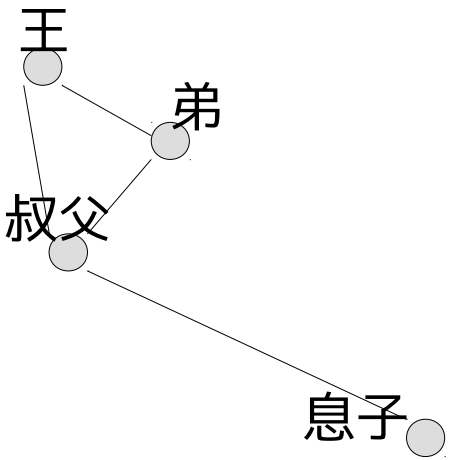
# ハンガリアン法による語の相似関係抽出
# Analogous word association and its detection by the Hungarian method.

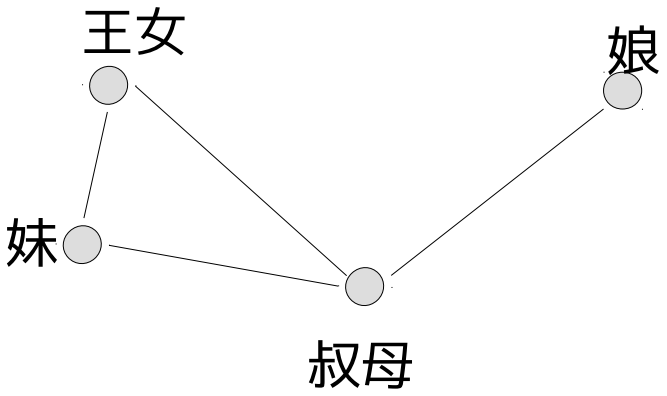
北海道大学 工学部
情報エレクトロニクス学科
コンピュータサイエンスコース
知識ベース研究室 B4平田泰樹

# 目的

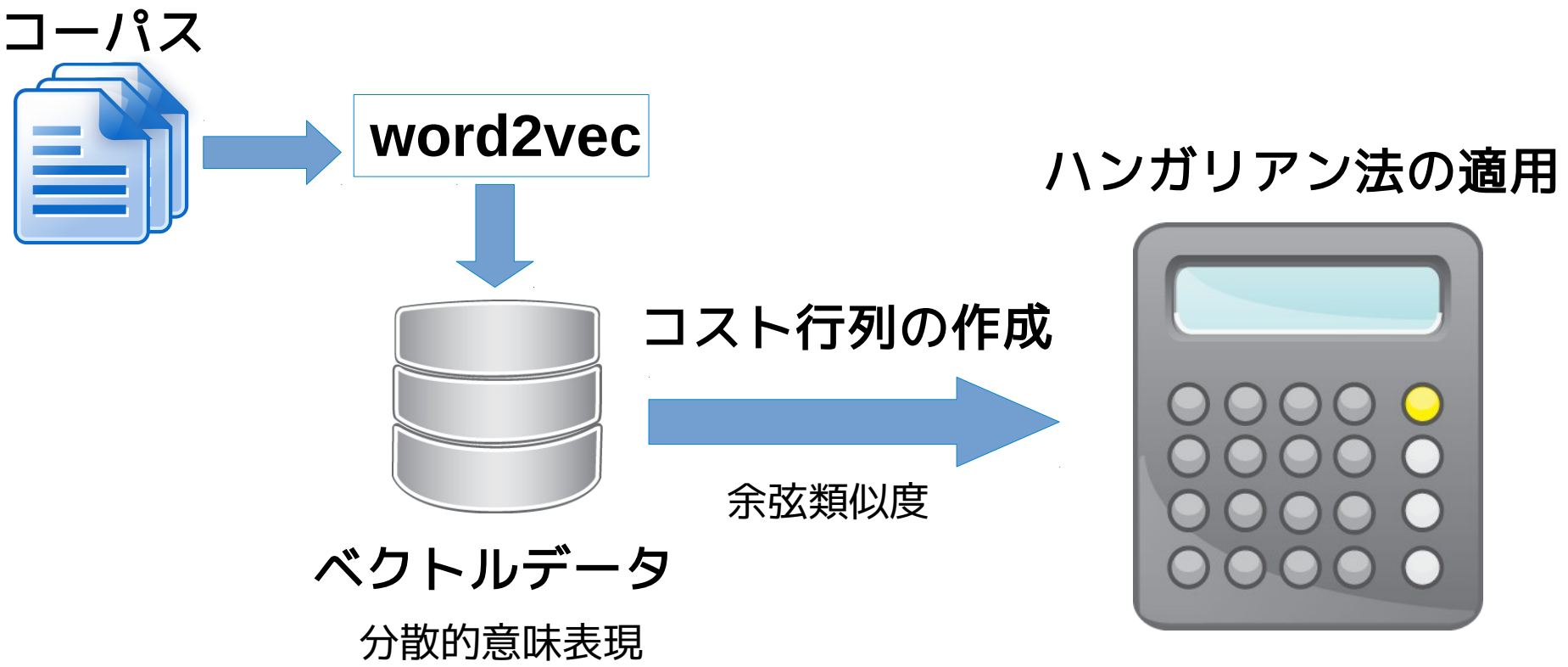
- 単語の分散的意味表現ベクトルから、可視化することなしに、単語の相似関係を見つける。
- 入力：一組の単語対
- 出力：入力された単語対の近傍単語 n 個で作った二つのグループの n 個の単語同士のマッチング

