情報処理学会NL研@九州大学
2008/11/26,27

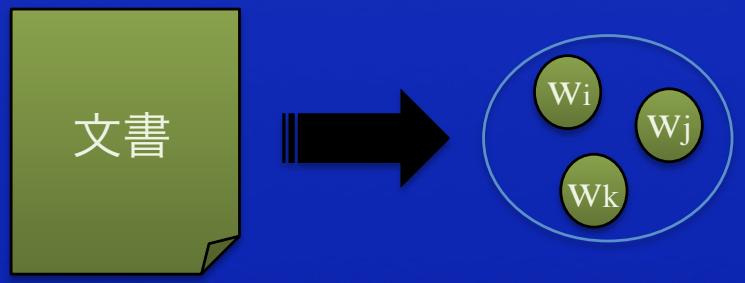
キーワード抽出の整数計画問題としての定式化

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　　橋本 泰一
内海 和夫　　石川 正道

東京工業大学 精密工学研究所
高村 大也

はじめに

- キーワード抽出
  - 文書から，その意味内容を端的に表す語(句)の集合を抽出する



- 要約への応用
  - キーワード抽出を要約の一形態とみなす
    - 論文へのキーワード自動付与
  - キーワード抽出を要約の要素技術とみなす
    - 要約文書生成の前処理として要約文書に含めるべき語を選別

キーワード抽出の先行研究

教師ありデータ

利用する

– 教師あり機械学習 ([Zhang+ 2006], [Sujian+ 2003]等)
  - キーワード候補を抽出する／しないの２値分類
  - 教師ありデータを準備する必要がある
    → 適用時の準備負荷(データ作成負荷)が高い

利用しない／できない

– 重み付けによるランキング　上位 **k**個を抽出

(1) 単語単体の特徴量　tfidf

(2) 単語間の特徴量(共起)で文書グラフ化 + PageRank等のグラフベース法 ([Mihalcea+ 2004])

