagenets, born in Le Mans, who within the city the cathedral of St. André was built. It was also the site of the first printing press in France, and so extended its influence. The Plantagenets, who were powerful symbols of the new regime, which halted the wine commerce with England and so deprived the city of its wealth.

In 1482 Bordeaux created a local parliament. However, it only regained its importance during the 16th century when it became a major trading centre for sugar and slaves from the *West Indies*, along with its traditional wine exports.^[5]

Bordeaux adhered to the Fronde, being effectively annexed to the Kingdom of France only in 1653, when the army of Louis XIV entered the city.

The 18th century saw the golden age of Bordeaux. Many downtown buildings (about 5,000), including those on the quays, are from this period. Victor Hugo found the town so beautiful he said: "Take *Versailles*, add *Antwerp*, and you have *Bordeaux*". Baron *Haussmann*, a long-time prefect of Bordeaux, used Bordeaux's 18th-century large-scale rebuilding as a model when he was asked by Emperor *Napoleon III* to transform a then still quasi-medieval *Paris* into a "modern" capital that would make France proud.

Towards the end of the Peninsula war on 12 March 1814, the Duke of Wellington sent William Beresford with two divisions and seized Bordeaux encountering little resistance. Bordeaux was largely anti-Bonapartist and the majority supported the Bourbons, so the British troops were treated as liberators.

In 1870, at the beginning of the Franco-Prussian war against Prussia, the French government temporarily relocated to Bordeaux from Paris. This recurred during the First World War and again very briefly during the Second World War, when it became clear that Paris would fall into German hands. However, on the last of these occasions the French capital was soon moved again to Vichy. In May and June 1940, Bordeaux was the site of the life-saving actions of the Portuguese consul-general, *Artur de Sousa Mendes*, who illegally granted thousands of Portuguese visas, which were needed to pass the Spanish border, to refugees fleeing the German Occupation.

From 1940 to 1943, the Italian Royal Navy (*Regia Marina Italiana*) established BETASOM, a submarine base at Bordeaux. Italian submarines participated in the *Battle of the Atlantic* from this base, which was also a major base for German U-boats as headquarters of 12th U-boat Flotilla. The massive, reinforced concrete U-boat pens have proved impractical to demolish and are now partly used as a cultural center for exhibitions.

Population (2016-01-01) ^[5]	78,623
• Density	2,800/km ² (7,200/sq mi)
Time zone	UTC+01:00 (CET)
• Summer (DST)	UTC+02:00 (CEST)
INSEE/Postal code	17300 ^[6] /17000
Elevation	0–28 m (0–92 ft) (avg. 4 m or 13 ft)

¹ French Land Register data, which excludes lakes, ponds, glaciers > 1 km² (0.386 sq mi or 7 acres) and river estuaries.

Kyoto History



Guillaume X to all in upheld the community. Guillaume was assisted by the city obtained major development of the

Plantagenet rule

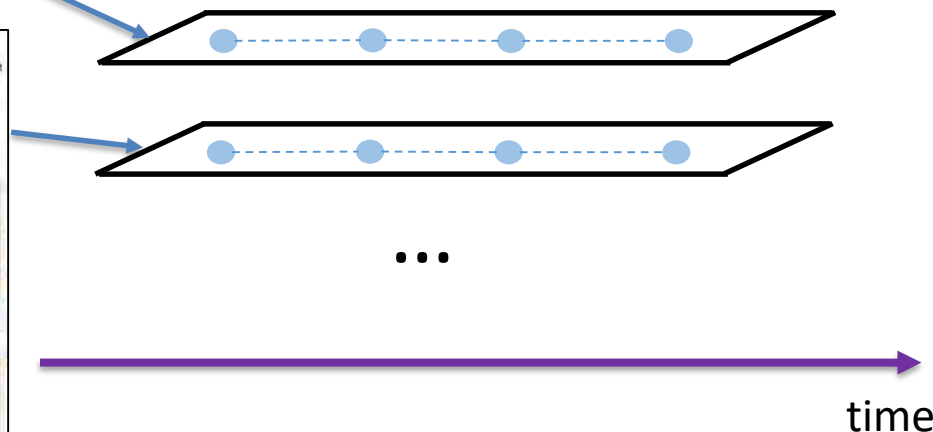
Eleanor married Henry putting La Rochelle During the Plantage



Edouard Manet: *Harbour at Bordeaux*, 1871



Rue Sainte-Catherine in 1905



実体のカテゴライズ Entity Categories

- 実体はおおむねカテゴリに分けられる(例: フランスの街、19世紀ドイツの作曲家、米国の野球選手など)。
Entities are usually grouped in categories (e.g., French cities, German composers in the 19th century, USA baseball players, etc.)
- 実体を整理し、分析する上でカテゴリに分けることは自然なことである。
Entity grouping is natural strategy for arranging and reasoning about entities
 - Wikipedia contains over 1 million categories

しかし、カテゴリ分けと実体の歴史は一緒に記述されていないことが多い。

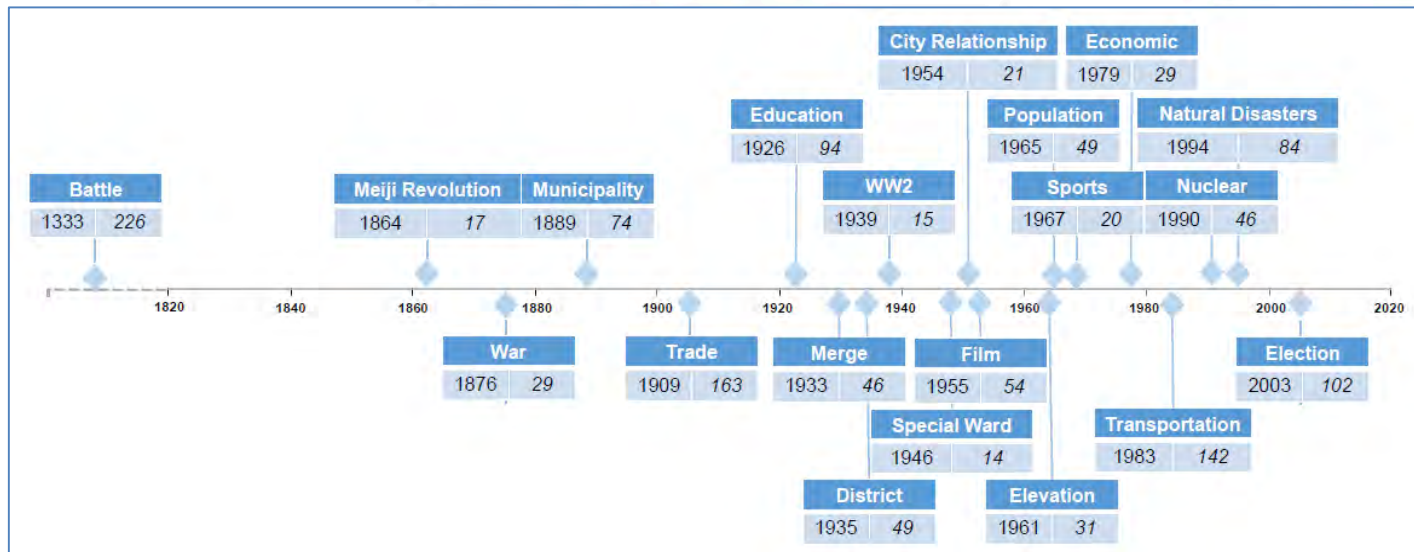
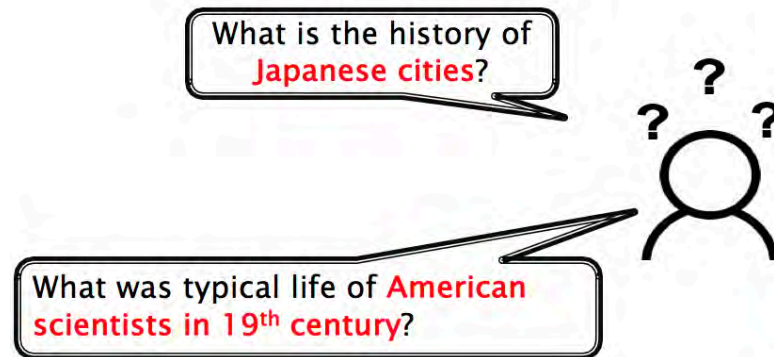
However, categorization and entity histories are usually not combined together