例）



# 実験システム

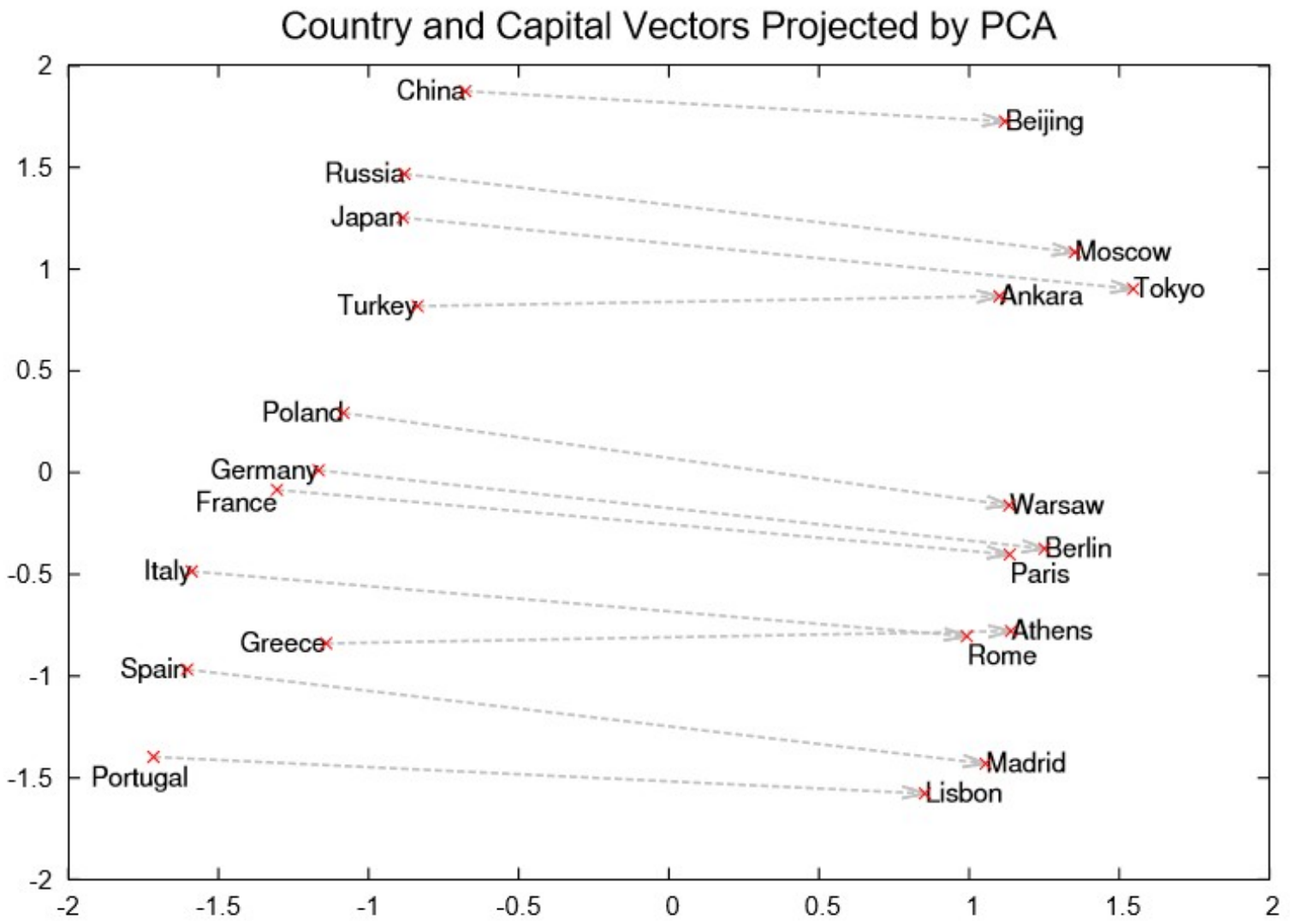


# word2vec とは

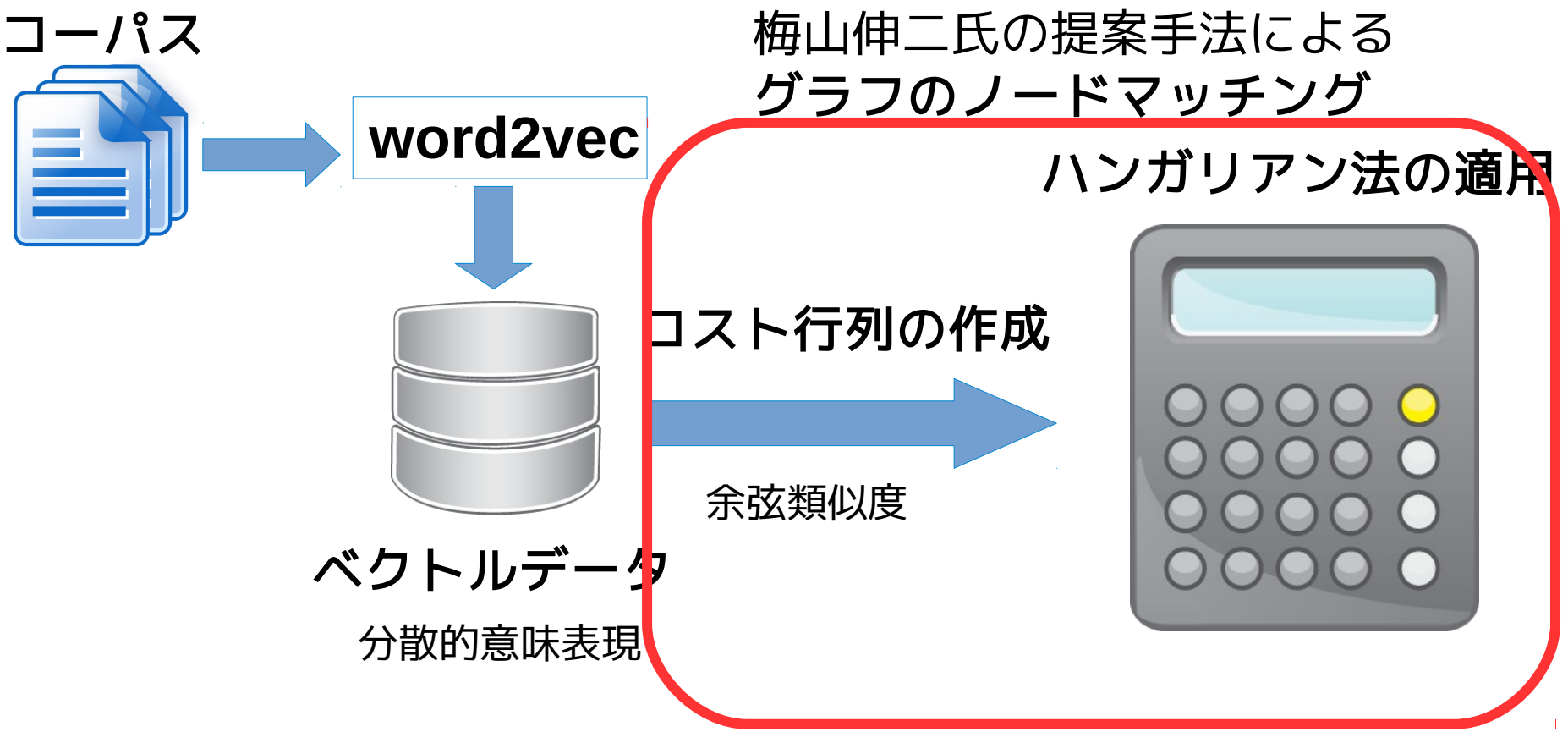
- **2013** 年に **Google** が公開した分散的意味表現の学習ツール。
- 形態素毎に分かち書きされたテキストデータを入力とし、テキストデータに含まれるすべての形態素の意味表現ベクトルを出力する。
- 前後の形態素を用いた出現予測を通じて意味表現ベクトルを学習する。

**“king” - “man” + “woman”**
**= “queen”**

# word2vec 出力の可視化



# 実験システム


コーパス
word2vec
ベクトルデータ
分散的意味表現
梅山伸二氏の提案手法による
グラフのノードマッチング
コスト行列の作成
余弦類似度
ハンガリアン法の適用