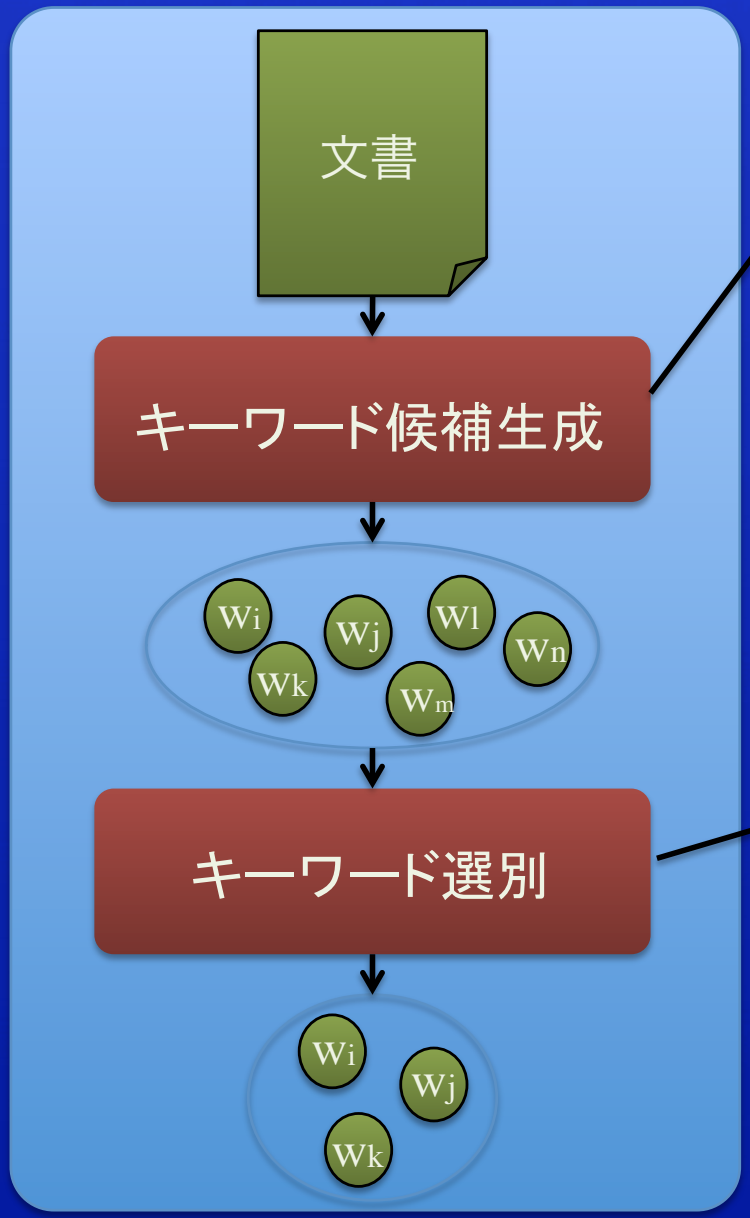
– (1)と(2) を同時に適切に考慮できる手法を提案する

本研究

提案手法

- 目次
  - キーワード抽出の全体的な流れ
  - 施設配置問題
  - 施設配置問題とキーワード選別
  - 評価実験

キーワード抽出の全体的な流れ



キーワード候補生成

- 入力文書中の単語列から
  キーワード候補を生成する
  - 被覆率を高く，もれなく
  - 文をChaSenで解析し，
    全ての名詞連続を候補として抽出