サブリサーチ①: 実体カテゴリの歴史の自動生成

SubResearch 1: Automatically Generating Histories of Entity Categories

ある実体の典型的な歴史はどのようなものか。

What was the typical history of a given entity category?

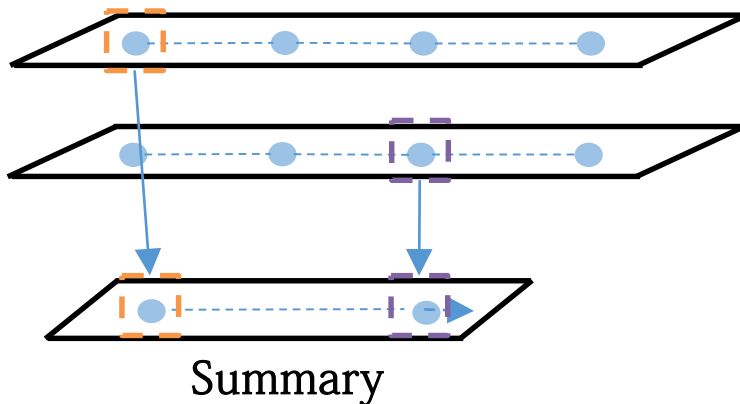


Example summary

模範ベース・プロトタイプベースの要約

Exemplar and Prototype-based Summary

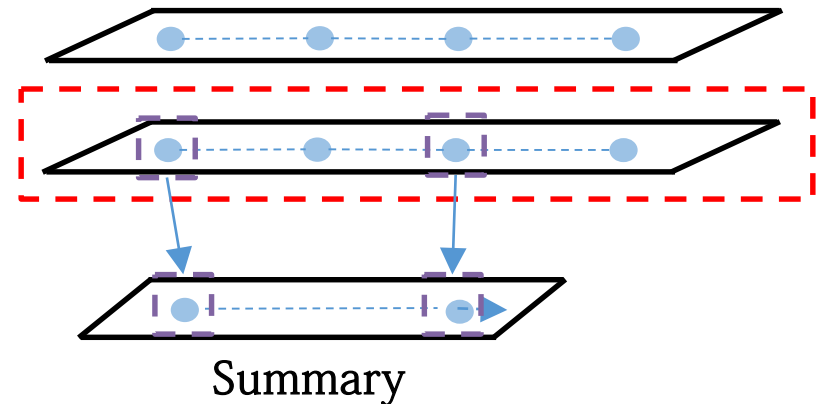
プロトタイプベースの要約 Prototype-based Summary



要約に含まれる出来事は別々の文書に由来する。

Summary events come from different entities

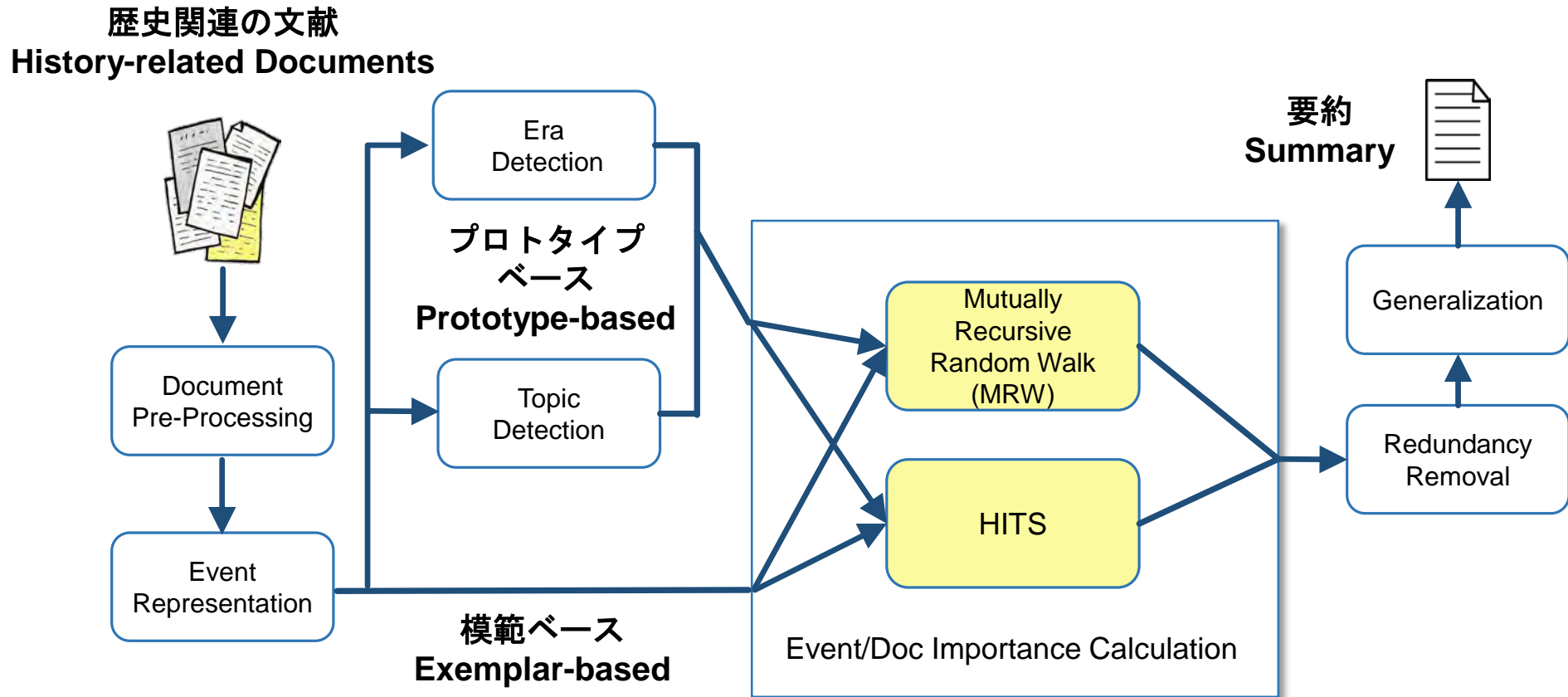
模範ベースの要約 Exemplar-based Summary



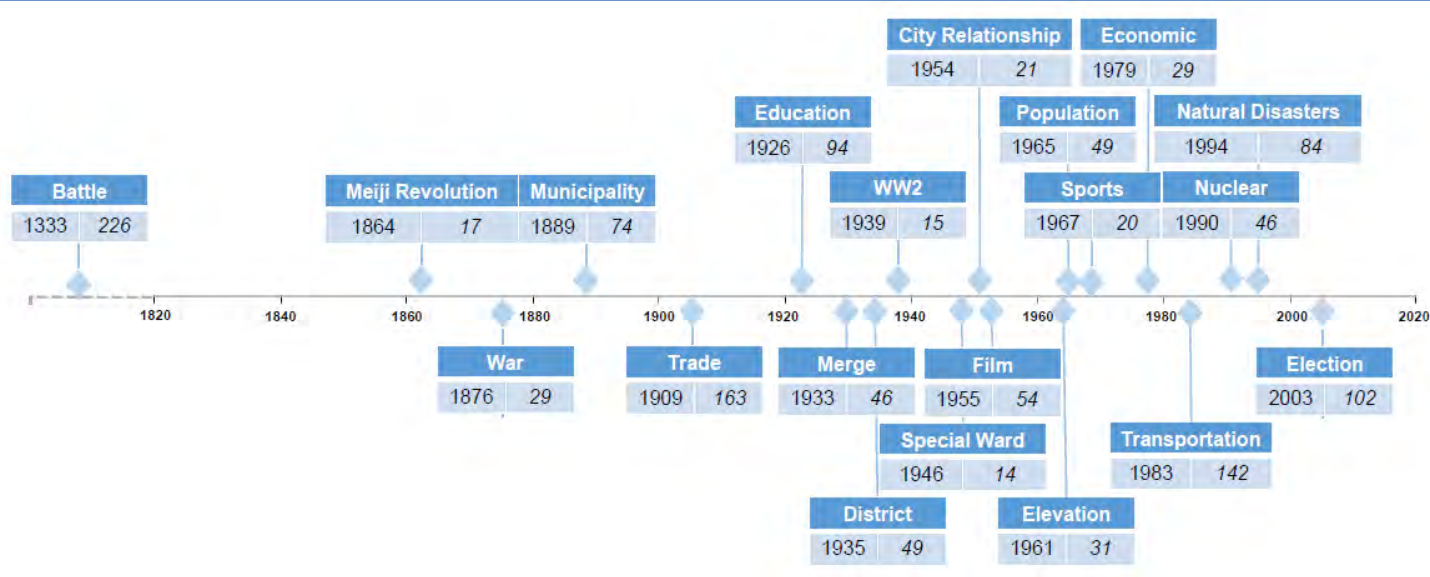
要約に含まれる全ての出来事は最も典型的な実体に由来する。

All summary events come from the most representative entity

システムの概要 System Overview



要約の例 Summary Example

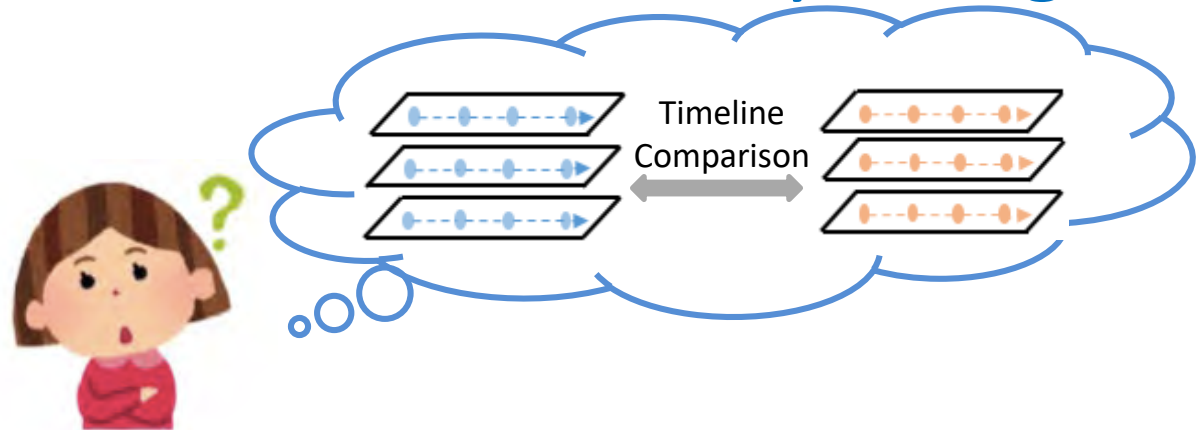


日本の550の市から
共有された歴史
Shared history of
550 cities in Japan

Event	Terms
Battle	battle, kamakura, fought, took, kumegawa, area, komaki, period, site, place, war, zenkunen, yasutsune, ultimately, ujigawa
Meiji Revolution	people, peasant, escape, christianity, another, rebellion, damage, raid, air, war, yokkaichi, went, weakened, toyotomi, subsequent
War	war, naval, japan, school, russojapanese, kiyohara, fujiwara, japanese, rebellion, navy, english, end, meiji, major, period
Municipality	system, within, municipality, establishment, modern, created, district, saitama, prefecture, restoration, gunma, town, creation, meiji
Trade	first, tea, made, tsuen, shop, service, held, festival, completed, yoshimitsu, world, waraji, uji, telephone, still
Education	school, established, confucian, high, william, welfare, vories, university, ueshiba, tsujido, teacher, taught, taizen, studies, science
Merge	merged, district, form, create, village, town, tkamachi, numakuma, nakaminato, incorporated, both, neighboring, urasaki, toyosu
