

言語処理学会  
第15回年次大会 (NLP 2009)

グラフカーネルに基づく  
非分かち書き文からの意味的語彙カテゴリ抽出

---

名古屋大学 大学院 情報科学研究科  
萩原正人 小川泰弘 外山勝彦

# 背景



- ▶ 固有表現, 固有表現間の関係
  - ▶ 意味解析において重要な知識源
  - ▶ 人手による構築・維持には大きなコスト
- ▶ ブートストラップによる半教師有り獲得
  - ▶ 語の意味的關係(is-a関係, part-of関係など)  
意味カテゴリ(大統領の名前, 国名など)
  - ▶ 少数のシードを入力, 文脈パターンを手がかりとして学習
    - ▶ *Snowbol* [Agichtein and Gravano 00]
    - ▶ *Espresso* [Pantel and Penacchiotti 06]
    - ▶ *Tchai* [Komachi and Suzuki 08]



- 英語などの分かち書きされた文や短い検索ログを対象
- 非分かち書き文に対して直接適用できない



# Monakaアルゴリズム [Hagiwara et al. 08]

シード

日本  
イタリア  
フランス

... で、協定に署名する。これを**受け**、**日本政府**は次期通常国会に協定の ...  
... 上で15日未明、17歳から39歳の**イタリア人男性**6人が車の中で頭を銃で ...  
... を題材にした ... など7点が**在日フランス大使館**の後援 ... して展示され ...

左側文脈

受け、#  
の#  
在日#

右側文脈

#政府  
#人男性  
#大使館

右側・左側文脈で挟む  
→ 挟み込み制約

インスタンス

イラク  
メキシコ  
コロンビア

... が15日に執行されたことを**受け**、**イラク政府**のダッバグ報道官は同日の...  
... 暮らす ... 動 ...  
... 位置す ... れ ...

正しく分かち書きされた重要語を獲得可能



# ブートストラップ法の拡張

---

*Espresso*

[Pantel &  
Penacchiotti 06]

英語・分かち書き文



*Monaka*

[Hagiwara et al. 08]

日本語・非分かち書き文



# ブートストラップ法の問題点

---

- ▶ 意味ドリフトが発生
  - ▶ ジェネリック・パターンにより, シードと関連の無い語が抽出
    - ▶ 右側文脈「#で」など
  - ▶ *Espresso*, *Monaka*などのブートストラップ法に不可避
- ▶ 定式化が不十分
  - ▶ 信頼度などスコアリングの根拠
  - ▶ ヒューリスティックに基づくフィルタリング
- ▶ 調整すべきパラメータが多い
  - ▶ 獲得するインスタンス数, パターン数, ...



# ブートストラップ法の拡張

*Espresso*  
[Pantel & Penacchiotti 06]

英語・分かち書き文



*Monaka*  
[Hagiwara et al. 08]

日本語・非分かち書き文



*(g-Espresso)*  
[Komachi et al. 08]

グラフ解析による定式化



*g-Monaka ??*





# 目的

---

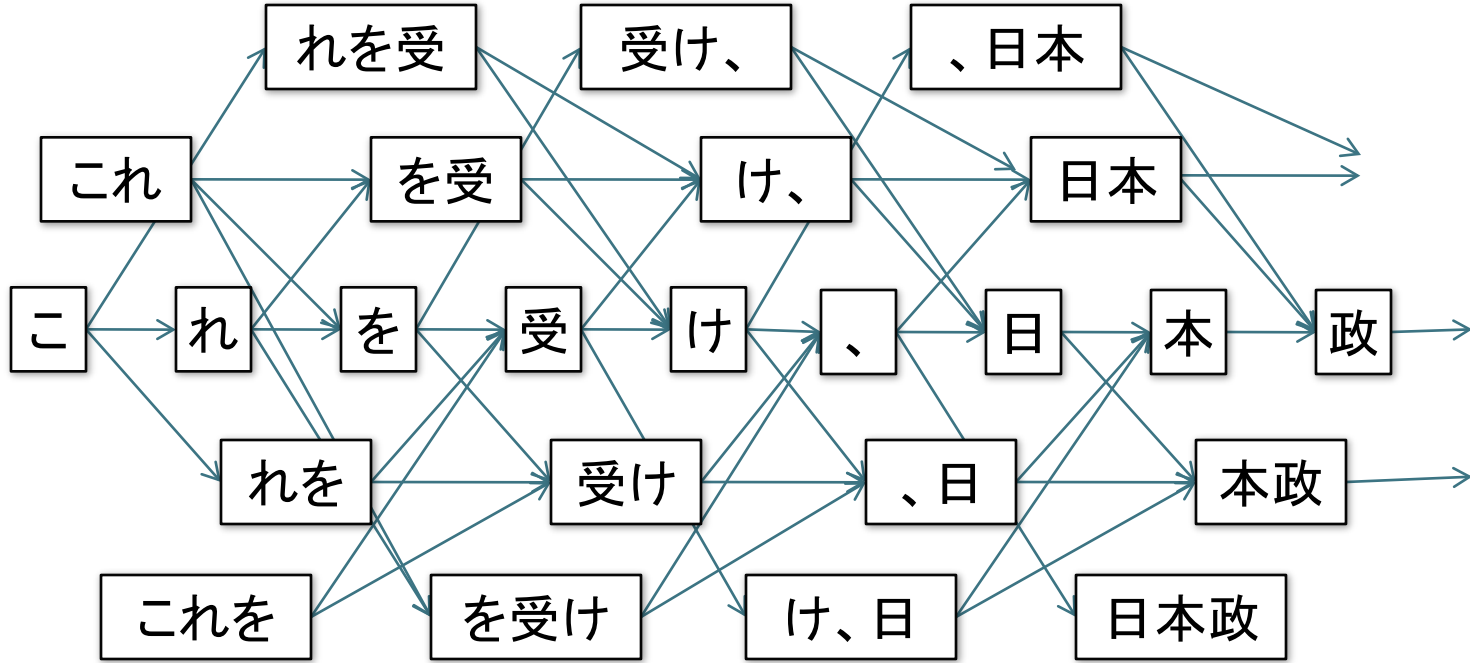
- ▶ 非分かち書き文からの意味カテゴリ抽出アルゴリズム *g-Monaka*を提案
  - ▶ グラフカーネルを用い, ブートストラップ法の問題に対処
  - ▶ 文字 $n$ グラムの隣接関係を有向グラフとして表現
  - ▶ 左側/右側文脈による挟み込み制約を表現
  - ▶ 分布類似度, *Espresso*, *Monaka*などの従来手法との比較



