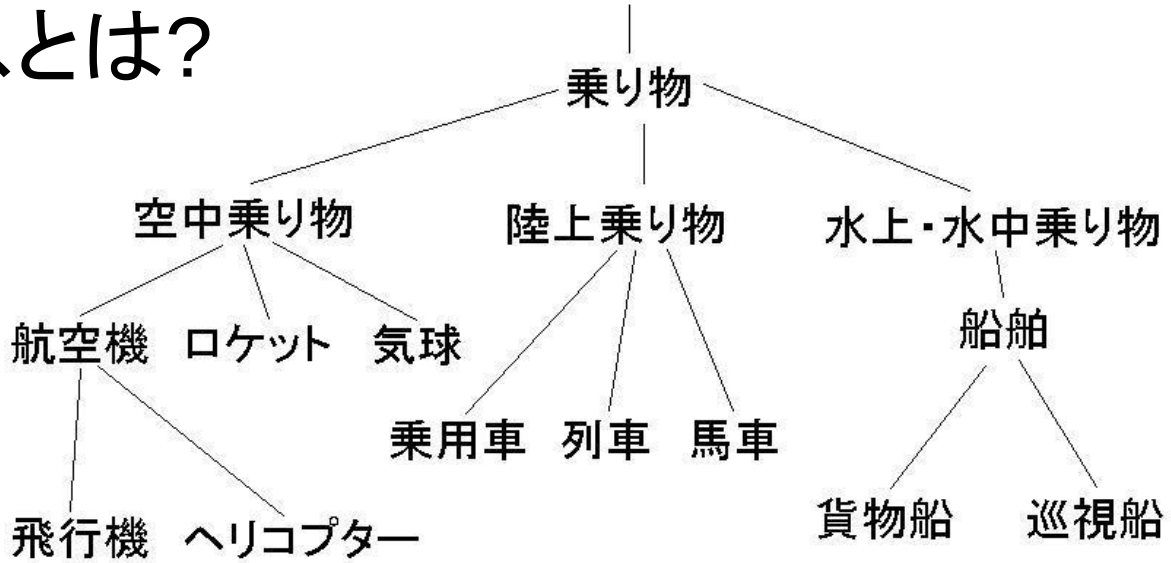
# 単語の共起関係と構文情報を利用した単語階層関係の統計的自動識別

大石 康智[1], 伊藤 克亘[2]
武田 一哉[1], 藤井 敦[3]

[1]名古屋大学大学院情報科学研究科
[2]法政大学情報科学部
[3]筑波大学図書館情報メディア研究科

# はじめに

- シソーラスとは?




  - 情報検索における検索質問の拡張
  - 機械翻訳
  - 既存のシソーラスにおける問題点
    - 人手による構築が主流
    - 多くの技術的な専門用語が未整理

# 自動構築のための従来研究と問題点

- 単一の国語辞典や百科事典の説明文の利用
  - 見出し語の説明量が少量
  - 辞書の改定が頻繁には行われない
- 単語間の並列関係を表す構文パターンの抽出
  (“～の一つ”, ”～は～である”, ”～のような”)
  - 構文パターンを網羅的に集める必要性
- 既存のシソーラスへの未知語の配置
  - 係り受け関係の類似度の算出
    - (飛行機, が, 飛ぶ)　(〇〇〇, が, 飛ぶ)
      未知語
  - 既存のシソーラス自体が人手による構築

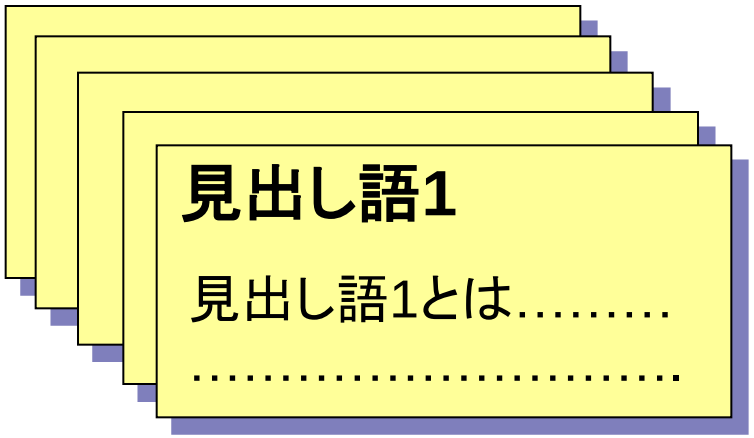
# 改善案

- Cycloneコーパス(事典的なコーパス)
  - Webを事典的に利用すること

⇒ 現在，約70万語以上の用語(見出し語)を取得

(1)Web検索エンジンを用い，
見出し語を含むWebページを網羅的に取得

(2)取得したページにおけるHTMLのタグ構造から
見出し語を含む段落を抽出

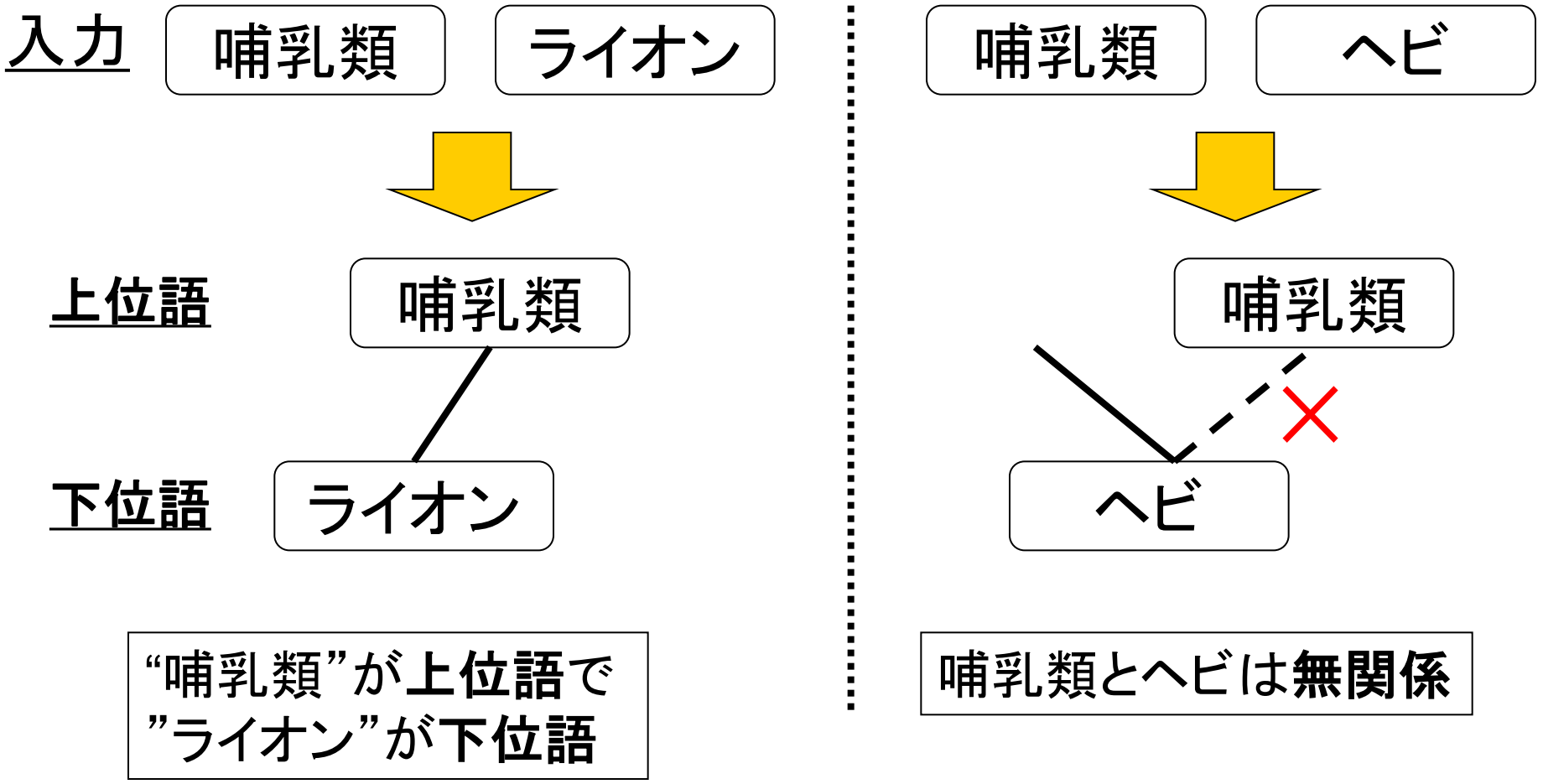



見出し語1の
複数の説明文

**問題点**
Webページの信頼性

# 本研究の目的

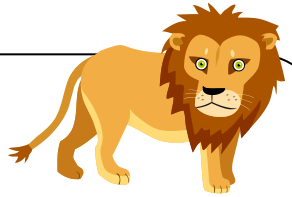
- 単語間の階層関係の自動識別



# 説明文の方向性を考慮した出現頻度モデル

- 説明文の方向性とは

**ライオン**



ネコ科の哺乳類。頭胴長2メートル内外、尾長90センチメートルほど。雄はたてがみがある。古くより「百獣の王」とよばれる。… 獅子(しし)。

**哺乳類**

脊椎動物門哺乳綱に属する動物の総称。大脳がよく発達し、……
恒温動物で、単孔類のみ卵生で、他はすべて胎生。……犬やネコなど。

“ネコ科の哺乳類” というように上位語 “ネコ”や ”哺乳類”を含む

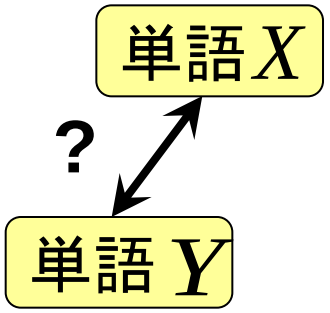
“犬やネコなど” というように必ずしも”ライオン”という下位語が説明に使われない