キーワード選別

- キーワード候補から抽出すべき
  キーワードを選別し，出力する

単語単体の特徴量 と 単語間の特徴量 を考慮

提案手法

- 目次
  - キーワード抽出の全体的な流れ
  - 施設配置問題

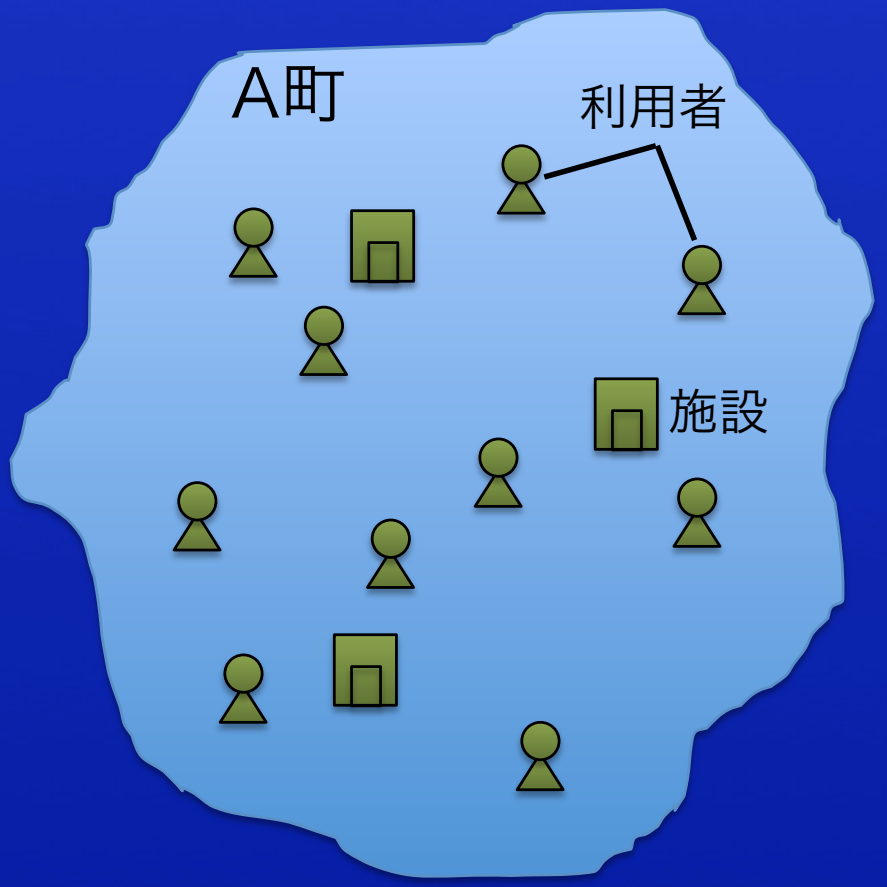
    単語単体の特徴量 と 単語間の特徴量 を考慮するにあたり，

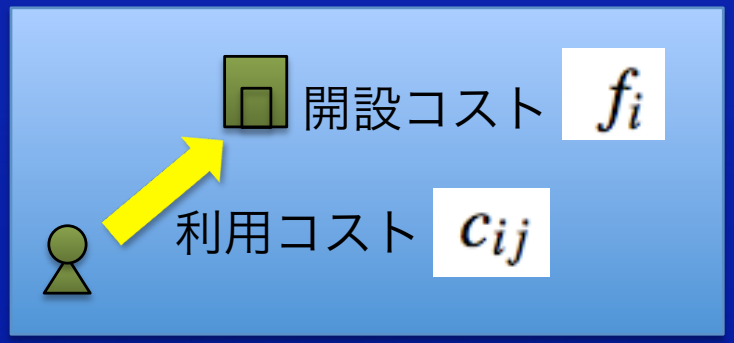
    最適化問題の下位問題である施設配置問題の観点から定式化する

  - 施設配置問題とキーワード選別
  - 評価実験

施設配置問題

- ある地域における学校や商業施設等の設立地計画をモデル化した問題
  - 施設をどこに？
  - いくつ？





開設コストと利用コストの総和を最小化する

施設配置問題（定式化）

開設コスト　利用コスト

$$\min \quad \sum_{i \in \mathcal{F}} f_i y_i + \sum_{i \in \mathcal{F}} \sum_{j \in \mathcal{D}} c_{ij} x_{ij}$$

目的関数

$$s.t. \quad x_{ij} \le y_i \qquad (i \in \mathcal{F}, j \in \mathcal{D})$$

利用する施設は開設されている

$$\sum_{i \in \mathcal{F}} x_{ij} = 1 \qquad (j \in \mathcal{D})$$

利用者は必ず利用施設を１つもつ

$$x_{ij} \in \{0,1\} \qquad (i \in \mathcal{F}, j \in \mathcal{D})$$

利用者の割当状況をあらわす変数

$$y_i \in \{0,1\} \qquad (i \in \mathcal{F})$$

施設候補の開設状況をあらわす変数

利用者　施設候補

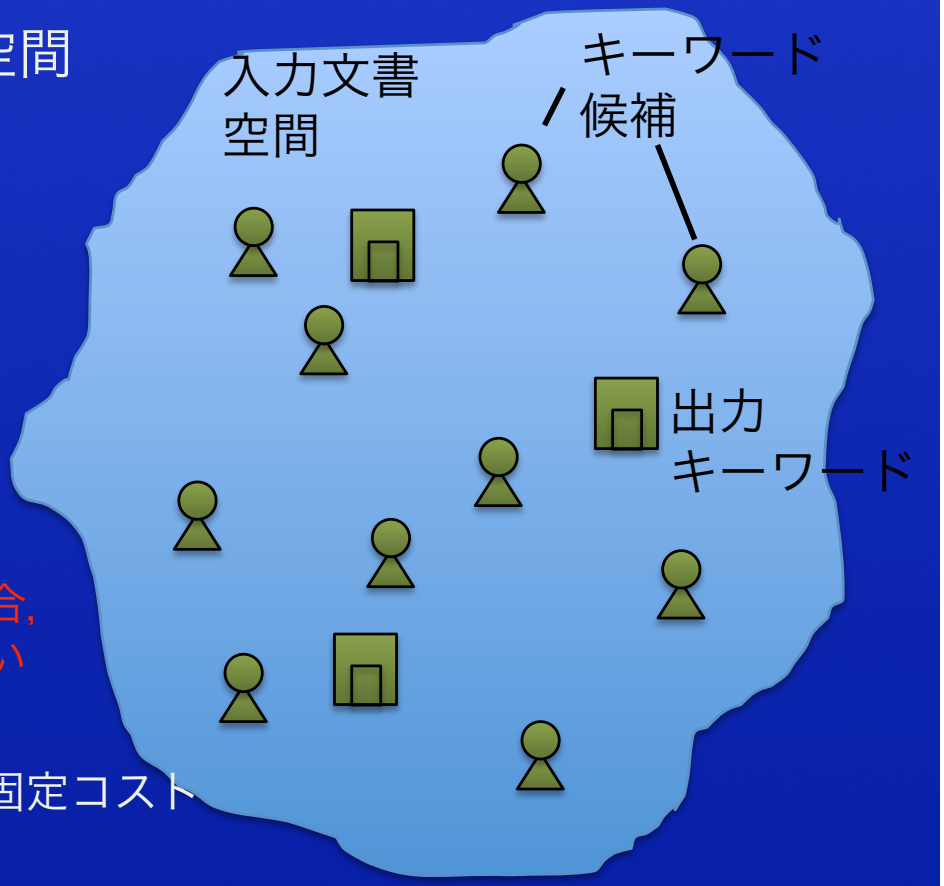
最終的に y_i =1となる候補を開設する

提案手法

- 目次
  - キーワード抽出の全体的な流れ
  - 施設配置問題
  - 施設配置問題とキーワード選別
    - 施設配置問題とキーワード選別を対応づける
  - 評価実験

