情報抽出と機械学習

鳥取大学工学研究科
知能情報工学科
村田　真樹

2014 年12月14日

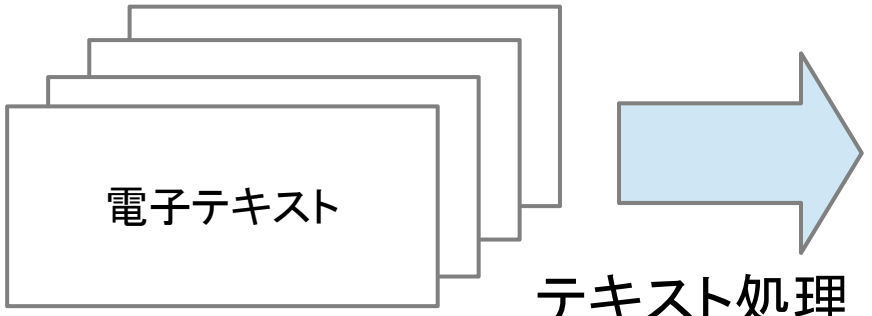
目次

1. 情報抽出（ネットワークの構築）
2. 機械学習を利用した文生成

1．ネットワークの構築

TF-IDFと不要ノード削除に基づく単語ネットワークの構築

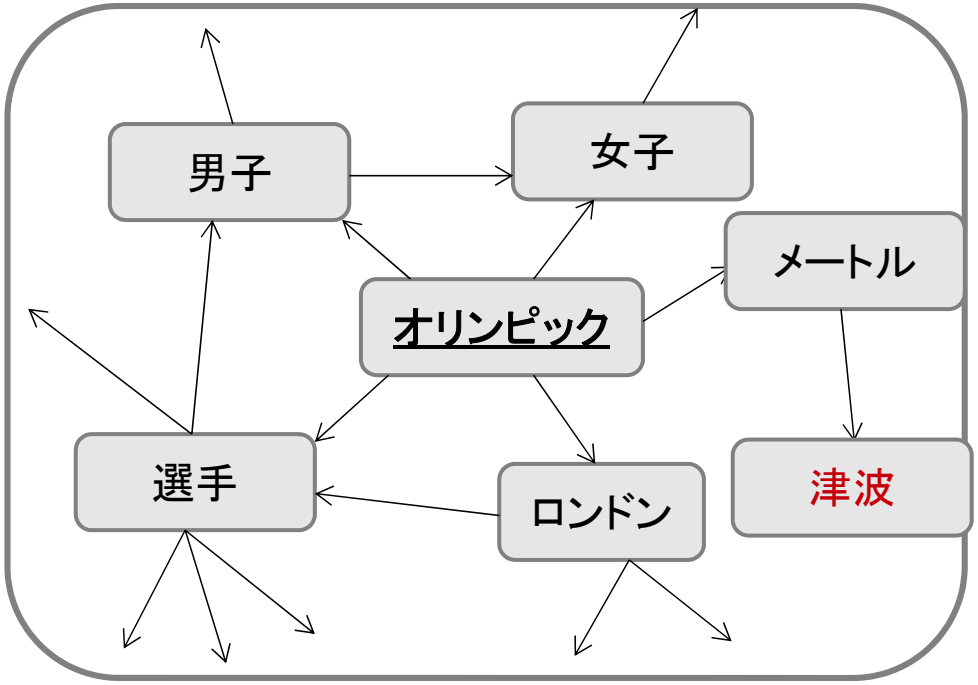