見出し語に対する複数の説明文において，
上位語は共通するが，下位語が，共通するとは限らない

# 説明文の方向性を考慮した出現頻度モデル

- 単語 $X$ が単語 $Y$ と階層関係をもつのか，無関係であるのかの指標 $|H(X|Y)|$

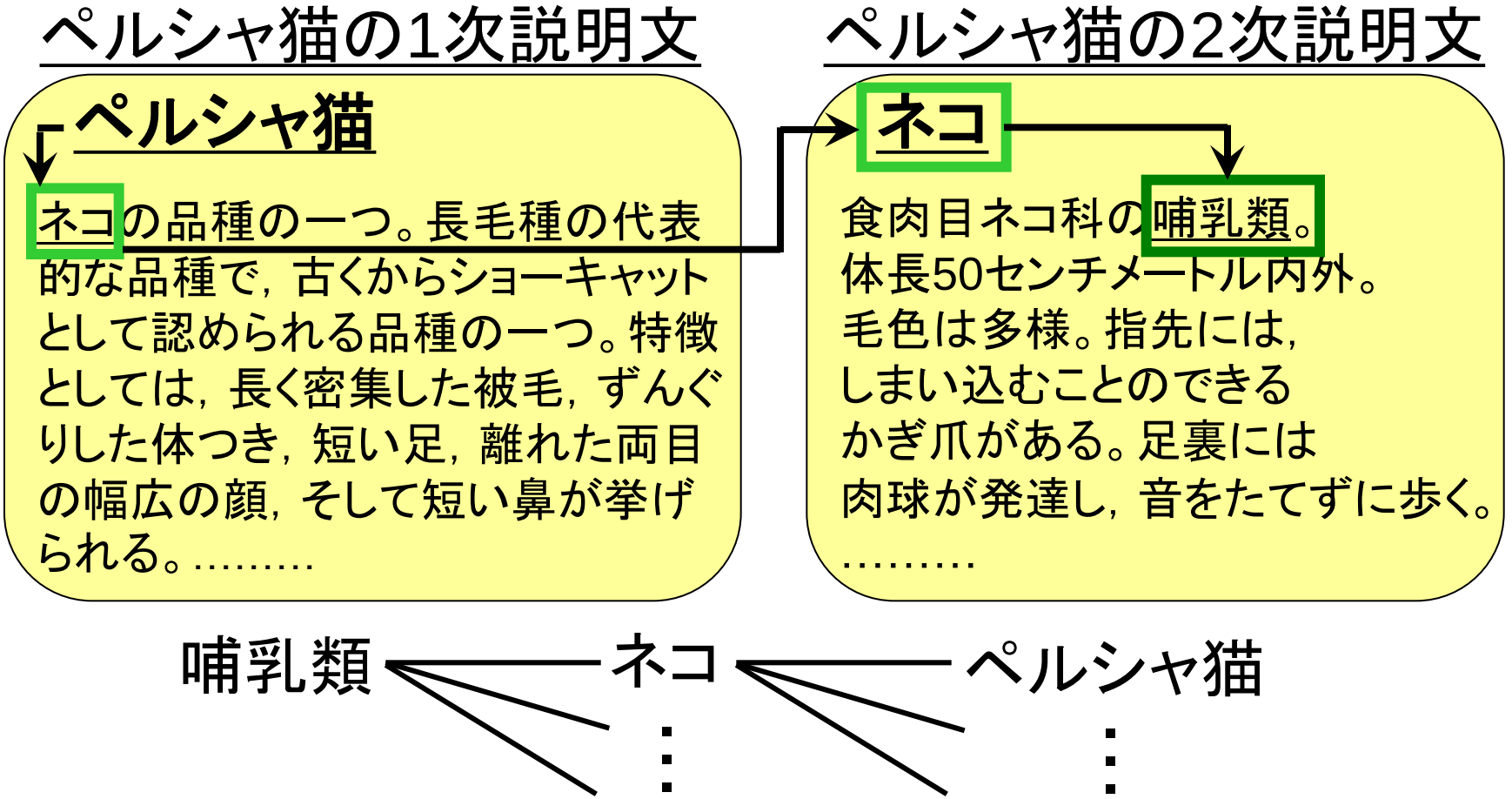
$$|H(X|Y)| = |C(X|Y) - C(Y|X)|$$




  - $C(X|Y)$は見出し語 $Y$とした説明文における単語 $X$ の出現確率
  - Cycloneコーパスは，1つの見出し語に対して，多くの観点に基づく複数の説明文を収集
    - $C(X|Y)$ と $C(Y|X)$ を計算
  - 階層関係をもつペアに対して $H(X|Y)$の符号に従って上位下位の決定

# 説明文の方向性を考慮した出現頻度モデル

- 大規模な階層関係を考慮するために
  - 単語の間接的な関係の利用



# 説明文の方向性を考慮した出現頻度モデル

- 説明文の展開
  - 説明文には出現しない単語の出現確率を，説明文を展開することにより間接的に推定
  - 見出し語 $W_j$ の説明文中の単語 $W_i$ の出現確率は

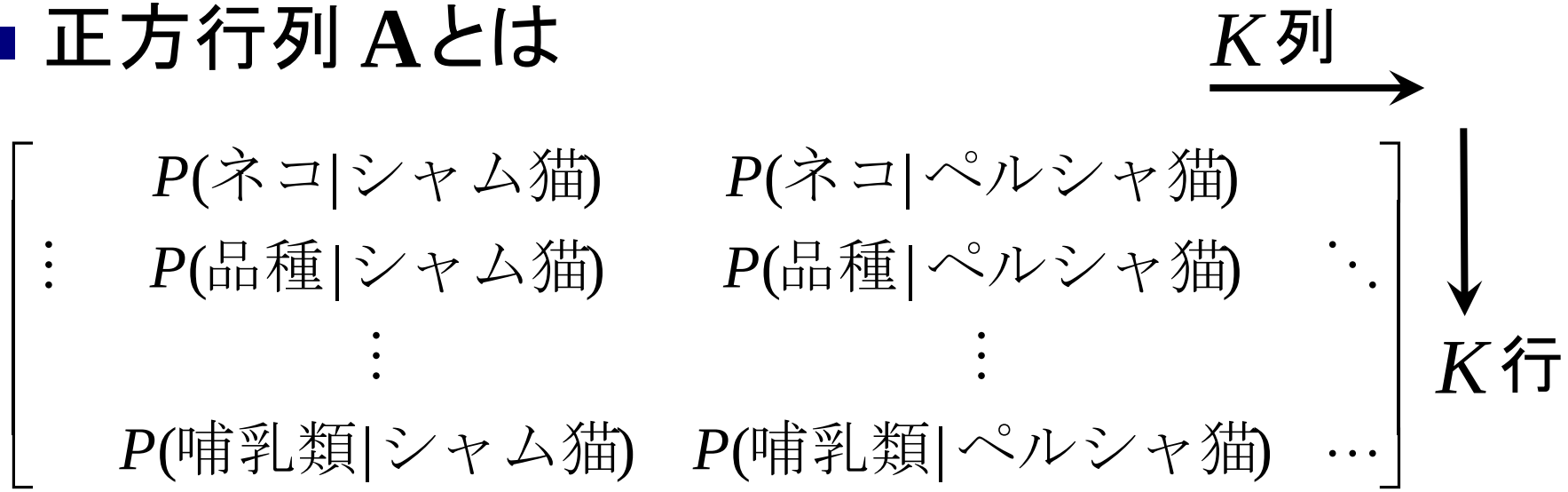
$$\mathbf{A}_{i,j} = P(w_i^{(1)} \mid w_j) = \frac{F(w_i \mid w_j)}{\sum_{k=1}^{K} F(w_i \mid w_j)} \quad (i, j = 1, \ldots, K)$$