District	takikawa, ebeotsu, becomes, continued, tend, village, district, hekikai, town, domain, began, area, period, yamagata, utashinai
WW2	training, center, military, imperial, navy, naval, japanese, industry, facility, built, army, air, production, nagoya, development
Special Ward	ward, became, tokyo, special, founded, city, district, former, shinj, shinagawa, sanbu, sanbe, nine, minamiadachi, metropolis
City Relationship	founded, city, relationship, established, ueno, sister, yamatotakada, wales, tkai, takaishi, sistercity, raised, nanao, mitaka, lomita
Film	year, story, film, festival, shibuya, sakura, record, narita, master, mai, every, appear, place, name, one
Elevation	elevated, city, status, ska, seba, village, town, surrendered, sekigawa, sashima, neighboring, matsumoto, kunitachi, kitaadachi
Population	public, housing, population, real, estate, development, trading, revenue, rapidly, rapid, debt, bubble, large, koku, construction
Sports	olympics, summer, hosted, host, winter, walk, sport, played, park, marathon, events, event, athletics, part, national
Economic	toyota, line, city, opened, nagoya, largest, economic, detroit, aichi, expanded, local, aircraft, became, plant, new
Transportation	expressway, road, line, junction, connected, station, tokaihokoriku, thoku, kaid, highway, established, train, tokyo, opened
Nuclear	fukushima, school, nuclear, evacuee, accident, city, status, student, public, problem, high, caused, rapid, housing, population
Natural Disasters	damage, earthquake, tsunami, thoku, suffered, caused, due, typhoon, rain, flooding, city, isewan, fishing, extensive, although
Election	mayor, motomiya, former, elected, plan, hall, first, party, ochiai, mayoral, kitamura, harue, city, office, woman

