コンパラブルな新聞記事からの固有表現の発見

新山 祐介

関根 聡

Computer Science Department

New York University

研究の背景

- 固有表現抽出：
  - 文脈、あるいは字句的な知識をもちいて、文章中に現れる固有表現をタグ付けする。
  - コーパスや辞書からあらかじめ学習（発見）しておく方法が主流。

研究の背景

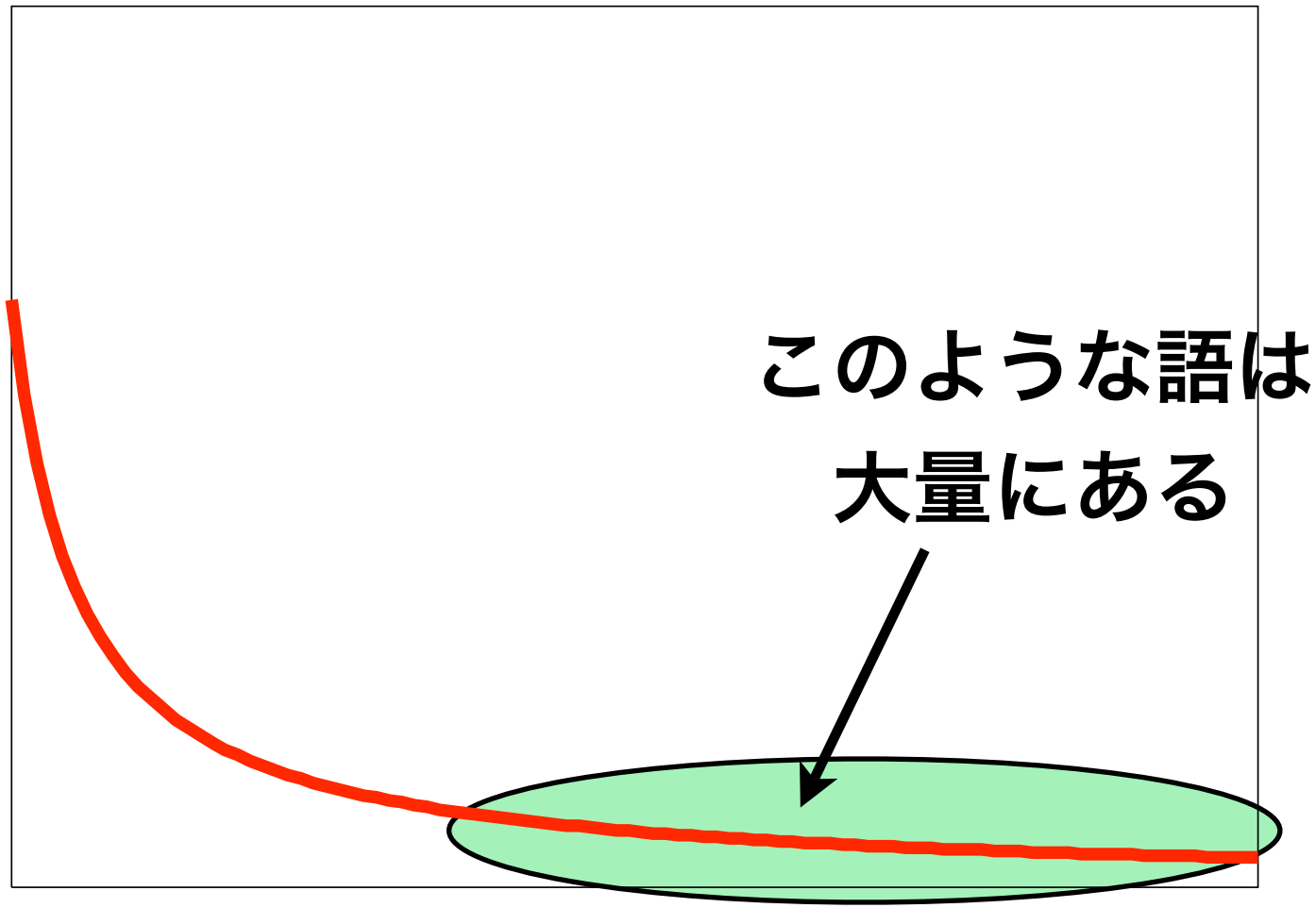
- 字句的な知識を構築する方法：
  - 機械学習を用いてコーパスから字句的な規則を自動的に学習する。
    - [Strzalkowski 1996]
    - [Collins 1999]
    - [Yangarber 2002]
  - 辞書を人手で構築する。

疎なデータの問題

- いかにして頻度の少ない単語の字句的な情報を学習するか？
  - 例. 「徳島 (Tokushima)」という地名は1995年の New York Times (14万記事) に１回しか現れない。

疎なデータの問題

- Zipfの法則



疎なデータの問題

- 頻度 (DF) の少ない単語がすべて固有表現であるとは限らない。
  - まれにしか使われない普通名詞
    - 例. `misgovernment`
  - タイプミス
    - 例. `uinversit`