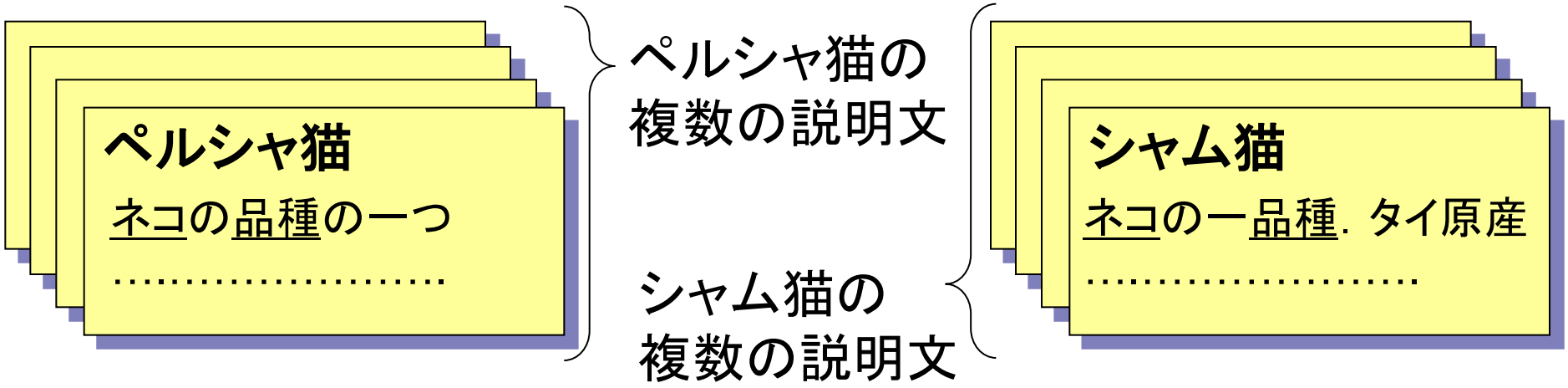
- $K$ は見出し語数，$F(w_i \mid w_j)$ は単語 $W_i$ の出現頻度