サブリサーチ②:実体カテゴリの歴史の自動比較

SubResearch 2: Automatically Generating Comparative Histories of Entity Categories



時系列順に整理された二つの文献コレクションを比較することで、
両者の共通点や違いを発見したいと望む利用者もいる。

Sometimes, users would like to compare two collections of timeline documents in order to discover commonalities and differences between them.

例1 Example 1

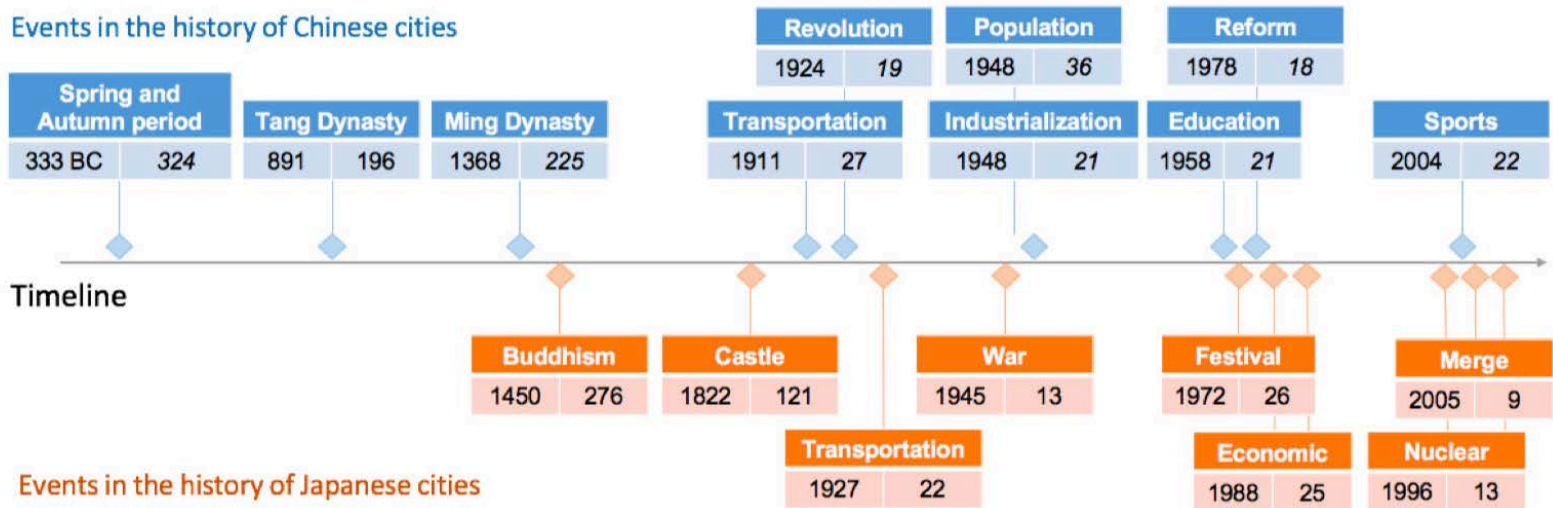
日本の都市の歴史と中国の都市の歴史の共通点は何か？
What are similarities between histories of Japanese cities and the histories of Chinese cities?

例2 Example 2

20世紀のヨーロッパの科学者と同世紀のアメリカの研究者の生活はおの
ように違うのか。
How different are lives of European scientists in the 20th century and those of American scientists at the same century?

例 Example

Events in the history of Chinese cities



Events in the history of Japanese cities

Table 6: Events in Chinese cities summary. For each event we show up to top 10 descriptive words due to space limit.

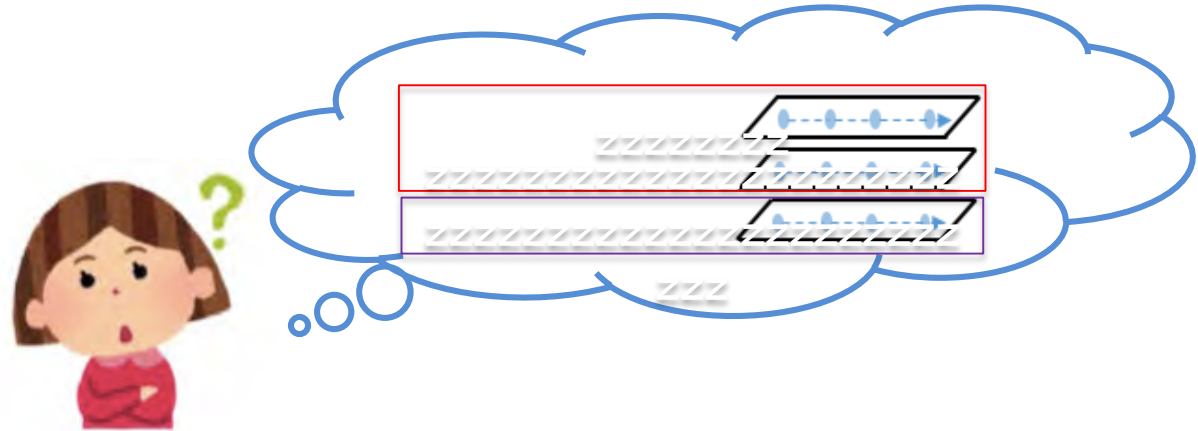
Event	Terms
Spring and Autumn period	qin, warring, chu, dynasty, capital conquered, zhou, subjugated, county, vassal
Tang Dynasty	emperor, capital, dynasty, city, kingdom tang, court, king, established, luoyang
Ming Dynasty	dynasty, kingdom, china, province, conquered ming, established, capital, empire, mongol
Transportation	railway, built, area, yangtze, kilometre line, north, completed, connecting, construction
Revolution	communist, rebellion, nationalist, revolt, army kmt, war, rebel, party, revolution
Industrialization	company, steel, iron, plant, installed oil, factory, production, cotton, mine
Population	population, million, per, estimated, urban reached, tripled, exceeded, xpc, increased
Education	university, medical, school, college, academy technology, harbin, institute, high, teachers
Reform	development, economic, growth, industry, investment port, zone, bank, billion, reform
Sports	game, host, asian, summer, sport games, fifa, expo, venue, olympics

Table 7: Events in Japanese cities summary. For each event we show up to top 10 descriptive words due to space limit.