施設配置問題 と キーワード選別 の対応づけ

- 地域 = 入力文書/キーワード候補空間
- 施設 = 出力キーワード
  - f_i 開設コスト → 出力コスト
    - 出力にかかる固定コスト
- 利用者 = キーワード候補
  - c_{ij} 利用コスト → 代表コスト
    - 候補Ａから候補Ｂが連想できる場合，ＢをＡで代表させ，Ｂは出力しない
    - ある候補を，(それを連想できる)別のある候補で代表させるための固定コスト



キーワード抽出
= 出力コストと代表コストの総和を最小化する

定式化

$$\min \quad \sum_{i \in \mathcal{D}} f_i y_i + \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}} c_{ij} x_{ij}$$

$$\text{s.t.} \quad x_{ij} \leq y_i \qquad (i, j \in \mathcal{D})$$

$$\sum_{i \in \mathcal{D}} x_{ij} = 1 \qquad (j \in \mathcal{D})$$

$$\sum_{i \in \mathcal{D}} y_i = k$$

$$x_{ii} = y_i \qquad (i \in \mathcal{D})$$

$$x_{ij} \in \{0, 1\} \qquad (i, j \in \mathcal{D})$$

$$y_i \in \{0, 1\} \qquad (i \in \mathcal{D})$$

キーワード選別のために
制約条件を追加する

出力キーワード数 (任意)

出力されるなら，代表先は自身

キーワード候補

(同じ集合に属していることに注意)

出力キーワード

最終的に y_i =1

となる候補を出力することで
キーワード抽出が実現できる

コスト

	f_i	c_{ij}
施設配置問題	開設コスト •建築費用	利用コスト •移動距離
キーワード抽出	出力コスト 出力利得 $\acute{f}_i = -f_i$ •出力した際に得られる利得 = tfidf等　単語単体の特徴量	代表コスト 代表利得 $\acute{c}_{ij} = -c_{ij}$ •連想できた際に得られる利得 = 共起確率等　単語間の特徴量

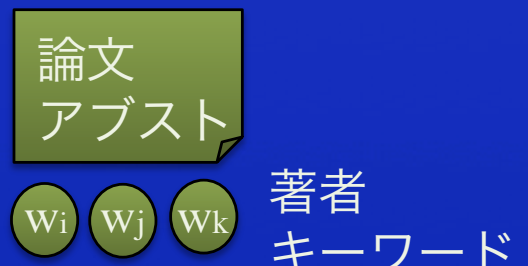
各コストの具体的な定義は予稿集を参照

提案手法

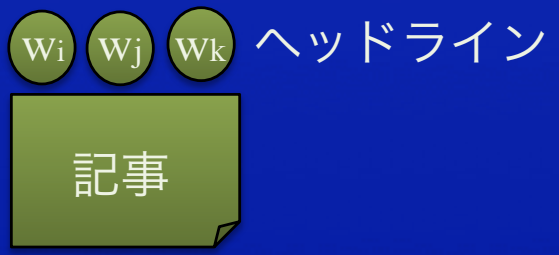
- 目次
  - キーワード抽出の全体的な流れ
  - 施設配置問題
  - 施設配置問題とキーワード選別
  - 評価実験

評価実験

- ２種類のデータセット
  - Paper
    - 正解 = 著者キーワード
    - 入力文書 = 論文アブストラクト
    - キーワード抽出を要約の一形態とみなした場合に対応



  - News
    - 正解 = ヘッドライン中の名詞連続
    - 入力文書 = 記事本文
    - キーワード抽出を要約の要素技術とみなした場合に対応



評価実験

- データセットの統計量

	文書数	平均キーワード数	上限値（F値）
Paper	100	3.9	0.917
News	100	3.8	0.787

評価実験

- solver
  - GLPK package
  - *http://www.gnu.org/software/glpk/*

評価実験

- 実験結果 (F値)

Paper	k = 3	k = 5
提案手法	0.278	0.330
f_i 単独	0.250	0.290
共起を利用した既存手法		
[Mihalcea+ 2004]	0.128	0.161
[Palshikar 2007]	0.108	0.132

News	k = 3	k = 5
提案手法	0.240	0.265
f_i 単独	0.221	0.256
共起を利用した既存手法		
[Mihalcea+ 2004]	0.152	0.172
[Palshikar 2007]	0.064	0.070