$P(w_i^{(1)} \mid w_j)$ : 見出し語 $W_j$ の1次説明文における単語 $W_i$ の出現確率

# 説明文の方向性を考慮した出現頻度モデル

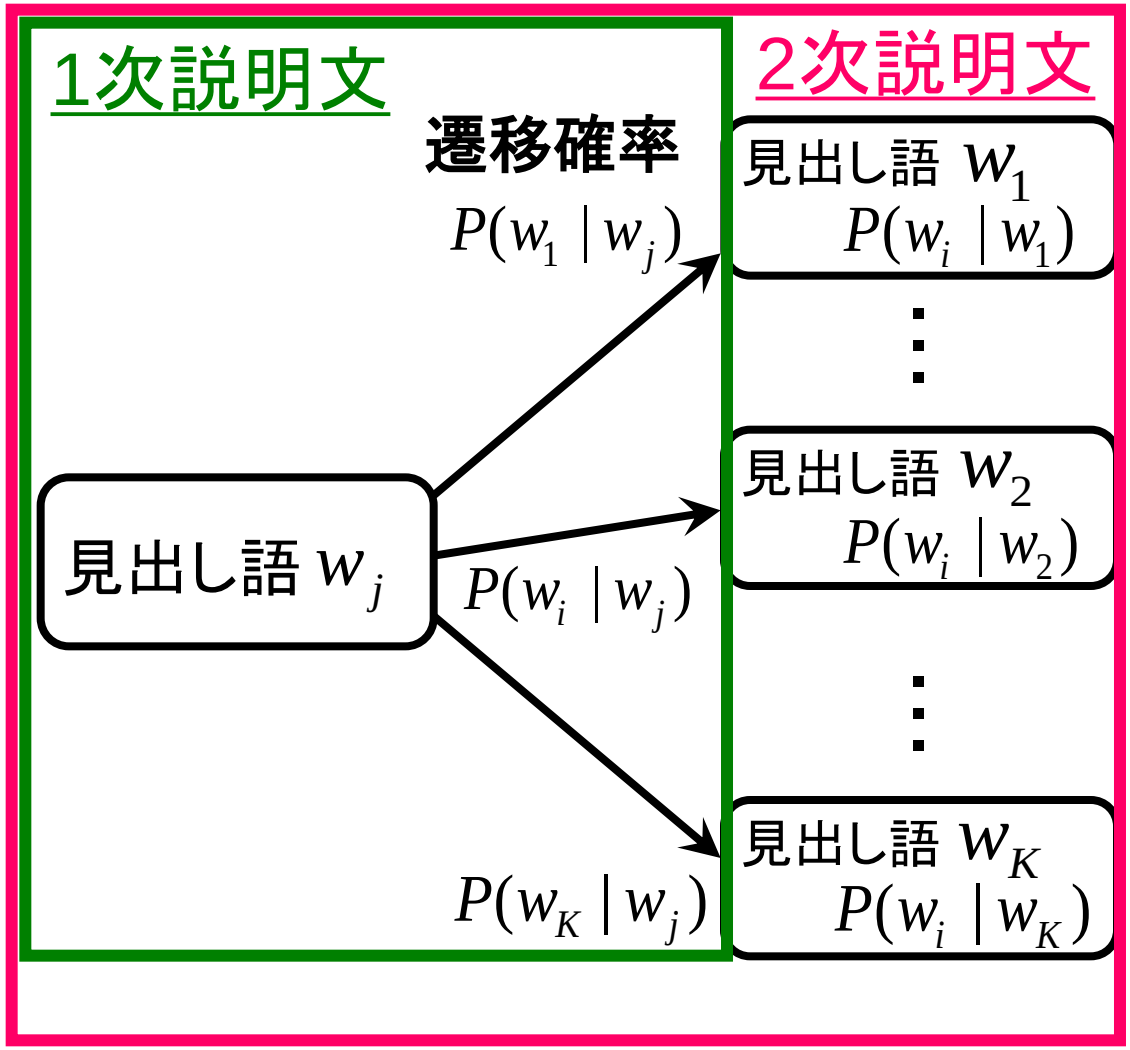
- 正方行列 **A**とは



# 説明文の方向性を考慮した出現頻度モデル

- 2次説明文の考え方



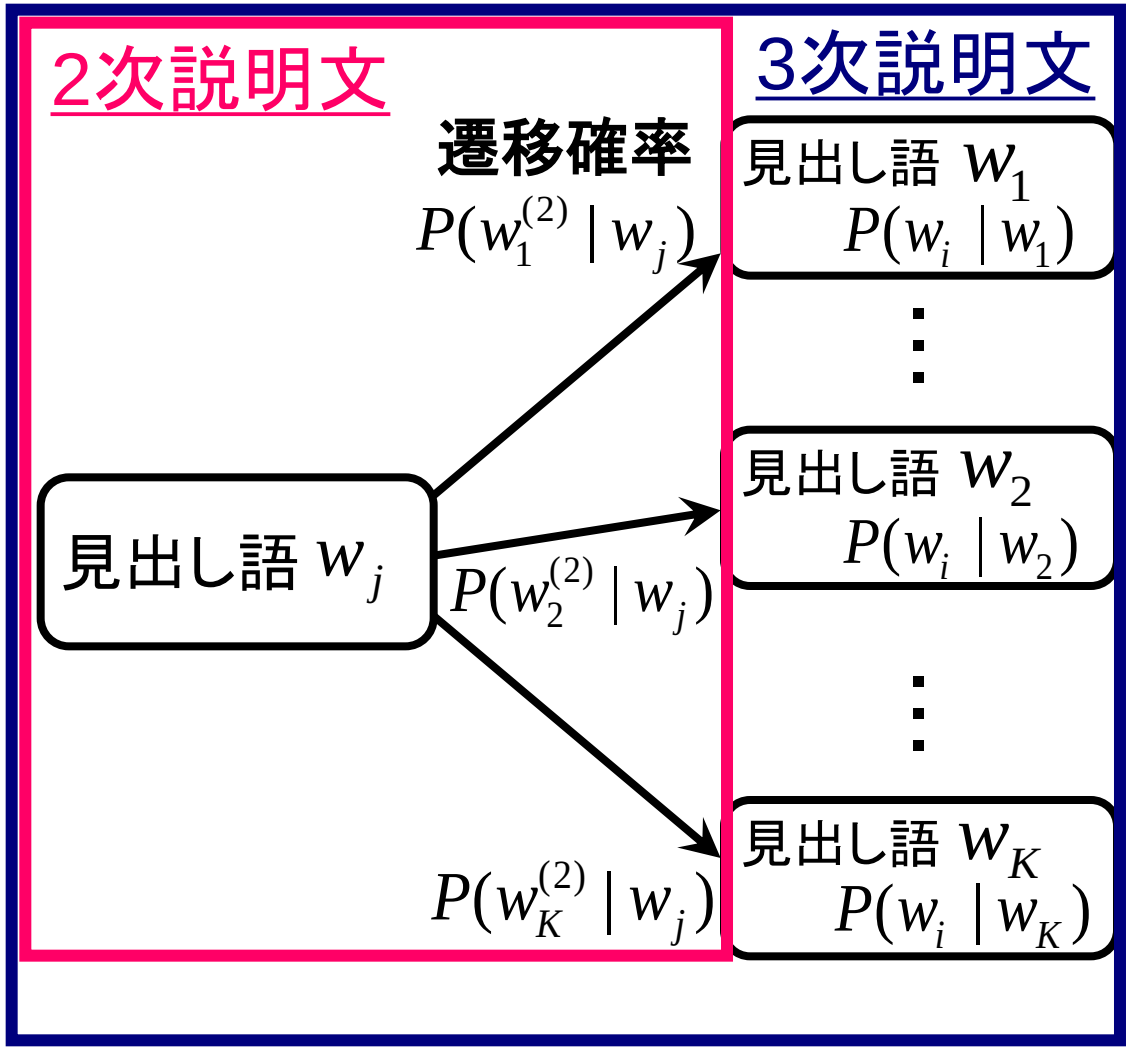

2次説明文

$$P(w_i^{(2)} \mid w_j) = \sum_k \underbrace{P(w_i \mid w_k)}_{\text{出現確率}} \underbrace{P(w_k \mid w_j)}_{\text{遷移確率}}$$

正方行列 $\mathbf{A}$ を用いると
2次説明文の全体は$\mathbf{A}^2$

# 説明文の方向性を考慮した出現頻度モデル

- 3次説明文の考え方




3次説明文

$$P(w_i^{(3)} \mid w_j) = \sum_k \underbrace{P(w_i \mid w_k)}_{\text{出現確率}} \underbrace{P(w_k^{(2)} \mid w_j)}_{\text{遷移確率}}$$

正方行列 $\mathbf{A}$ を用いると
3次説明文の全体は $\mathbf{A}^3$

# 説明文の方向性を考慮した出現頻度モデル

- n次説明文は正方行列 $\mathbf{A}^n$
- 拡張説明文
  - $1 \sim N$ 次説明文の線形結合

$$\mathbf{C} = \sum_{n=1}^{N} [\alpha_n \mathbf{A}^n]$$

  - 問題設定における $H(X \mid Y) = H(w_i \mid w_j)$ を算出

$$H(w_i \mid w_j) = \mathbf{C}_{i,j} - \mathbf{C}_{j,i}$$

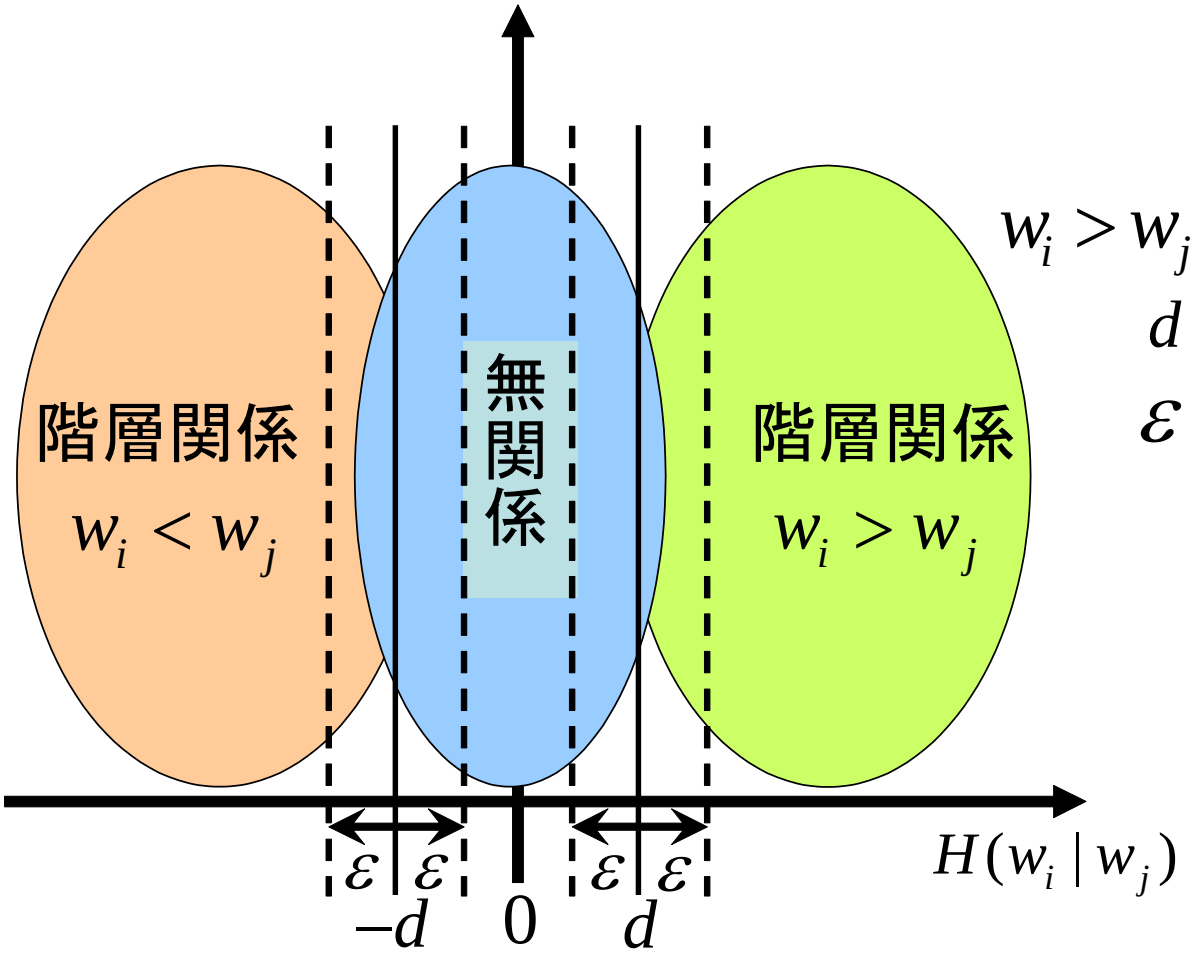
  - $\alpha_n$ は $\left| H(w_i \mid w_j) \right|$ の値によって $w_i$ と $w_j$ が
    - 階層関係であるのか
    - 無関係であるのか

正しく識別できるように線形判別分析により決定

# 説明文の方向性を考慮した出現頻度モデル

- $H(w_i \mid w_j)$の値を利用した2クラス分類

$$H(w_i \mid w_j) = \mathbf{C}_{i,j} - \mathbf{C}_{j,i}$$




$w_i > w_j$: $w_i$は$w_j$の上位語

$d$ : 識別境界

$\varepsilon$ : 識別境界におけるマージン.
$H(w_i \mid w_j)$の誤差による識別誤りを防ぐために設ける

# 説明文の方向性を考慮した出現頻度モデル

- $H(w_i \mid w_j)$の値を利用した3クラス分類

$$H(w_i \mid w_j) = \mathbf{C}_{i,j} - \mathbf{C}_{j,i}$$

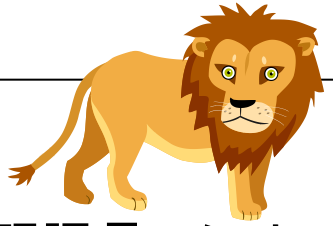
$$\begin{cases} H(w_i \mid w_j) > d + \varepsilon & (w_i \text{ は} w_j \text{ の上位語}) \\ H(w_i \mid w_j) < -(d + \varepsilon) & (w_i \text{ は} w_j \text{ の下位語}) \\ -d + \varepsilon \le H(w_i \mid w_j) \le d - \varepsilon & (w_i \text{ と} w_j \text{ は無関係}) \end{cases}$$

今回は $\varepsilon$ を0とする
すなわち識別境界におけるマージンを考慮しない

# さらに上位下位の識別性能を向上させるために

- 局所的な構文情報に基づく統計モデル
  - 単語の周辺に頻出する形態素に着目
    - 単語に後接する形態素は?

## ライオン



ネコ科の哺乳類。頭胴長2メートル内外、尾長90センチメートルほど。雄はたてがみがある。古くより「百獣の王」とよばれる。… 獅子(しし)。

“ネコ”に後接する”科”
“哺乳類”に後接する”。”

## 哺乳類

脊椎動物門哺乳綱に属する動物の総称。大脳がよく発達し、……
恒温動物で、単孔類のみ卵生で、他はすべて胎生。… 犬やネコなど。

“犬”に後接する”や”
“ネコ”に後接する”など”

見出し語に対して，上位語なのか下位語なのか?