# g-Monakaアルゴリズム

## ▶ 隣接関係のモデル化

▶ 文字 $n$ グラムの隣接関係 → 重み付き有向グラフ



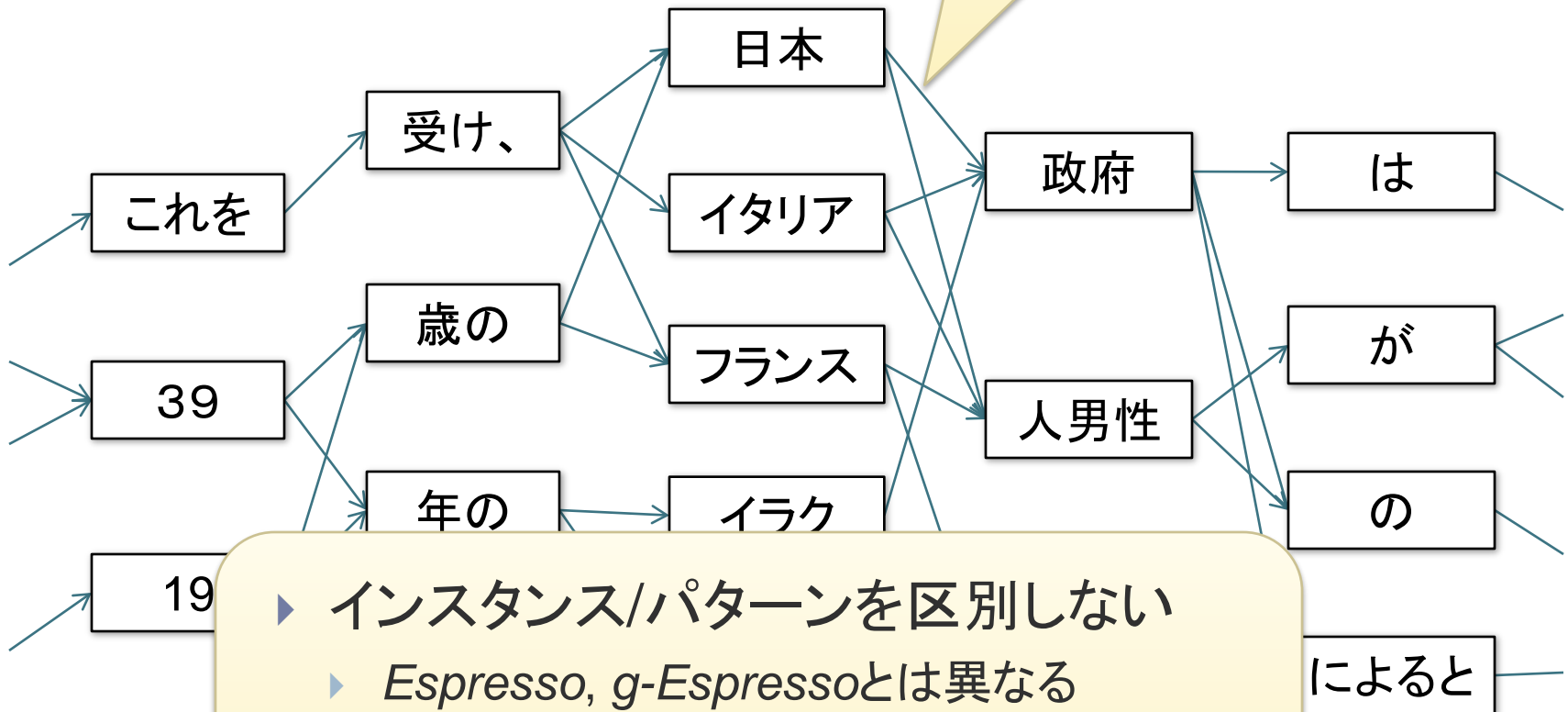
... これを受け、日本政府は次期通常国会に協定の承認案を提出する考えた ...

