# 名詞の共起頻度に着目した専門用語自動抽出手法の提案

○木村優介*　馬場睦也**　刘钊宇**　波多野賢治*
* 同志社大学 文化情報学部　** 同志社大学大学院 文化情報学研究科

## 1. 背景


ある文書

用語抽出 → 抽出される専門用語の例
「知識」，「システム」，「学習者」 等

応用先： テキストの難易度推定[2]


テキストに出現する専門用語からテキスト自体の難易度を推定

■ 専門用語抽出は人手と時間のかかる作業であるため，
自動化することは意義がある[1]

【研究目的】
難易度推定のために専門用語の自動抽出手法を提案

## 2. 関連研究

■ 専門用語かどうかをスコアで示す指標 FLR[1]

- 専門用語の多くは複合名詞→単名詞，複合名詞を専門用語候補語とする
- 表1の特徴を用いている
- スコアが高ければ高いほど専門用語であることを示す
- 15000語の候補語抽出の際には*F*値約0.6で抽出できる

表1：専門用語の特徴[3]

| | 意味 | 各性質の計測方法 |
|---|---|---|
| ユニット性 | コーパス内での安定的な使用 | 出現頻度 |
| ターム性 | ある言語的単位の持つ分野固有の概念への関連性の強さ | 複合名詞の構造（単名詞の連接頻度） |

- ユニット性・ターム性がお互いに補完しあうことで高い精度で抽出できる

$$FLR(CN) = f'(CN)\left(\prod_{i=1}^{L}(FL(N_i)+1)(FR(N_i)+1)\right)^{\frac{1}{2L}}$$

$N_i$ :複合名詞$CN$ を構成する$i$ 番目の単名詞
$CN$ :単名詞$N_1$，$N_2$，…，$N_L$の順で連接する複合名詞
$f'(CN)$ :複合名詞$CN$ が単独で出現した回数
$L$ :複合名詞$CN$ を構成する単名詞の数
$FL(Ni)$ :$N_i$の左側に隣接した単名詞の延べ語数
$FR(Ni)$ :$N_i$の右側に隣接した単名詞の延べ語数

■ 佐藤らによる専門用語の特徴

| 専門用語の特徴[4] |
|---|
| **1．特定分野で広く，それなりに使われている** |
| **2．一般語ではない** |
| **3．定義や説明が存在する** |
| **4．関連する専門用語が存在する** |

本研究では上記の3，4番から，専門用語（名詞）は共起して出現する他の専門用語が存在すると仮定し，「名詞の共起性」があると考えた

【関連研究の問題点】
FLRは専門用語に関係すると考えられる名詞の共起性を扱っていない

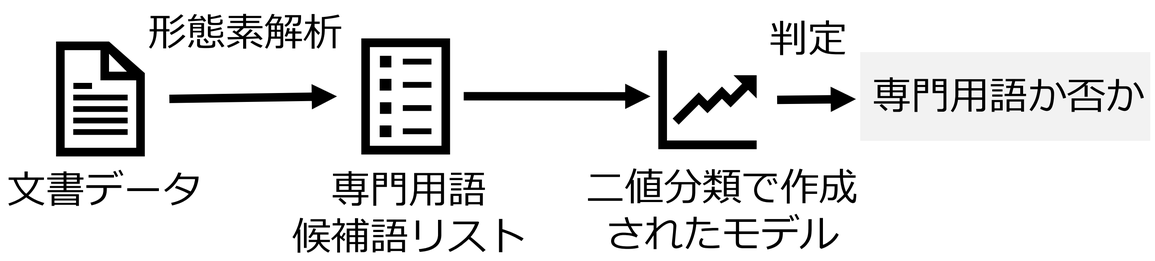
## 3. 提案手法

名詞の共起に関係する変数を使用し，
二値分類によるモデルで専門用語か否かの判定を行う

■ 正解専門用語のうち約86%が名詞
また，正解専門用語のうち約65%が複合名詞

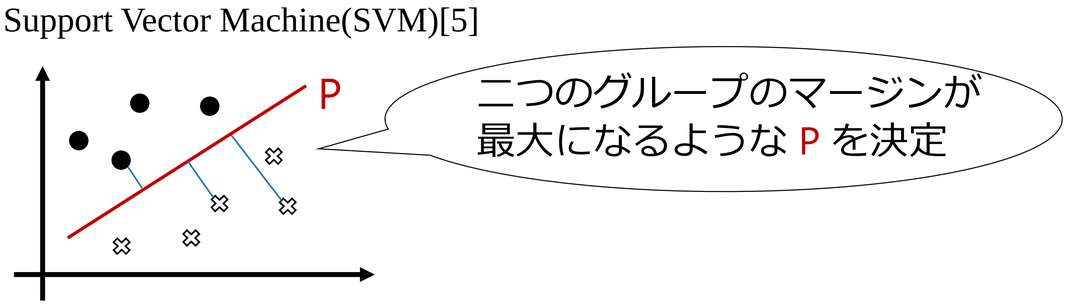
→ 専門用語候補語を複合名詞を構成する単名詞と複合名詞とする

■ 提案手法の流れ




■ 二値分類のモデル

- Support Vector Machine(SVM)[5]




他の二値分類のモデルも同様に用いて，
最も性能が高いものを選択する

■ 二値分類のモデルに用いる変数

- 単名詞の連接頻度数
（ある単名詞がほかの単名詞とどれだけ結びつくか）
- 専門用語候補語の出現頻度
- 単名詞か否か（単名詞であれば 1，そうでなければ 0）
- 文字長
- 専門用語候補語を構成する単名詞数
- 専門用語候補語のIDF： $idf(t) = \log\frac{N}{df(t)} + 1$
(idf(t) : ある候補語 t のIDF値，N：全文書，
df(t)：ある候補語 t が出現する文書の数)
- 同文書内および一文での名詞との共起頻度（名詞の共起性）

## 4. 評価方法

■ 使用データ・正解データ

NTCIR-1用語抽出研究用テストコレクションに含まれる人工知能分野の論文約1800件を本研究での使用データとし，そのデータから人手で専門用語を判断，抽出したデータを正解データとして使用

■ FLRによってスコアリングされた専門用語候補語を本研究の手法で専門用語かどうか判定
→FLRの再現率・適合率と比較

## 5. 今後の課題

- 提案手法の変数を考えられる限り増加
- 提案手法で示された有効な変数を用いてスコアリング手法の考案

## 6. 参考文献

[1]中川裕志，湯本紘彰，森辰則（2003）．『出現頻度と連接頻度に基づく専門用語抽出』『自然言語処理』, 10 (1),27-45.
[2]内山清子，鈴木崇史，相澤彰子（2010）．『専門用語の専門度の指標に関する一考察』『言語処理学会第16 回年次大会』
[3] Kyo Kageura and Bin Umino（2001）．『Methods of Automatic Term Recognition --- A Review ---』『Terminology』, 3
[4] 佐藤理史，佐々木靖弘 （2002）．『ウェブを利用した関連用語の自動収集』『情報処理学会研究報告自然言語処理』，2003（4），pp. 57-64.
[5]V.N.Vapnik（1998）．『Statistical Learning Theory』『Wiley』．

連絡先：bip0098@mail4.doshisha.ac.jp