頻度の少ない表現を正確に学習することが固有表現抽出の精度向上には不可欠。

コンパラブルコーパスを使う

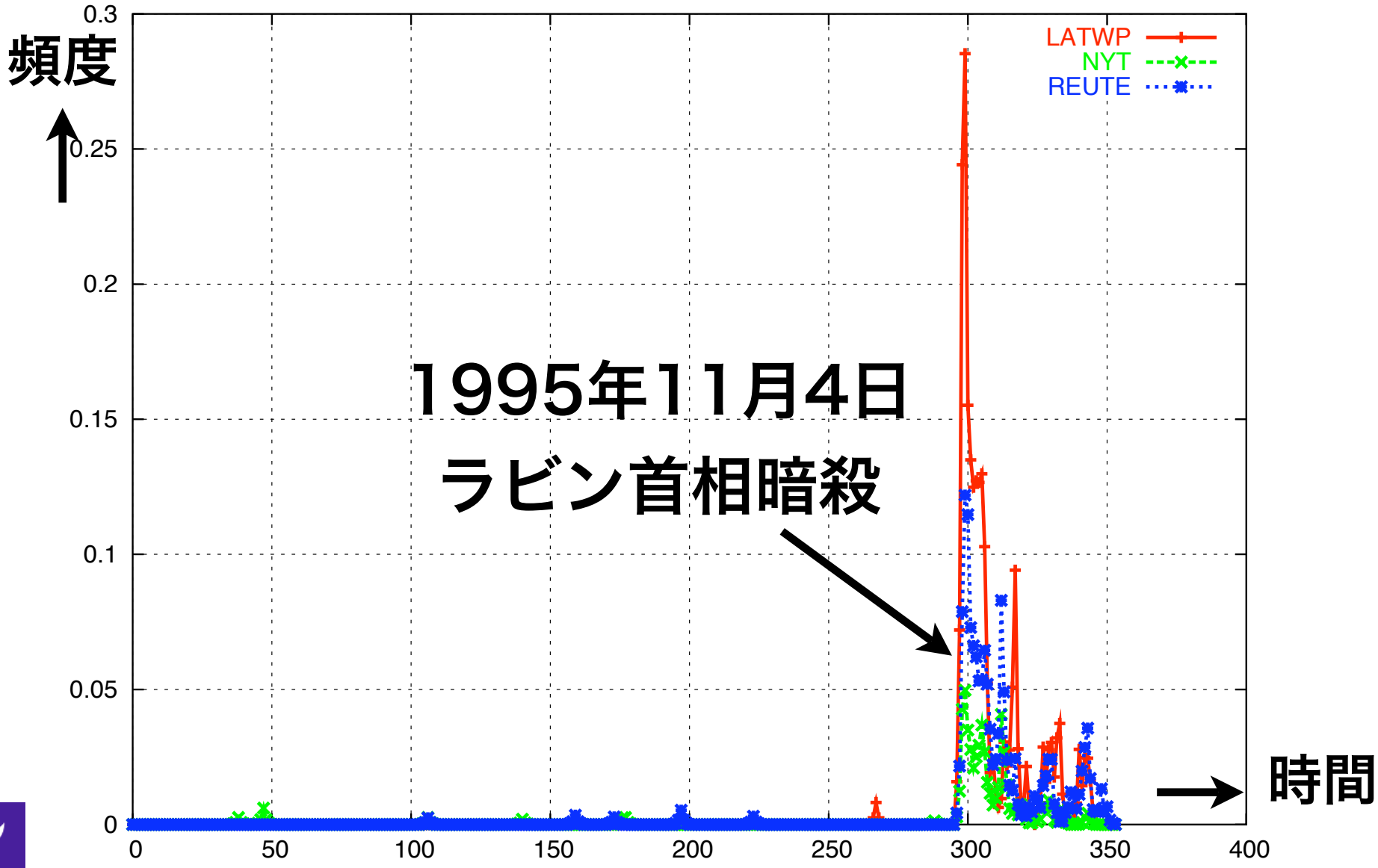
- コンパラブルコーパスとは
  - ほぼ同じ内容をもつ別々の文書どうしが対応づけられたコーパス
    - 例. 同一の英語文書の複数の日本語訳
    - 例. 同一の事件の複数人による記述
  - 文書レベルでの言い換え
  - ほとんどの固有表現は言い換えしにくい。(例外. New York → Big Apple)

