自然言語処理入門
「人間にできることが計算機にできないわけがない！！」

東京大学　情報基盤センター
（総合文化研究科、情報学府　兼担）
中川裕志
nakagawa@r.dl.itc.u-tokyo.ac.jp
http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/

自動文書要約

- 一見、難しそうな処理だが、大変古くから研究されてきた。
- 1953年には既にIBMのLuhnによって、単語の重要度を定義し、重要な単語を多く含む文を文書から抽出するという方法で、かなり質のよい要約文が作られていた。
- Luhnは、
  - 中程度の頻度の単語が現れること
  - 文書の先頭に近いほうが望ましい
- という2点を考慮してスコアの高い文を抽出した。
- 現在でも基本的には踏襲。

自動要約の応用分野

- ある分野のサーベイの自動生成
  - cf. e-science
- 会議議事録の自動生成
- 携帯端末への要約テキスト表示
- 音声表示（要約しないと読み上げでは長い時間がかかる）
- 高齢者や児童への手短かつ分かり易い表現（言い換えも含む）
- 字幕の自動生成
- ニュースやドラマなどのビデオコンテンツの要約(skimming)

要約例

- 新幹線の車両ドアの上の液晶ディスプレイでのニュース表示
- インターネットに配信されているメールマガジンの見出し
- iモードのニュースは通常のニュースの要約
- などなど
- 次のものは作り物の例

要約： いろは金融本郷支店での強盗殺人・放火事件で捜査本部が犯人のモンタージュ公開。逃走車両？目撃情報も。

いろは金融放火：捜査本部が犯人のモンタージュを公開

文京区本郷の13階建て雑居ビルの最上階にある消費者金融「いろは金融本郷支店」で1日に発生した強盗殺人・放火事件で、警視庁捜査本部は犯人のモンタージュを公開した。また、「出火直後にビルの前から黒色の軽乗用車かワゴン車に男が乗り込み、走り去った」との目撃情報があり、警視庁は犯人の逃走車両の疑いもあるとみて行方を追っている。

調べでは、同支店はこの日、通常通り午前10時に営業を開始。同10時45分ごろ、男は持っていた18リットル石油缶の半分くらいをブリキ缶に入れた灯油のようなものをいきなりカウンター越しにまいた。その際に叫んだ「金を出さんかい。出さんと火つけるぞ」は、なまりがあったという。

従業員によると、男とは面識がないという。60～65歳、身長170センチ台前半で中肉。白髪交じりの短髪でサングラスをかけていた。

一方、従業員とみられる10人の焼死体はいずれも店舗中央のカウンター近くで発見され、6人が内側で折り重なり、4人が外側付近に倒れていた。店舗出入り口はエレベータ付近と浦階段の2カ所だった。けがをした従業員によると、放火された火は一度爆発し、すぐに黒煙が広がったという。事務所には窓ガラスを破って使うことのできる避難器具はあったが、裏階段はロックされており10人は迫る炎と煙にまかれ、逃げ場を失ったとみられる。

要約の機能

- Indicative
  - その文書を読むべきかどうかの判断材料を与える
- Informative
  - 要約を読むだけで、おおよその内容が分かる
- Evaluative
  - 要約者の評価も加わった要約（重要なポイントの強調など）

人間の要約専門家はどうやっているか

- 表層情報としては以下を利用：
  - タイトル、見出し、キーフレーズ、位置情報
- 深層情報としては以下を利用：
  - 談話構造、修辞構造、意味内容（目的、方法、結果、結論）、その分野の知識
- 上記により重要文を抽出し、編集、再構成
  - トピックの文を抜き出し、前後の辻褄合わせ
  - トピック文をさらに変換
- 理解した意味から自分で作文することは少ない（とても大変で時間がかかる）

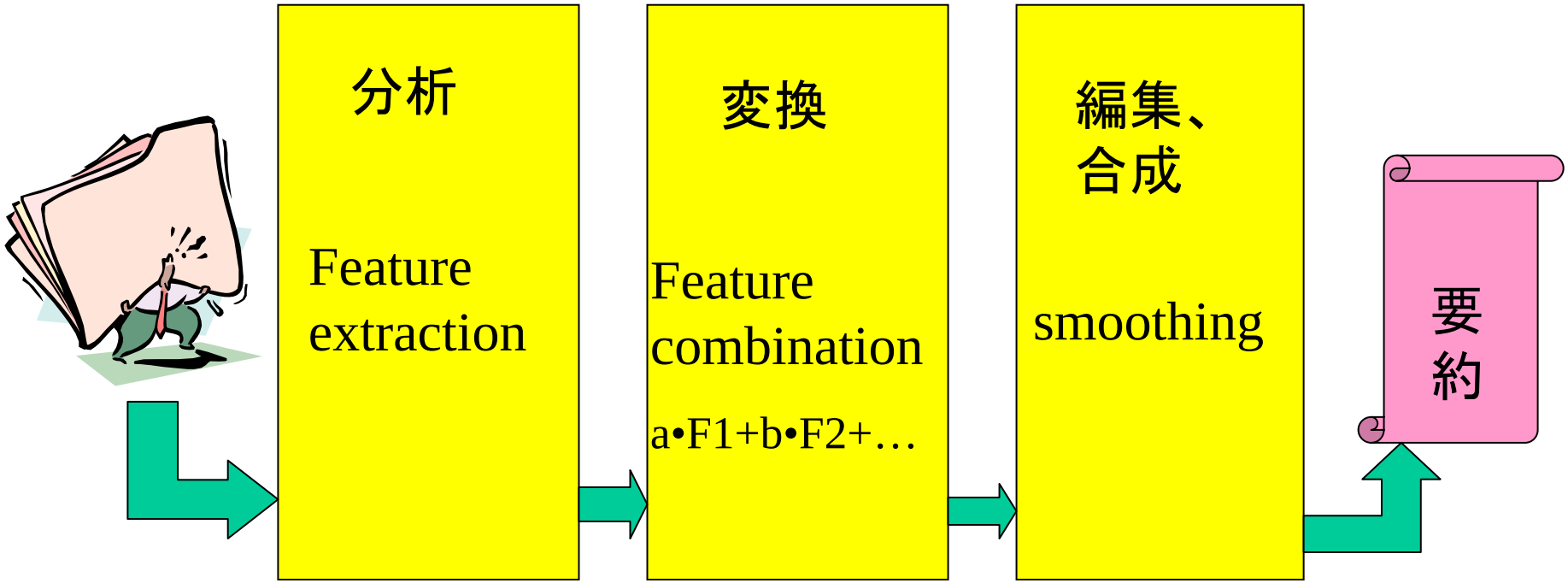
要約のパラメタ

- 圧縮率(compression rate: C)
  - C=length(要約テキスト)/length(原テキスト)
- Semantic Informativeness: SI
  - テキストT の内容を命題Mi (i=1,2,..)の重みつき集合とする。

$$SI = (1 - \frac{length(S)}{length(T)}) \times (\frac{weight(M(S))}{weight(M(T))})$$

  - *S*: 要約テキスト、*T*:原テキスト
  - *M(S)*:要約テキストの命題集合、*M(T)*:原テキストの命題集合。当然、$M(S) \subset M(T)$

自動要約システムの構成(shallow)



- これらモジュールの機能は以下のものの関数
  - 圧縮率
  - 読者: Generic or User focused
  - 機能: indicative or informative or evaluative
  - 結束性: fragment, or connected text

文を選択するためのfeature

- 以下の feature は数値的な重みで表わされる
- 位置
  - 先頭からの文字数or単語数、段落、section、タイトルなど特殊なsection, section の深さ
- テーマ単語（重み）
  - 文章を特徴つける単語、複合語など
  - tf×idf などで重みの大きいターム，など
- 特徴的言い回し
  - 「まとめると」、in summary
  - 「重要な」「特に」、important, in particular

文を選択するためのfeature

- 付加ターム（重み）
  - タイトル、headline、先頭段落に現れる単語
  - 利用者のプロファイルや質問文に現れる単語
- 文の長さ（適当な長さあり。長すぎるのはカット）
- 結束性(cohesion)
  - 同一表現あるいは synonymy, hypernymy, 反復
  - 参照、省略、照応、接続
- 談話構造
  - 修辞構造、話題構造
  - 文書の形式

Featureの線形結合による文の重み付け

- 文をuとする
- uにおける重みW(u)を計算
  - テーマ単語や、付加タームは、uに現れた、相当する個々の単語、タームの重みの総和
  - 特徴的言い回しも複数あれば総和
- W(u)=a×位置(u)+b×特徴的言い回し(u)
  +c×テーマ単語(u)+d×付加ターム(u)
  +e×文の長さ(u)
- W(u)の大きい文から順番に要約文として選択