# *g-Monaka*アルゴリズム

接続の強さ(PMI)によって  
重み付け

## ▶ 隣接関係のモデル化

- ▶ 文字 $n$ グラムの隣接関係 → 重み付き有向グラフ



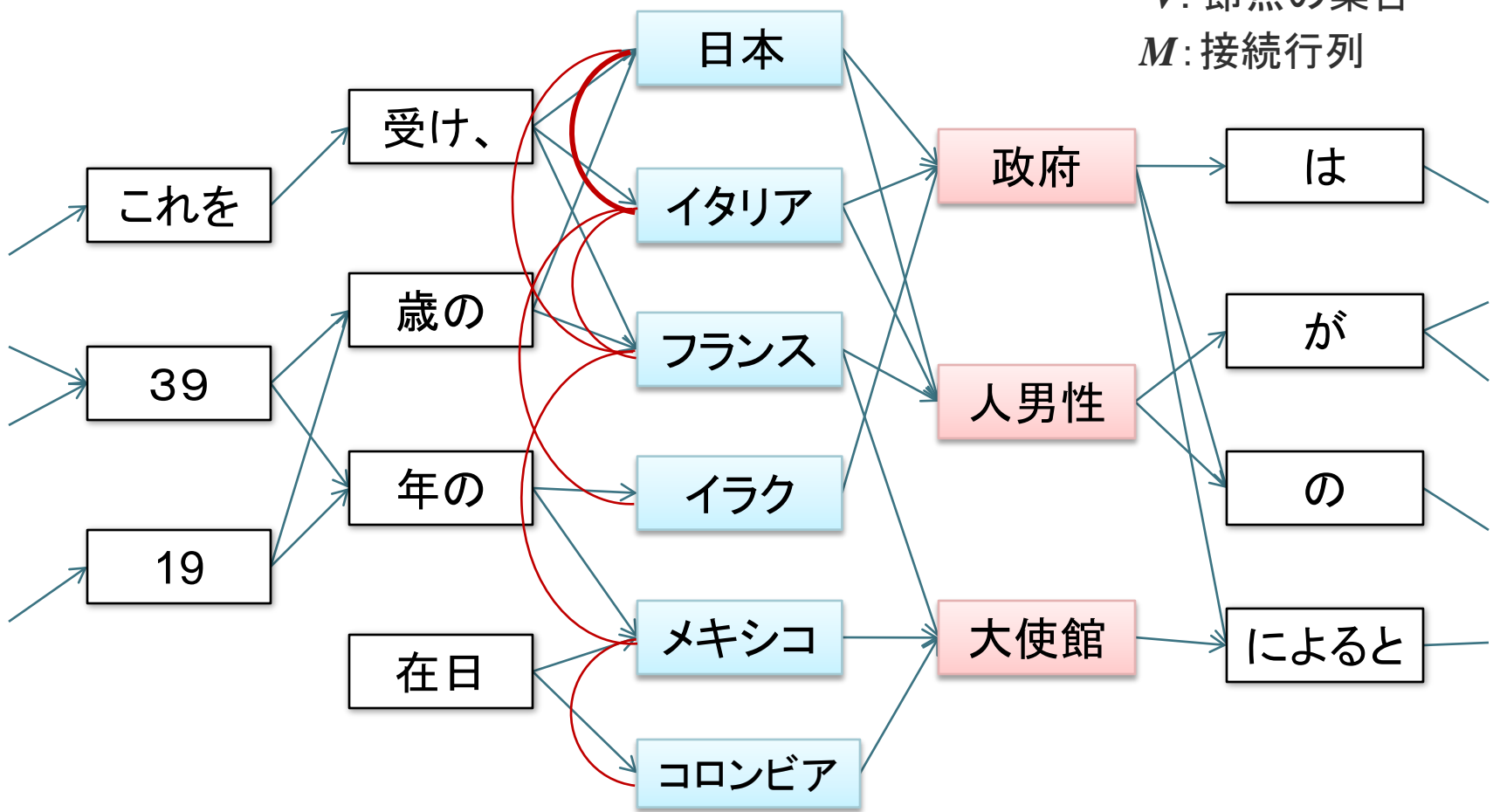
- ▶ インスタンス/パターンを区別しない
  - ▶ *Espresso*, *g-Espresso*とは異なる
  - ▶ 文書間の引用グラフと同じ構造
  - ▶ 種々の引用解析手法が適用可能