既存手法は英語データに関する論文であるので，参考数値である

評価実験

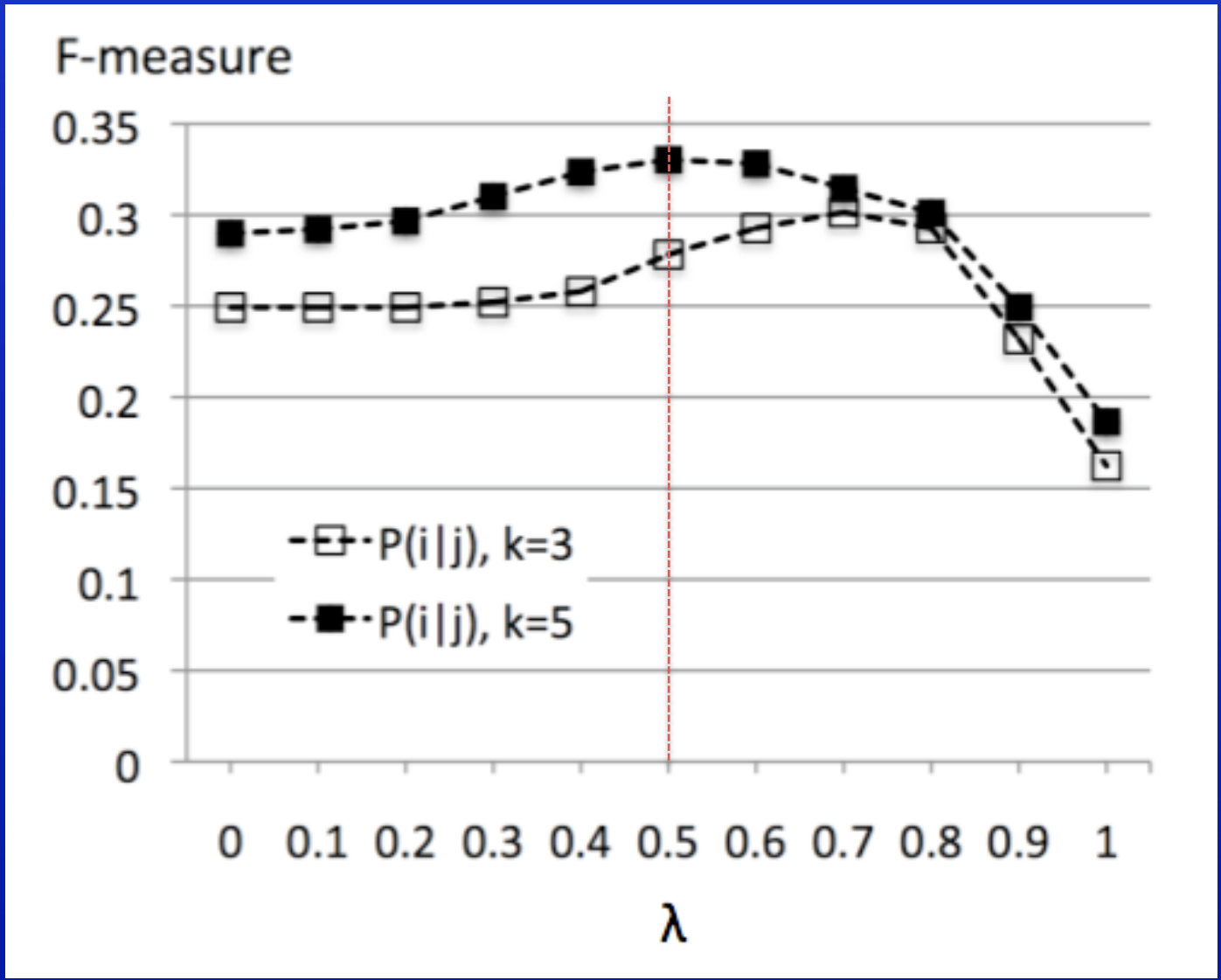
- 考察　コストのバランス調整
  - 現状：出力コストと代表コストは等しい重み
    → どちらを重視するのがよいか?

$$\min \quad \sum_{i \in \mathcal{D}} f_i y_i + \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}} c_{ij} x_{ij}$$

$$\min \quad (1-\lambda) \sum_{i \in \mathcal{D}} f_i y_i + \lambda \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}} c_{ij} x_{ij}$$