Feature の線形結合による文の重み付け

- Kupiec（1995）の実験
  - 科学技術分野論文188の全文と要約（平均3文）のペア
  - 要約に含まれるべき文を計算された重みの順に選択
  - 要約文が全文のどの文にマッチするかを知らなければならない。
- これによれば、位置が最も強力な feature で単独で33%の再現率
- 位置＋特徴的言い回し＋長すぎる文のカット が最高性能 44%の再現率

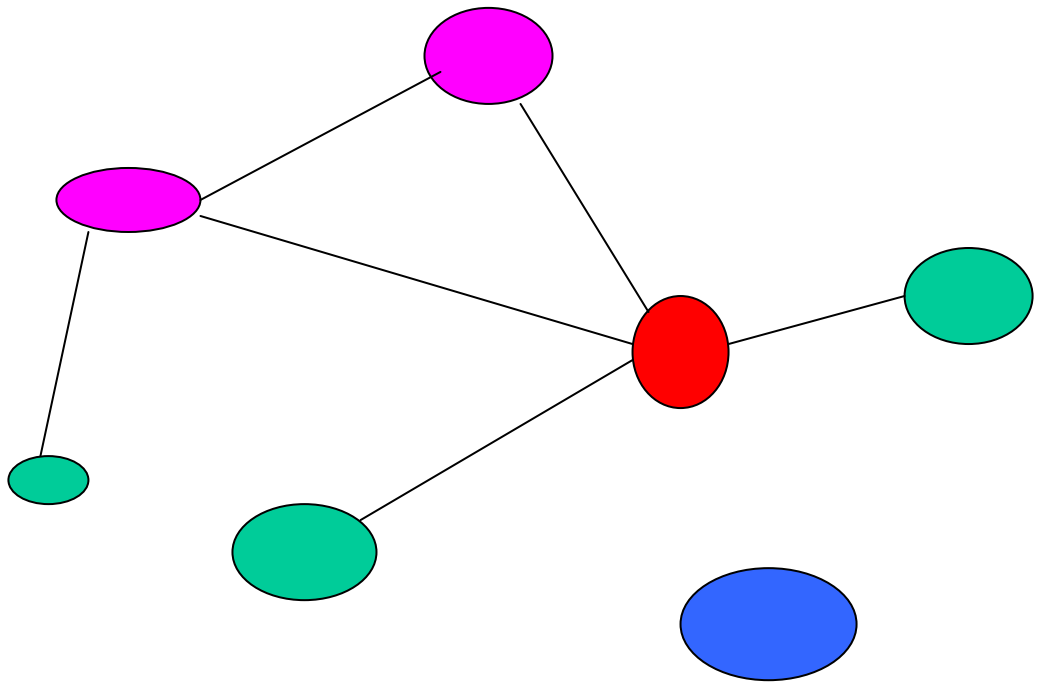
Feature へ掛ける重みの学習

- W(u)の定義式のa,b,c,d,e などの重みを最適化する方法
- 機械学習による
  - 例えば、人手でつけた正解の要約文の集合を使う。
  - 正解の要約文をうまく抽出できるような重みを統計的な機械学習で求める
  - ベイス統計、C4.5、SVM　など、高度な理論やソフトが使えるようになってきた。

談話のfeature

- テキストにおける結束性 cohesion
- 文法的結束性
  - 照応
  - 省略
  - 接続
- 語彙的結束性：下記の単語が現れる文は結束しているので、同時に要約に入れる。
  - 同義語(synonymy)
  - 上位語(hypernymy)
  - 繰り返し

- 強く結束している文の集合を要約文として選択
- 結束性としては、
  - 同じ語彙（類義も含む）、照応関係を含む、など。



抽出文の再編集

- 結束性向上のための浅いスムース化
- 照応
  - 照応の対象物（代名詞）で始まる文を削除
  - あるいは、代名詞や省略のある文の直前のいくつかの文を要約に含める。
- ギャップを埋める
  - 重要度の低い文を選ばれた文の間に埋め込む
  - 並列な内容のN文のうち、後半が選ばれたなら、それより前の文も追加
- 個別表現の短縮
  - 総理大臣→首相

複数テキストの要約

- 関連するテキストの自動収集
- 関連するテキストからの情報抽出
  - 重要個所の抽出
  - テキスト間の共通点の検出
  - テキスト間の相違点の検出
- テキスト間の文体の違いを考慮した要約文書の生成

情報検索されたテキスト群の要約の基準

- Q 検索要求
- R システムのよって検索されたテキスト群から抽出された文の集合。抽出は単一文書要約の場合と同じ方法でよい。
- S 既に選択されたRの部分集合
- Maximal Marginal Relevance: *MMR*
  - 以下の式を満足するようにDiを順番にRから選択してSに追加していく。

$$MMR(Q,R,S)$$

$$= \mathbf{argmax}_{D_i \in R \setminus S} [\lambda \times sim(D_i, Q) - (1-\lambda) \max_{D_j \in S} sim(D_i, D_j)]$$

統計的方法

- 原テキストと要約の組が大量にある場合
- t 原テキスト、s　要約
- $s' = \mathrm{argmax}_s P(s|t) = \mathrm{argmax}_s P(t|s)P(s)$
- $P(s)$は文としての確からしさ（bigramなどで近似）
- $P(t|s)$は原テキスト中の単語が要約に現れる確率

要約の展開

- 要約対象の拡大
  - 文書、Webページ
  - 音声発話
    - （書き起こし、あるいは音声認識結果）
  - マルチメディア
    - 映像の要約（ skimming)
- 携帯端末表示を目指す要約

言い換え

研究の背景

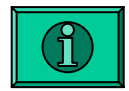
- 言い換えの多くの応用
  - 機械翻訳の前処理
  - 平易な表現への言い換え
  - 文章を短縮する

多量の言い換えパターンが必要

応用例：携帯端末上への表示を

- 表示内容を携帯向けにするとは？
  - 狭急：
    - 狭い画面と高い時間圧の下でのブラウズ
  - ひと目：
    - １画面に収まるコンテンツ→要約、言い換え
  - 簡単操作：
    - 悪くとも単純なスクロールで閲覧できる

  →表のリスト化
- 言語リソースとして何を使うか？
  - すでに i-モード 用ページ有り。
  - これを参考にしない手はない。
  - この言語データから要約や言い換えの戦略を学習する。



言い換え

- 変換: 全角－半角の変換、
- (文脈利用による)表現の簡略化、
  - 「午後2時」→「14時」「14:00」「1400」
  - 東京都文京区本郷 → 文京区本郷、 本郷
  - 小泉総理大臣→首相 (可能な文脈は？)
  - 階層的表現の上位文脈の継承:
    - 本図書館開館時間は10時→開館10時
- 本研究の最終的な狙いは

言い換えを利用したテキスト圧縮

言い換え（体言止めの類）

- 体言止め、助詞止め
  - 衆議院で審議に入る → 衆院審議入、衆院審議へ
  - 体言止め変換はどういう条件で可能か
  - 不完全文で許される助詞の使用法とは

- 助詞省略
  - 「似顔絵を公開」→「似顔絵公開」
  - 助詞の省略はどういう条件で可能か
- 助詞、記号での置換
  - 「放火の疑いもある」→「放火か」あるいは「放火？」
  - 助詞、記号で置換できる表現は何か。
- →これらの問題を言語学的に解決するのは難しい！

データ収集

- 対応コーパスの作成
  - 通常のインターネットWebページ（例えば新聞記事）を収集
  - i-mode など携帯端末向けのページを収集
  - 通常記事は1日200記事。i-モード 記事は1日40記事程度
  - 現在、毎日新聞で、通常記事とi-モード記事をインターネットに配信しているが、i-モード 記事は1日で消えてしまう。
  - 毎日収集して蓄積しておく

Web記事の性質

- 政治、経済、国際、社会の4ジャンルを収集
- キーワード、タイトル、本文により構成
- タイトルは20文字程度
- 本文は数百文字、文字数が多い場合は段落によりまとめられている

Web記事の例

＜ 金 融 ＞	キーワード
為替（東京） １４日終値１＄＝１２３円２３銭	タイトル
１４日の円相場の終値は・・・ ・・・(中略)・・・ ・・・が影響している。	本文