コンパラブルコーパスを使う

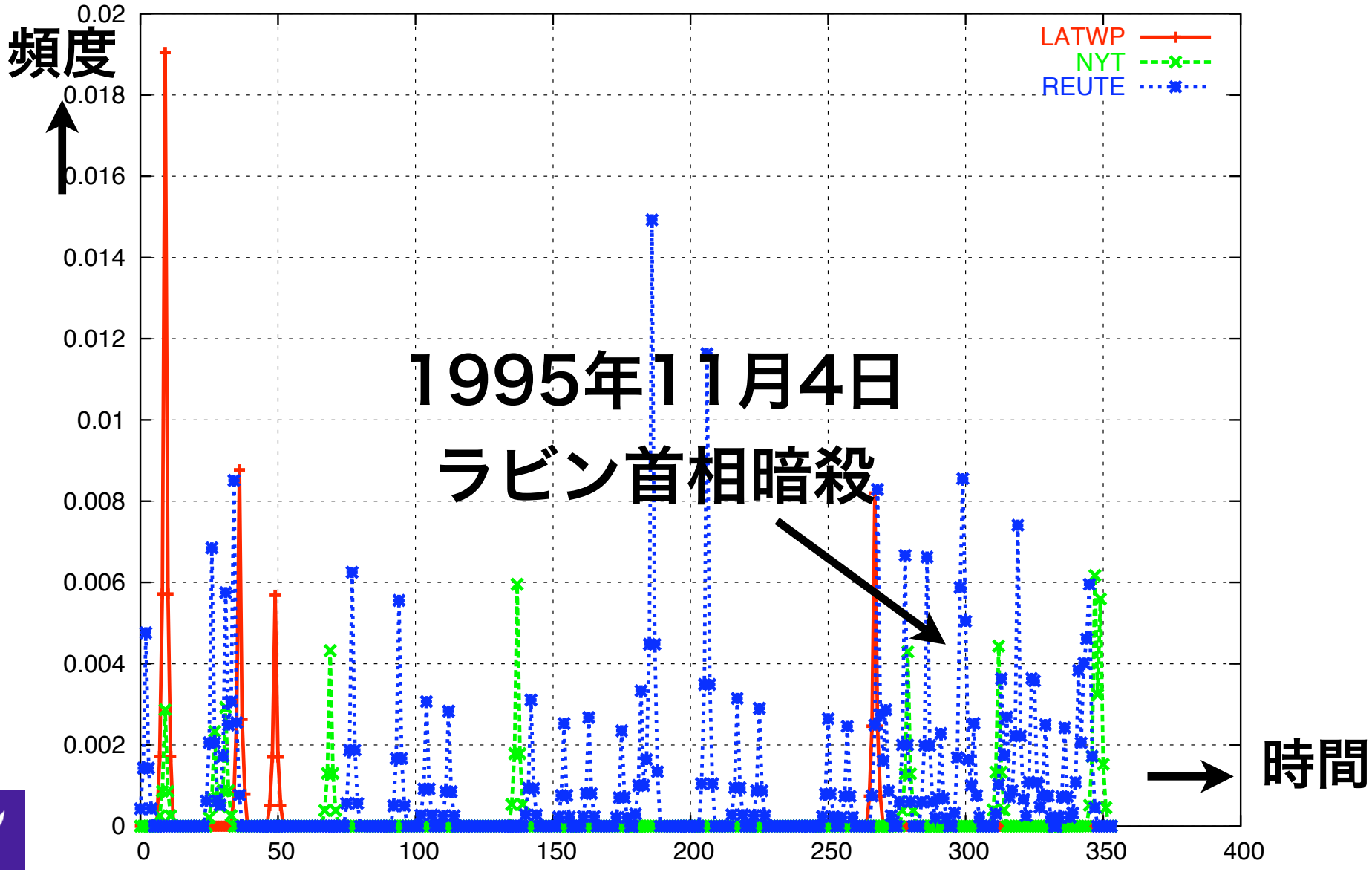
- 複数の異なる新聞社による同一日の新聞記事は、コンパラブルな記事を多く含んでいる。
- たとえ異なる新聞社の間でも、固有表現は同一日に等しく現れるのではないか？

ある単語が固有表現ならば、
その出現頻度は複数の新聞で同期するはず。

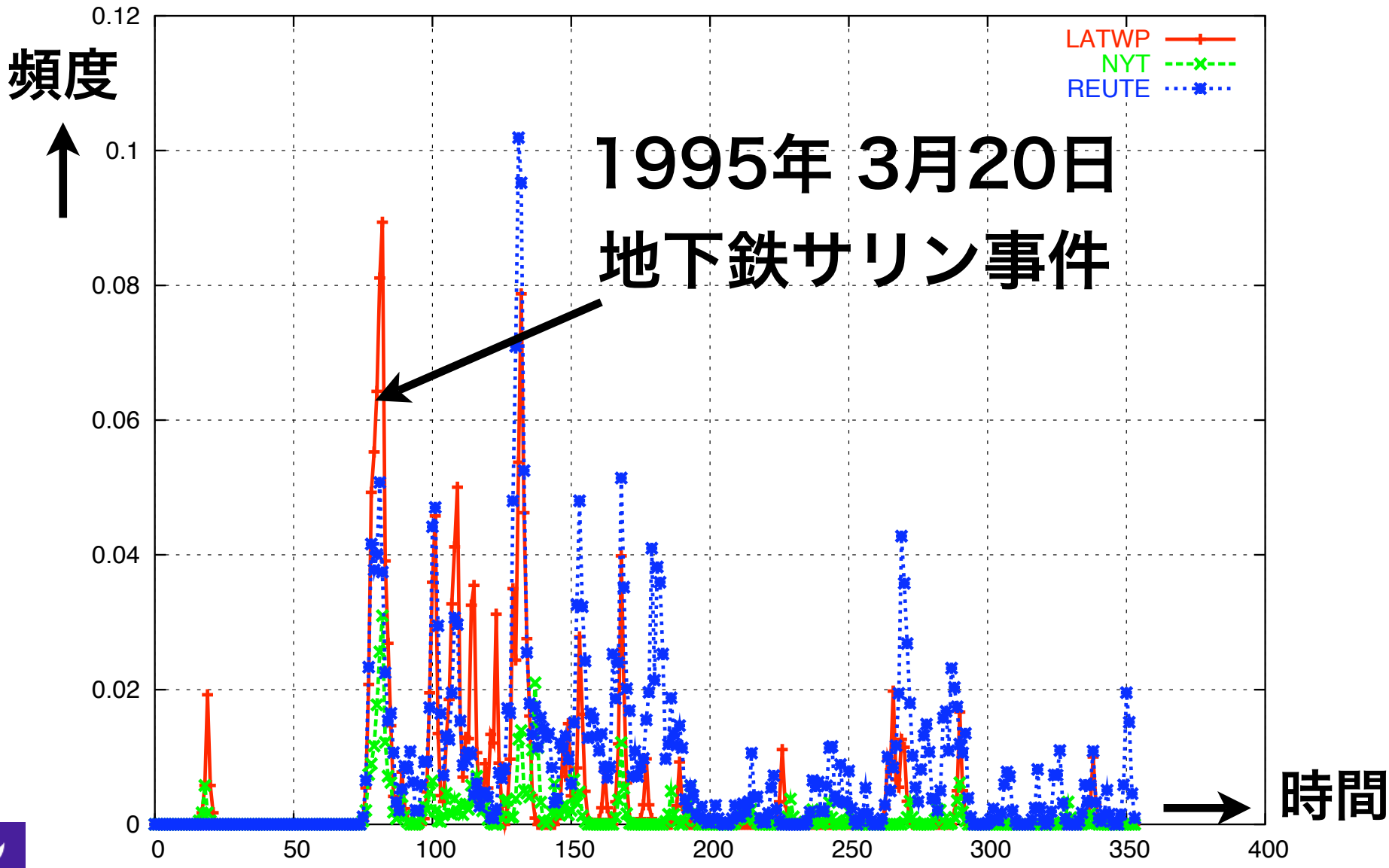
“Yigal” の出現頻度 (95年)