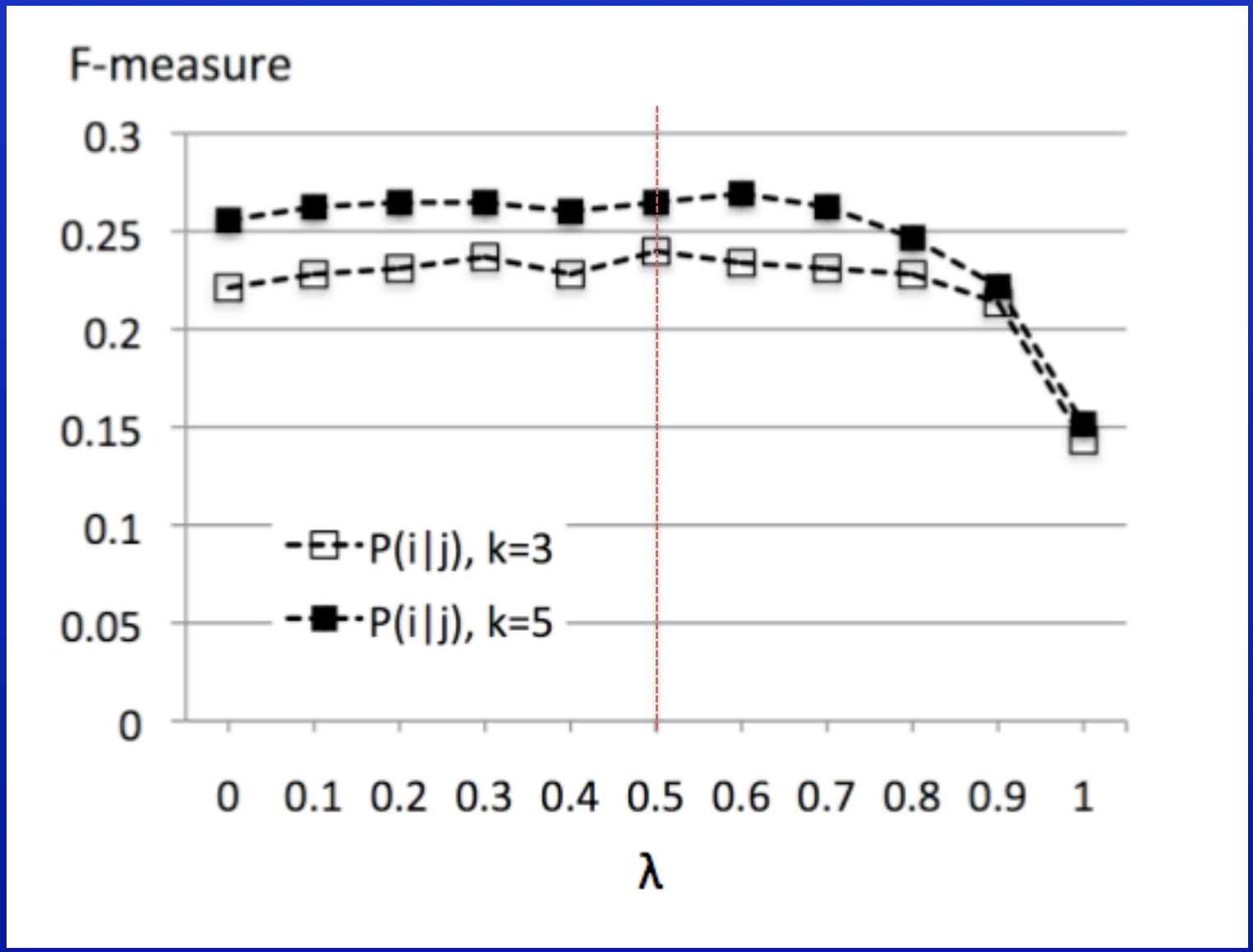
評価実験

Paper


F-measure
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
P(i|j), k=3
P(i|j), k=5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
λ

評価実験

News


F-measure
0.3
0.25
0.2
0.15
0.1
0.05
0
P(i|j), k=3
P(i|j), k=5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
λ

まとめ

- 新しいキーワード抽出手法を提案
- 提案手法の特徴
  - 教師ありデータを用いない手法
    - 適用が容易
  - キーワード抽出(選別)を施設配置問題の観点から定式化
    - 単語単体の特徴量と単語間の特徴量を共に考慮
    - 文書内のキーワード候補の相互関係を考慮した大域的最適化

- 今後の課題
  - 出力キーワード数の自動推定
  - キーワード候補生成の精錬
  - 教師あり機械学習手法との融合 など