携帯記事の性質

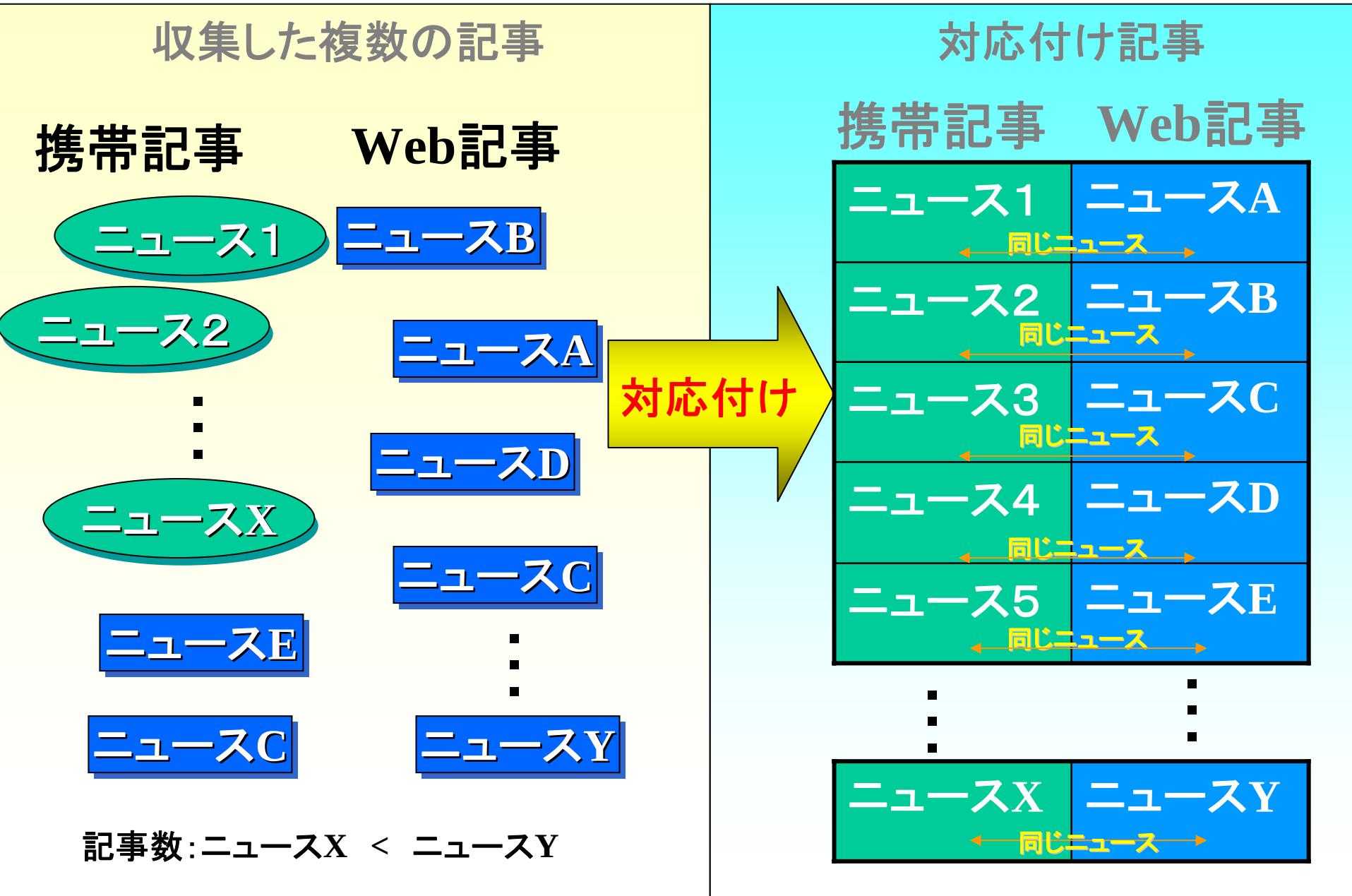
- 政治、経済、国際、社会の4ジャンルを収集

特徴:

一つの記事の文字数が約50文字程度に

抑えられている

対応付け記事


収集した複数の記事
携帯記事
Web記事
ニュース1
ニュース2
ニュースX
ニュースB
ニュースA
ニュースD
ニュースC
ニュースE
ニュースC
ニュースY
対応付け
記事数：ニュースX ＜ ニュースY
対応付け記事
携帯記事
Web記事
ニュース1
ニュースA
同じニュース
ニュース2
ニュースB
同じニュース
ニュース3
ニュースC
同じニュース
ニュース4
ニュースD
同じニュース
ニュース5
ニュースE
同じニュース
ニュースX
ニュースY
同じニュース

対応付けの方法

- 同日、同ジャンルの各新聞記事を比較
- 名詞（未知語を含む）に注目

  名詞が一致した記事中の位置による重み付け

  数字列も名詞として扱う

このようにして得た一致名詞数の大きい記事のペアが対応するとした。

配点の具体例

・携帯記事

金融･･･14日の終値１＄＝１２３円１４銭。円が売られ、ドルが買い戻された。

（携帯記事の本文）

各記事の名詞を抽出する

・Web新

＜金融＞

（キーワード） 配点W1（×３点）

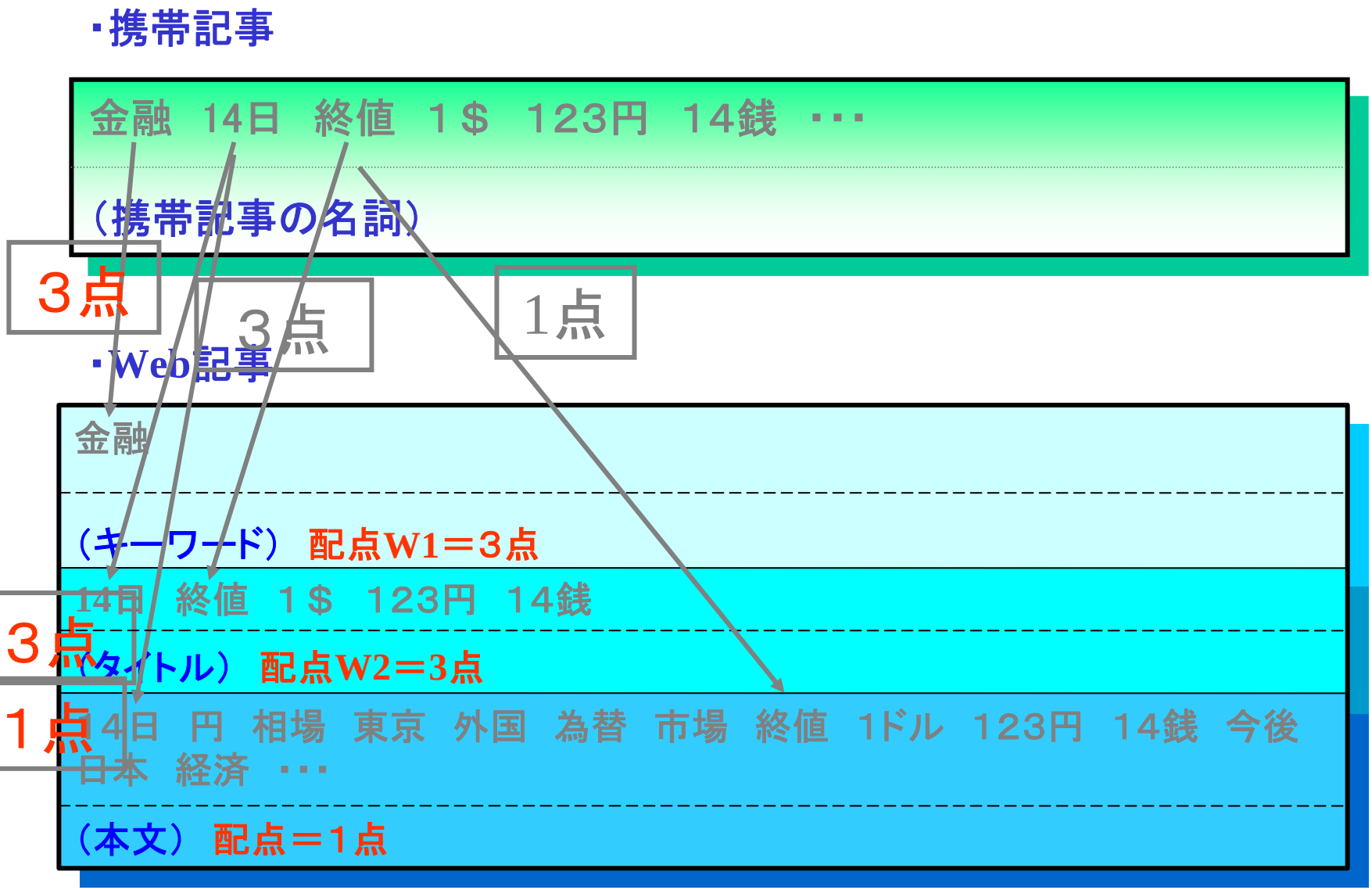
為替（東京） 14日終値１＄＝１２３円１４銭

（タイトル） 配点W2（×３点）

１４日円相場（東京外国為替市場）の終値は1ドル１２３円１４銭となった。「今後の日本経済の見通しが定まらない」との声が強まり、海外を中心に円が売られ、ドルが買い戻された。