Event	Terms
Buddhism	temple, period, shrine, year, buddhist history, area, site, built, nara
Castle	castle, built, constructed, building, shrine tower, canal, completed, reconstruction, construction
Transportation	line, station, railway, opened, rail main, route, train, shinkansen, service
War	war, air, world, raid, army japanese, bombing, naval, base, imperial
Festival	festival, held, event, anniversary, every matsuri, dance, annual, firework, celebrated
Economic	billion, gdp, population, million, employment city, industry, greater, increase, economy
Natural Disasters	earthquake, tsunami, damage, suffered, typhoon caused, struck, magnitude, killed, city
Nuclear	city, nuclear, evacuee, accident, fukushima student, public, problem, caused, rapid
Merge	district, merged, town, village, amalgamated city, mitsugi, numakuma, absorbed, incorporated
Election	mayor, election, elected, city, party mayoral, assembly, government, politics, succeeded

サブリサーチ③:歴史に基づく実体カテゴリの自動生成

SubResearch 3: Automatically Generating Entity Categories based on their Histories



実体を歴史上の関連性に基づいて大きなカテゴリーに分類することを望む利用者もいる。

Sometimes, users may want to group entities in large categories according to their historical correspondence

例1 Example 1

日本の都市の歴史的成り立ちの類似性に基づいたサブカテゴリを作るにはどうしたら良いか。

How to find meaningful sub-categories of Japanese cities based on the similarities of their historical developments?

例2 Example 2

20世紀ヨーロッパの科学者を、伝記の共通パターンにより分類するにはどうしたら良いか。

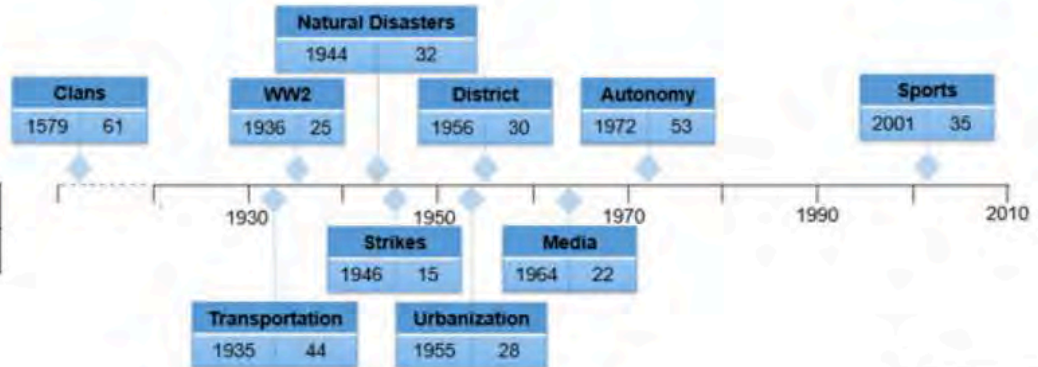
How to divide European scientists alive in the 20th century based on the common patterns in their biographies?

例 Example

Subcategory-1



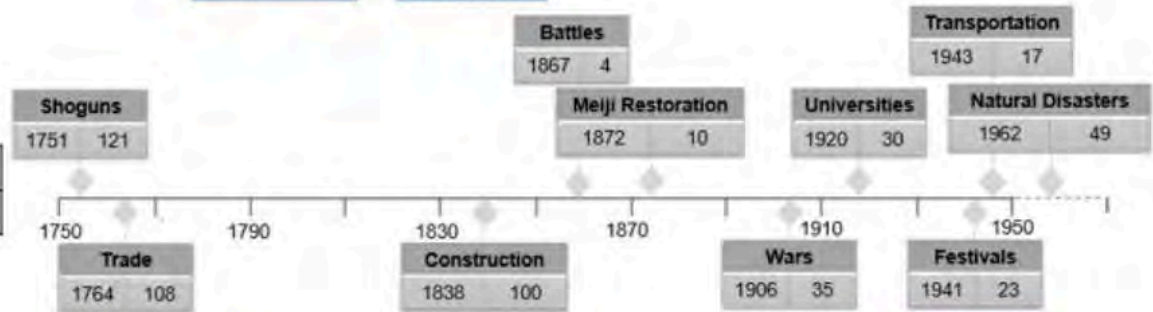
Exemplar	Matsubara
Members	Tokyo, ... , Nagoya (461)



Subcategory-2



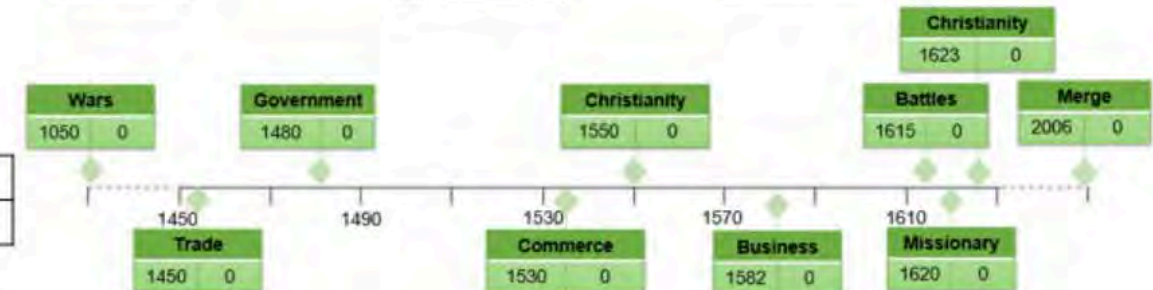
Exemplar	Dazaifu
Members	Kyoto, ... , Yamato (66)



Subcategory-3



Exemplar	Oshu
Members	Sakai, Nagasaki (2)



面白さに基づくアーカイブ収集: ニュースアー
カイブから面白そうなコンテンツを推薦する
INTERESTINGNESS-ORIENTED ARCHIVAL
RETRIEVAL: RECOMMENDING INTERESTING
CONTENT FROM NEWS ARCHIVES

アジェンダ Talk Schedule

1. はじめに Introduction
2. 異なる時代における類似物(時間的アナログ)の検出
Temporal Analog Detection
 - 時間を超えた類似性の説明
Across-time Term Similarity Explanation
3. 時間を超えた比較の要約
Across-time Comparative Summarization
4. 歴史に基づく実体のグループ化と要約
History-based Entity Summarization
5. 面白さに基づくアーカイブからの情報検索
Interestingness-oriented Archival Retrieval
6. 現在との関連性を志向する文献検索に向けて
Towards Present-relevance Oriented Document Search
7. 結び Conclusions

