fit2008@慶応大学湘南キャンパス
20080902

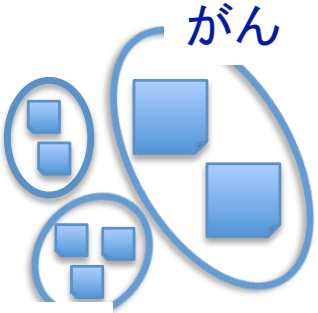
新聞記事からの社会課題に対する技術的対策情報の抽出

東京工業大学 統合研究院
イノベーションシステム研究センター
乾　孝司　　内海 和夫
橋本 泰一　　石川 正道

奈良先端科学技術大学院大学
情報科学研究科
村上 浩司

研究の背景

- 学内研究プロジェクトの立案支援
  - 学内研究者情報（論文，特許）の提供
  - 社会的課題についての情報の提供
    - 新聞記事のクラスタリングがベース技術
    - 課題記事クラスタの構築[内海ら2007][橋本ら2008]
      - 課題に対する技術的対策情報の整理
      - 技術的対策情報の抽出

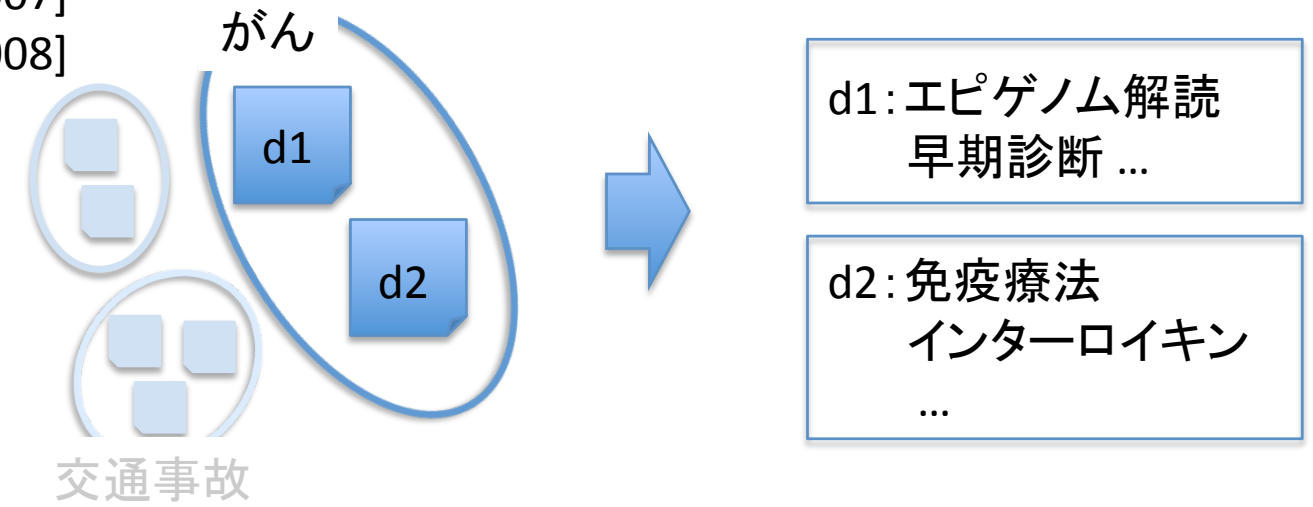


問題設定

入力：ある社会課題についての記事群

出力：技術的対策情報（対策用語）

[内海ら2007]
[橋本ら2008]



エピゲノム解読：遺伝子の後天的変化を追跡する技術
インターロイキン：免疫反応に関連する細胞間相互作用を媒介する蛋白性物質

対策用語抽出の手続き

(記事単位で処理する)

1. 対策用語候補を生成
    - 文節同定＋付属語等の排除
2. 対策用語候補をスコアリング (後述)
3. スコア上位 n 個を抽出

スコアリング

- 対策用語への２つの要件
  - 課題との関連が強く,
  - 技術とも関連が強い.