（Web記事の本文） 配点（×１点）

名詞の抽出と配点の方法



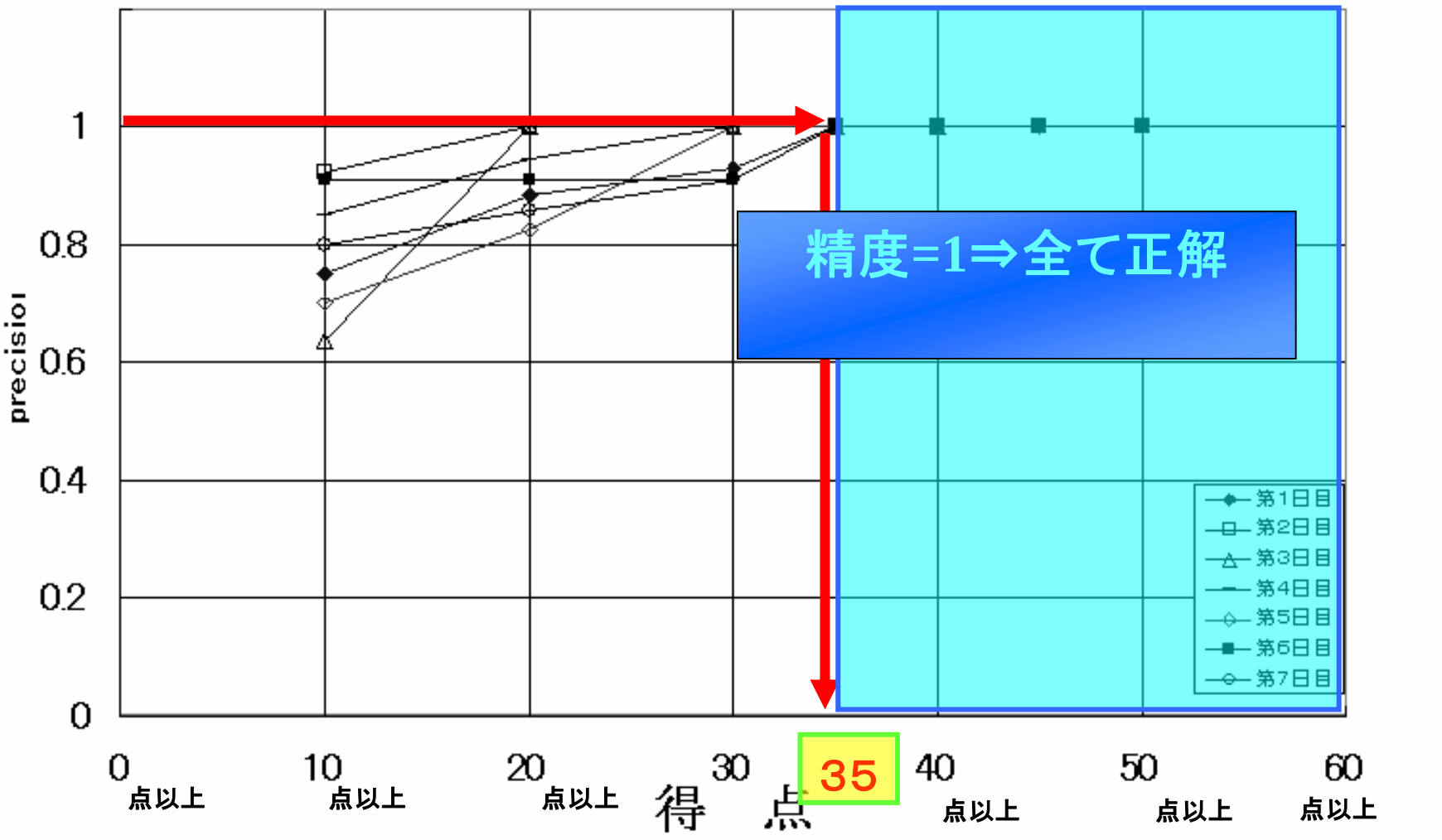
対応付けの精度

- 自動的に対応付けしたデータの正確性の調査
- 調査記事数：605記事（携帯端末向け新聞7日分）

$$precision(\text{精度}) = \frac{(\text{抽出した対応付け正解記事数})}{(\text{抽出した対応付け全記事数})}$$

対応付けした結果を人手で正解か不正解かを評価

日付別の精度


precision
1
0.8
0.6
0.4
0.2
0
0
10
20
30
35
40
50
60
点以上
得　点
精度=1⇒全て正解
第1日目
第2日目
第3日目
第4日目
第5日目
第6日目
第7日目

記事対応付け結果

正解の定義：
精度35点以上で全て正しい対応付データ

携帯端末向け記事の調査記事数　605記事
正確に対応付けられた記事数　　481記事

正解記事の抽出率　76．19%
（誤りは対応記事が存在しない場合のみ）

収集したデータのまとめと研究の方向

- ３年間分、約90000記事対の対応付けコーパス→今後の研究の基礎となる言語リソースができた。
- このデータを利用し、要約パタン、言い換えパタンを学習する
- 分かっていることは、
  - 新聞記事は最初の段落が重要
- 要約については時間があれば紹介

例

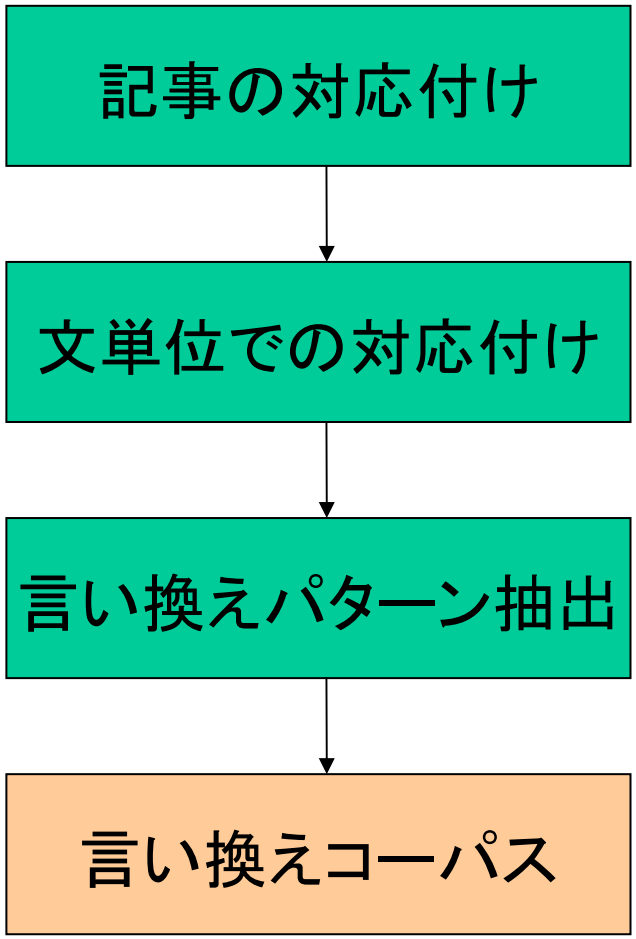
本法案が衆議院本会議で審議が始まった。

本法案、衆議院本会議で審議。

HIVの母子感染防止に有力な方法が分かった。

HIVの母子感染防止に有力な方法が判明。

研究のアプローチ



- 携帯記事とWeb記事から対応付け
- 文と文を対応付ける
- 文字走査による言い換え抽出

使用した記事データ

- 対応付けコーパス[大森2003]
  - 携帯端末向けに作成された記事(携帯記事)からパソコン向けに作成された記事(Web記事)の内容が同じものを対応付けた記事集合
  - 2001年4月26日～2003年11月30日
  - 合計48075対応記事