発想支援、分析に使える



電子テキスト群

テキスト処理

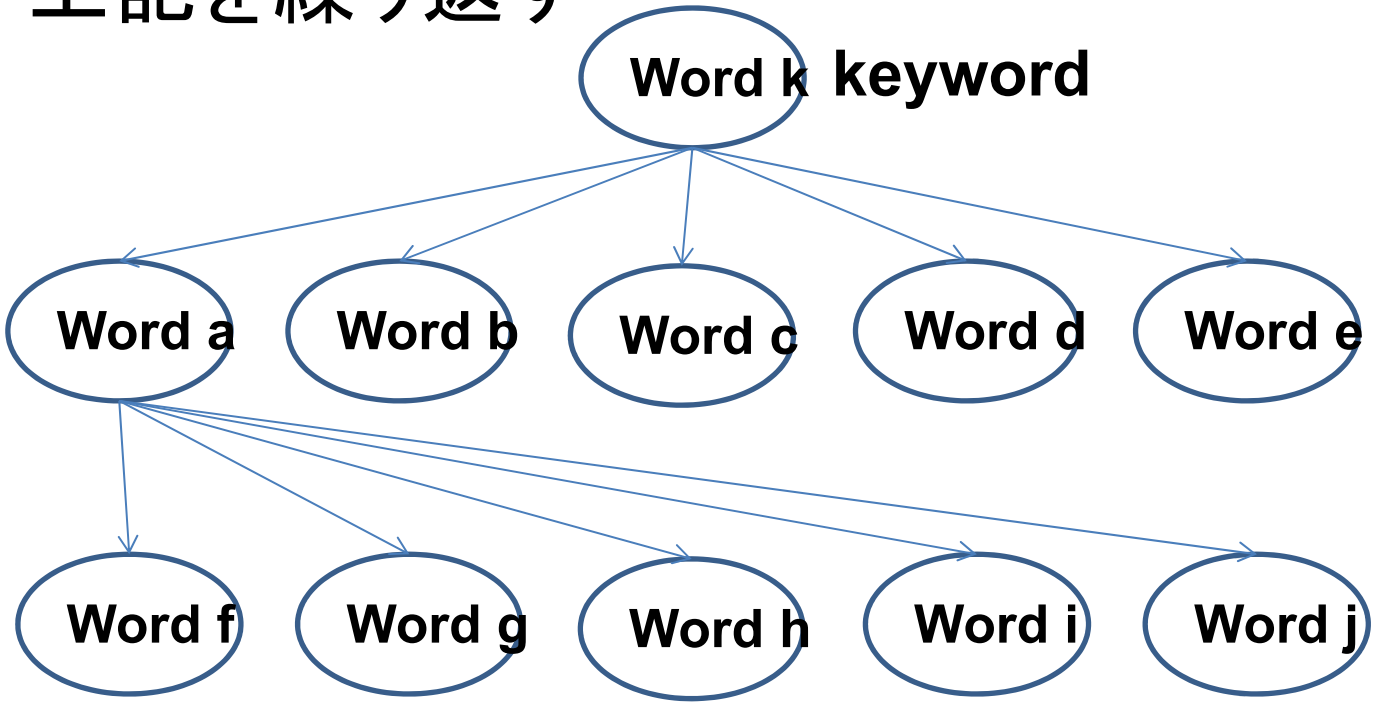
例）キーワード：オリンピック



問題点：関連のない単語群を含むネットワークへ発展

ネットワークの基本的な構築手法

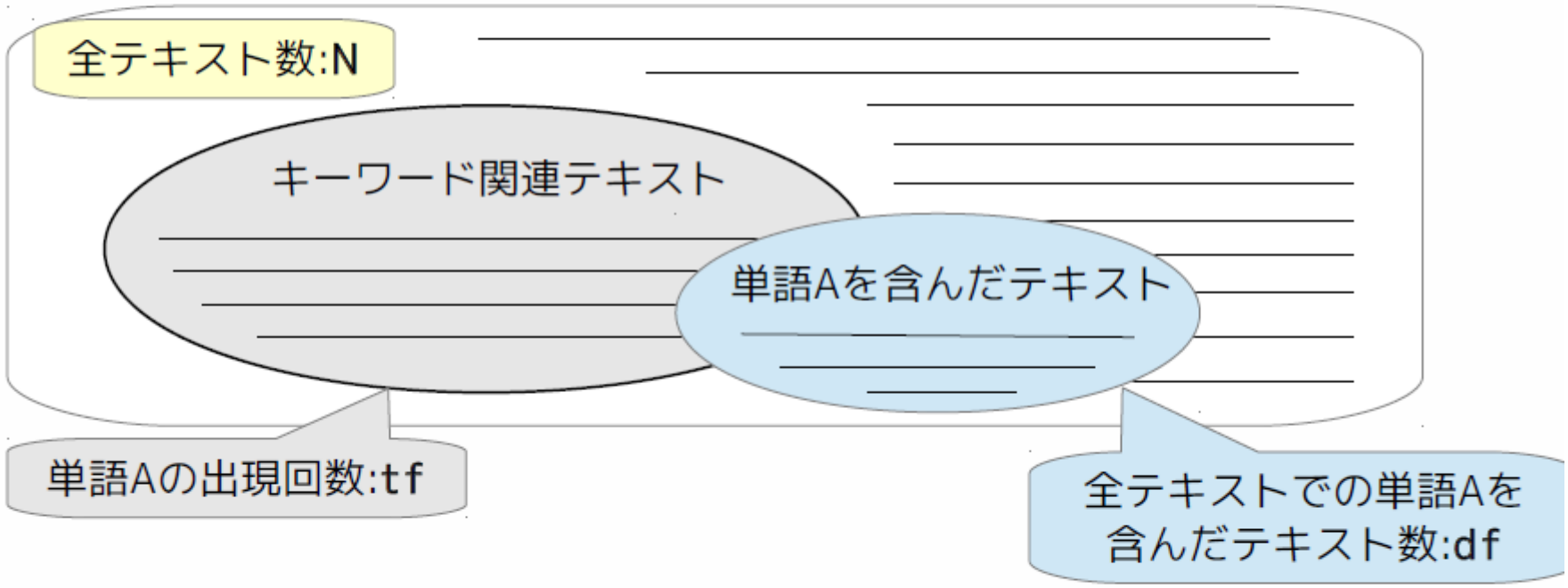
1. キーワードを含む文書で求めたtf-idfが大きい5つの単語を取り出す
2. 上記を繰り返す