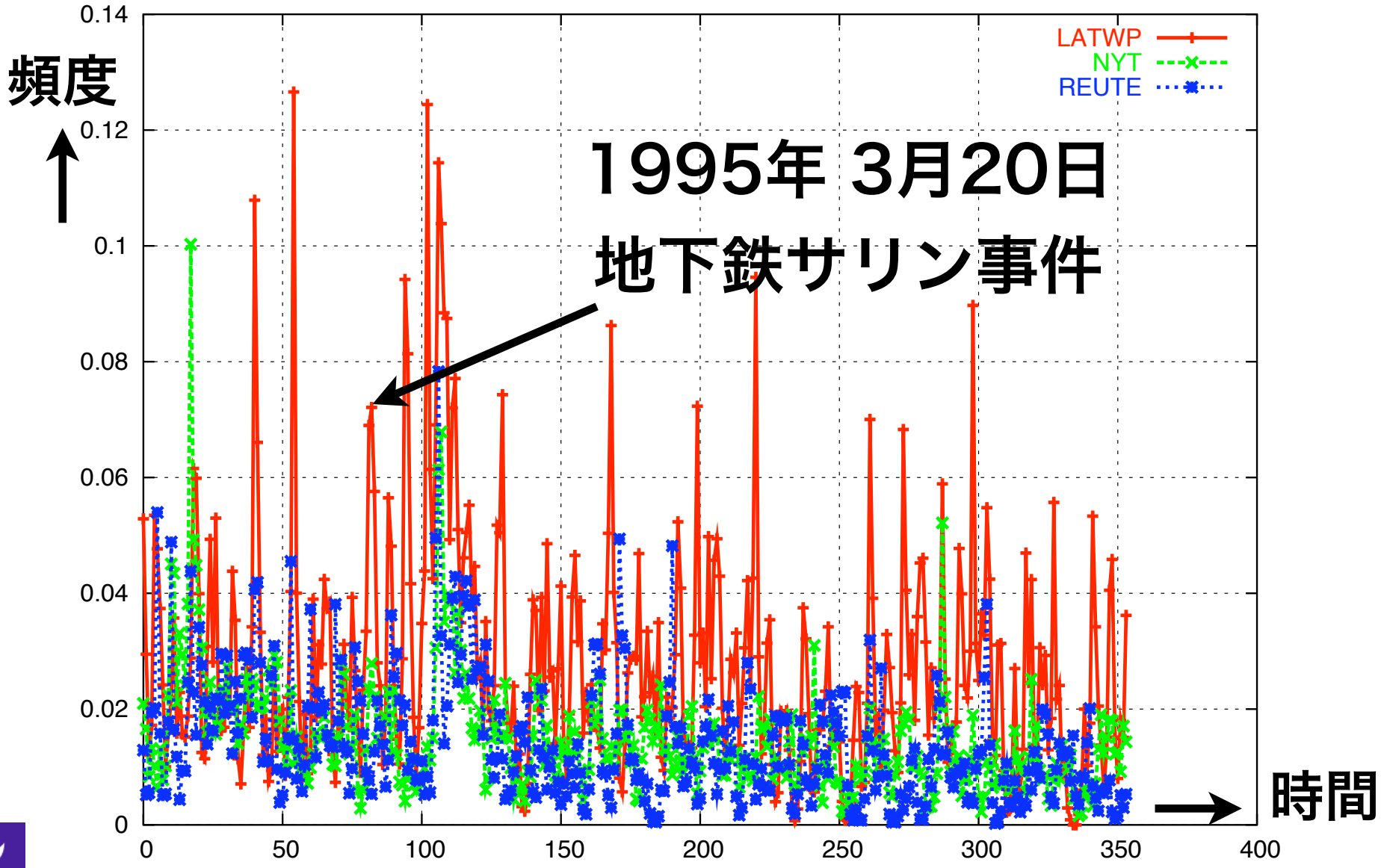
“assasination” の頻度



“Shoko” の頻度 (95年)