# Umeyama 1988.

$$\min_{\varphi} \sum_{ij} \left( w_L\left(v_i, v_j\right) - w_R\left(v_{\varphi(i)}, v_{\varphi(j)}\right) \right)^2$$

- 重み付きグラフの同型性を調べるために、ノード同士の距離の近さ遠さを保持したベクトル表現にし、上記の最適化問題を内積総和最大化問題に帰着しハンガリアン法で解いた。
- 帰着の過程で二つのグラフが同型もしくはほぼ同型であることを仮定している。

# ハンガリアン法

- コスト行列を元に、最適な組み合わせを見つける。（割当問題）**(a~c: 作業員、 A~C: 仕事 )**

| | A | B | C |
|---|---|---|---|
| a | 250 | 180 | 190 |
| b | 230 | 190 | 200 |
| c | 240 | 170 | 210 |

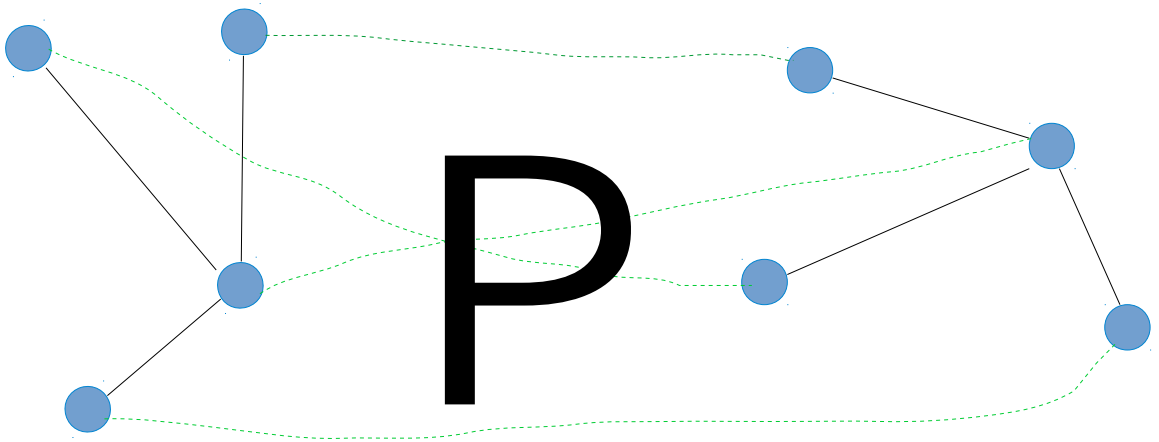
# ハンガリアン法

- 各行、各列の最小値をそれぞれの行、列の全要素から引き、０となった要素を通る最小本数の線を引く。

| | A | B | C |
|---|---|---|---|
| a | 30 | 0 | 0 |
| b | 0 | 0 | 0 |
| c | 30 | 0 | 30 |

# 抽出可能なグラフ構造

- マッチングが取れるのはほぼ同型なグラフ。（マッチング問題を内積総和最大化問題に帰着する過程でグラフがほぼ同型であることを仮定しているため）
- 方向と距離の関係を内積総和の最大化で対応づけているため、回転拡大操作に対して頑健。



- 本研究では、 **word2vec** の出力ベクトル集合の、方向性の対応を最大限保持した一対一頂点対応を、余弦類似度最大化により求めた。
- **word2vec** の出力ベクトルにおいて、ベクトルの長さが持つ意味合いが不明瞭であるため、本研究では内積総和の最大化ではなく、余弦類似度の最大化を解いた。

# 実験データ

- **日本語版 Wikipedia**
  - MeCab で形態素解析
- **形態素解析したデータを word2vec で学習し、ベクトルを出力する（学習時のハイパーパラメータは下記の通り）**
  - 適用手法：**CBoW**（対象単語の周辺語彙から対象単語の出現を予測）
  - 学習するベクトルの次元数：**200**
  - 文脈窓：**8**
  - 負例サンプリング数：**25**
  - 階層型ソフトマックス：利用しない
  - 学習の反復回数：**15**

# 予備実験 1

- 梅山氏の提案手法を適用することの妥当性検証のため
- 対応関係の抽出を想定する単語群を用意してノードマッチングを適用

| 集合 X | 集合 Y |
|---|---|
| 叔父 | 妹 |
| 王 | 祖母 |
| 老人 | 王女 |
| 父 | 雌 |
| 兄 | 老婆 |
| 祖父 | 花子 |
| 弟 | 姉 |
| 息子 | 叔母 |
| 雄 | 娘 |
| 太郎 | 母 |