単語の抽出：TF-IDF法

・TF-IDFを利用した単語の抽出： $w = tf * \log \frac{N}{df}$

単語A：ノード候補



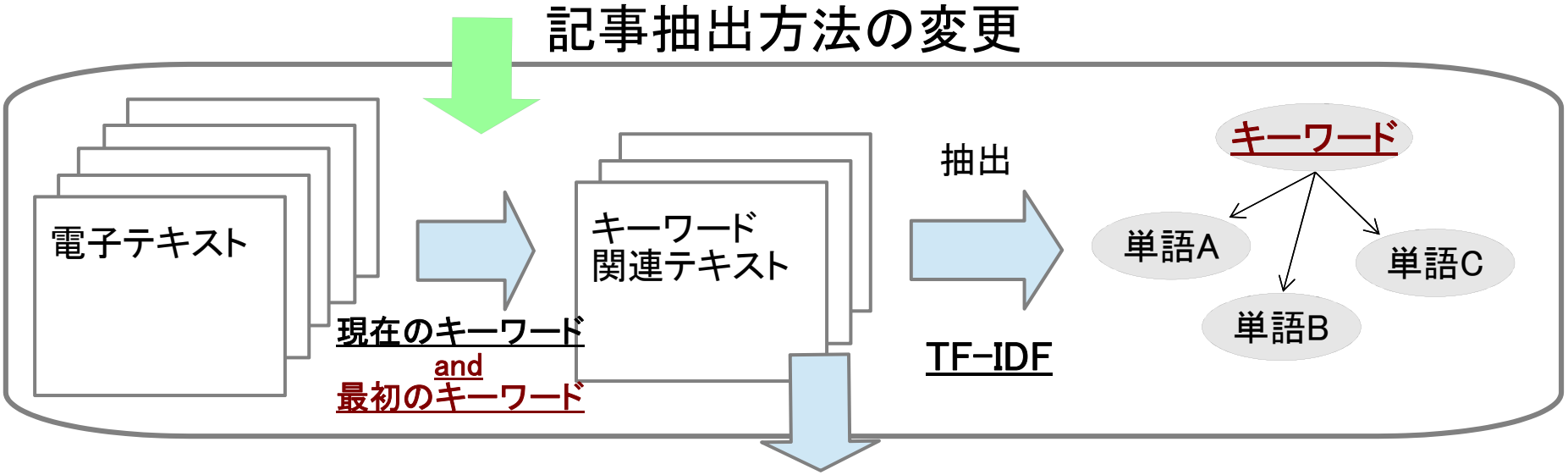
・df ：高い→頻度は多いが一般的な単語は除外

・TF-IDF上位5単語をキーワードからノード

不要ノードの削除

- テーマ限定抽出法
- テーマ関連抽出法

テーマ限定抽出法