“terror” の頻度


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
頻度
LATWP
NYT
REUTE
1995年 3月20日
地下鉄サリン事件
0
50
100
150
200
250
300
350
400
時間

単語の“固有表現らしさ”の測定

- ある単語 w の時系列的な分布の類似度を測定する。
  - ２社の新聞記事 A, B を日付ごとに揃え、各日付における単語 w のドキュメント頻度 (DF) を並べたベクトル F_AとF_Bを作成する。
  - 類似度 = ２つのベクトルの *cosine*

単語の“固有表現らしさ”の測定

２つの新聞記事 A, B から作成したベクトル：

$$F_A(w) = [df_A(w,1), ..., df_A(w,365)]$$

$$F_B(w) = [df_B(w,1), ..., df_B(w,365)]$$

類似度：

$$sim(w) = \frac{F_A(w) \cdot F_B(w)}{|F_A(w)||F_B(w)|}$$

実験環境

- 1995年の Los Angels Times および Reuters の記事１年分（計13万記事）を使用。
  - 双方の新聞に含まれている単語で、ドキュメント頻度が100以下のものを抽出。
  - ごく簡単なトーカナイザのみを使用。
  - 分布の類似度を単語のスコアとしてランキングする。

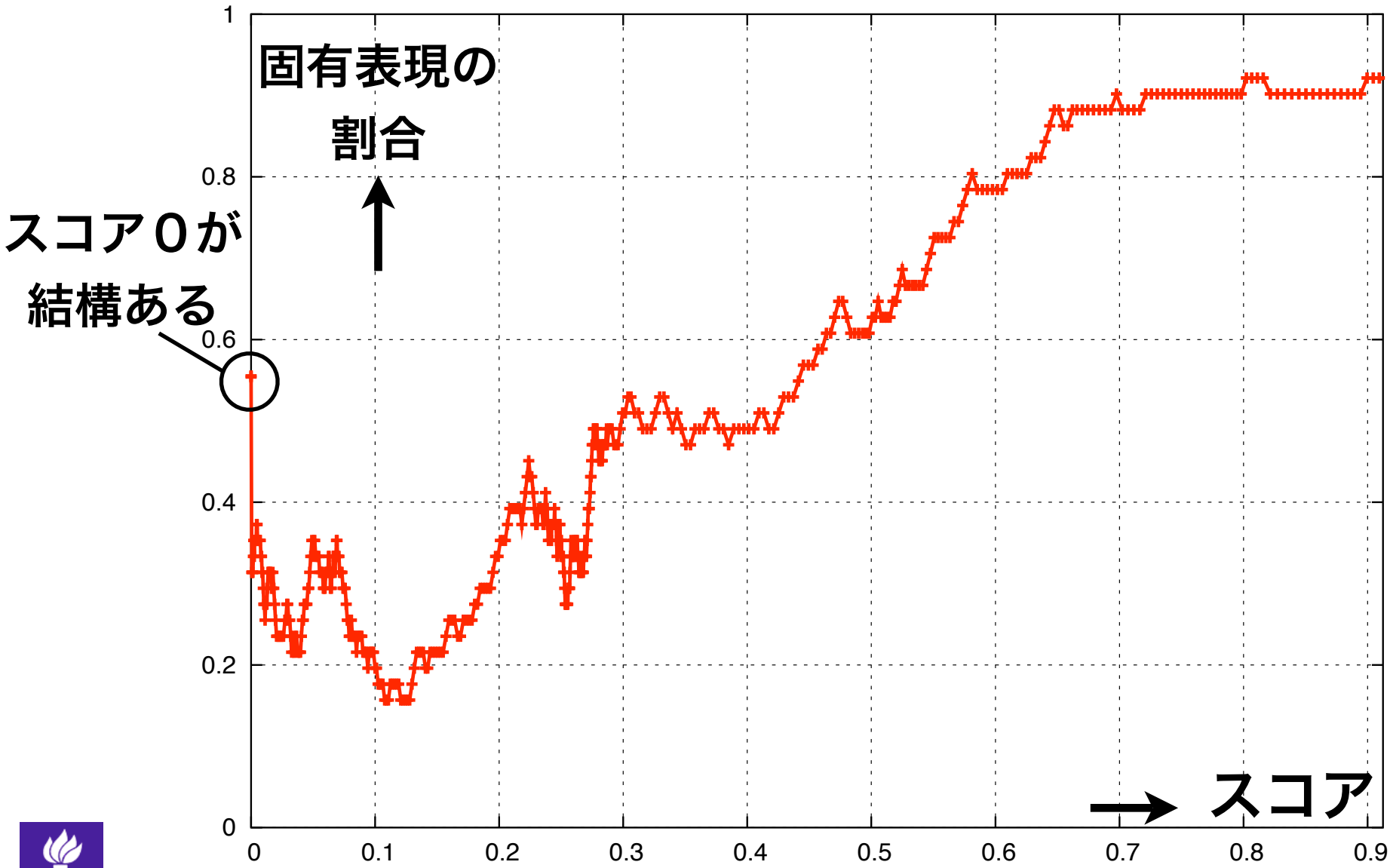
評価方法

- 無作為に抽出した単語 966個を IREXの基準に従い、PERSON などの４つのカテゴリに人手で分類。
- これらの単語を本手法のスコアによってランキングする。
- 高ランクの単語にどれくらいの割合で固有表現が含まれていたかを測定。