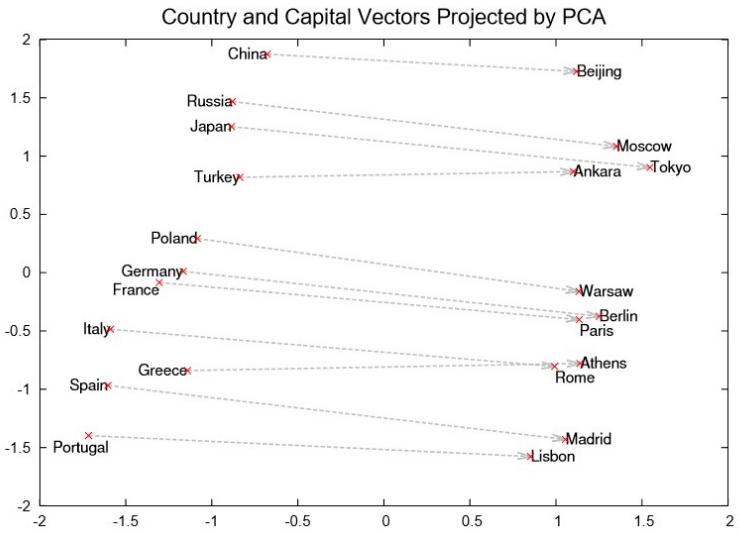
適用

| 集合 X | 集合 Y |
|---|---|
| 叔父 | 叔母 |
| 王 | 王女 |
| 老人 | 老婆 |
| 父 | 母 |
| 兄 | 姉 |
| 祖父 | 祖母 |
| 弟 | 妹 |
| 息子 | 娘 |
| 雄 | 雌 |
| 太郎 | 花子 |