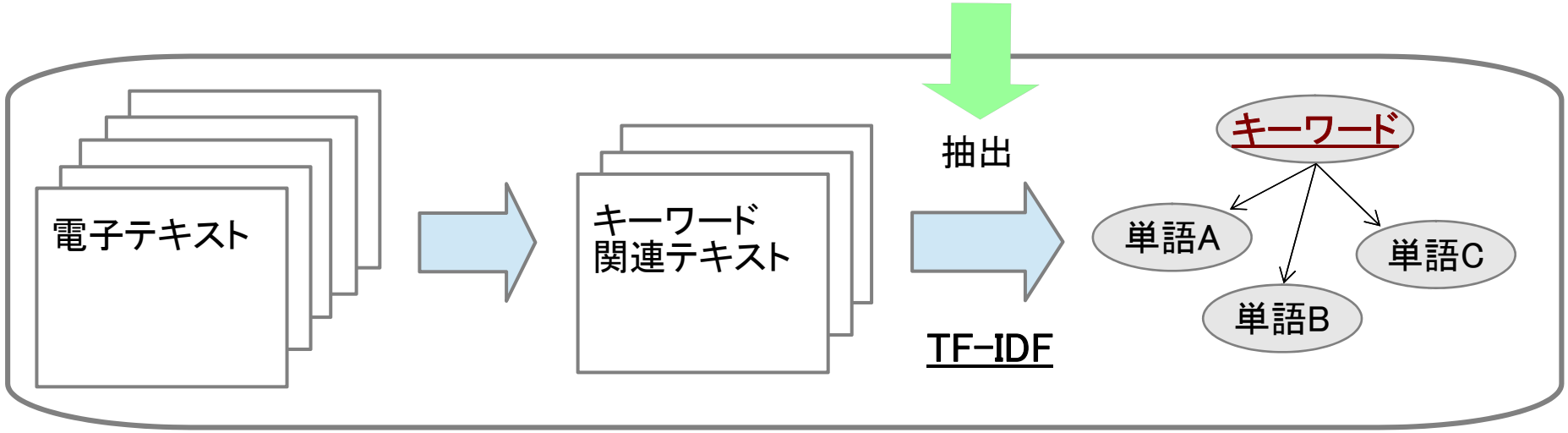


最初のキーワードに関連した単語が得られやすくなる

例（オリンピック）

テーマ関連抽出法

閾値による関連度の低い単語の削除



関連度：$\frac{df_{k,t}}{df_k} / \frac{df_t}{N}$

kがある時のtの出現率　　tの一般的出現率

$df_{k,t}$ k と t の共に出現した記事数

df_k k の出現した記事数