対応記事の例

携帯記事

１１月の月例経済報告で政府の景気認識を示す基調判断を下方修正へ…内閣府。下方修正は3カ月ぶり。

Web記事

月例経済報告：基調判断、１１月は下方修正へ　3ヶ月ぶり

　内閣府は３０日、１１月の月例経済報告で政府の景気認識を示す基調判断を下方修正する方針を固めた。これまでの「引き続き悪化」から表現を引き下げる。同日発表された９月の完全失業率が急上昇したほか、２９日発表の９月鉱工業生産が一段と落ち込んだため。下方修正は3カ月ぶりとなる。

携帯文の文末品詞

品詞		頻度	頻度/合計[%]
名詞	サ変接続	34312	38.8
		15875	18.0
助詞		14484	16.4
動詞		16186	18.3
助動詞		6671	7.6
その他		805	0.9
合計		88333	100.0

対応文の抽出

- 携帯記事とWeb記事で文と文の対応付けを行う
- 名詞の出現頻度によって行う

- 携帯記事から抽出した文(携帯文)
- Web記事から抽出した文(Web文)

対応文の抽出

携帯記事

１１月の月例経済報告で政府の景気認識を示す基調判断を下方修正へ…内閣府。下方修正は3カ月ぶり。

Web記事

内閣府は３０日、１１月の月例経済報告で政府の景気認識を示す基調判断を下方修正する方針を固めた。これまでの「引き続き悪化」から表現を引き下げる。同日発表された９月の完全失業率が急上昇したほか、２９日発表の９月鉱工業生産が一段と落ち込んだため。下方修正は3カ月ぶりとなる。

対応文抽出の結果

- 88333組の対応文を抽出

- 対応文の抽出精度は92.8%
  （ランダムに抽出した500文を人手で評価した結果）

言い換えパターンの抽出方法

1. 携帯文から候補集合の作成
2. 候補集合の単語を文末に含む携帯文とそれに対応するWeb文を抽出
3. Web文で文末からの文字列走査

候補集合

- 携帯文の最後にくる一単語を抽出
- 特殊な文末表現は削除
- 「へ」「か」など助詞で終わる場合は二単語
- 頻度の降順でならべる
- 頻度が一回の単語は除く

4566種類の単語を抽出

候補集合の例

抽出単語	頻度	抽出単語	頻度	抽出単語	頻度
した	2370	表明	913	求める	469
高	2002	死亡	718	合意	429
安	1892	決定	648	検討	426
」と	1652	いる	625	批判	424
発表	1382	いた	600	られる	417
示す	1358	ため	520	開始	400
れる	1302	方針	505	協議	390
逮捕	1098	見通し	501	語る	375
する	1054	強調	485	要請	365
会談	992	判明	477	発言	361