# 予備実験２

- **word2vec の学習精度検証のため**
- **国名 - 都市名の対応例でのマッチングを確認**

**（Distributed Representations of Words and Phrases and their Compositionality.T.Mikolov,etc…,2013.）**

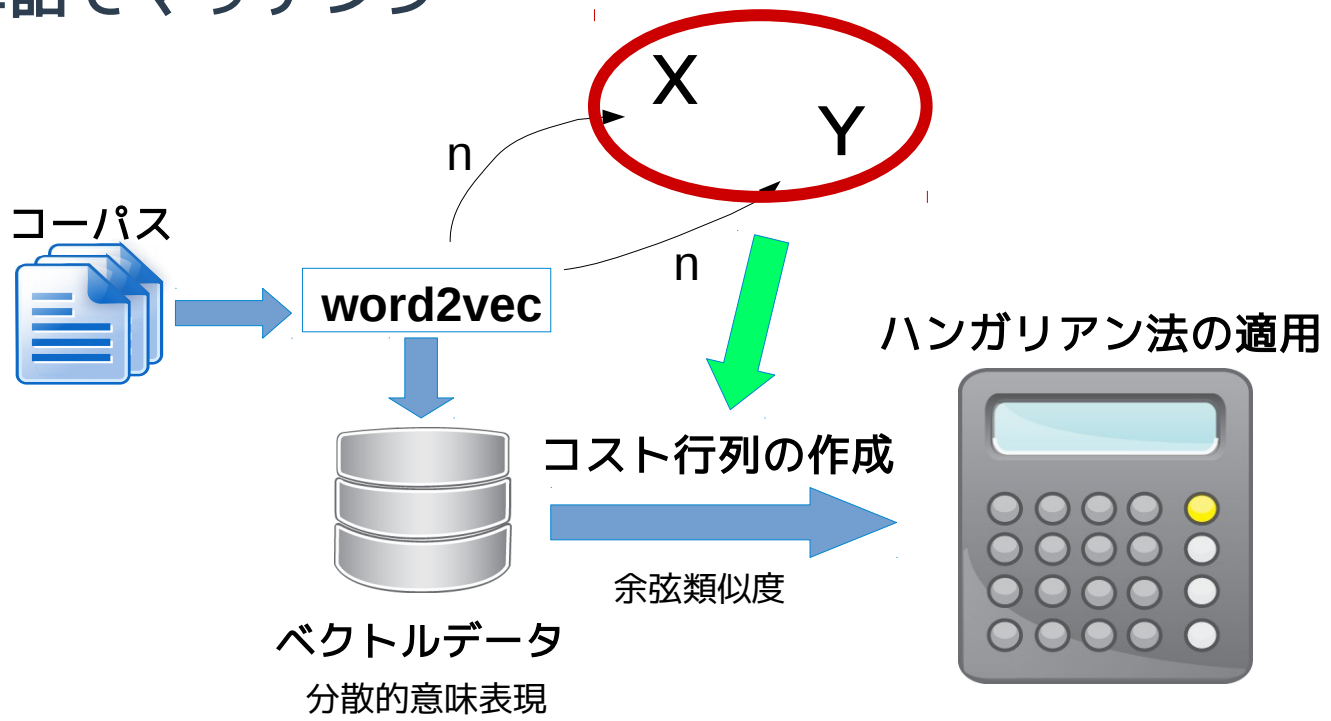



比較

| 集合 X | 集合 Y |
|---|---|
| 中国 | 北京 |
| ポーランド | ワルシャワ |
| ギリシャ | アテネ |
| ポルトガル | リスボン |
| ドイツ | ベルリン |
| スペイン | マドリード |
| フランス | パリ |
| ロシア | モスクワ |
| トルコ | アンカラ |
| 日本 | 東京 |
| イタリア | ローマ |

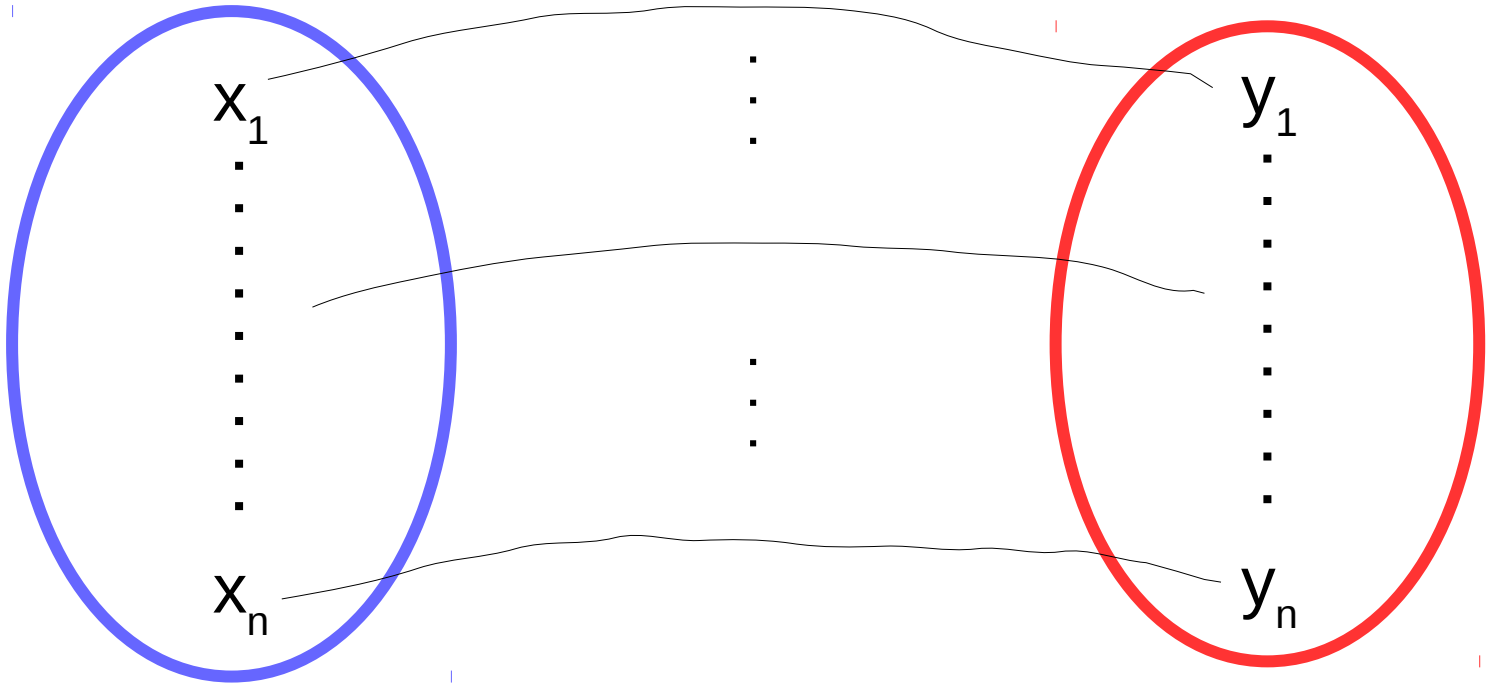
# 実験

- 入力：一組の単語対
- 出力：単語対の近傍単語 n 個ずつのグループで単語対と同じ関係性にある単語でマッチング



# 実験 1.1

- 入力として与えた単語対 ($w_x$,$w_y$) の近傍 n 個の単語で作ったグループ内の単語同士でのマッチング



# 結果 1.1

- 与えた単語対の関係を作れる単語が各グループに揃わない
- 得られる単語に偏りがある

王,女王

| 集合 X | 集合 Y |
|---|---|
| は王 | 王 |
| 皇帝 | 国王 |
| 女王 | ヴィクトリア |
| 国王 | エリザベス女王 |
| に王 | 王妃 |
| 王妃 | 王配 |
| 君主 | 王子 |
| 大王 | 王女 |
| 伯 | エリザベス 1 世 |
| 聖王 | エリザベス 2 世 |

