# シラバスからの専門用語抽出手法の検討

芳鐘 冬樹† 井田 正明† 宮崎 和光† 野澤 孝之† 喜多 一‡
大学評価・学位授与機構† 京都大学 学術情報メディアセンター‡

## 1 はじめに

大学における授業計画を紹介するシラバスには，授業の内容を表す多くの専門用語が含まれている．それらの専門用語を適切に抽出できれば，大学ごとの傾向の分析や，科目の関連性の分析（あるいは科目分類の支援）などに活用が可能である．ただし，抽出に際しては，シラバス特有の言語表現や，抽出源とする項目の選択に注意を払わねばならない．

本研究では，結束性に注目する抽出手法とパターンベースの抽出手法を，収集・XML 変換したシラバスデータに適用し，それぞれのパフォーマンスを検証する．それぞれの手法の有効性と問題点を考察したうえで，シラバス特有の問題を考慮した抽出手法について検討する．

## 2 データ

近年，多くの大学が Web 上でシラバスの公開を行っている．本研究では，情報系学科を対象に，収集した７大学のシラバス（2002 年度）を専門用語の抽出源データとした．抽出に先立ち，まず，Web から収集した html あるいは pdf 形式のファイルを，井田[1]が提案する XML 形式のフォーマットに変換した．さらに「科目名」「授業概要」「授業計画」「履修により達成される目標」の４項目を取り出し，それらを抽出源データとした．

自動抽出のパフォーマンス評価の際に用いる専門用語の正解集合は，上記の４項目のデータから手作業で抽出した．（抽出支援 GUI ツールを開発・使用した．抽出作業は情報系博士課程の学生による．）抽出源とする４項目のデータと正解集合の基本的数量を表１に示す．なお，本研究では，形態素解析は茶筌[2]を用いて行っている．

表１ シラバスデータの基本的数量

| | |
|---|---|
| 大学（学科）数 | 7 |
| 科目数 | 465 |
| 形態素数（延べ） | 97660 |
| 形態素数（異なり） | 5370 |
| 正解専門用語数（延べ） | 22421 |
| 正解専門用語数（異なり） | 7119 |

## 3 専門用語抽出手法

これまで提案されてきた統計的な抽出手法のほとんどは，はじめに名詞列を候補語として抽出した後で，出現頻度(TF)[3]，連接頻度[4]などに基づいて専門用語か否かを判定している[5]．それらは，(1)代表性（対象分野をどこまで代表しているか：分野内頻度）に基づくもの，(2)識別性（対象分野を他の分野とどこまで識別しうるか：分野間の偏差）に基づくもの，(3)結束性（対象分野において，重要な概念を表している語はコロケーションが強いという理論）に基づくものに大別される[6]．(1)は一般語との区別を必ずしも考慮したものではないこと，(2)は他分野のデータが必要であることから，本研究では，(3)をベースにする手法である湯本[4]の手法を適用することとした．ただし，湯本の手法は（他の多くの統計的手法と同様）基本的に名詞列だけを抽出対象としているため，それ以外の品詞パターンの専門用語（例：深い知識）は取りこぼされてしまう．

本研究は，科目の傾向分析等への応用を想定した再現率重視の立場から，統計的な抽出手法をパターンベースの抽出手法で補うという方策をとる．（多少一般語のノイズが入り抽出精度が落ちても，一般語は科目間の偏差が小さく，分析の際にある程度は吸収されるので，むしろ再現率を重視する．）さらに，シラバス特有の言語表現を考慮したフィルタリングを行うことで，精度面での改善も試みる．

An Examination of Methods for Extracting Technical Terms from Syllabus Data
†Fuyuki Yoshikane, Masaaki Ida, Kazuteru Miyazaki, and Takayuki Nozawa
National Institution for Academic Degrees and University Evaluation
‡Hajime Kita
Academic Center for Computing and Media Studies, Kyoto University

## ４　抽出結果

### ４．１　各手法に基づく結果

まず，(a)湯本の手法(b)パターンベースの手法，それぞれ抽出パフォーマンスを測定した．(a)では，次の式により計算される重要度に従って専門用語を抽出する．

$$LR(CN)=(\Pi_{i=1,\ldots,L}\,(LN(N_i)+1)(RN(N_i)+1))^{1/2L}$$

ここで，$N_i$は，抽出候補の名詞列 $CN$ が含む $i$ 番めの単名詞($i \leqq L$)を，$LN(N_i)$（$RN(N_i)$）は，$N_i$ が単名詞と左方（右方）に連接する頻度を表す．本研究では，この手法を実装したシステム[7]を利用した．(b)については，Takeuchi[8]が定義する８つの基本用語品詞パターンにマッチするものを専門用語として抽出した．