# 局所的な構文情報に基づく統計モデル

- 構文ベクトルの作成
  - 単語に後接する形態素の確率ベクトル

(1)説明文中の見出し語の上位語・下位語に後接する形態素を抽出

(2)頻度の高いもの上位 $W$ 種類に対する構文ベクトルを作成

$$\mathbf{x}_{w_i|w_j} = (x(s_0), x(s_1), \ldots, x(s_W))$$

$s_k$ は $w_i$ に後接する形態素のうち $k$ 番目に頻出するもの

$s_0$ は上位の $W$ 種類以外の残りの形態素

$$x(s_k) = \frac{F(s_k)}{\sum_{l=0}^{W} F(s_l)}$$

$F(s_k)$ は $w_i$ に後接する形態素 $s_k$ の頻度

# 構文ベクトル $\mathbf{x}_{w_i|w_j}$ の例

- 単語に後接する形態素の確率ベクトル
  - 見出し語 $w_j$ :ライオン，上位語 $w_i$ :哺乳類

$\mathbf{x}_{\text{哺乳類|ライオン}} = (x(s_0), x(\text{の}), x(\text{は}), \ldots, x(\text{など}), \ldots)$

$= (\ 0.1\ ,\ 0.3\ ,\ 0.2\ , \ldots, 0.002\ , \ldots)$

  - 見出し語 $w_j$ :哺乳類，　下位語 $w_i$ :ネコ

$\mathbf{x}_{\text{ネコ|哺乳類}} = (x(s_0), x(\text{の}), x(\text{は}), \ldots, x(\text{など}), \ldots)$

$= (\ 0.1\ , 0.05\ ,\ 0.1\ , \ldots,\ 0.2\ , \ldots)$

# 局所的な構文情報に基づく統計モデル

- 拡張説明文に基づいた構文ベクトル $\mathbf{X}_{w_i|w_j}$ の算出
  - $\mathbf{x}_{w_i|w_j}$ が疎であるため識別に有効でないため

(1)見出し語 $w_j$ のn次説明文における単語 $w_i$ の出現確率は $P(w_i^{(n)} \mid w_j)$

$$\mathbf{x}_{w_i|w_j}^{(n)} = P(w_i^{(n)} \mid w_j) \cdot \mathbf{x}_{w_i|w_j}$$
$$= (\mathbf{A}^n)_{i,j} \cdot \mathbf{x}_{w_i|w_j}$$

n次説明文での $w_i$ に対する構文ベクトル

(2)1～N次説明文まで重み付き加算

$$\mathbf{X}_{w_i|w_j} = \sum_{n=1}^{N} [q_n \mathbf{x}_{w_i|w_j}^{(n)}] \qquad q_n = 1 \quad (n = 1, \ldots, N)$$

# 局所的な構文情報に基づく統計モデル

- 構文ベクトル $\mathbf{X}_{w_i|w_j}$ を利用した単語階層関係の上位下位推定
  - ロジスティック回帰分析の利用

$$\Pr(w_i > w_j \mid \mathbf{X} = \mathbf{X}_{w_i|w_j}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_i|w_j})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_i|w_j})}$$

$$\Pr(w_i < w_j \mid \mathbf{X} = \mathbf{X}_{w_i|w_j}) = \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_i|w_j})}$$

ここで

$$\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_i|w_j} = \beta_0 + \beta_1 x(s_0) + \beta_2 x(s_1) + \cdots + \beta_W x(s_W)$$

$x(s_0), x(s_1), \cdots, x(s_W)$ :単語 $w_i$ に後接する形態素の出現確率

⇒ 学習データを利用して $\boldsymbol{\beta}$ を推定

# 局所的な構文情報に基づく統計モデル

- 推定したβを利用した出力の評価方法

$$\Pr(w_i > w_j \mid \mathbf{X}_{w_i|w_j}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_i|w_j})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_i|w_j})}$$

$$\Pr(w_i < w_j \mid \mathbf{X}_{w_i|w_j}) = \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_i|w_j})}$$

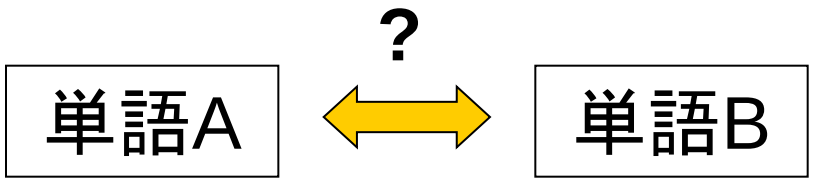
$$\Pr(w_i > w_j \mid \mathbf{X}_{w_j|w_i}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_j|w_i})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_j|w_i})}$$

$$\Pr(w_i < w_j \mid \mathbf{X}_{w_j|w_i}) = \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_{w_j|w_i})}$$

最も確率の大きいものを階層関係の判定結果とする

# 評価実験

- 2つの単語の意味関係の自動識別
  - 階層関係(どちらが上位でどちらが下位か?)
  - 無関係



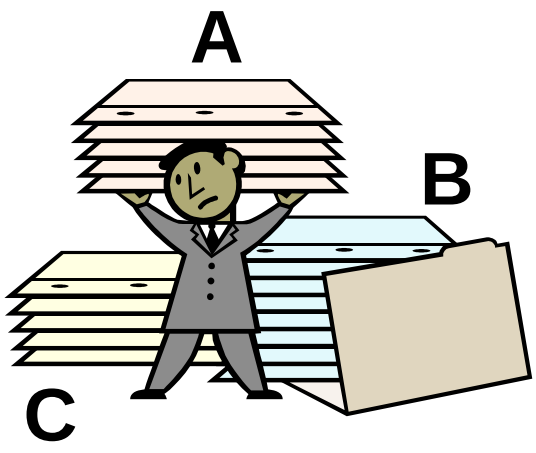

- 使用データ
  - Cycloneコーパスの
    コンピュータ関連の見出し語(2074語)
- 正解データの作成
  - JICST科学技術シソーラス1992年度版の利用
    - 見出し語2074語のうち，172語を記述

# 評価実験

- 無関係データの作成
  - 2074語からランダムに500組を抽出
  - 手動で判定したところ497組の無関係データの取得
- 見出し語に対する説明文の精度
  - Cycloneコーパスは Web文書を利用しているため信頼性に偏りがある

手動で以下のように判定

- **A** (見出し語を正しく説明している)
- **B** (見出し語を部分的に説明している)
- **C** (見出し語を説明していない)



# 使用データの一覧

コンピュータ関連の見出し語2074語の説明文の精度とテストセットデータ

| 判定 | 見出し語 | 平均説明文数 | テストセット | |
|---|---|---|---|---|
| | | | 正解データ | 無関係データ |
| A | 1624 | 6.62 | 136組 | 301組 |
| A+B | 1803 | 10.4 | 168組 | 366組 |
| ALL | 2074 | 80.7 | 206組 | 497組 |

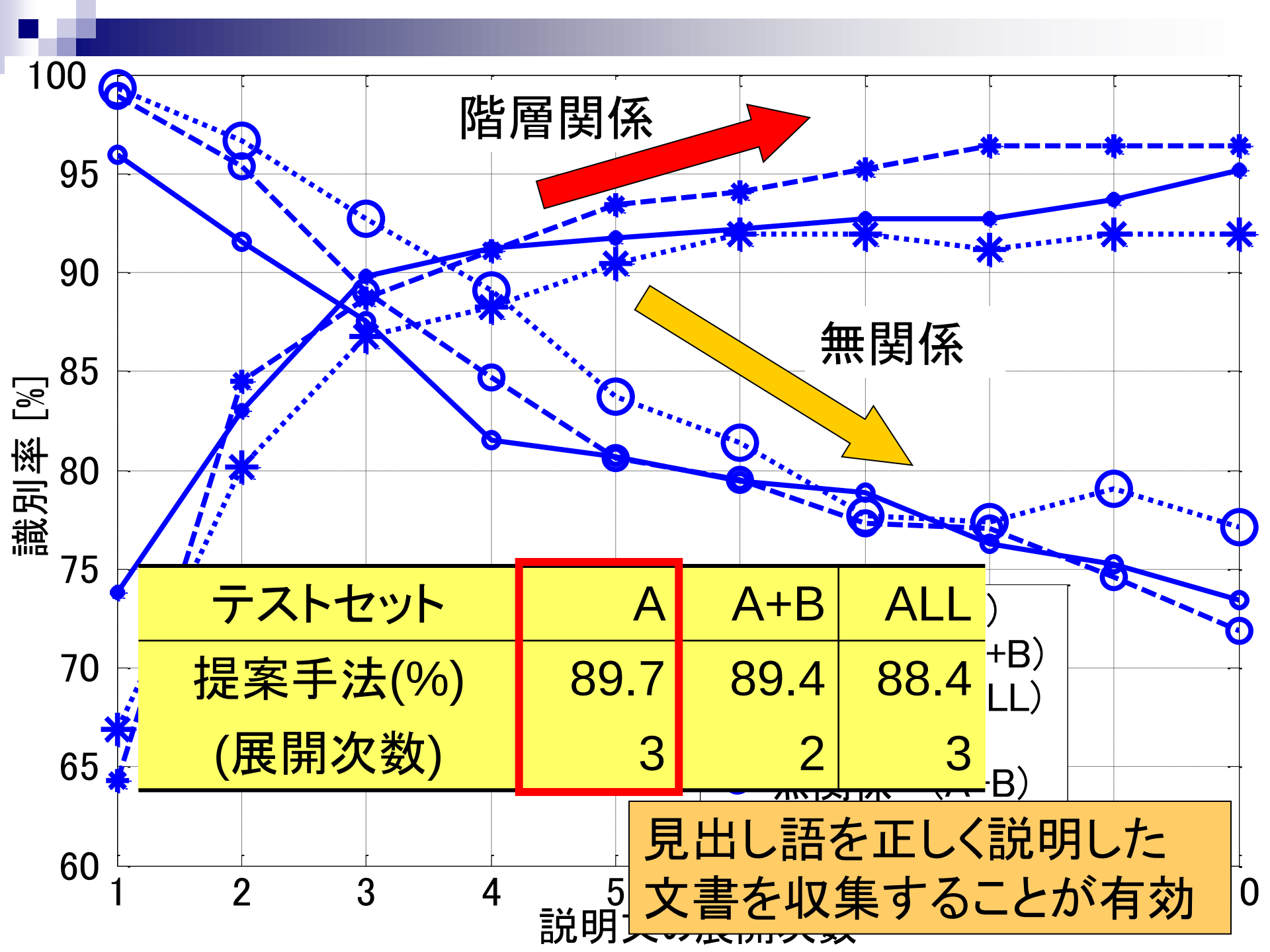
- JICSTシソーラスは見出し語172語を記述
- テストセットに含まれていないその他の見出し語の関係については，今回調査は行わない