c.f　tf-idf 法

スコアリング

- 対策用語への２つの要件
  - 課題との関連が強く，→ 課題関連度 指標を導入
  - 技術とも関連が強い．→ 技術関連度 指標を導入

c.f　tf-idf 法

課題関連度 × 技術関連度

でスコアリング．

スコアリング：課題関連度

- 課題は記事クラスタとして表現されている
- クラスタ内の特徴語は課題と関連が強い

→ クラスタ内の特徴語の指標（カイ２乗値）を利用する

スコアリング：技術関連度

- 技術を表す言語パタンに着目
- パタンと近い位置にある語は技術との関連が強い

$$tr(t,d) = \sum_{t_j^d \in ins(t,d)} tr_{intra}(t_j^d)\, tr_{inter}(t_j^d)$$

頻度２以上の場合は足し算

文内関係に基づく関連度

文間関係に基づく関連度

X を → 開発する

X の研究を → 手掛ける　（「→」は係り受け関係）

スコアリング：技術関連度

- 文内関係に基づく関連度

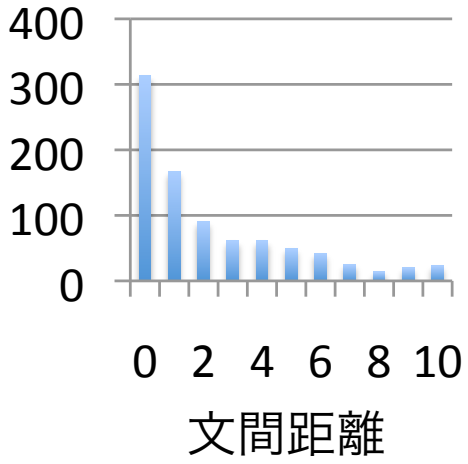
$$tr_{intra}(t_j^d) = \begin{cases} 100 & (t_j^d \text{ が言語パタン内の項} \\ & \quad or \text{ その修飾要素}) \\ 1 & (otherwise). \end{cases}$$

X を → 開発する

X の研究を → 手掛ける　（「→」は係り受け関係）

スコアリング：技術関連度

- 文間関係に基づく関連度



対策用語を含む文と
言語パタンを含む文との文間距離のヒストグラム

指数回帰式は,

$$y = e^{-0.77x}$$

$$tr_{inter}(t_j^d) = \max_{p_k^d \in P^d} \exp(-0.77 \times r\_pos(p_k^d, t_j^d))$$

パタンが複数ある場合は最近隣を選択

文間距離を求める

スコアリング：まとめ

- 課題関連度
  - クラスタ内の特徴語の指標
    - カイ２乗値
    - （同じ語であれば，記事間で同じ値）
- 技術関連度
  - 言語パタンとの位置関係に基づく指標
    - 新規で設計 $tr(t,d) = \sum tr_{intra}(t_j^d) tr_{inter}(t_j^d)$
    - （同じ語でも，記事ごとに異なる値）

抽出実験

- データ：日本経済新聞 2005年
  - がんクラスタ/ 生活習慣病クラスタ
  - 専門家が人手で正解キーワードを作成
- 抽出用語数
  - 文数が１０未満の記事の場合 → 文数
  - １０以上 → １０で固定
- tf-idf スコアリングとの比較

抽出実験

- 結果（F値）

	がん	生活習慣病
提案手法 (課 × 技)	0.532	0.608
課題関連度	0.473	0.558
技術関連度	0.487	0.512
tf-idf	0.470	0.509

抽出例

- 正しく正解できた　抗がん剤 がん治療法 薬物送達システム ＤＤＳ 中性子線照射 ホウ素製剤 がん細胞
- 正解できなかった　死滅 微小カプセル 副作用
- 誤って抽出した　患部　集中的

- 正しく正解できた　心臓病 動画 診断精度向上 冠状動脈
- 正解できなかった　コンピューター断層撮影装置 新型ＣＴ 早期発見
- 誤って抽出した　患者 血流 心筋梗塞

まとめ

- 新聞記事からの技術的対策情報の抽出
  - 課題関連度と技術関連度
  - tf-idf 法よりも抽出精度が向上
- 今後の課題
  - 評価実験規模の拡大
  - 誤りへの対応
    - 「従来技術」を誤って抽出してしまう