# g-Monakaアルゴリズム

▶ 右側文脈による類似度  $\Leftrightarrow$  書誌結合  $A_R = \frac{1}{|V|^2} M M^T$

$V$ : 節点の集合  
 $M$ : 接続行列



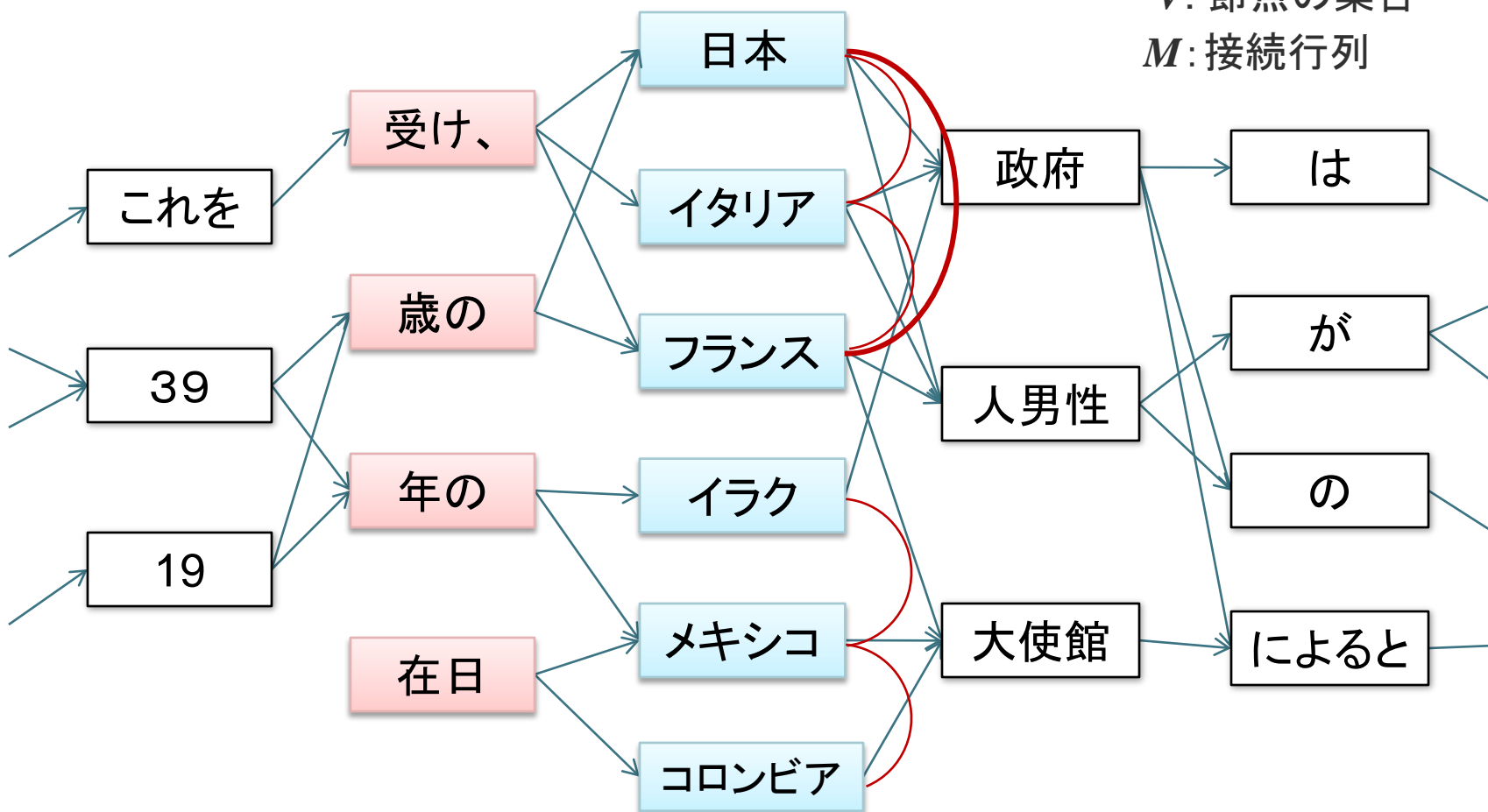


# g-Monakaアルゴリズム

▶ 左側文脈による類似度  $\Leftrightarrow$  共引用

$$A_L = \frac{1}{|V|^2} M^T M$$