アイデア Idea

- 平均的な利用者にとって、アーカイブの利用はあまり一般的ではない。

The usage of archives is not very popular for average users

- 平均的な利用者にとってアーカイブをより魅力的に見せるには、**コンテンツを推薦**する機能が便利ではないかと思われる。

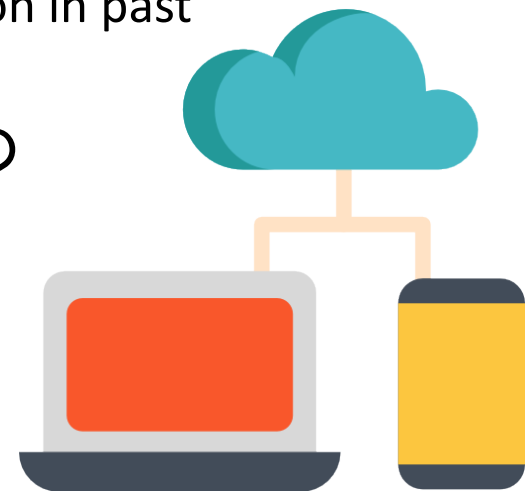
Special kind of **content recommendation** could be useful for archives to increase their attractiveness for average users

- アーカイブにアクセスした人は、過去の文献から**面白い**情報を発見し、文化資産を鑑賞することができる。

visitors could discover “**interesting**” information in past documents to appreciate our heritage

- 面白さを志向するメカニズムは、アーカイブの検索エンジンに組み込むことができる。

Interestingness-oriented mechanisms can be incorporated into search engines operating on archives



コンテンツの面白さの多面性

Varying Aspects of Content Interestingness

- 面白いコンテンツとはどのようなものか？

What can be considered as interesting content?

- 利用者の関心や趣味に関連するコンテンツ
Content related to user interests and hobbies
- 現在起きている重要な出来事に関連するコンテンツ
Content relevant to ongoing important events
- 魅力的なストーリーを伝えるコンテンツ
Content describing captivating stories
- 重要な教訓を教えてくれるコンテンツ
Content that can teach us valuable lessons
- etc.

面白さ＝意外性
Interestingness = Unexpectedness



時間の経過に起因する驚き
Surprise resulting from passage of time

例 Example

- アイスカッターとは何か？ What is Ice cutter ?

1800s

「アイスカッター」は、冷蔵庫の普及以前、氷を凍った池や川から切り出していた頃には人あるいは職業を指す言葉であった。

Ice cutter was a person (or a job). Before the widespread use of refrigerators, ice was cut from frozen lakes and rivers by men.

now

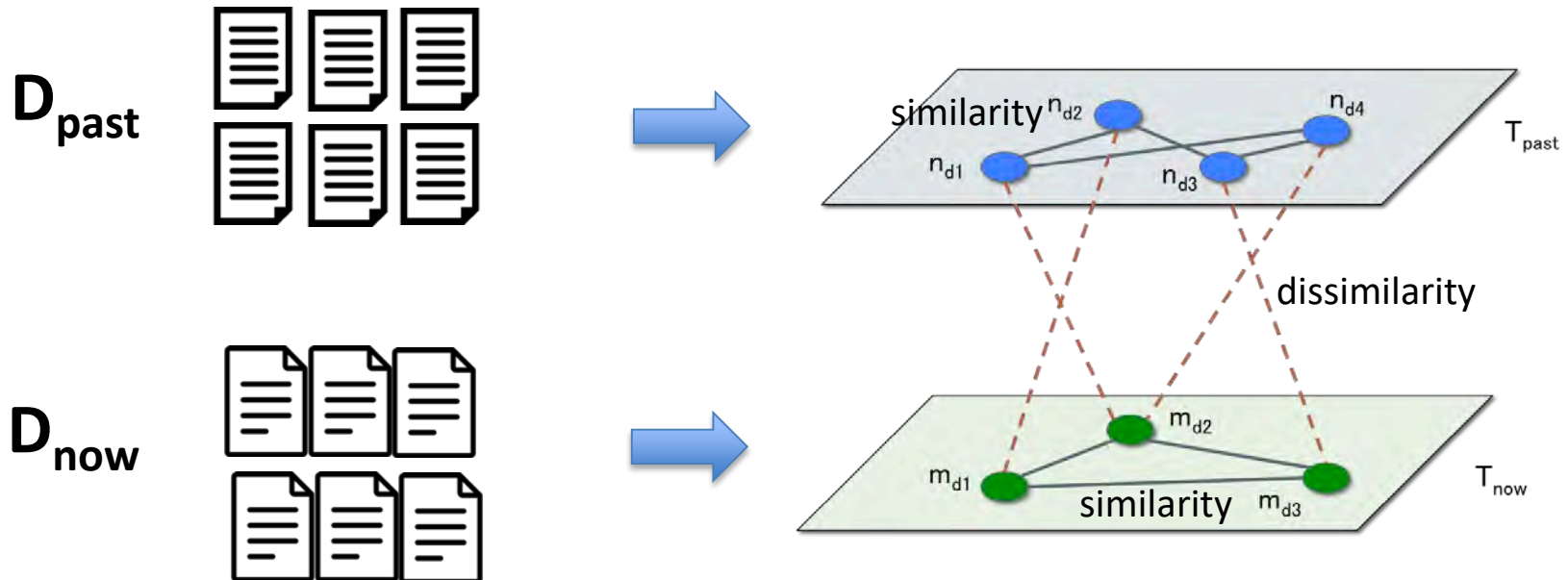
相互再帰的ランダムウォーク(1/2)

Mutually Recursive Random Walk (MRRW) Method (1/2)

仮説1: 過去のコンテンツは、現在ありふれているコンテンツと異なっていて過去のコンテンツの中にあってはありふれている時に面白いと感じられる。

Hypothesis 1: Past content is interesting if it is different from content common in present and if it was common in the past

- 相互再帰的ランダムウォークは①過去における重要性、②真新しさと物珍しさを反映すると予測される。
- MRRW is expected to reflect:
 - *past importance*
 - *novelty and unfamiliarity*



相互再帰的ランダムウォーク(2/2)

Mutually Recursive Random Walk (MRRW) Method (2/2)

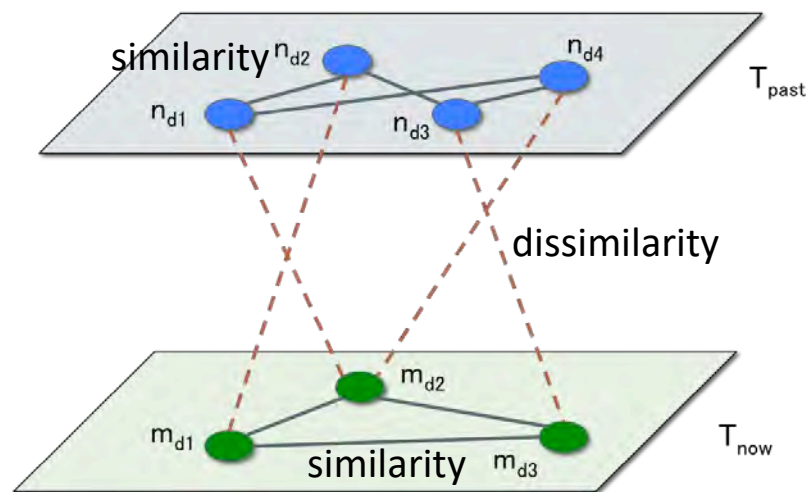
- ノードの値は以下の式によって補強される:
- Reinforced by:

$$\begin{cases} S_P = (1 - \alpha)S_P + \alpha \cdot L_{PP}L_{PN}S_N \\ S_N = (1 - \alpha)S_N + \alpha \cdot L_{NN}L_{NP}S_P \end{cases}$$