候補集合に対するWeb文

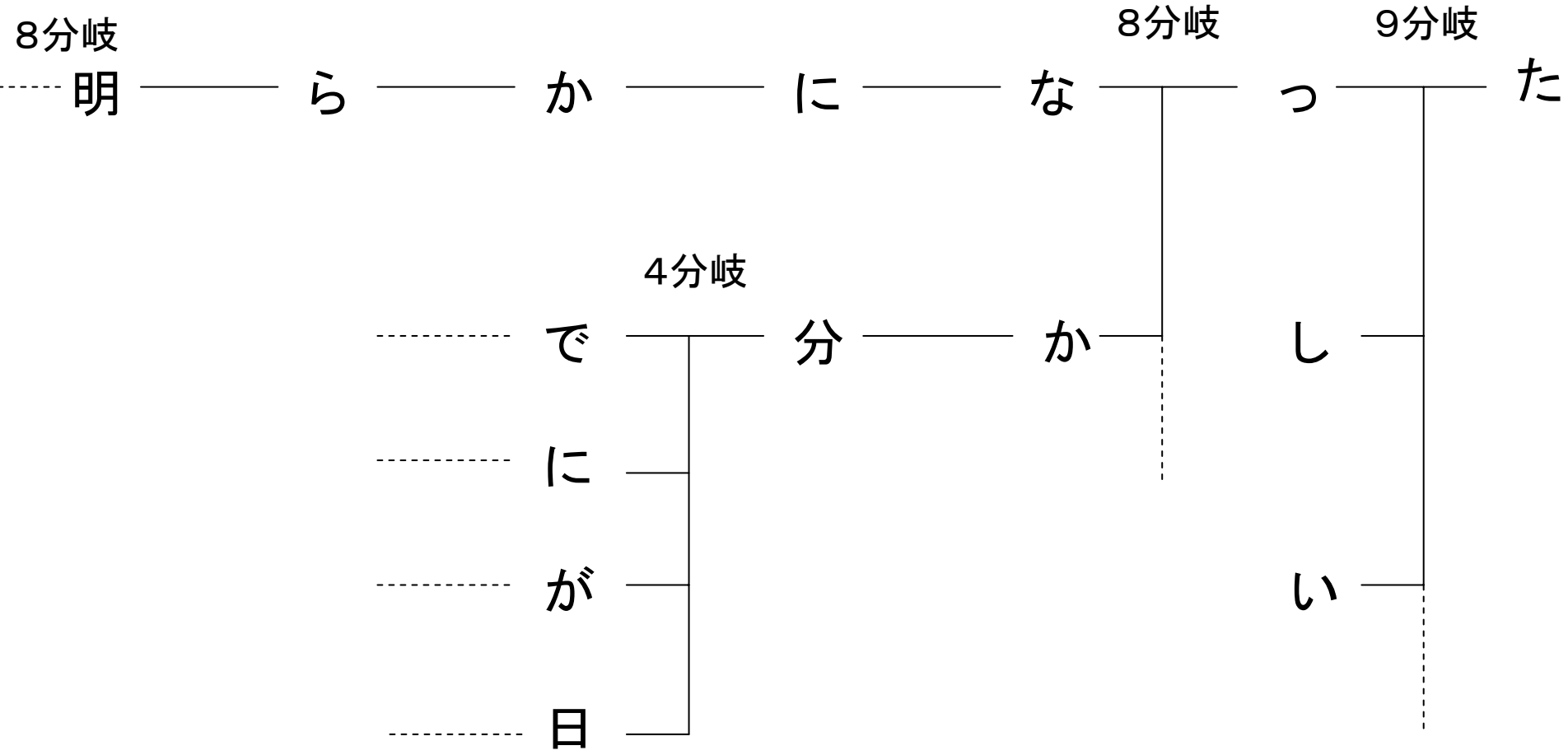
判明

・・・・・・・対象にした調査で分かった

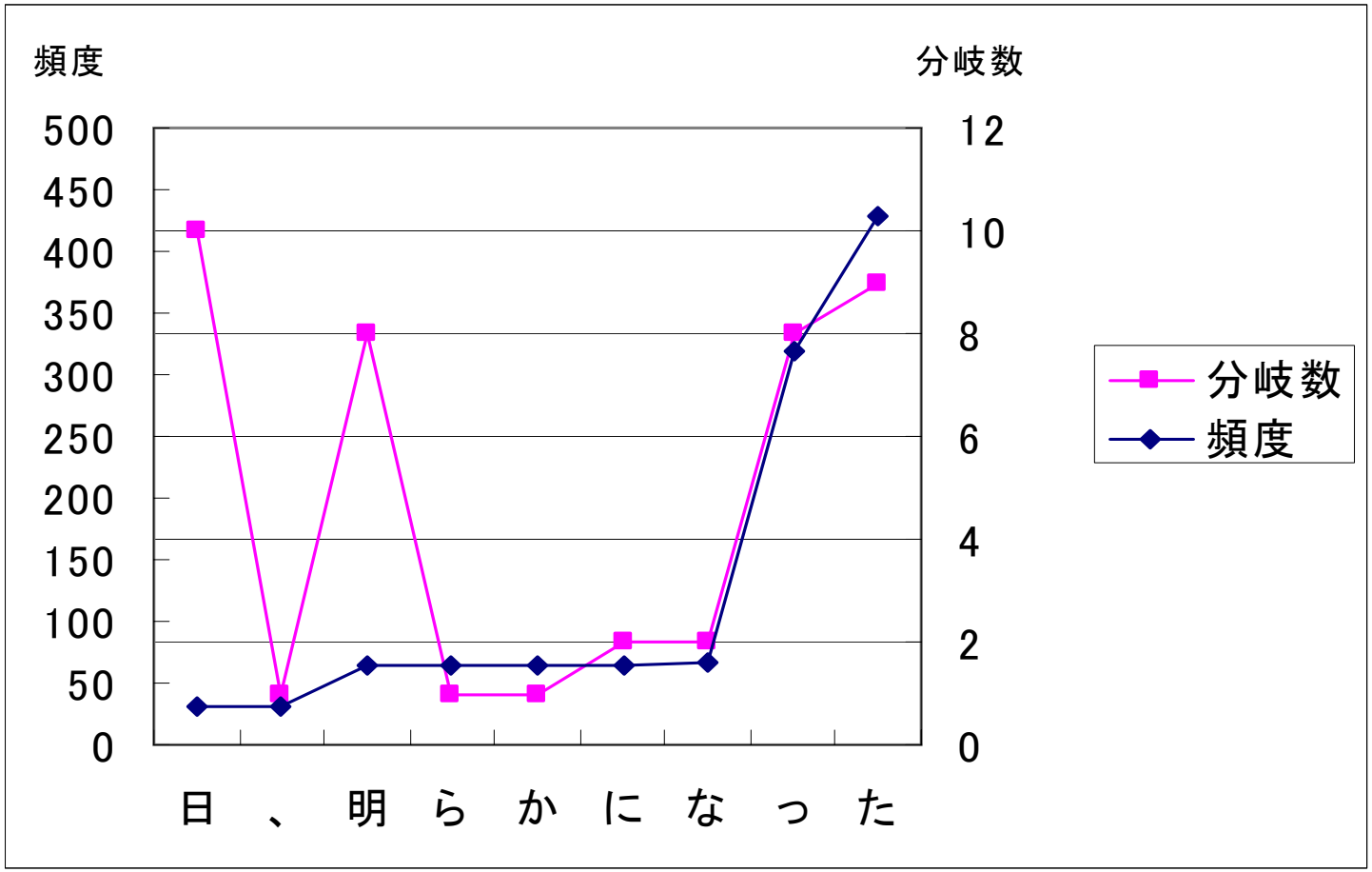
・・・・・・・集めていたことが、９日、分かった

・・・・・・・ 接見した弁護士などの話で明らかになった

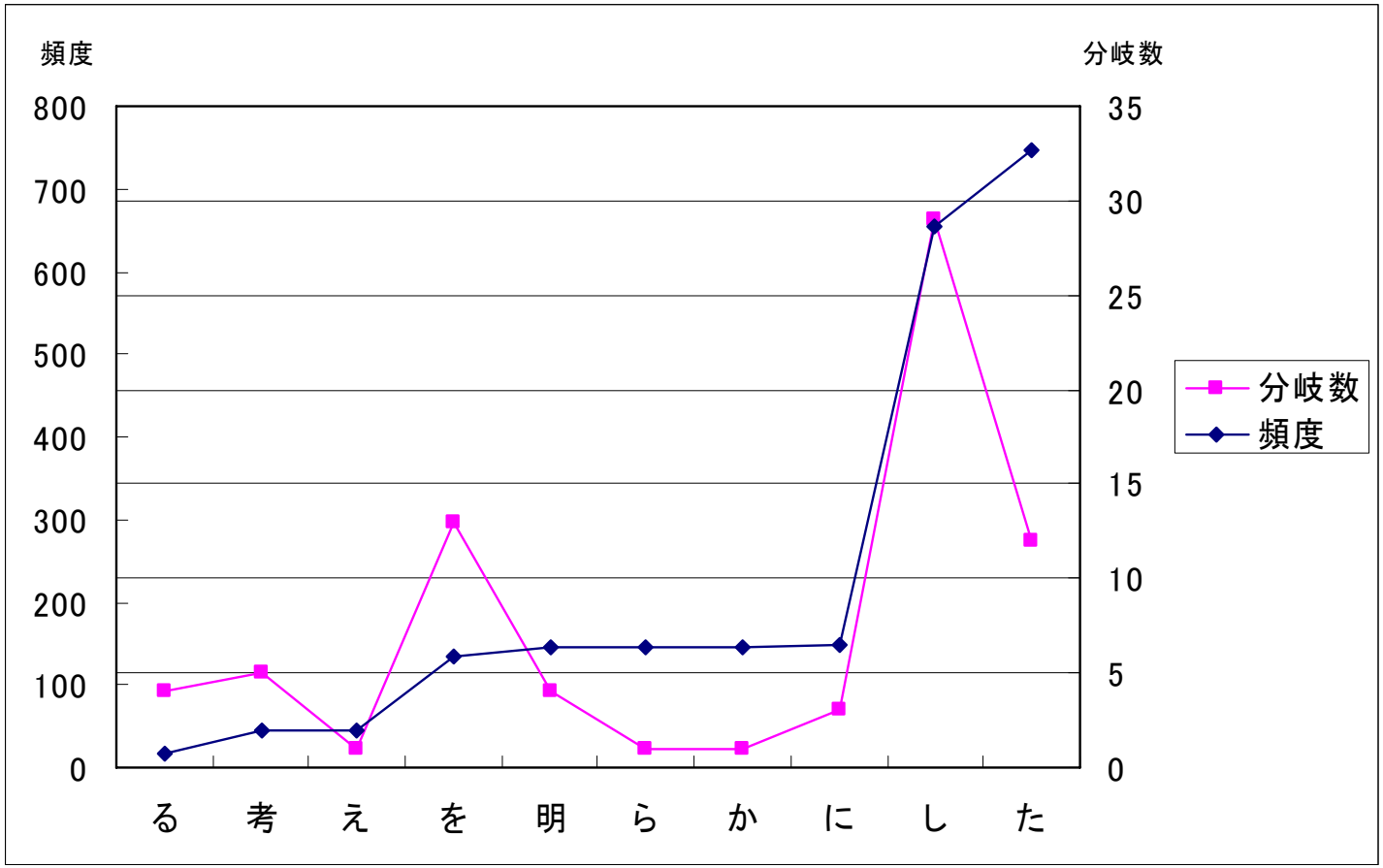
文末からの文字列走査[判明]



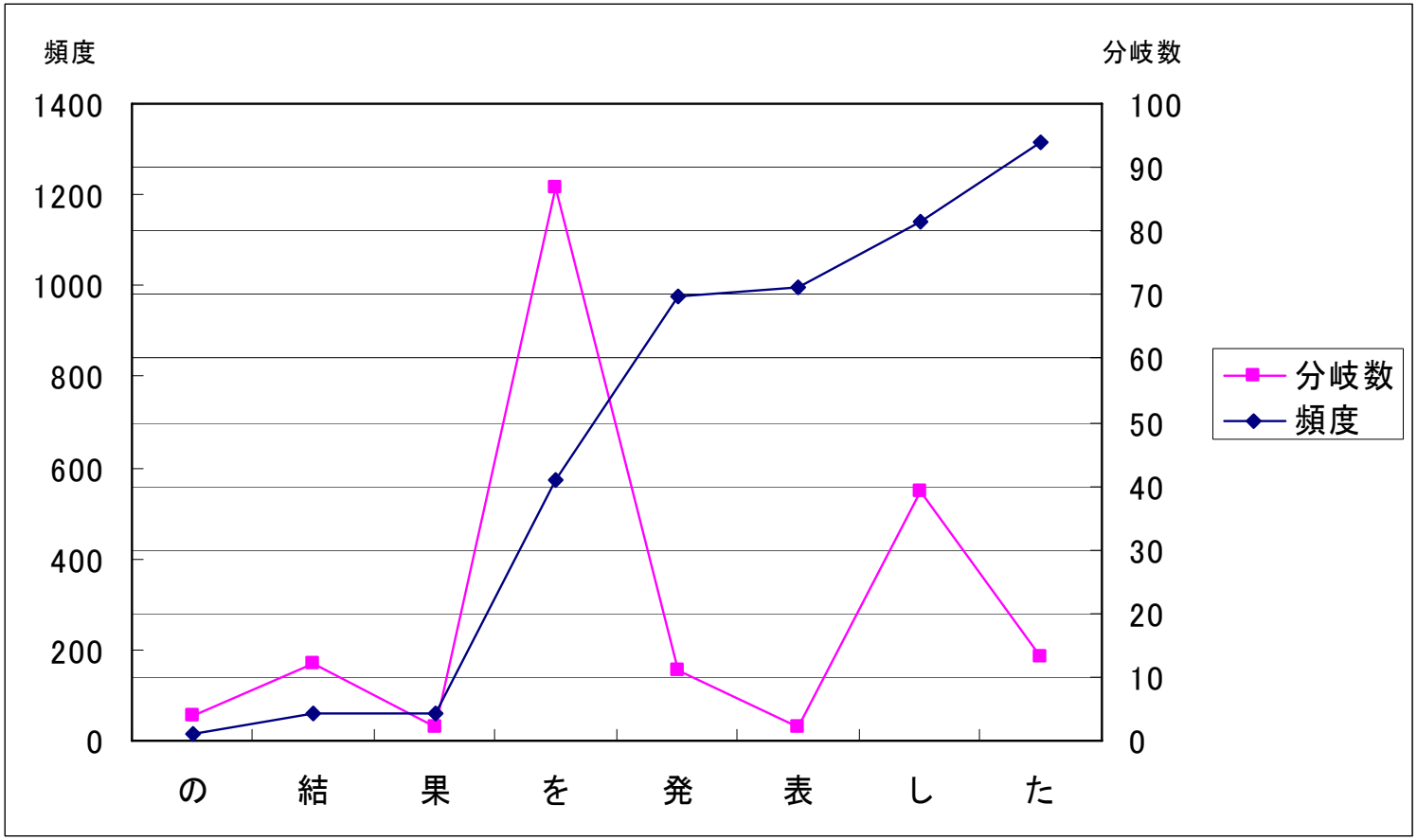
分岐数と出現頻度[判明]