(a)(b)それぞれについて，前述の４項目に適用したときの結果と，「科目名」を除いた３項目に適用したときの結果を表２に示す．科目名は，授業内容を特定するスペシフィックな表現に必ずしもなっておらず，専門用語の抽出源として不適切とも考えられるため，科目名を除いた場合のパフォーマンスも検証した．

表２　各手法のパフォーマンス

| | 再現率 | 精度 | F 値 |
|---|---|---|---|
| (a) ４項目 | 0.69 | 0.62 | 0.65 |
| (a) ３項目 | 0.68 | 0.62 | 0.65 |
| (b) ４項目 | 0.47 | 0.37 | 0.42 |
| (b) ３項目 | 0.46 | 0.38 | 0.42 |

品詞パターンだけを手掛かりにすると，一般語のノイズが混ざり精度は低くなりがちである．今回，基本パターンのみを使うことで精度低下の抑制を図ったが，再現率・精度ともに (a)より低い結果となった（バランス指標の F 値[1]も同様）．また「科目名」の有無は抽出パフォーマンスにほとんど影響しないことも確認できた．

### ４．２　手法の組合せ・調整

前に述べたとおり，(a)は名詞列だけを候補語とするため，形容詞＋名詞など，他のパターンの専門用語は抽出できないという問題がある．そこで，本研究では(a)を(b)で補うことで再現率の向上を目指した．(a)と(b)の抽出結果をマージした結果(c)を表３に示す．

(a)を(b)と単純にマージした場合，(a)単独と比較して再現率は 10%程度高いが，一方，精度は20%程度低くなる．この再現率を保ちつつ，精度を改善するために，以下に述べる調整を行った．

表３　２つの手法の組合せ

| | 再現率 | 精度 | F 値 |
|---|---|---|---|
| (c) ４項目 | 0.79 | 0.42 | 0.55 |
| (c) ３項目 | 0.77 | 0.43 | 0.55 |
| (d) ４項目 | 0.78 | 0.52 | 0.63 |
| (d) ３項目 | 0.77 | 0.53 | 0.62 |

i) ８つのパターンのうち，動詞＋名詞など，一般語のノイズが非常に多いパターンは採用しない．
ii)「講義」「学生」など，シラバス 特有の頻出単名詞を含む候補語はストップワードとする．

上記の制約を(c)に加えたのが表中の(d)である．80%程度の再現率を保持しながら，50%程度まで精度を上げることができた．バランス指標である F 値では(a)に若干劣るものの，網羅性重視のアプリケーションにおいては，本研究で示した手法の組合せと調整が有効であると考えられる．

謝辞　本研究を遂行するにあたりご協力いただいた大学評価・学位授与機構「大学評価情報の構造解析と評価プロセスへの応用の研究会」参加者の皆様に謝意を表します．

[1] F 値 = (2・再現率・精度)/(再現率+精度)

### 参考文献

[1] 井田, 宮崎, 芳鐘, 喜多: シラバス XML データベースシステム構築に関する考察, 情報処理学会第 65 回全国大会 2A-6, pp. 4-247-4-248, 2003
[2] 松本, 北内, 山下, 平野, 松田, 浅原: 日本語形態素解析システム『茶筌』version 2.0 使用説明書第 2 版, NAIST Technical Report, NAIST-IS-TR99012, 奈良先端科学技術大学院大学, 1999
[3] 小西: 自動構築型知識に基づく専門用語形成システム, 情報処理学会論文誌, Vol. 30, No. 2, pp. 179-189, 1989
[4] 湯本, 森, 中川: 出現頻度と連接頻度に基づく専門用語抽出, 情報処理学会第 145 回自然言語処理研究会, pp. 111-118, 2001
[5] 辻, 芳鐘: 専門用語として普及しそうな語の自動抽出, 第 51 回日本図書館情報学会研究大会 発表要綱, pp. 105-108, 2003
[6] 影浦: 自動専門用語抽出の諸問題, TP&B フォーラム, 1997
[7] 東京大学中川研究室, 横浜国立大学森研究室:専門用語自動抽出システム, http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/
[8] Takeuchi, Kageura, Koyama, Daille, Romary: Pattern based term extraction using ACABIT system, IEICE Technical Report, NLC2003-20, pp. 31-36, 2003