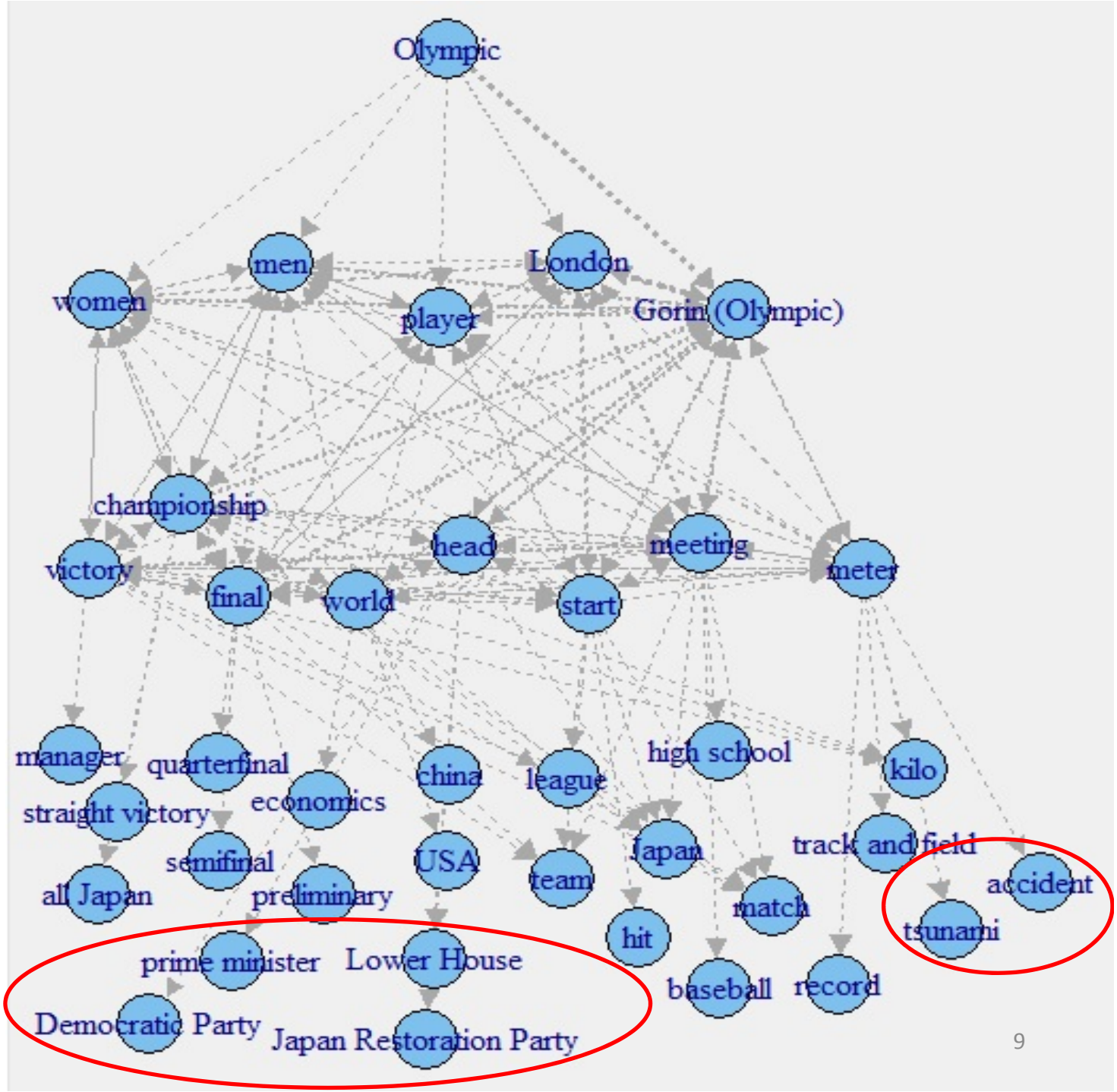
df_t : t の出現した記事数

N : 全記事数

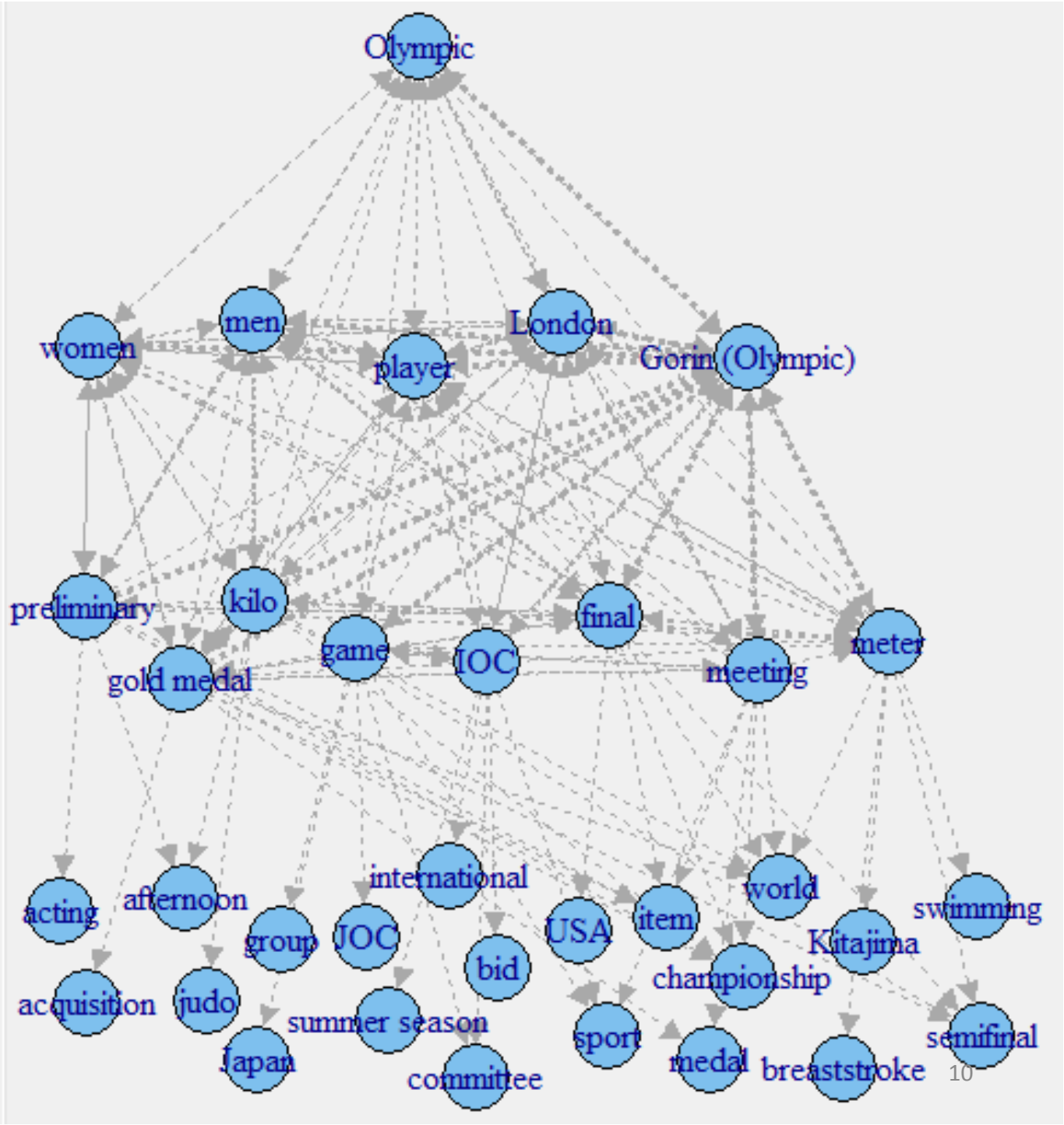
k：最初に設定したキーワード
t：ノード候補になっている単語

実験結果

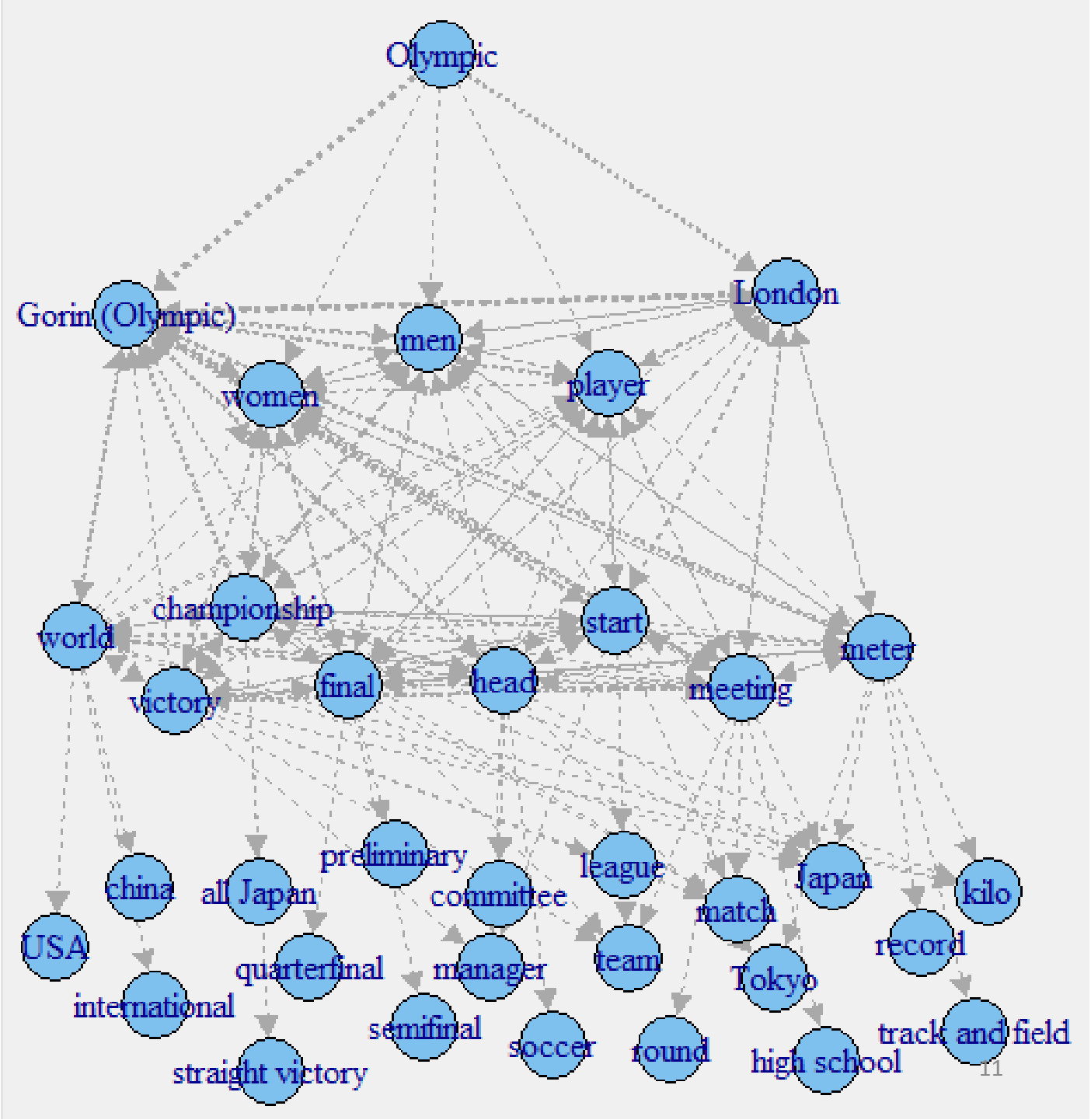
不要ノードの削除をしない場合(ベースライン)