北海道,札幌

| 集合 X | 集合 Y |
|---|---|
| 道内 | 北海道 |
| 釧路 | 釧路 |
| 札幌 | 函館 |
| 十勝 | 帯広 |
| 札幌市 | 岩見沢 |
| 道東 | 旭川 |
| 十勝支庁 | 札幌市 |
| 東北地方 | 仙台 |
| 道南 | 小樽 |
| 道北 | 新潟 |

マッチング率
２組/10組
およそ20%

# 実験 1.2

- 入力：一組の単語対
- 出力：単語対の近傍単語 20 個ずつのグループでマッチング

## （北海道，札幌）

| 集合 X | 集合 Y |
|---|---|
| 道内 | 北海道 |
| 釧路 | 釧路 |
| 札幌市 | 札幌市 |
| 札幌 | 函館 |
| 十勝 | 帯広 |
| 青森県 | 青森 |
| 根室 | 室蘭 |
| 道東 | 北見 |
| 九州 | 福岡 |
| 十勝支庁 | 旭川市 |

| 釧路市 | 岩見沢 |
|---|---|
| 東北地方 | 仙台 |
| 胆振 | 江別 |
| 函館 | 名古屋 |
| 道南 | 旭川 |
| 空知 | 丘珠 |
| 道北 | 苫小牧 |
| 道外 | 小樽 |
| 留萌 | 新潟 |
| 沖縄県 | 東京 |

## （東京，東京タワー）

| 集合 X | 集合 Y |
|---|---|
| 大阪 | 通天閣 |
| 新宿 | サンシャイン 60 |
| 名古屋 | お台場 |
| 日比谷 | スカイツリー |
| 関西 | 日本電波塔 |
| 札幌 | 六本木ヒルズ |
| 横浜 | ランドマークタワー |
| 赤坂 | 大阪タワー |
| 静岡 | FCG ビル |
| 神奈川 | ビル街 |

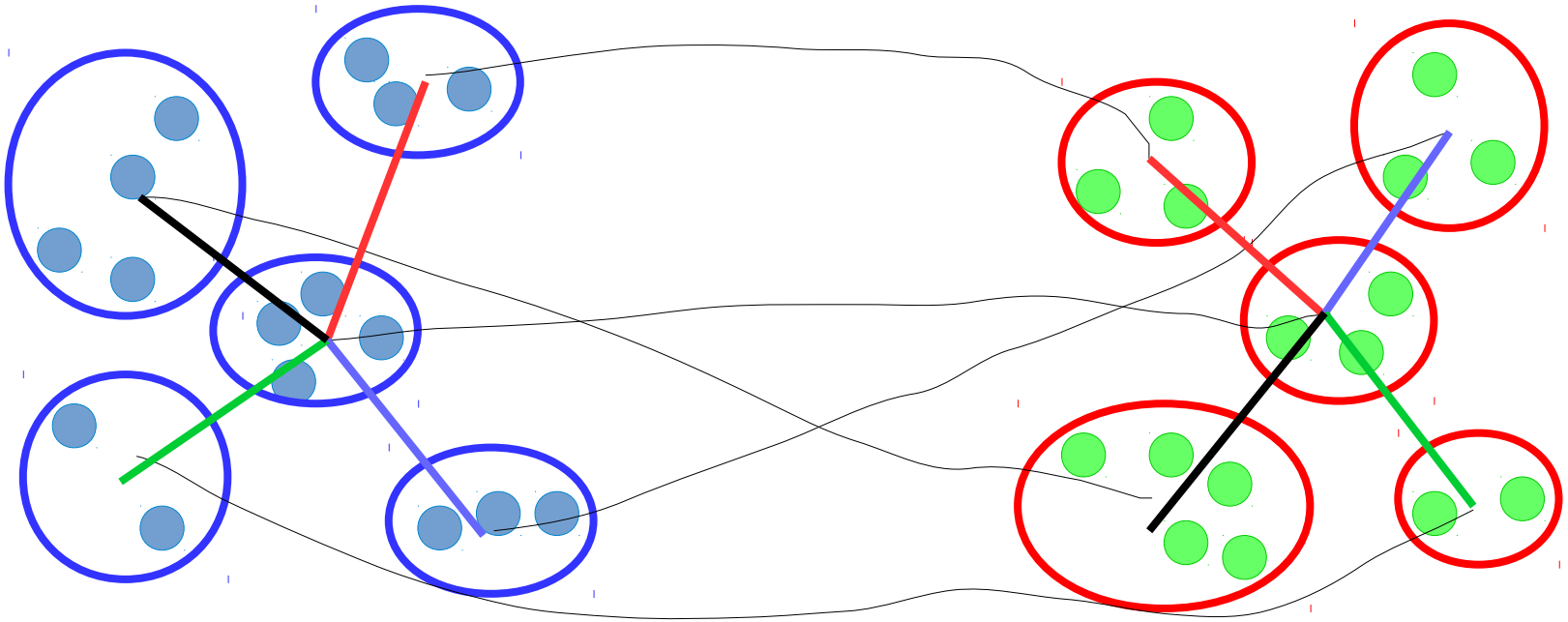
| 福岡 | 東京スカイツリー |
|---|---|
| 麹町 | 名古屋テレビ塔 |
| 神戸 | 屋上 |
| 都内 | オカンとボクと、時々、オトン |
| 埼玉 | 電波塔 |
| 京都 | 銀座和光 |
| 愛知 | テレビ塔 |
| 青山 | タワー |
| 渋谷 | 展望台 |
| 金沢 | 鉄塔 |