$V$ : 節点の集合  
 $M$ : 接続行列

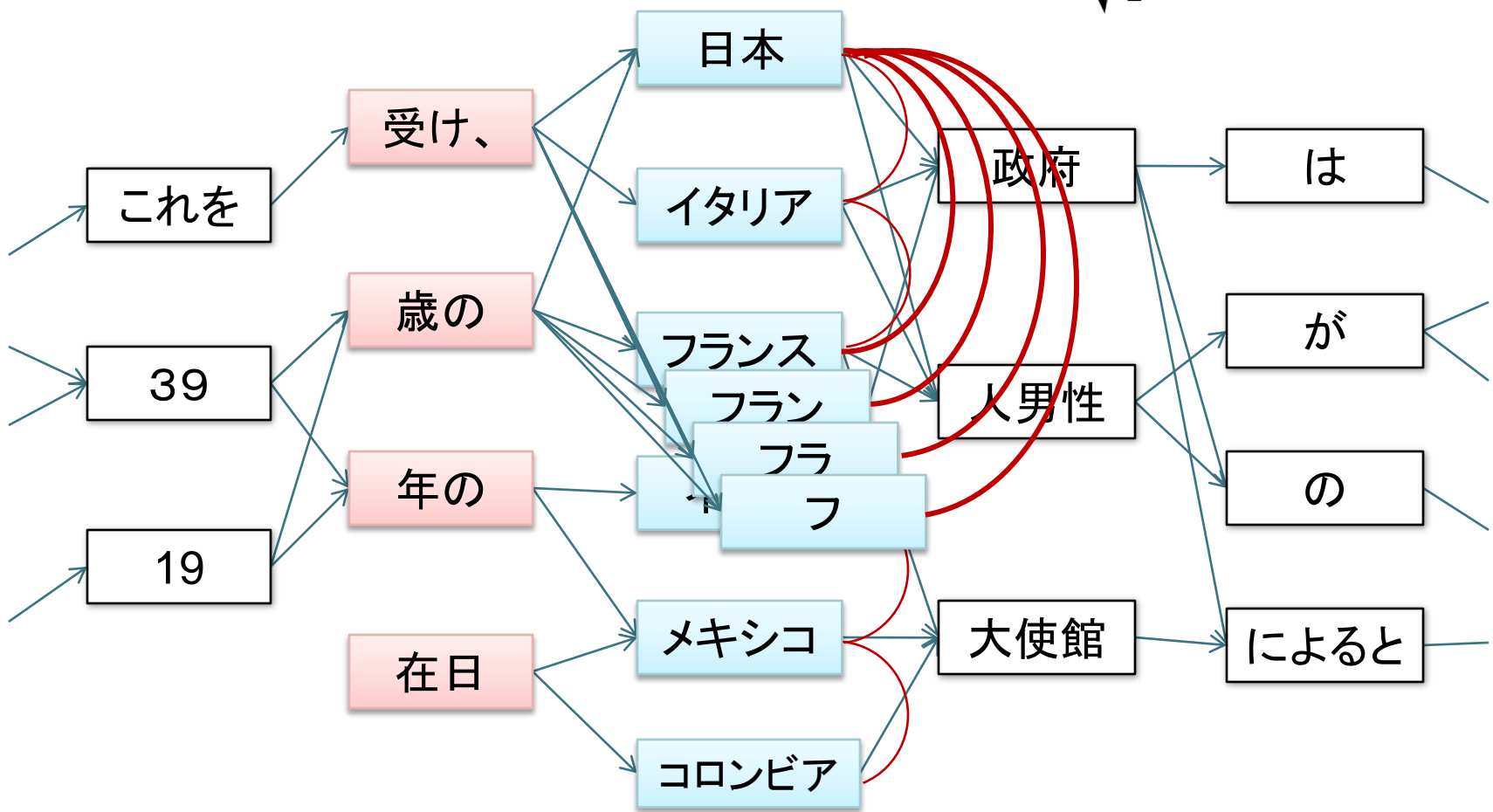




# g-Monakaアルゴリズム

▶ 右側文脈と左側文脈の両者を考慮

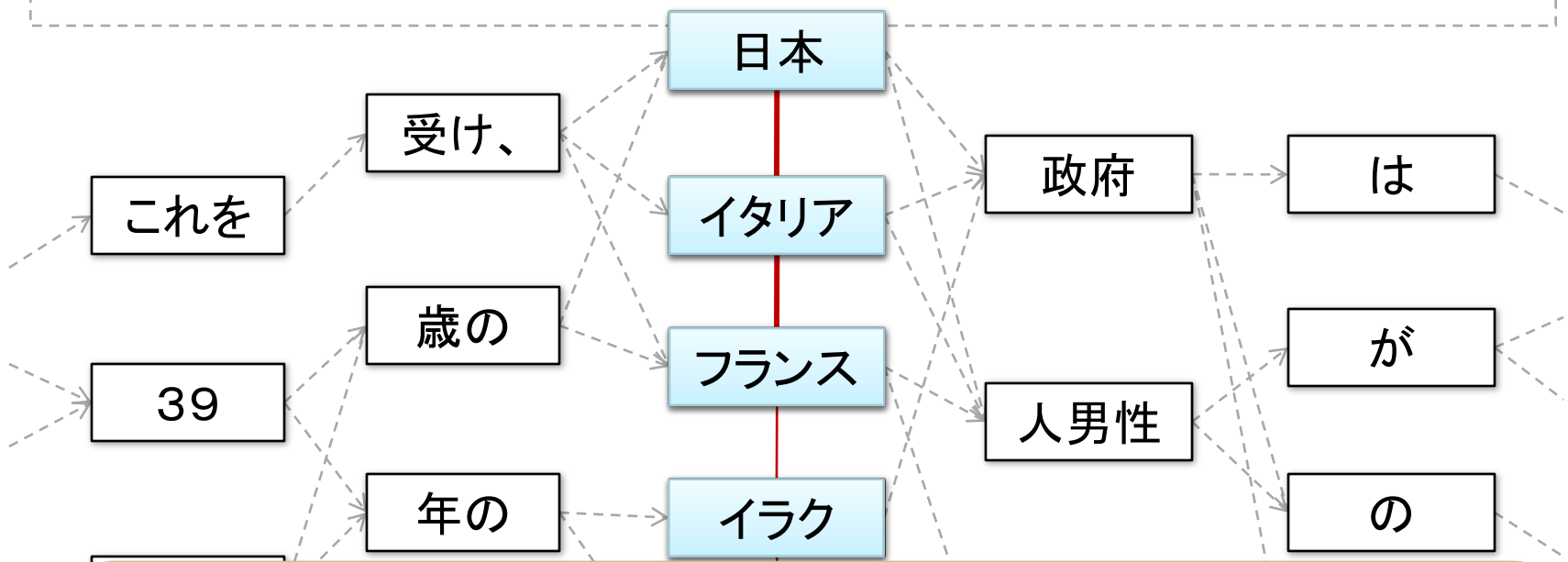
▶ 両類似度を一般化平均により結合  $A(i,j) = \sqrt[m]{\frac{1}{2}(A_R(i,j)^m + A_L(i,j)^m)}$