テーマ限定抽出法を利用



テーマ関連抽出法を利用



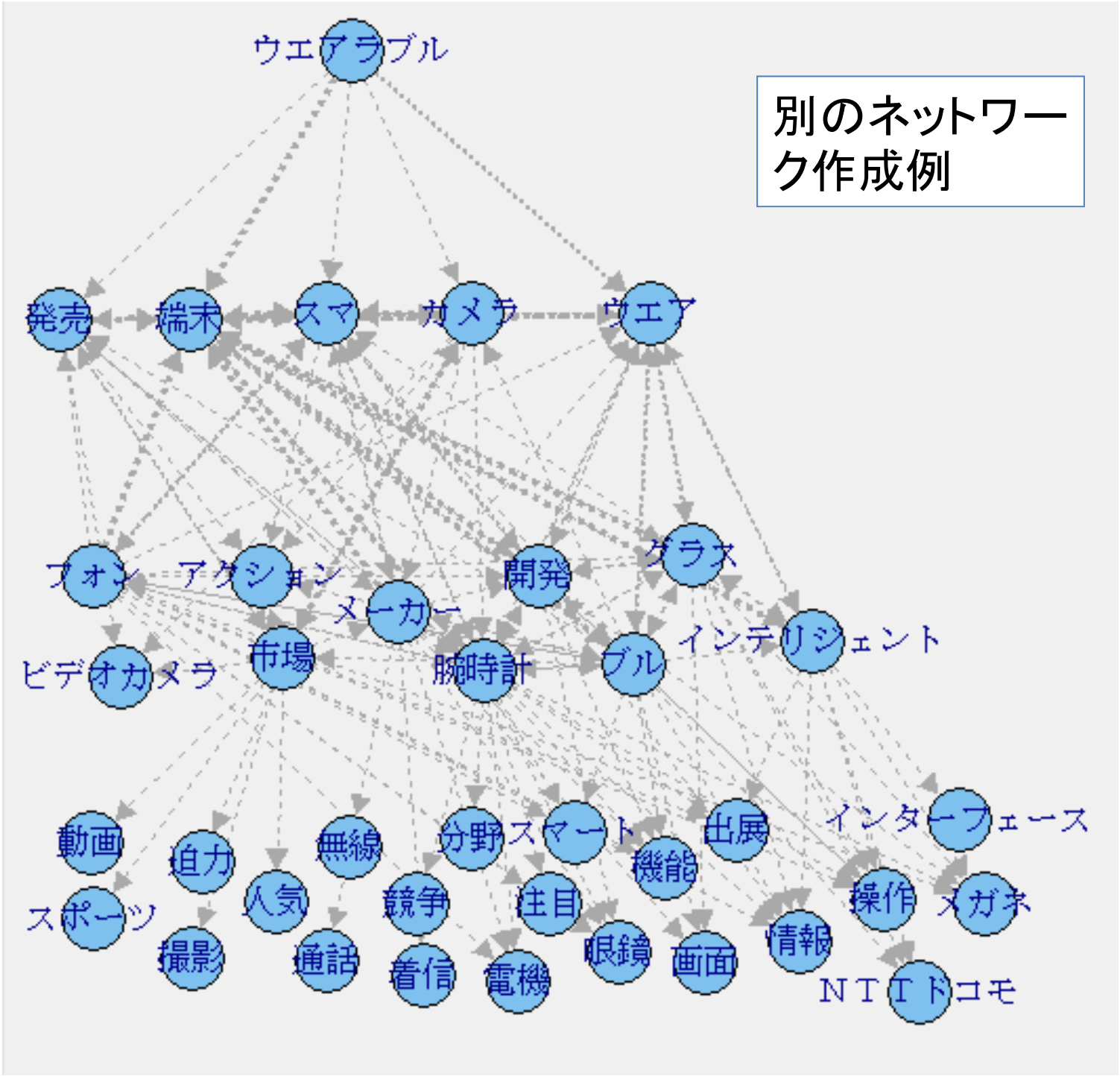
評価1：不要な単語のノードを含まない割合

- ４つのネットワークで実験

データ	方法		
	ベースライン	テーマ限定抽出	テーマ関連抽出
オリンピック	0.83 (29/35)	1.00 (35/35)	1.00 (35/35)
地震	1.00 (52/52)	1.00 (39/39)	1.00 (47/47)
ギリシャ	0.93 (64.5/69)	1.00 (24/24)	0.93 (62.5/67)
オウム	0.65 (50/77)	1.00 (24/24)	0.67 (48/72)
合計	0.84 (195.5/233)	1.00 (122/122)	0.87 (192.5/221)

評価2：ベースライン手法で得られた単語のカバー率

データ	方法	
	テーマ限定抽出	テーマ関連抽出
オリンピック	0.86 (12/14)	1.00 (14/14)
地震	0.95 (18/19)	1.00 (19/19)
ギリシャ	0.67 (16/24)	1.00 (24/24)
オウム	0.83 (15/18)	1.00 (18/18)
合計	0.81 (61/75)	1.00 (75/75)


別のネットワーク作成例
ウエアラブル
発売
端末
スマ
カメラ
ウエア
フォン
アクション
メーカー
開発
グラス
インテリジェント
ビデオカメラ
市場
腕時計
ブル
動画
迫力
無線
分野
スマート
出展
インターフェース
スポーツ
人気
競争
注目
機能
操作
メガネ
撮影
通話
着信
電機
眼鏡
画面
情報
NTTドコモ

２．機械学習を利用した文生成

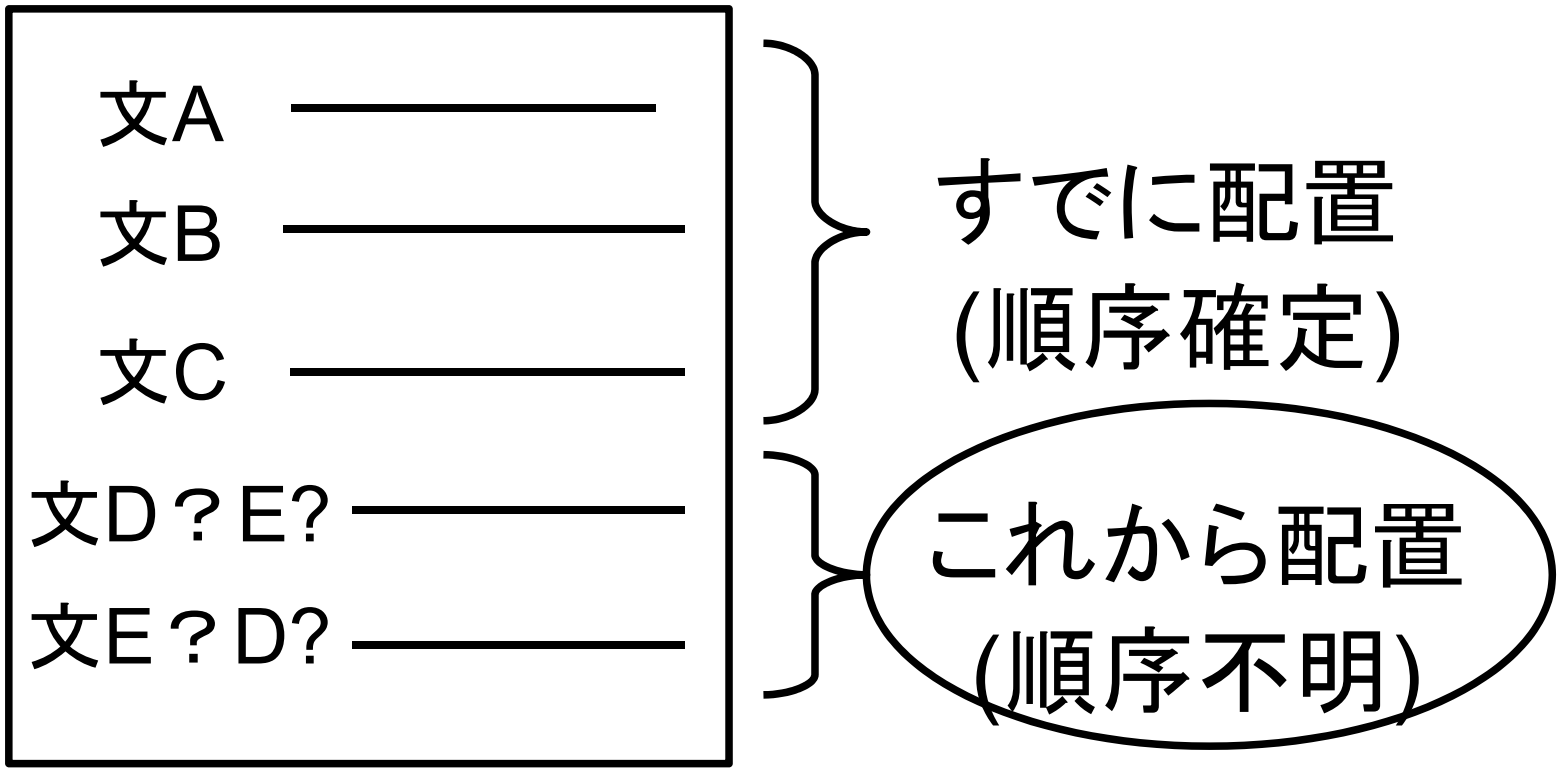
- ①機械学習を利用した文の順序推定
- ②機械学習を利用した段落の順序推定
- ③機械学習を利用した単語の使い分け