# 結果 1.2

- **精度の上昇は見られなかった。**
  - 与える単語対が近い位置関係にあり、それぞれの近傍として取得する単語が重複したことにより、同一単語でのマッチングが増えた。
  - 与えた単語対の関係を作れる単語が各グループに揃わなかった。
- **近傍単語の特徴にばらつきがある。**

# 実験２

- 取得する近傍単語数を増やし、それぞれを同数のクラスターに分割した。
- 分割したクラスター同士の相似関係抽出を行った。



# 実験２

| クラスター No.(北海道) | クラスター No.(札幌) |
|---|---|
| 0 | 0 |
| 1 | 2 |
| 2 | 1 |
| 3 | 4 |
| 4 | 3 |

(北海道:0, 札幌:0) = (後志, 倶知安・岩見沢市・余市),(十勝郡, 十勝)

(北海道:1, 札幌:2) =なし

(北海道:2, 札幌:1) =(十勝支庁, 帯広市),(道北, 稚内・名寄市),(空知支庁, 旭川空港・滝川市),...

(北海道:3, 札幌:4) =(道東, 釧路),(道南, 苫小牧),(札幌, 中島公園),...

(北海道:4, 札幌:3) =(東北地方, 仙台),(九州, 福岡),(北海道南部, 函館),...

# 実験２ 考察

- 実験１では単語単位で１対１対応を求めており、制約が厳しかった。
- 実験２では似通った単語を１つのクラスターとし、その対応を見た。
- ” いくつかの単語”対”いくつかの単語”で対応を取ったため、対応関係を取りやすかった。
- 人力での確認のため、評価尺度がやや不明瞭

# 実験結果まとめ

- ベクトルの加算減算で抽出できる対応関係についてはハンガリアン法の適用で取得可能であることが期待できる。
- 近くにある単語を単純に **n** 個取ったグループ同士の相似関係は抽出が難しい。
- 近傍単語をクラスタリングし、クラスタ同士での対応を見ると、ささやかながら相似関係が見える。
- 今回の実験は全体的に検証が不十分。

# 参考文献


[1] J.R.Firth. A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis, pp.1-32, 1957.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean Distributed Representations of Words and Phrases and their Compositionality. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf, NIPS 2013.

[3] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://taku910.github.io/mecab/, 2013.

[4] 日本語 WordNet http://nlpwww.nict.go.jp/wn-ja/

[5] ダヌシカ ボレガラ, 岡直観, 前原貴憲 機械学習プロフェッショナルシリーズ ウェブデータの機械学習. 株式会社 講談社, 2016.

[6] 宮崎修一 グラフ理論入門 基本とアルゴリズム 森北出版, 2015.

[7] 喜多陵. 記述の構造類似性に基づく法的観点と判例のマッチング. Master's thesis, 北海道大学大学院情報科学研究科, 2014.

[8] 梅山伸二 An eigendecomposition approach to weighted graph matching problems. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, pp.695-703, Sep, 1988.

ご静聴ありがとうございました。