# g-Monakaアルゴリズム

- ▶ 類似度の無向重み付きグラフ  $G_A$  の構築
  - ▶ ブートストラップ:  $G_A$  上の伝播モデル



## グラフカーネルの適用

ノイマンカーネル:  $K_\beta(A) = A + \beta^2 A^2 + \beta^4 A^4 + \dots = A \sum_{n=0}^{\infty} \beta^n A^n$

正則化ラプラシアンカーネル:  $R_\beta(A) = A \sum_{n=0}^{\infty} \beta^n (-L)^n, \quad L = D - A$



# 性能比較実験 – 意味カテゴリ抽出タスク

---

- ▶ 正解セット
  - ▶ 世界の国・地域名 e.g. アルゼンチン, エクアドル, キルギス, ...
  - ▶ 日本の現行法令名 e.g. 独占禁止法, 刑法, 教育基本法, ...
  - ▶ 国内外の自動車メーカー名 e.g. トヨタ, 日産, クライスラー, ...
- ▶ シード
  - ▶ 3~5個無作為に選択
- ▶ コーパス
  - ▶ 2007年版毎日新聞コーパス 1面, 2面, 3面, 国際, 経済
  - ▶ 文字  $n$  グラム ( $1 \leq n \leq 8$ )
    - ▶ 出現頻度20未満と30,000以上は削除 異なり数 306,919個
- ▶ 比較手法
  - ▶ (1) 分布類似度 ( $m = 1.0, 0.1$ )
  - ▶ (2) *Filtered Espresso*
  - ▶ (3) *Monaka*
  - ▶ (4) *g-Monaka* (提案手法)



# 結果：獲得インスタンスの比較

手法	獲得されたインスタンス(上位50個)
分布類似度 ( $m = 1.0$ )	韓国, 韓, 日本, 中国, 米国, 米, 、韓国, ロシア, 北朝鮮, イラン, バングラデシュ, イラク, 英国, 韓国の, インド, パキスタン, ドイツ, 英, 北, アルゼンチン, 同国, 政府, フランス, トルコ, 、日本, イスラエル, イラ, 欧州, 北朝, 韓国政府, ミャンマー, 東, 今, 地, 国内, エクアドル, 大統領, 民主党, 世界, フィリピン, 台湾, 東京, 自, 欧, ロシ, ペルー, アフガン, 日本の, 、中国, 外
Monaka	韓国, 日本, 中国, 米国, 米, ロシア, イラン, 北朝鮮, バングラデシュ, イラク, 韓, 英国, パキスタン, インド, ドイツ, アルゼンチン, 英, 政府, 同国, トルコ, フランス, 、韓国, イスラエル, 韓国の, 欧州, ミャンマー, 国内, 大統領, 台湾, フィリピン, 地, エクアドル, 民主党, 韓国政府, アフガン, ペルー, 、日本, 世界, 外, アフリカ, 北京, 地元, キルギス, タイ, コ, 首相, 各国, 地域, 地方, 今回
g-Monaka	韓国, 中国, 日本, 米国, ロシア, 米, 北朝鮮, イラン, 英国, 韓, イラク, パキスタン, ドイツ, インド, 英, イスラエル, 欧州, アフガン, ミャンマー, フランス, 韓国政府, 台湾, トルコ, 同国, 韓国の, ペルー, フィリピン, 欧米, カナダ, 政府, 、韓国, 香港, 朝鮮, 外国, 両国, 民主党, スーダン, 欧, オーストラリア, 地元, 中国、韓国, マカオ, アフリカ, ブッシュ米, 各国, パレスチナ自治, インドネシア, 国内, ベトナム, 北京