当初の重要性
Initial
importance

値はレイヤー内及びレイヤー間の伝播によりアップデートされる

The score will be updated by within- and between-layer propagation



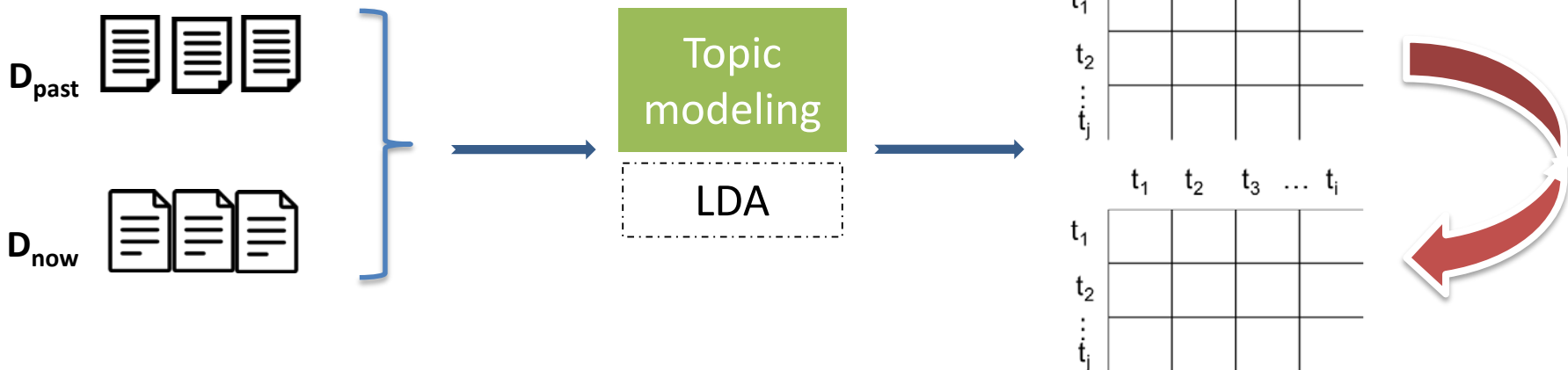
トピック共起法における分岐

Divergence in Topic Co-occurrence Method

仮説2: 過去のコンテンツは、現在はあまり一般的ではない複数のトピックの組み合わせである時に面白いと感じられる。

Hypothesis 2: Past content is interesting if it is about the common combination of topics which are uncommon in present

- 相互再帰的ランダムウォークは意外性と驚きを反映すると予測される。
- Expected to reflect:
 - *Unexpectedness and surprise*



トピック共起の比較

comparison of topic co-occurrences

実験方法

Experimental Settings



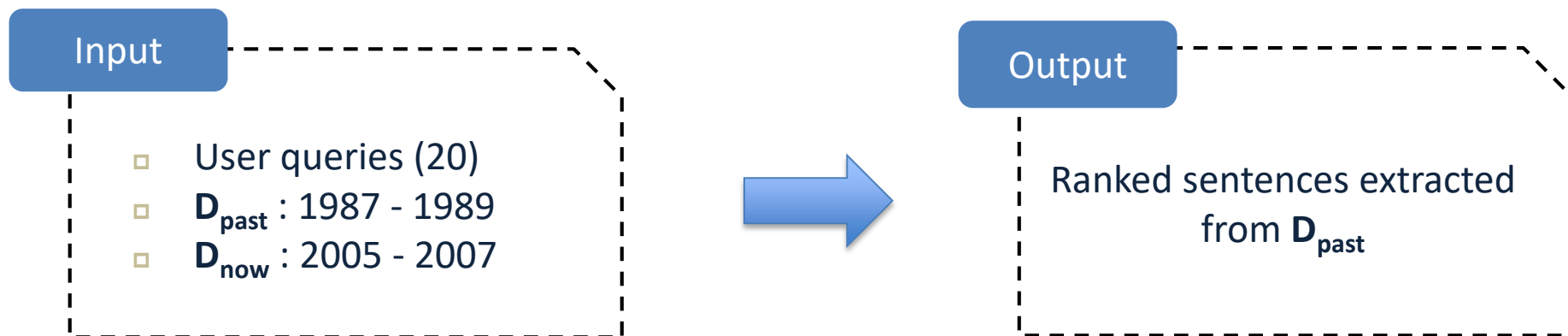
New York Timesの記事アーカイブを使用する:
New York Time corpus used as target archive:

- About 1.7mln articles published between 1987-2007

クラウドソーシング (Figure Eight社) を利用し、1文あたり5つの判断基準により評価を行う。(各検索結果の上位15件の記事から1,800個の文を使用。)

Evaluation by crowdsourcing (figure-eight) with 5 judgments per sentence (1,800 sentences from 15 top-ranked results by each method)

Category	Query
Economy	Currency, Economy, Trade, Market
Places	Japan, Florida, Los Angeles, New York
Politics	Election, President, Nomination, Poll
Sports	Basketball, Team, Olympics, Sport
Technology	Machine, Computer, Plane, Technology



結果 Results

平均逆順位 (MRR) スコア (高い程良い)

Mean Reciprocal Rank (MRR) scores (the higher the better)

	Economy	Places	Politics	Sports	Tech	<i>Average</i>
Random	47.50	10.49	35.00	23.96	27.08	28.81
Centroid	41.67	43.75	10.83	22.62	25.83	28.94
TF-IDF+MRRW	19.58	47.92	25.00	64.58	75.00	46.42
Co-occurrence	5.20	18.94	8.33	44.58	70.83	29.58

Method

Example Sentence

MRRW

LEAD: The chairman of the Florida Seminole tribe was acquitted today of state charges of killing an endangered Florida panther.

Divergence
in Topic Co-
occurrence

Of the 715 apartment fires in Moscow last month, 90 were blamed on exploding television sets, a statistic the Soviet press has viewed as an alarming commentary on soviet technology.

現在との関連性を志向する
文献検索に向けて
**TOWARDS PRESENT-RELEVANCE
ORIENTED DOCUMENT SEARCH**

アジェンダ Talk Schedule

1. はじめに Introduction
2. 異なる時代における類似物(時間的アナログ)の検出
Temporal Analog Detection
 - 時間を超えた類似性の説明
Across-time Term Similarity Explanation
3. 時間を超えた比較の要約
Across-time Comparative Summarization
4. 歴史に基づく実体のグループ化と要約
History-based Entity Summarization
5. 面白さに基づくアーカイブからの情報検索
Interestingness-oriented Archival Retrieval
6. 現在との関連性を志向する文献検索に向けて
Towards Present-relevance Oriented Document Search
7. 結び Conclusions

アーカイブの検索 Archival Search

- 専門家(例: 歴史家)は自分が何を探したいかを知っている。

Professionals (e.g., historians) know what they want to find

- 検索意図が明確に定義されている。
They have well-defined search intent

Clinton China trade conflict 🔍



アーカイブの検索 Archival Search

- 一般的な利用者は検索意図があまり定義されていないかもしれない。

General users may have **less defined search intent**


- ー 現在に関連する文献を探そうとするかもしれない。
例1: 現在の出来事の背景となった出来事/人物/場所
例2: 現在の出来事に類似した出来事/人物/場所
They may try to find documents related to the present.