機械学習を利用した文生成

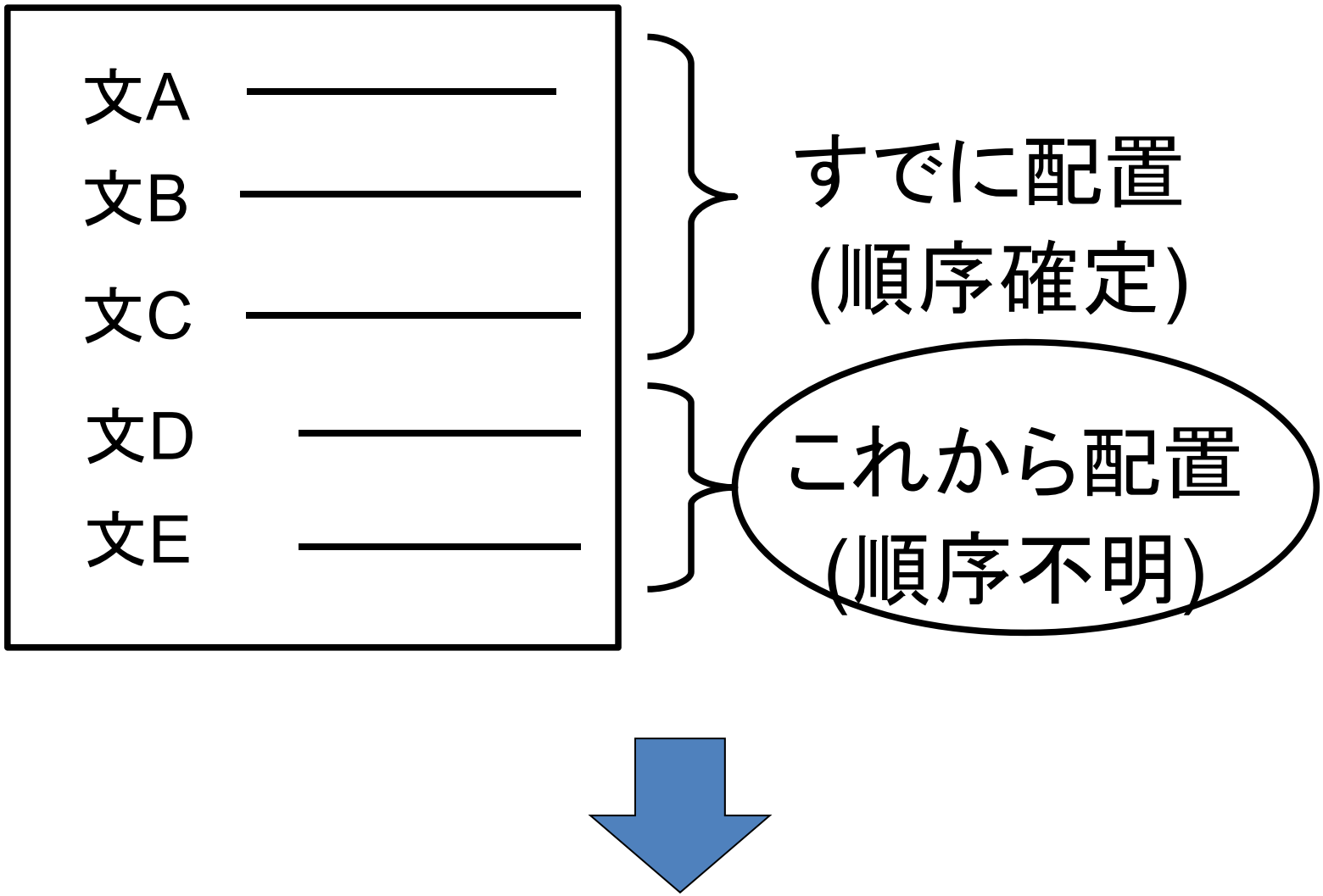
- 基本的に文生成には教師あり機械学習が使いやすい
- 生コーパスがそのまま学習データに使える
  - コーパスの文が正しい文と考える
  - 生成の目標(正解データ)は普通の文

- 問題点(今は考えない)
  - コーパスの文自体が間違っている場合
  - コーパスの文以外の言い方も正しい場合

①機械学習を利用した文の順序推定



順序推定



正解の文の順序： 文D → 文E

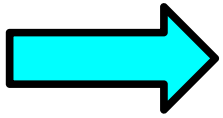
提案手法で用いる素性

f1 :出現する単語と品詞
f2 :出現する品詞のみ
f3 :主語の有無
f4 :体言止めの有無
f5 :文を助詞「は」で区切りその以前の自立語
f6 :文を助詞「は」で区切りその以後の自立語
f7 :2文で使用されている助詞の対
f8 :2文での単語の共起数
f9 :1文目の助詞「は」以後の自立語と
　2文目の助詞「は」以前の自立語の共起数
f10:判定する2文以前に並べた文で出現する単語と品詞
f11:判定する2文の直前文の体言止めの有無
f12:判定する2文の直前文の主語省略の有無
f13:判定する2文と直前文の自立語の一致数

実験結果

- 確率手法(先行手法) との比較
  - テストデータを提案手法と先行研究手法で推定

	提案手法	確率手法
case1(最初のみ)	**0.76**	0.60
case2(連接全て)	**0.72**	0.58
case3(全通り)	**0.72**	0.57