# 結果：獲得インスタンスの比較

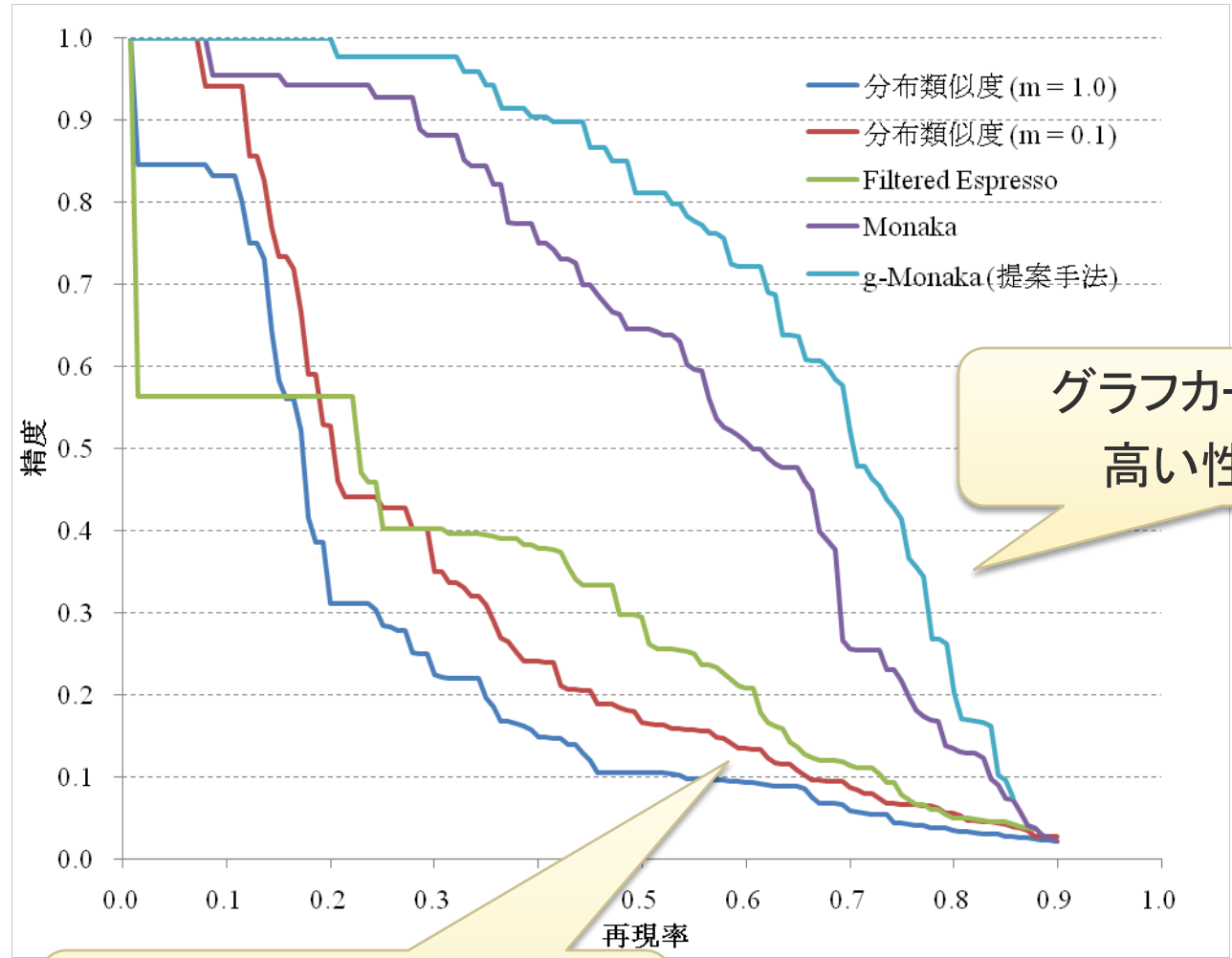
手法	獲得されたインスタンス(上位)
分布類似度 ( $m = 1.0$ )	韓国, 韓, 日本, 中国, 米国, デシュ, イラク, 英国, <u>韓国</u> の, チン, 同国, 政府, フランス, トルコ, <u>日本</u> , イスラエル, <u>イラ</u> , 欧州, <u>北朝</u> , 韓国政府, ミャンマー, 東, 今, <u>地</u> , 国内, エクアドル, 大統領, 民主党, 世界, フィリピン, 台湾, 東京, <u>自</u> , 欧, <u>ロシ</u> , ペルー, アフガン, <u>日本の</u> , <u>中国</u> , <u>外</u>
Monaka	韓国, 日本, 中国, 米国, 韓, 英国, パキスタン, イ, フランス, <u>韓国</u> , イスラエル, <u>地</u> , 台湾, フィリピン, <u>地</u> , エクアドル, 民主党, 韓国政府, アフガン, ペルー, <u>日本</u> , <u>世界</u> , <u>外</u> , アフリカ, 北京, <u>地元</u> , キルギス, タイ, <u>コ</u> , 首相, <u>各国</u> , <u>地域</u> , <u>地方</u> , 今回
g-Monaka	韓国, 中国, 日本, 米国, ロシア, 米, 北朝鮮, イラン, 英国, 韓, イラク, パキスタン, ドイツ, インド, 英, イスラエル, 欧州, アフガン, ミャンマー, フランス, 韓国政府, 台湾, トルコ, 同国, <u>韓国</u> の, ペルー, フィリピン, 欧米, カナダ, 政府, <u>韓国</u> , 香港, 朝鮮, 外国, <u>両国</u> , 民主党, スーダン, 欧, オーストラリア, 地元, 中国、韓国, マカオ, アフリカ, ブッシュ米, <u>各国</u> , パレスチナ自治, インドネシア, <u>国内</u> , ベトナム, 北京