Ex1. Events/figures/places that are background for present events

Ex2. Events/figures/places similar to present events

- 時間の経過により、過去の文献の現在との関連度にはばらつきがある。

Due to time passage, past documents have **varying level of relation to the present**

US trade conflict 

US trade president 



例 Example

1988年の文献A Document A from 1988 [3]

Hale Stores, has moved to coordinate its major businesses by naming Ira Neimark, chairman and chief executive of its Bergdorf Goodman subsidiary, to the additional post of vice president of merchandise development for the Neiman-Marcus Group.

1987年の文献B Document B from 1987 [4]

When it comes to merchandising, Donald J. Trump and Mody Dioum could not be much farther apart. But they agree on one thing: The holiday season was hard on Fifth Avenue street peddlers. And as the avenue stepped back to normal yesterday, it appeared that the city's recent crackdown on merchandise peddlers was still having an effect.

[3] <https://www.nytimes.com/1988/04/12/business/credit-markets-neiman-shifts-key-executives.html>

[4] <https://www.nytimes.com/1987/01/06/nyregion/anti-peddler-drive-pleases-fifth-ave-merchants.html>

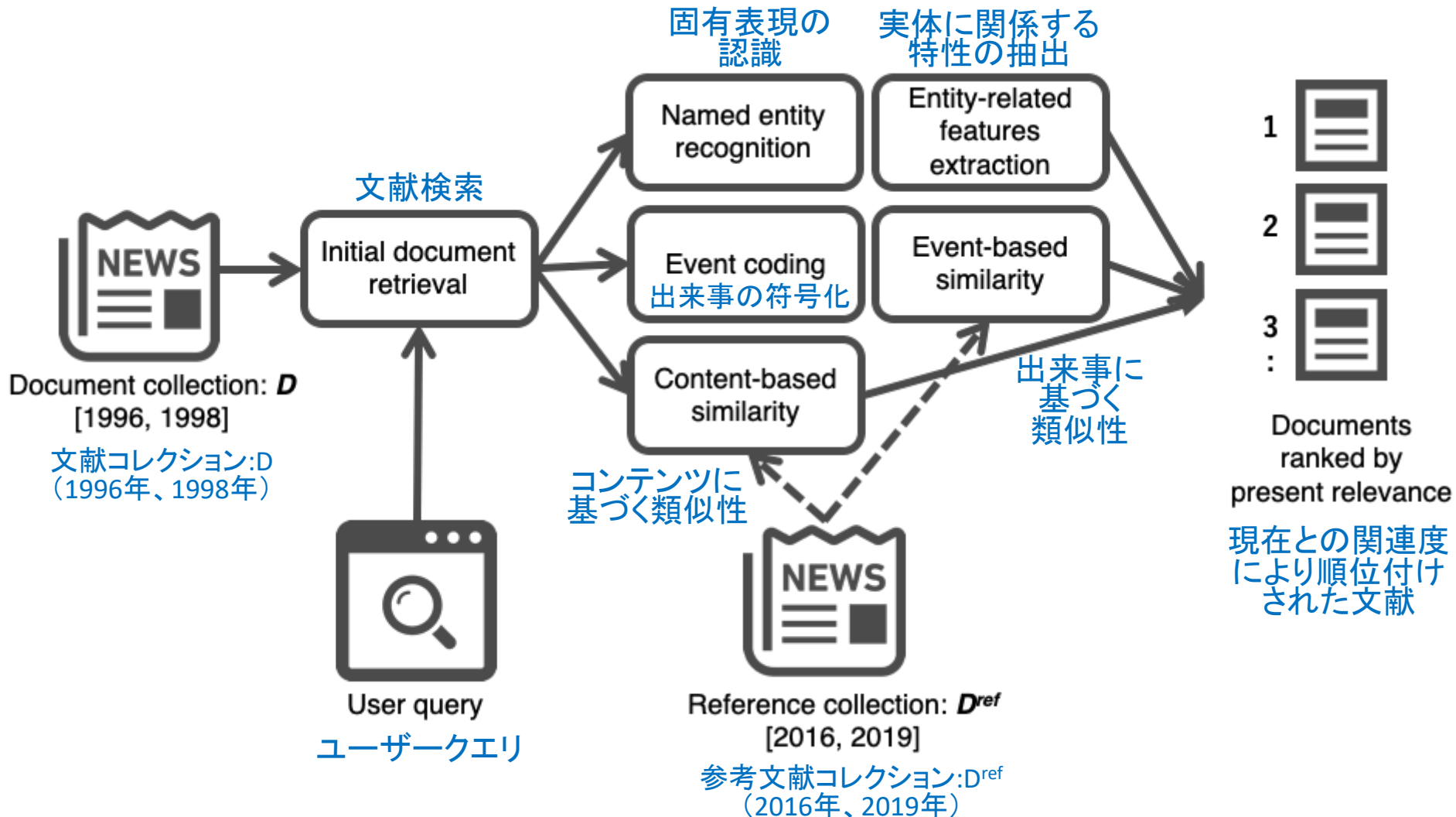
どうすれば過去の文献から現在と関係するものを見つけられるでしょうか？

How can we find past documents
which are related to present?

アプローチの考え方 Approach Idea

- *現在との関連性*は、多くの分野にまたがる複雑な構造である。
Present relevance is a complex construct that involves many aspects
- 一つの技術や機能により*現在との関連性*を把握することは難しい。
We believe that it is difficult to capture *present relevance* with a single technique or feature
- したがって、文献と*現在との関連*を示すことができるかもしれない一連の機能を次に示す。
We then propose a range of features that likely indicate present relevance of documents

アプローチの概要 Approach Overview



結び & 今後の課題 Conclusions & Future Work

情報アクセスの新しい手法と長期間に渡るニュースアーカイブからのナレッジ抽出 Novel Ways of Information Access and Knowledge Extraction from Long-term News Archives

1. はじめに Introduction
2. 異なる時代における類似物(時間的アナログ)の検出
Temporal Analog Detection
 - ー 時間を超えた類似性の説明
Across-time Term Similarity Explanation
3. 時間を超えた比較の要約
Across-time Comparative Summarization
4. 歴史に基づく実体のグループ化と要約
History-based Entity Summarization
5. 面白さに基づくアーカイブからの情報検索
Interestingness-oriented Archival Retrieval
6. 現在との関連性を志向する文献検索に向けて
Towards Present-relevance Oriented Document Search (Future)
7. 語義変化の分析
Word Meaning Evolution Analysis (Future)

ご清聴ありがとうございました。

Thank you for the attention

Q & A