頻度
分岐数
500
450
400
350
300
250
200
150
100
50
0
12
10
8
6
4
2
0
分岐数
頻度
日 、 明 ら か に な っ た

分岐数と出現頻度[表明]


頻度
分岐数
800
700
600
500
400
300
200
100
0
35
30
25
20
15
10
5
0
分岐数
頻度
る
考
え
を
明
ら
か
に
し
た

分岐数と出現頻度[発表]


頻度
分岐数
1400
1200
1000
800
600
400
200
0
100
90
80
70
60
50
40
30
20
10
0
分岐数
頻度
の
結
果
を
発
表
し
た

抽出語の重要度計算

- 抽出したWeb文文末から以下の式で重要度を計算し、選別を行う

$$Weight(string) = 分岐数(string) \times 出現頻度(string) \times \log(length(string) - 1)$$

評価

- 候補集合の上位10種類のサ変名詞について評価を行う

  発表、逮捕、会談、表明、死亡

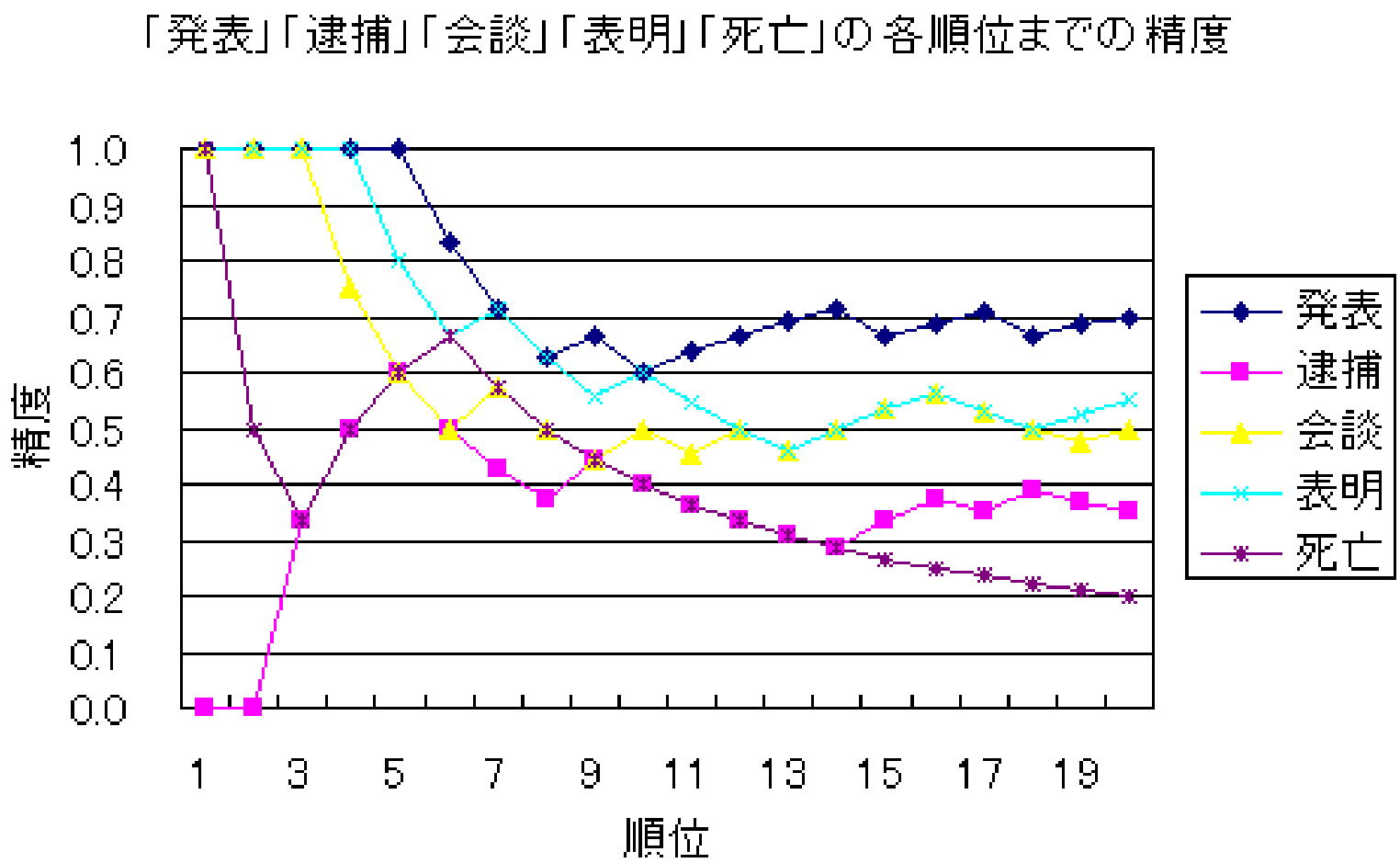
  決定、強調、判明、合意、検討

- それぞれの重要度20位までの語の正解率をとった

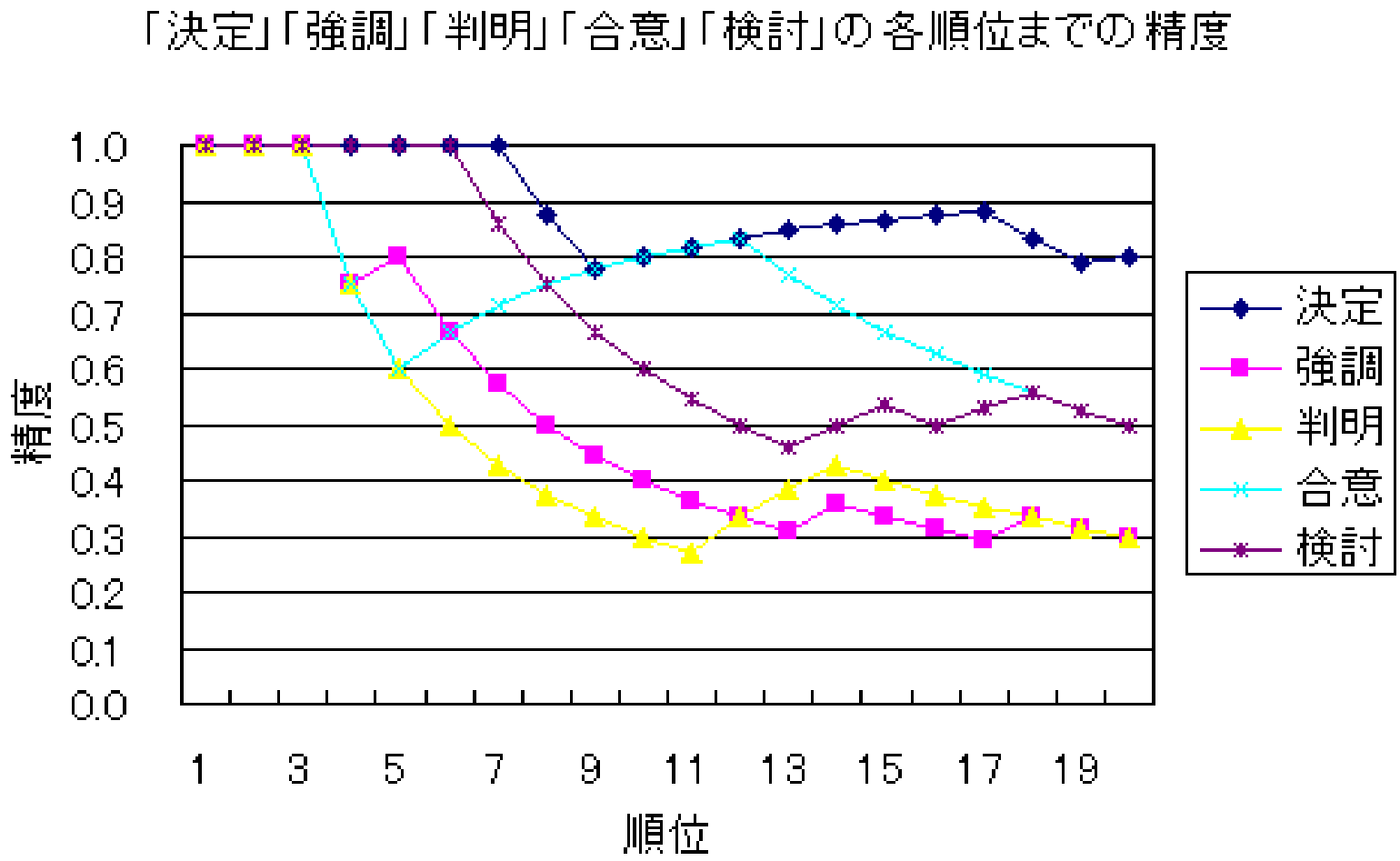
評価基準

- 正解とする判断方法は以下の様にした
  - 言い換え語を置き換えてみて意味が大きく変わらない
  - ２人が正解かどうか判断して、判断が異なる場合は３人目が判断し多数決で決める

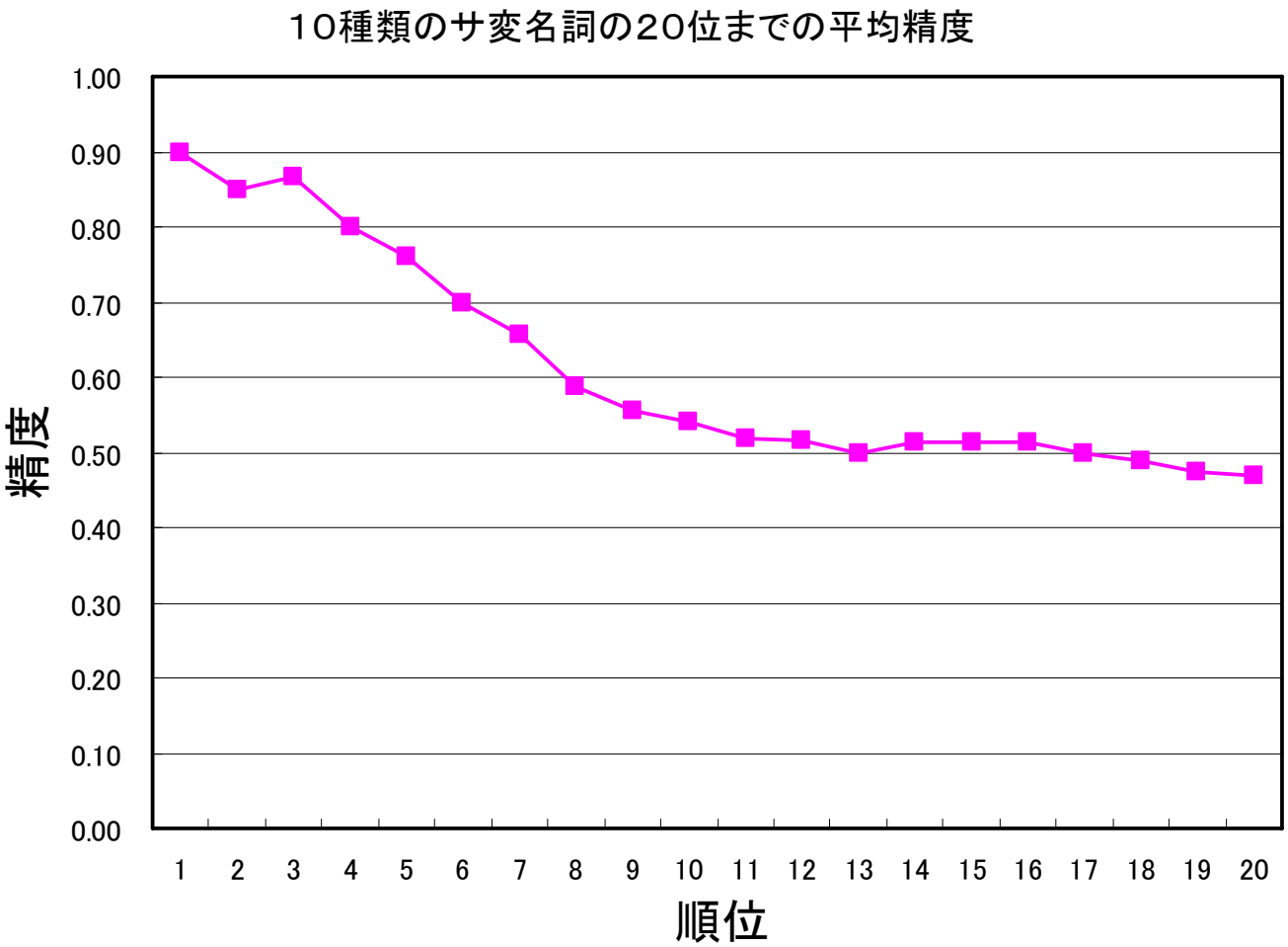
重要度20位までの正解率（1）



重要度20位までの正解率(2)



各順位の精度の平均値[a×b×c]



実際の抽出例

候補集合	抽出されたパターン				
	1位	2位	3位	4位	5位
発表	を発表した	発表した	と発表した	すると発表した	したと発表した
逮捕	容疑で逮捕した	の疑いで逮捕した	逮捕した	で逮捕した	を逮捕した
会談	と会談した	会談した	で会談した	大統領と会談した	首相と会談した
表明	を表明した	表明した	を明らかにした	する意向を表明した	する考えを明らかにした
死亡	死亡した	人が死亡した	人が負傷した	死亡したと発表した	間もなく死亡した