分がち書きの正しくない  
インスタンスが抽出

ジェネリック・インスタンスが混入



# 結果: 精度-再現率グラフによる性能比較

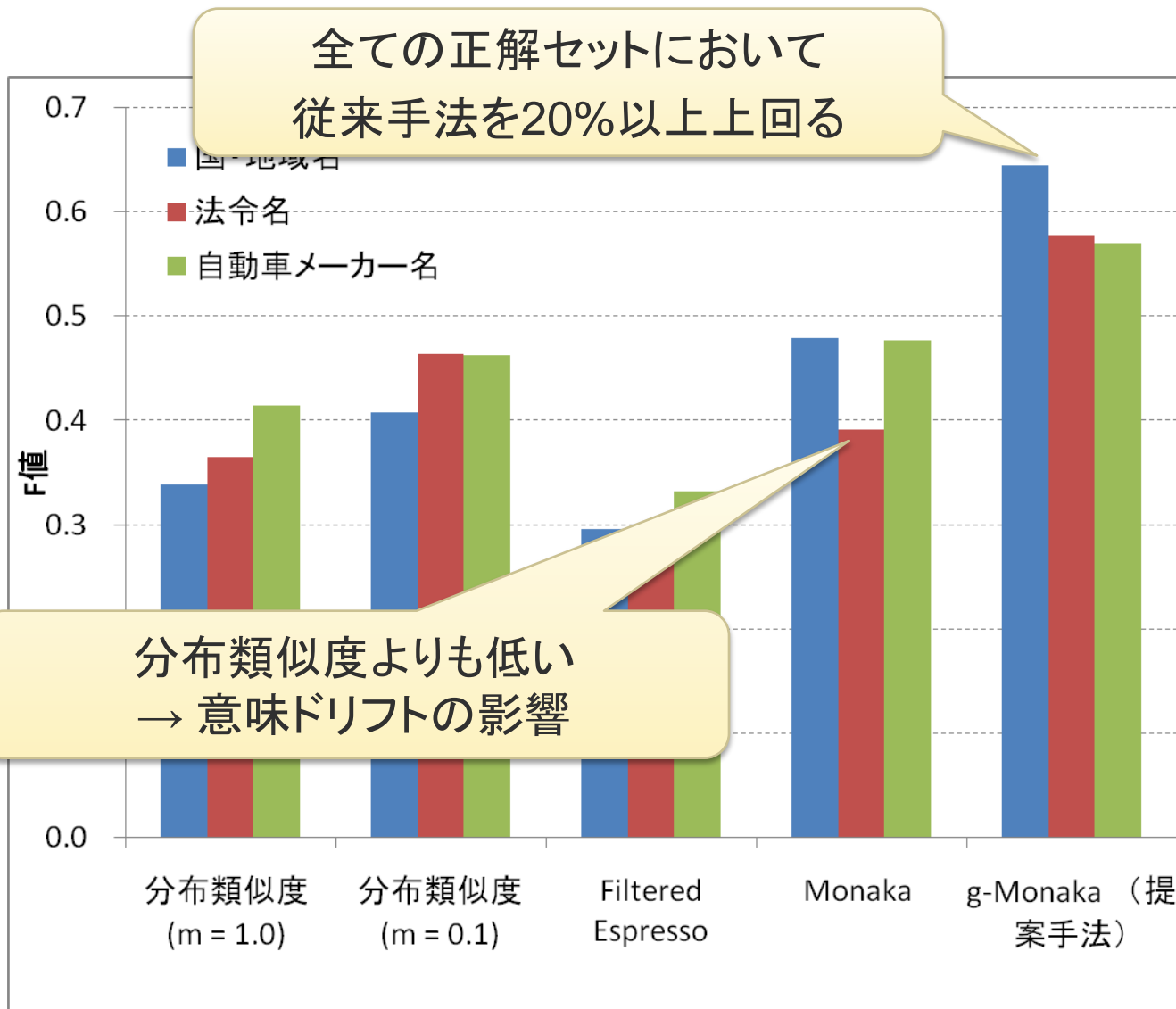


グラフカーネルにより  
高い性能を維持

挟み込み制約の無い場合  
性能の低下が顕著



# 結果:F値(最大値)による性能比較





# まとめ

---

- ▶ 非分かち書き文からの意味カテゴリ抽出アルゴリズム *g-Monaka*を提案
  - ▶ 文字種・形態素情報を使用しない
  - ▶ 文字 $n$ グラムの隣接グラフ上における引用解析としてブートストラップを定式化
  - ▶ グラフカーネルを用い, 従来手法よりも高い精度を実現
- ▶ 今後の課題
  - ▶ 左右の文字列を用いた抽出手法との比較
    - ▶ KnowItAll [Etzioni et al. 04], SEAL [Wang and Cohen 07]
  - ▶ 他の非分かち書き言語への適用
    - ▶ 中国語, タイ語