得られた単語

スコア	単語	種類	頻度
1.000	sykesville	LOCATION	3
1.000	puellet	PERSON	2
1.000	deschodt	PERSON	2
0.912	rechner	PERSON	8
0.894	minutello	PERSON	3
0.894	peresic	PERSON	10
0.843	yigal	PERSON	90
0.817	pauli	PERSON	6
…	…	…	…
0.679	australopithecus	NE?	9
0.655	earvin	PERSON	11
0.616	alpirez	PERSON	97
0.578	edt	NE?	77
0.572	moskowitz	PERSON	13
0.535	modernizations	noun	7

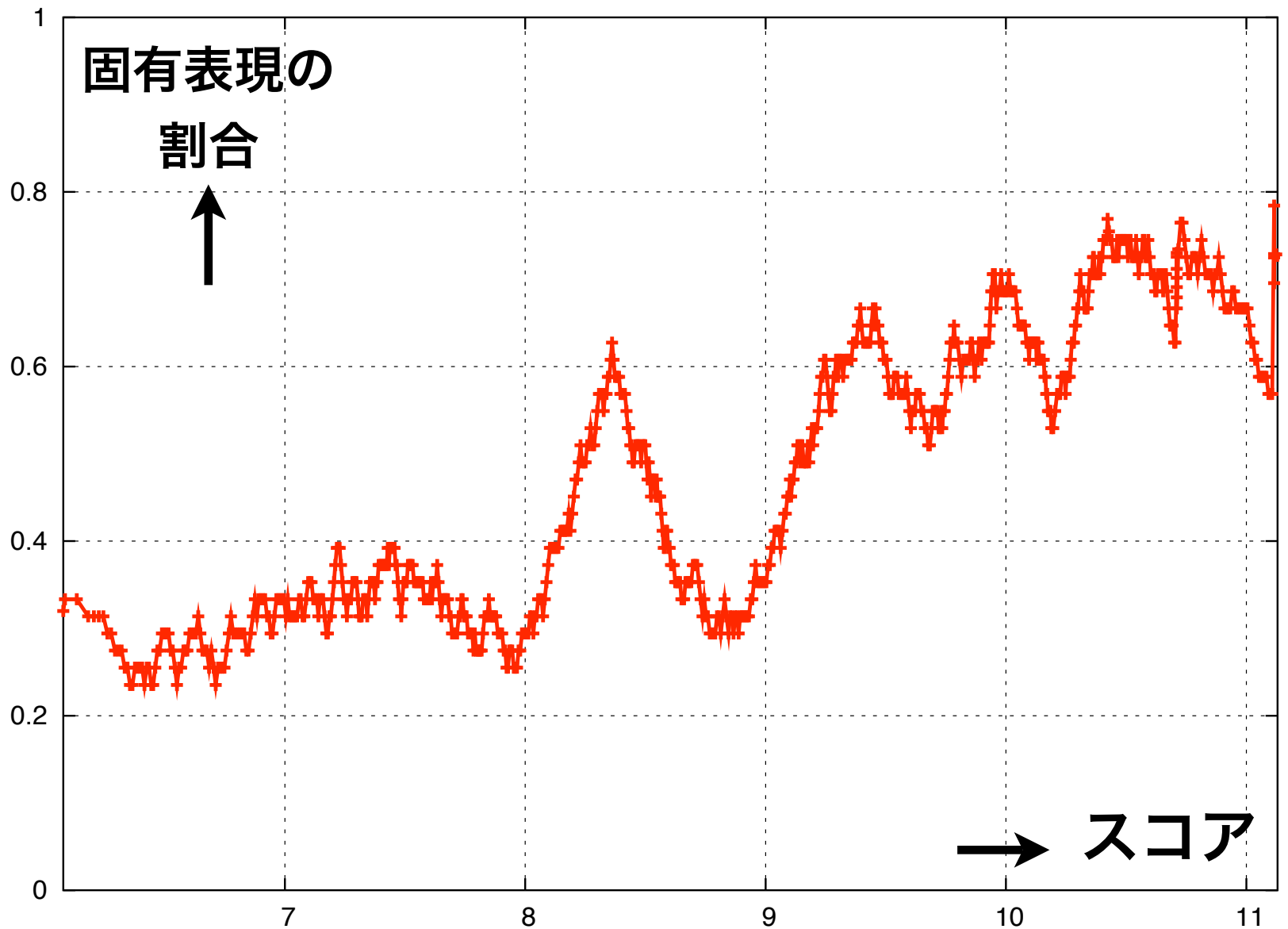
本手法でランキングした場合



スコアの低かった単語

スコア	単語	種類	頻度
0.000	unicameral	adj	9
0.000	gruffly	adv	6
0.000	kingly	adj	4
0.000	langauge	(typo)	3
0.000	windsocks	noun	2
0.000	metformin	PRODUCT	2
0.000	catoctin	LOCATION	2
0.000	runny	adj	12