# 2つの単語の階層関係・無関係の2群識別実験

- 説明文の方向性を考慮した出現頻度モデルの評価
  - 線形判別分析により，拡張説明文の重み $\alpha_n$ を学習
  - 従来手法との比較
    - 指数重み手法(鈴木, 2003)

$$\mathbf{C} = \lim_{n \to \infty} b(a\mathbf{A} + a^2\mathbf{A}^2 + \ldots + a^n\mathbf{A}^n)$$

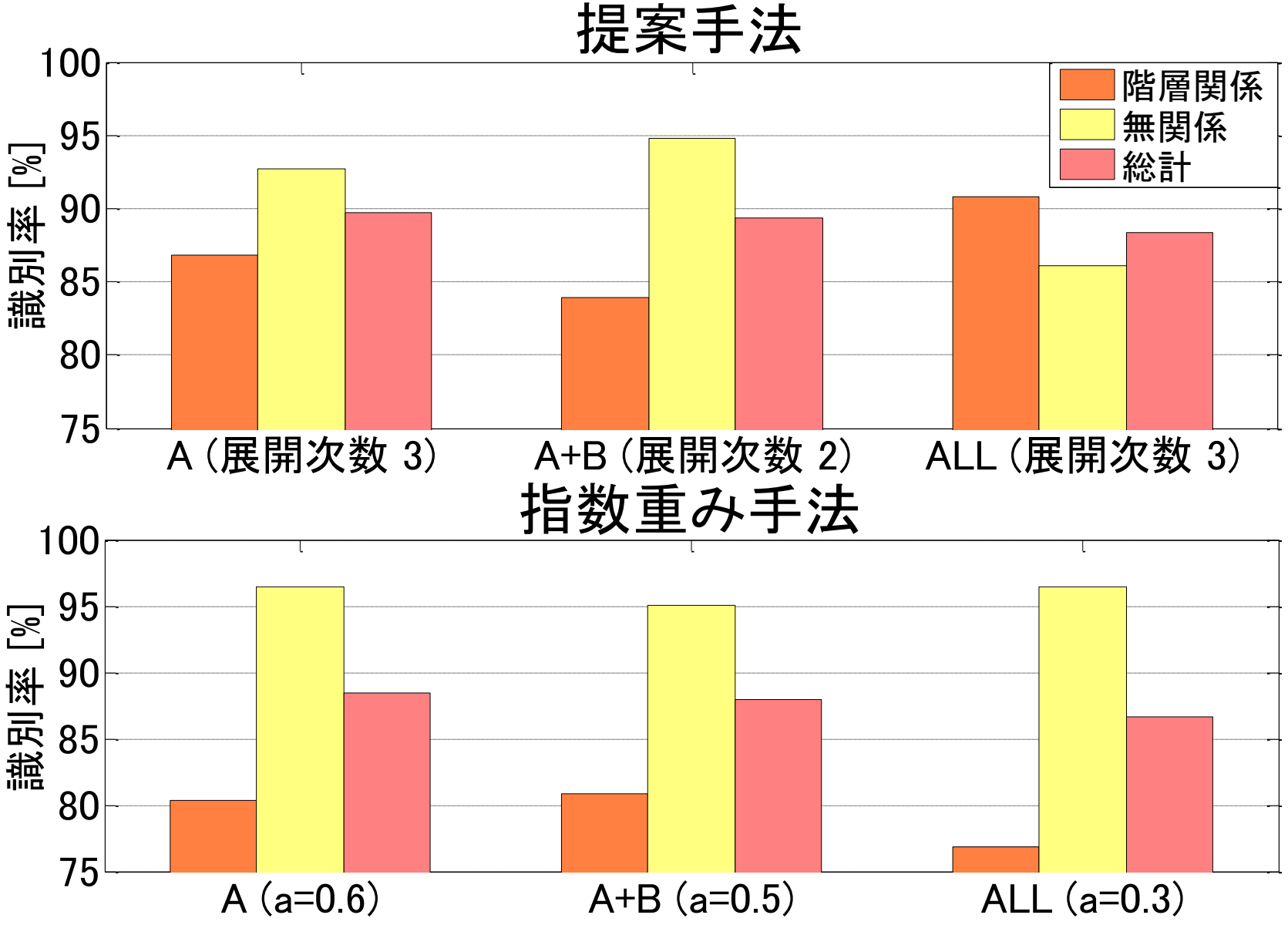
  - 評価方法
    - テストセットA, A+B, ALLに対する 4-fold クロスバリデーション
    - 説明文の展開次数 $N$ を変化