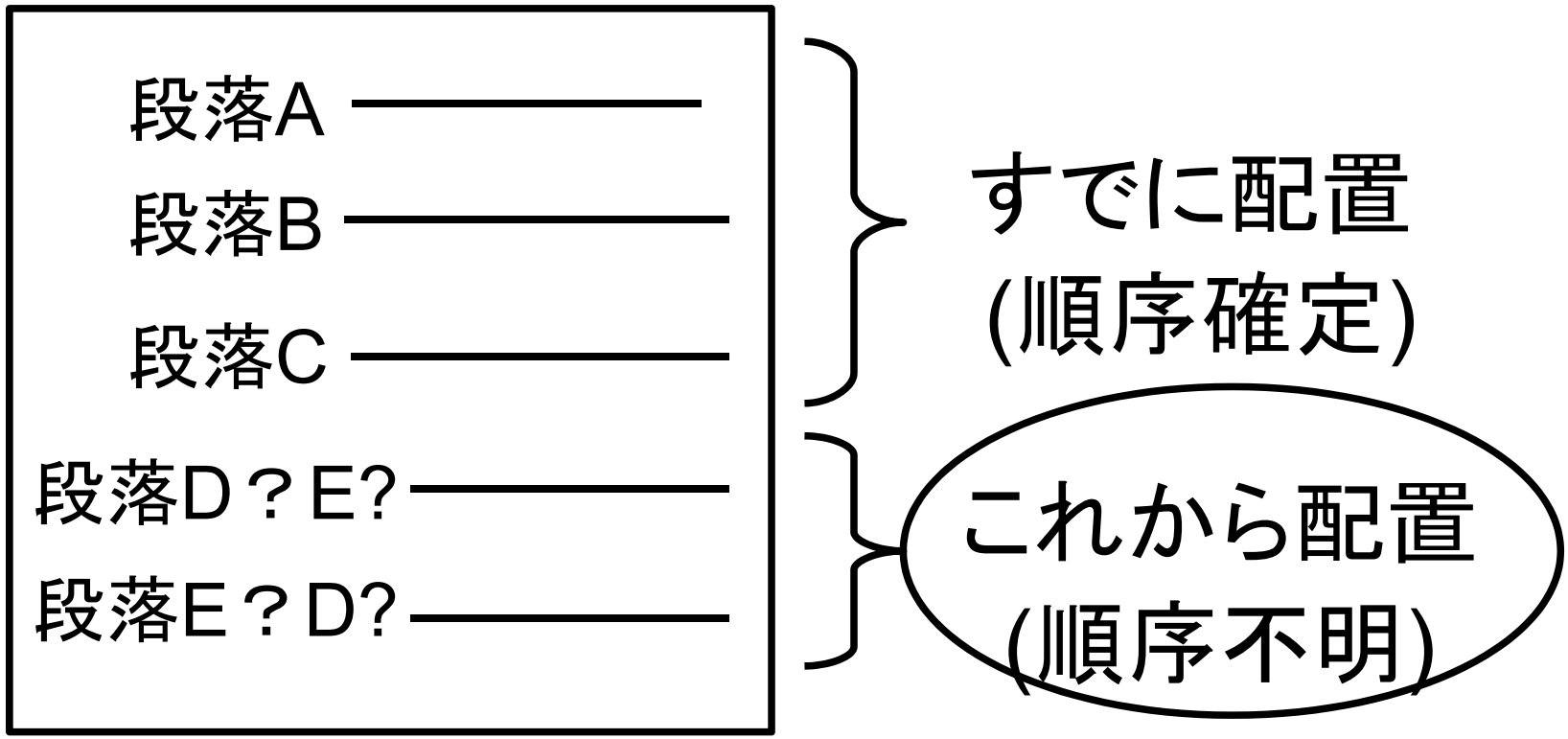


どのcaseでも
提案手法＞確率手法

SVMを利用

単語の連鎖を利用

②機械学習を利用した段落の順序推定



順序推定

機械学習を利用した段落の順序推定結果

- 記事中の最初の2段落
  - −0.85

SVMを利用

- 記事中の2段落連続
  - −0.65

③機械学習を利用した単語の使い分け

教師あり機械学習を利用して、類義語のうちどちらの語が文中にあったのかを推定

文中の情報を素性として使用

例:「衣料」と「衣類」

- 衣料の防虫科学研究では国内の第一人者。
- リュックサックに寝袋、衣類を詰めての貧乏旅行だ。

↓

- Xの防虫科学研究では国内の第一人者。 → 分類先:衣料
- リュックサックに寝袋、Xを詰めての貧乏旅行だ。 → 分類先:衣類

Xとした部分にどちらの語があったのかを推定する。

→ 正しく推定できた場合は文脈になんらかの特徴がある

→ 使い分けが必要な可能性

実験結果：正解率の比較

４５個の類義語対での正解率の平均

	提案手法（機械学習）	ベースライン手法
正解率の平均	0.86	0.70

提案手法＞ベースライン手法

提案手法自体が同義語の使い分けに有用

ベースライン手法　--- 常に頻度の高い方の語を出力

機械学習には最大エントロピー法を利用

２．機械学習を利用した文生成（まとめ）

- ①機械学習を利用した文の順序推定
- ②機械学習を利用した段落の順序推定
- ③機械学習を利用した単語の使い分け

- 基本的に文生成には教師あり機械学習が使いやすい
- 生コーパスがそのまま学習データに使える
  - コーパスの文が正しい文と考える
  - 生成の目標（正解データ）は普通の文

出典

- １．情報抽出（ネットワークの構築）
  - Yuta Doen, Masaki Murata, Ryuta Otake, Masato Tokuhisa, Qing Ma,Construction of Concept Network from Large Numbers of Texts for Information Examination Using TF-IDF and Deletion of Unrelated Words, SCIS & ISIS, pp.1108-1113, 2014.

- ２．ー①機械学習を利用した文生成
  - Yuya Hayashi, Masaki Murata, Liangliang Fan, Masato Tokuhisa, Japanese Sentence Order Estimation using Supervised Machine Learning with Rich Linguistic Clues, International Journal of Computational Linguistics and Applications, Vol.4, No.2, pp.153-167, 2013.

出典

- ２．―②機械学習を利用した段落の順序推定
  - Satoshi Ito, Masaki Murata, Masato Tokuhisa, Qing Ma, Order Estimation of Japanese Paragraphs by Supervised Machine Learning, SCIS & ISIS, pp.1096-1101, 2014.

- ２．―③機械学習を利用した単語の使い分け
  - Masaki Murata, Yoshiki Goda, Masato Tokuhisa, Automatic Selection and Analysis of Synonyms in Japanese Sentences Using Machine Learning, The 2014 International Conference on Artificial Intelligence, 2014.