特徴語抽出の精度改善に向けた反復度と条件付き確率の比較

菊地 真人，山内 達登，Bui Tuan Thanh，梅村 恭司
豊橋技術科学大学情報・知能工学系

1. 研究背景

➢ **本研究で扱う特徴語抽出について**

- ある文書からその文書を特徴づける語を抽出
- 応用： 情報検索，文書要約，文書分類，etc.
- 語の重要度を下記の二つの統計量で推し量って語を抽出する

例）ワードクラウド

https://blog.web.nifty.com/engineer/2217

➢ **(A) 反復度**

- “ある文書における特徴語はその文書で繰り返し出現する”という仮定に基づいた統計量
- 特徴語やその定義などの教師情報が事前に与えられなくても使用できる

➢ **(B) 教師情報を用いた統計量（条件付き確率）**

- 教師情報を用いないケース（例：反復度の使用）より有効と考えられる
- しかし特徴語の定義は主観的で，教師情報が事前に与えられることは実用的に多くない

- 特徴語抽出の実験で，(A) と (B) のふるまいを比較し反復度の実用性を検証
- 論文のアブストラクトに出現する語のうち，どの語がタイトルに現れるかを予測する問題を解く

2. 特徴語抽出の問題設定

➢ **使用するコーパス：NTCIR-1 コーパス**

- 332,921 種類の論文タイトル，アブストラクトの組

ID	タイトル	アブストラクト
01	電気回路演習…	大学等での基礎的な電気回路演習を…
02	講義方式「半…	従来の講義方式による半導体工学用…
03	複合メディア…	本報は、大学工学部電気系学科にお…

➢ **本研究における特徴語の定義**

- タイトルに現れる語を対応するアブストの特徴語とする
- 語の単位は文字バイグラムとする

ID-01 特徴語：電気，気回，回路 …
ID-02 特徴語：講義，義方，方式 …

➢ **実験手順と評価尺度**

特徴語の総数

データ	組数	バイグラム数	正解数
学習	110,972	31,491,771	4,113,198
開発	100	24,965	3,165
評価	100	23,661	3,167

（0）NTCIR-1 コーパスを基に実験データを作成

（1）学習データが含むバイグラムの頻度を計数

（2）評価データにあるアブストのバイグラムについて，（3.で示す）統計量をそれぞれ計算

（3）統計量ごとに推定値の降順でバイグラムを順位付けし下記の指標を求める

- スピアマンの順位相関係数：反復度と条件付き確率の順位リストの相関
- R-適合率：順位リストの上位R件に含まれる特徴語の割合（Rは評価データ中の正解数）

3. 使用する統計量

➢ **反復度**

- **学習データのアブストのみから計算する**
- アブストにおける文字バイグラム w の繰り返しやすさ

$$\mathrm{adapt}(w) \approx \begin{cases} \dfrac{df_2(w)}{df(w)}, & (df_2(w) \geq \theta_{\mathrm{ad}}) \\ 0, & (\mathrm{otherwise}) \end{cases}$$

$df_2(w)$：文字バイグラム w が 2 回以上出現するアブストラクトの数
$df(w)$：文字バイグラム w が 1 回以上出現するアブストラクトの数
θ_{ad}：$df_2(w)$ のしきい値

➢ **条件付き確率**

- **学習データのアブストに加えてタイトル（教師情報）も使って計算する**
- あるアブストに w が出現した条件下で，対応するタイトルにも w が出現する確率

$$p(e_T(w)|e_A(w)) \approx \begin{cases} \dfrac{f_{T\wedge A}(w)}{f_A(w)}, & (f_{T\wedge A}(w) \geq \theta_{\mathrm{cp}}) \\ 0, & (\mathrm{otherwise}) \end{cases}$$

$f_{T\wedge A}(w)$：アブストラクトとタイトルで w を共に含む論文数
$f_A(w)$：アブストラクトで w を含む論文数
θ_{cp}：$f_{T\wedge A}(w)$ のしきい値

※ 開発データのR-適合率が最大となった $\theta_{\mathrm{ad}} = \theta_{\mathrm{cp}} = 15$ をしきい値として採用

4. 実験結果＆今後の課題

➢ **スピアマンの順位相関係数**

- 反復度と条件付き確率の順位リストの相関
- 0.691 と高い相関を示した

➢ **R-適合率**

- 上位R件（3,167件）に含まれる特徴語の割合
- 反復度：0.219，条件付き確率：0.474

➢ **バイグラムと推定値の例**

IDF（逆文書頻度）等と組み合わせるとよい？

◎ **反復度はアブストだけでもある程度の特徴語を抽出できる（左図）**

× **反復度はどのアブストにも繰り返し現れやすい，文の要素（Not 特徴語）も多く抽出（右図）**

バイグラム	反復度	条件付き確率	正解／不正解
水ト	0.607	0.557	3/0
重合	0.636	0.529	3/0
汚損	0.589	0.518	9/0
ポリ	0.576	0.496	11/4
放電	0.570	0.484	73/20

バイグラム	反復度	条件付き確率	正解／不正解
た。	0.751	0	0/183
る。	0.669	0	0/183
して	0.565	0.015	0/124
は、	0.555	0	0/106
こと	0.531	0	0/113

➢ **今後の課題**

- 反復度の改良（特徴語のみを抽出できるようにしたい）
- TF-IDFといったスタンダードな手法との性能比較，ふるまいの差異の分析
- 様々な特徴語の定義における実験および評価