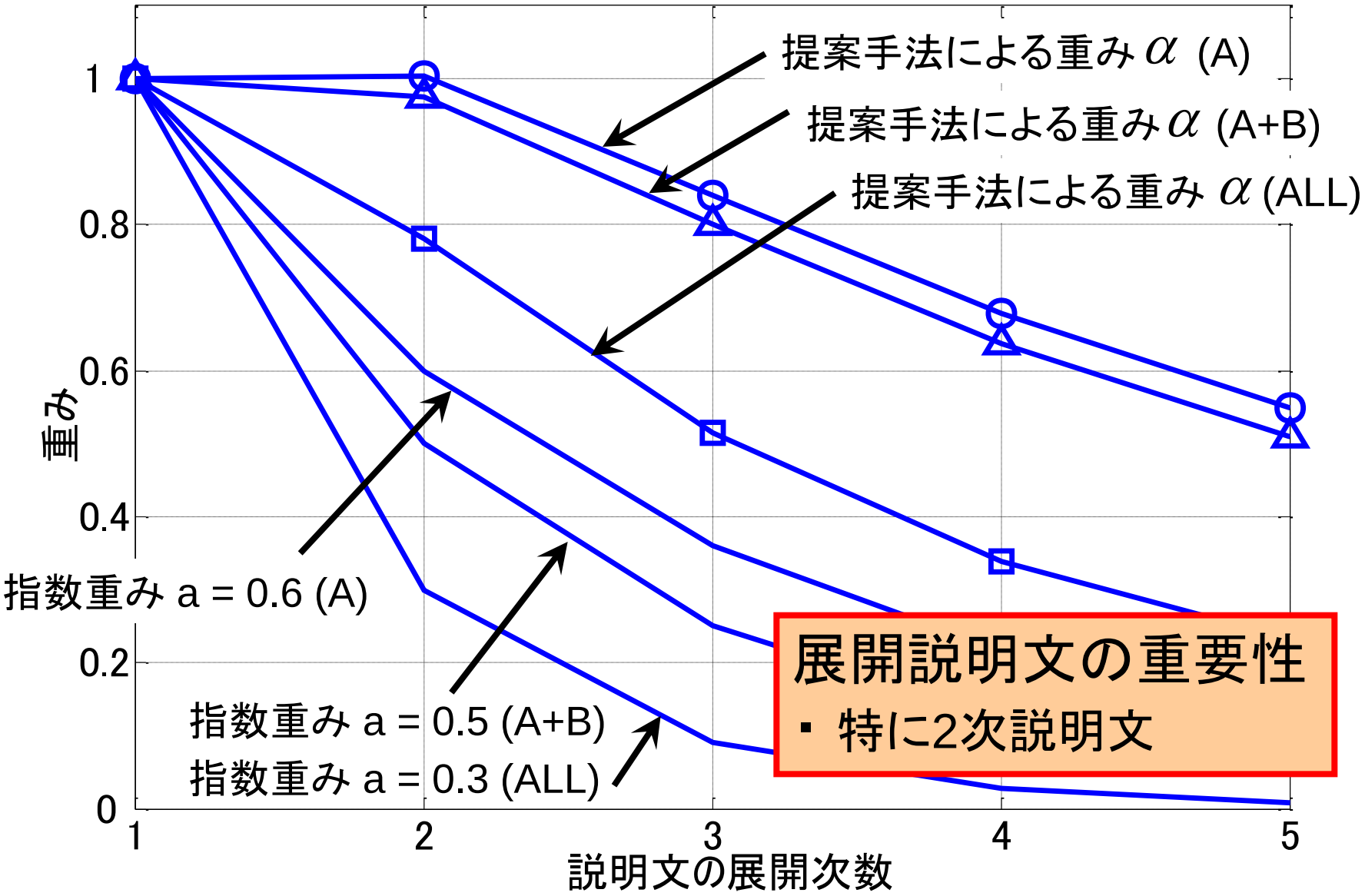

| テストセット | A | A+B | ALL |
|---|---|---|---|
| 提案手法(%) | 89.7 | 89.4 | 88.4 |
| (展開次数) | 3 | 2 | 3 |

見出し語を正しく説明した
文書を収集することが有効

# 指数重み手法との比較(1) - 最も性能が高い次数 -



# 指数重み手法との比較(2) - 5次までの重み -




**展開説明文の重要性**

- 特に2次説明文

# 階層関係をもつ単語間の上下判定のための識別実験

- どちらが上位語でどちらが下位語か?