決定	を決めた	することを決めた	を決定した	ことを決めた	決定した
強調	を強調した	強調した	と強調した	考えを強調した	を改めて強調した
判明	分かった	で分かった	明らかになった	の調べで分かった	日、明らかになった
合意	することで合意した	で合意した	合意した	ることで合意した	ことで合意した
検討	を検討していることを明らかにした	を検討している	検討していることを明らかにした	検討に入った	検討を始めた

候補集合に対する1位の言い換え

言い換え先	言い換え元
発表	を発表した
逮捕	容疑で逮捕した
会談	と会談した
表明	を表明した
死亡	死亡した
決定	を決めた
強調	を強調した
判明	分かった
合意	することで合意した
検討	を検討していること を明らかにした

言い換え先	言い換え元
批判	を批判した
開始	を開始した
協議	協議した
要請	を要請した
発言	を示した
指摘	」と指摘した
調査	で分かった
予定	する予定
確認	を確認した
開催	で開かれた
示唆	を示唆した

抽出語詳細[発表]

を発表した
発表した
と発表した
すると発表した
したと発表した
結果を発表した
声明を発表した
計画を発表した
を明らかにした
調査結果を発表した

明らかにした
ことを明らかにした
たことを明らかにした
したことを明らかにした
する声明を発表した
となった
を正式に発表した
ることを明らかにした
見通しを発表した
になった

抽出語詳細[会談]

と会談した
会談した
で会談した
大統領と会談した
首相と会談した
外相と会談した
について意見交換した
で一致した
を表明した
と相次いで会談した

との認識で一致した
について協議した
を確認した
意見交換した
と首相官邸で会談した
協議した
を示した
問題などについて協議した
合意した
などについて意見交換した

形態素解析との違い

- 形態素解析を用いてもほぼ同様の言い換え語候補を得ることができる

抽出語の違い[発表]

１文字ずつ処理し、
言い換え語を抽出した場合

を発表した
発表した
と発表した
すると発表した
したと発表した
結果を発表した
声明を発表した
計画を発表した
を明らかにした
調査結果を発表した

形態素解析を用いて
言い換え語を抽出した場合

を発表した
と発表した
発表した
すると発表した
たと発表した
したと発表した
結果を発表した
声明を発表した
計画を発表した
を明らかにした

形態素単位で処理した場合の各順位での精度の平均値