IDFでランキングした場合



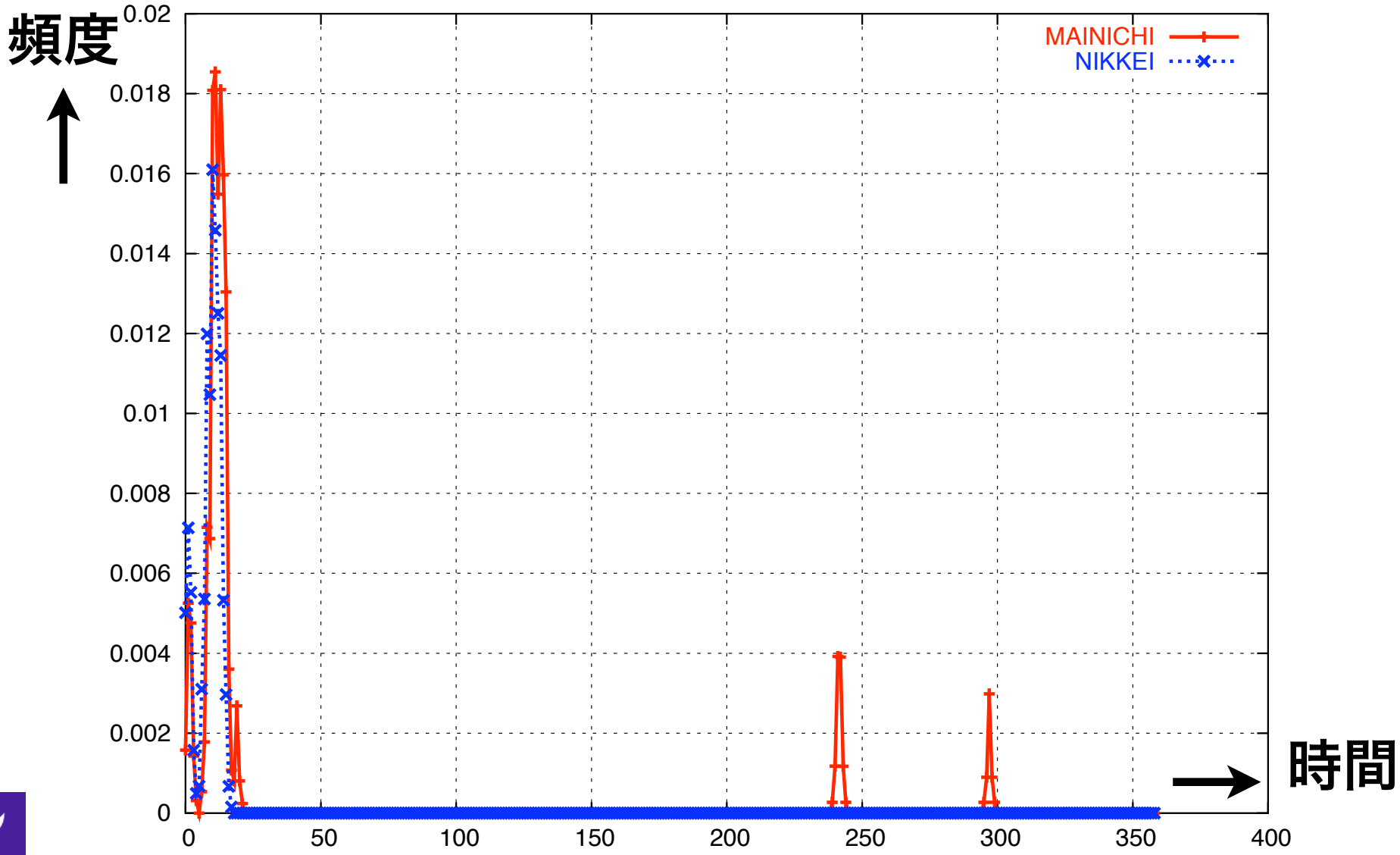
日本語への適用

- 1995年の毎日、日経の記事１年分(29万記事）を使用。
- 形態素区切りの問題を避けるために、連続する漢字列のみを単語として評価した。

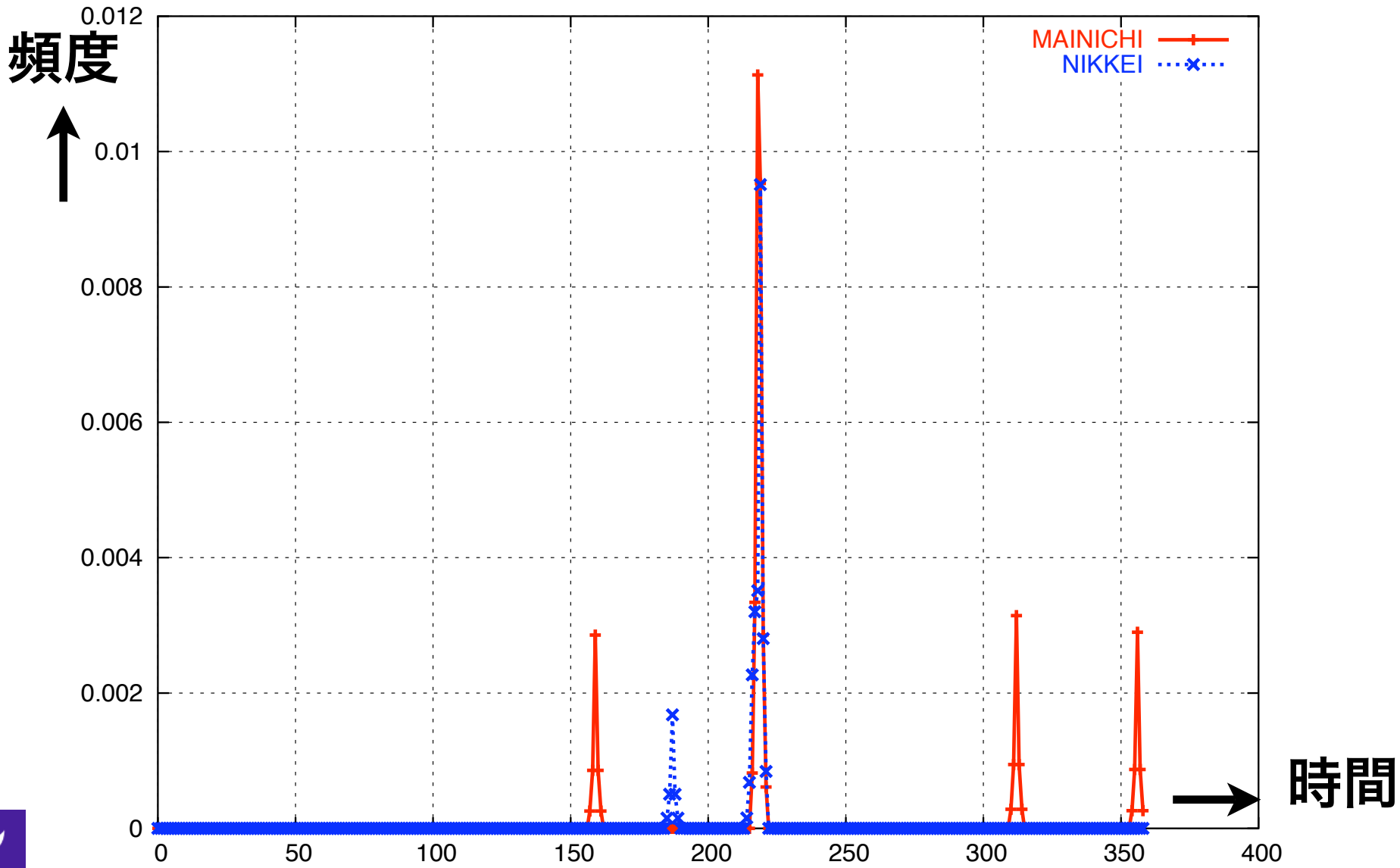
漢字列を評価した場合

スコア	単語	種類	頻度
1.000	優太	PERSON	2
1.000	藤木孝男	PERSON	2
1.000	先発後発	-	3
1.000	子供地球基金顧問	ORGANIZATION	2
1.000	検定前	-	3
1.000	奥富喜一氏	PERSON	2
0.972	貫田宗男	PERSON	11
0.938	田畑良幸助役	PERSON	8
0.886	村山富市委員長	PERSON	54
0.848	所掌事項	-	4
0.808	群馬県上野村	LOCATION	10
0.763	佐藤正久三等陸佐	PERSON	10

“村山富市委員長”


頻度
時間
MAINICHI
NIKKEI
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
50
100
150
200
250
300
350
400

“群馬県上野村”


頻度
0.012
0.01
0.008
0.006
0.004
0.002
0
0
50
100
150
200
250
300
350
400
MAINICHI
NIKKEI
時間

複合語への適用

- ２つの連続する単語対の頻度を測定し、その分布の類似度をはかる。
- “the” などのストップ語は除いた。
- 高い類似度 (分布の一致) はあまりみられなかった。

単語対を評価した場合

スコア	単語対	種類	頻度