$$H(w_i \mid w_j) = \mathbf{C}_{i,j} - \mathbf{C}_{j,i}$$

$H(w_i \mid w_j) > d$

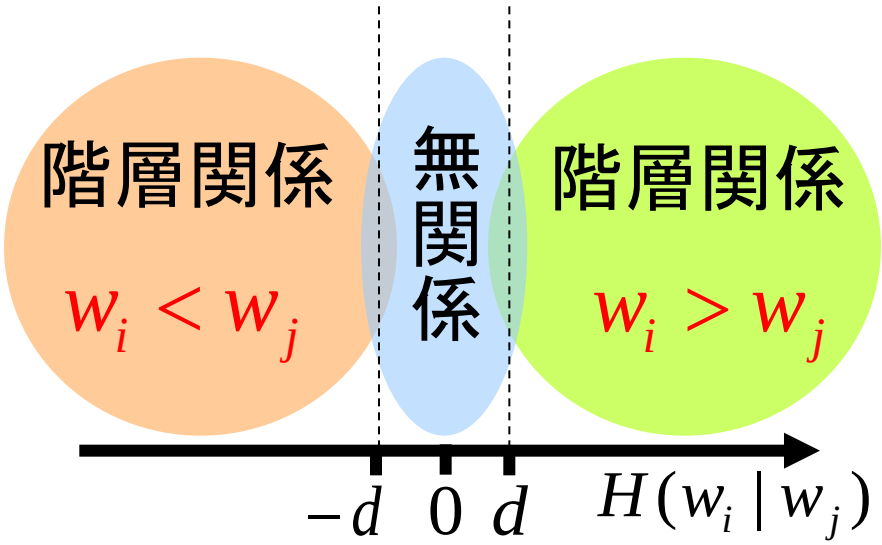
⇒ $w_i$ は $w_j$ の上位語

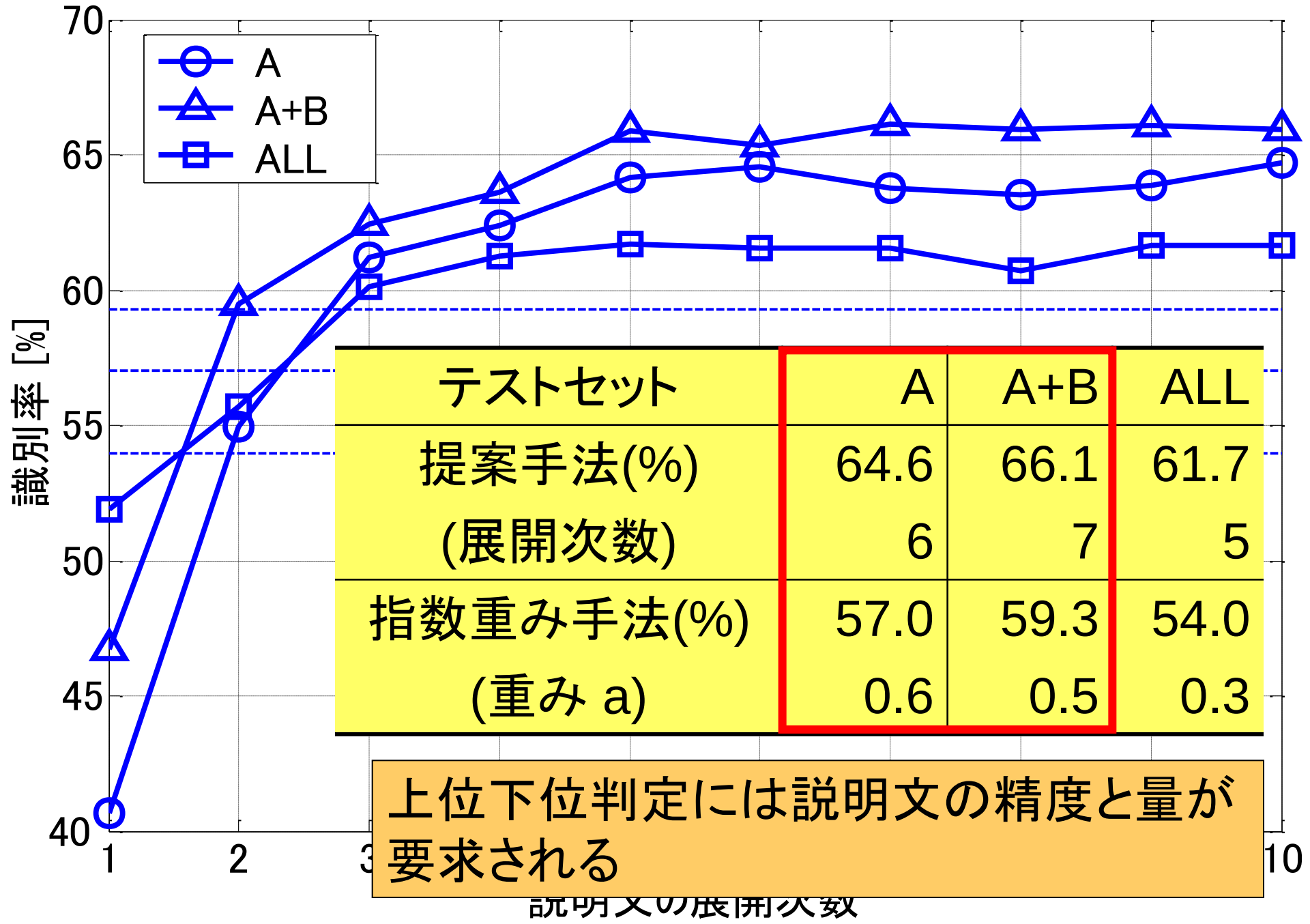
$H(w_i \mid w_j) < -d$

⇒ $w_i$ は $w_j$ の下位語

$-d \le H(w_i \mid w_j) \le d$

⇒ $w_i$ と $w_j$ は無関係

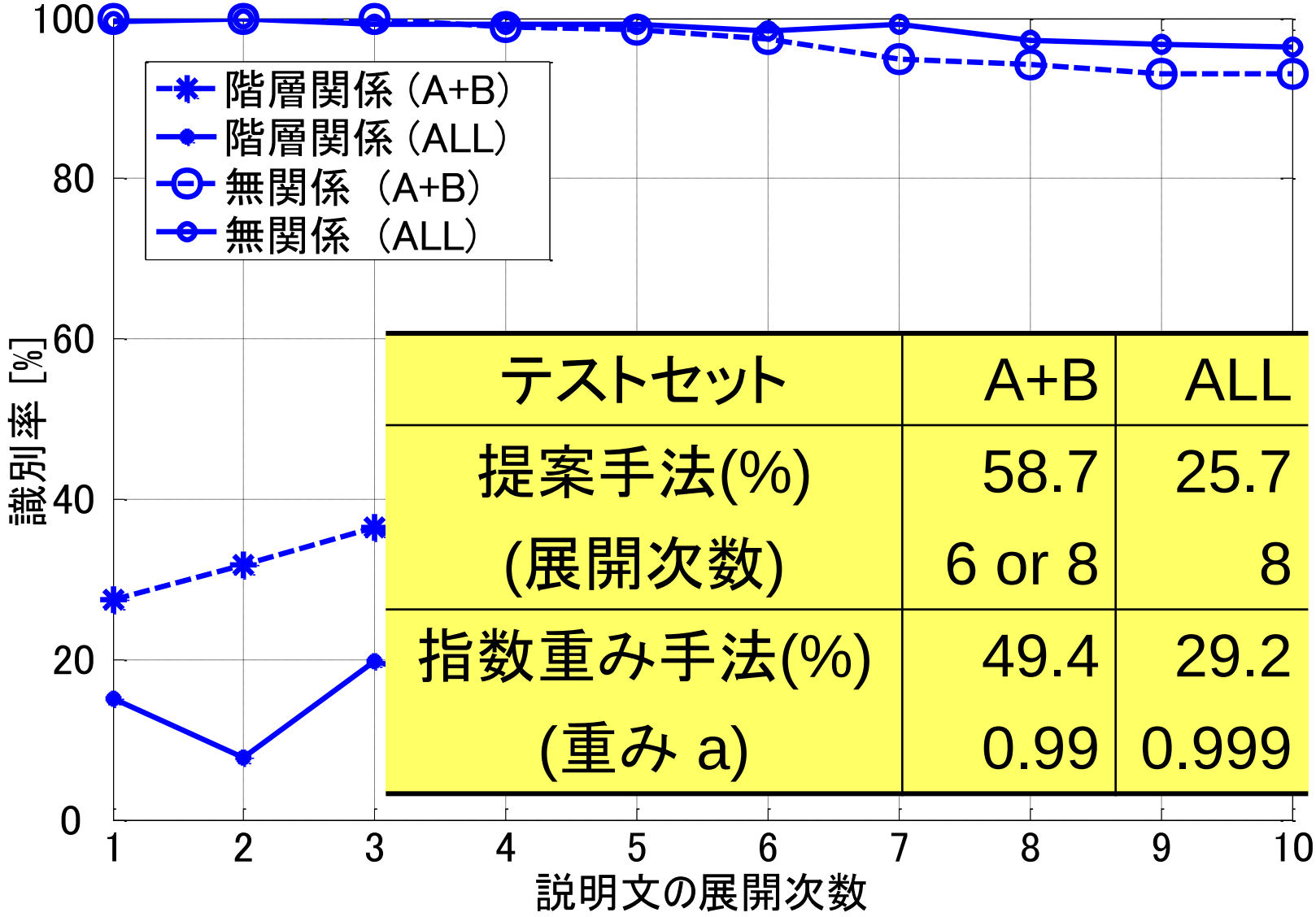





| テストセット | A | A+B | ALL |
|---|---|---|---|
| 提案手法(%) | 64.6 | 66.1 | 61.7 |
| (展開次数) | 6 | 7 | 5 |
| 指数重み手法(%) | 57.0 | 59.3 | 54.0 |
| (重み a) | 0.6 | 0.5 | 0.3 |

上位下位判定には説明文の精度と量が要求される

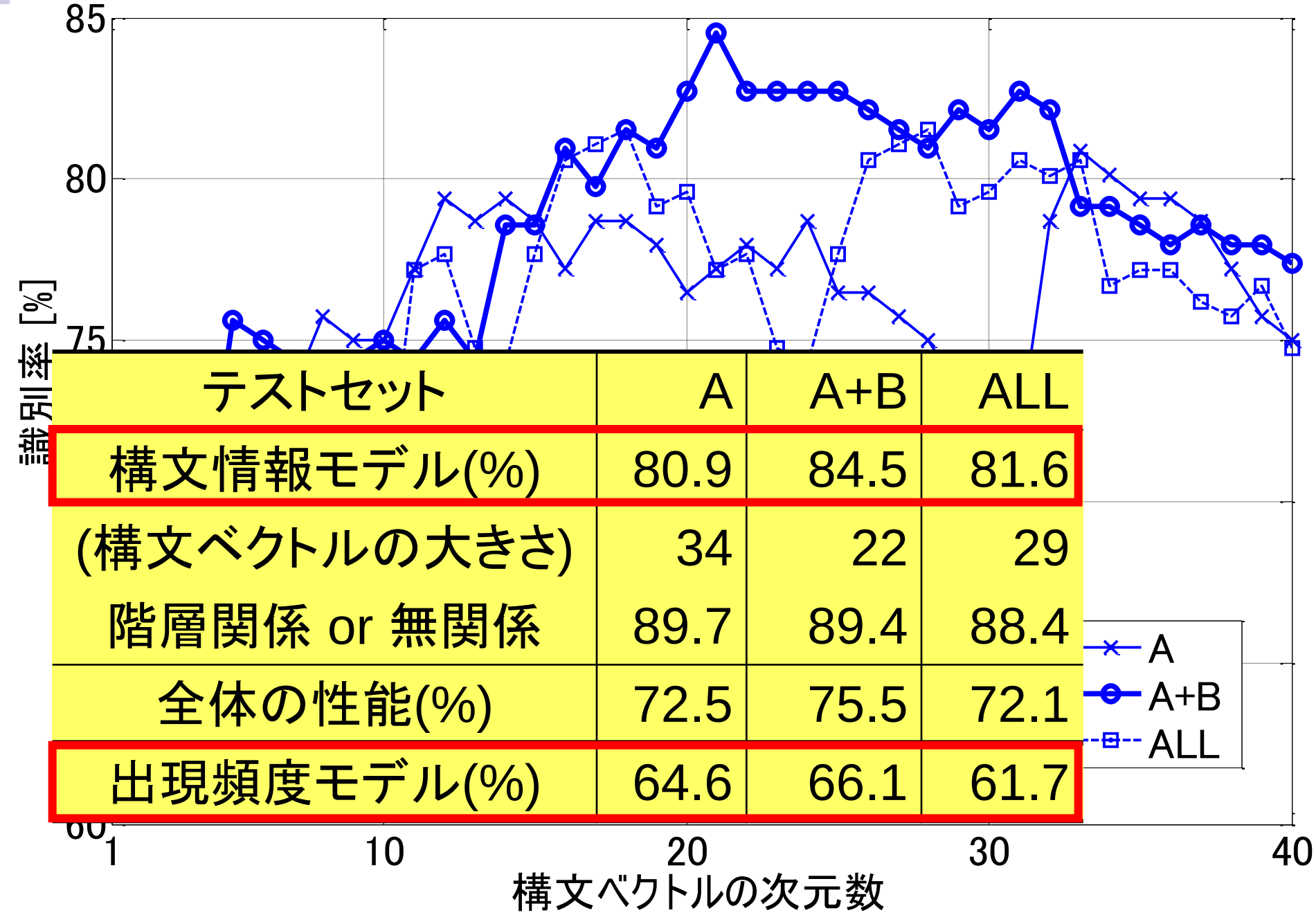
# 単一の説明文の利用による単語階層関係･無関係の識別性能




| テストセット | A+B | ALL |
|---|---|---|
| 提案手法(%) | 58.7 | 25.7 |
| (展開次数) | 6 or 8 | 8 |
| 指数重み手法(%) | 49.4 | 29.2 |
| (重み a) | 0.99 | 0.999 |

# 構文情報を利用した上位下位判定の評価

- 階層関係をもつと識別されたテストデータ
  - 構文ベクトルの算出
    - $\mathbf{X}_{w_i|w_j}$ :見出し語 $w_j$ の説明文における単語 $w_i$ に後接する形態素の出現確率
    - $\mathbf{X}_{w_j|w_i}$ :見出し語 $w_i$ の説明文における単語 $w_j$ に後接する形態素の出現確率
  - ロジスティック回帰分析による評価




| テストセット | A | A+B | ALL |
|---|---|---|---|
| 構文情報モデル(%) | 80.9 | 84.5 | 81.6 |
| (構文ベクトルの大きさ) | 34 | 22 | 29 |
| 階層関係 or 無関係 | 89.7 | 89.4 | 88.4 |
| 全体の性能(%) | 72.5 | 75.5 | 72.1 |
| 出現頻度モデル(%) | 64.6 | 66.1 | 61.7 |

# まとめ

- Cycloneコーパスを利用した単語間の階層関係を判定するための統計的推定手法の提案
  - 説明文の方向性を考慮した出現頻度モデル
  - 局所的な構文情報に基づく統計モデル
- JICSTシソーラスに記述されたコンピュータ関連の見出し語の関係のうち75.5%の階層関係を検出
- 単語間の階層関係・無関係を識別するためには，
  - 精度の高い説明文を数多く集めることの重要性
  - 実験データとして，Cycloneコーパスの有効性

# 今後の展開

- 単語間の意味的な関係として
  同義語，関連語の概念を導入すること
- 2つの単語は，
  - 階層関係である
  - 同義・関連関係である
  - 全く意味的な関係をもたない

  3カテゴリの識別問題への拡張
- 性能の向上
- 識別方法の検討