0.738	goran klintmalm	PERSON	37
0.696	everyday miracles	noun	16
0.639	particularly agitated	verb	21
0.504	detective tom	PERSON	111
0.483	commander robert	PERSON	67
0.458	angry voices	noun	59
0.430	rabin funeral	PERSON	17
0.430	thai nation	LOCATION	82
0.395	forestry association	ORGANIZATION	73
0.374	outer banks	LOCATION	60
0.356	t global	ORGANIZATION	48
0.353	stanford linear	ORGANIZATION	12

結論

- 単語分布の類似度と固有表現の割合には一定の相関があることを確認した。
- 再現率は十分ではない。
  - スコア０の単語が結構ある。
  - 素性として使うことにより、固有表現抽出のさらなる精度向上が期待できる。

今後の課題

- 類似度の計算方法の改善
  - 報道の時期にズレがある場合の補正。
  - 突出した出現頻度を重視する。
    - 現在の手法だと、“the” などもスコアが高くなってしまう。
- コンパラブルコーパスの特性を利用した固有表現抽出器の製作。

固有表現の数

- 全テストデータ:
  - 469 / 966 (48%が固有表現)
- スコアが 0.6 以上の単語:
  - 80 / 90 (適合率: 89%, 再現率: 17%)
- スコアが 0 の単語:
  - 239 / 452 (53%が固有表現)
  - 同じ日に現れないものが多い。

再現率と適合率の関係