5D-4

機関リポジトリを活用した部局別英語学術表現リストの作成

田中 省作[1]　冨浦 洋一[2]　宮崎 佳典[3]　小林 雄一郎[4]　徳見 道夫[5]
1 立命館大学文学部　2 九州大学大学院システム情報科学研究院　3 静岡大学情報学部
4 日本学術振興会 / 立命館大学　5 九州大学大学院言語文化研究院

1 はじめに

科学論文の作成や読解に求められる英語（English for Academic Purpose: EAP）には，EGP（English for General Purpose）とよばれる一般的な英語とは異なる表現や語法が数多くある．さらに，EAPは分野によっても大きく異なり，分野に依拠した学術表現リストの作成は重要な課題の一つである．本研究は，近年，多くの研究機関で整備されつつある，自組織の研究者らが執筆した論文・記事などの著作物を電子的に蓄積・公開しているデータベース「機関リポジトリ」（Institutional Repository，以後適宜IRと略記する）を，そのような学術表現リスト作成の言語資源として活用する．IRには多くの場合，組織構造が反映されたかたちで著作物・資料が蓄積されており，当該機関が扱う研究分野に依拠した，従来よりも粒度の細かい英語学術表現リストの効率的な作成が可能となる．

2 機関リポジトリの言語資源としての活用

2.1 利点

IRは当該機関の関係者がかかわった著作物・資料のデータベースである．それらは必然的に，当該機関で推進されている分野ものが集中することになる．そのような資料に基づいた表現リスト等は，所属する研究者らにとって自身の研究環境に特化したものとなり，英語論文執筆の際に大きな助けとなり得る．IRの資料だけではなく，それらの中で参照されているような文献等を外部に求めることで，当該機関の扱うテーマの深くかかわる周縁的な言語資料も期待される．

また，多くのIRが，当該機関の組織構造を反映したかたちで資料を蓄積している．代表的なIRシステムであるDSpaceは，”community” という概念によって資料を束ねている*．このような組織情報を活用すれば，IR内の資料を組織別に表現リスト作成等に活用することもできる．

2.2 問題点

IRの現状には幾つかの問題がある．IRはまだ歴史が浅く，対象となるべき著作物が全て蓄積されているとは限らない．たとえば，IR整備が比較的進んでいるといわれる九州大学においても英語著作物は5,838点（2012年7月時点）である†．IRには登録されていない著作物が，CiNiiのような外部データベースには蓄積・公開されているようなこともある．ハーベスティングなどによるIRの補完も考え得る．本研究では，まず，現状のIRを素直に利用した事例を報告する．

3 部局別学術表現リストの作成法

3.1 方針

本研究で指向する学術表現リストは，英語科学論文を読んだり書いたりするのに有用な英語表現のリストで，[1]が目指すものとほぼ一致する．[1]は，有用な学術表現の特徴として次のような6項目を挙げ，その抽出法を提案している．(1) 高頻度で出現する，(2) 論文に特有の語彙を含む，(3) 短すぎない，(4) 意味的まとまりの列である，(5) 省略表示を含む，(6) 様々な種類の表現と連接する．本研究では，最終的には関連分野の英語識者がチェック・編纂することを念頭に，[1]の抽出法を，スコ

Making a List of Expressions for Academic English of each Section Using Institutional Repository
†Shosaku TANAKA, College of Letters, Ritsumeikan University
‡Yoichi TOMIURA, Graduate School of Information Science and Electrical Engineering, Kyushu University
‡Yoshinori MIYAZAKI, Faculty of Informatics, Shizuoka University
‡Yuichiro KOBAYASHI, Japan Society for the Promotion of Science / Ritsumeikan University
‡Michio TOKUMI, Faculty of Languages and Cultures, Kyushu University

*この community は，大学でいえば概ね学部・研究科の部局に対応づけられていることが多い．
†異なり数で，このなかには学位論文，通常の学術資料とは多少性格の異なる刊行物（たとえば，学位審査報告書や学内学会のニューズレター）なども含まれている．

アを一元値にするなど簡易化し，英語学術表現リストの作成を試みる．

3.2 手順

標的とする機関の IR に含まれる英語著作物を事前に下部組織別に分け，それぞれで次のような処理を施し，組織別の学術表現リストを生成する．

1. **浅い句構造の同定**: 構文解析を施し，句構造を同定する．なお，ここで注目する句構造は [1] に倣い，補文（LC）と入れ子をもたない最小の基本名詞句（NC）である．各語は動詞の分詞形を除き原形表記に統一した後に，名詞・動詞といった浅い品詞レベルで細分化する．冠詞や数字は DT や CD といった具合に記号化している．たとえば，“This paper shows that ...” は，
   $[_{\mathrm{NC}}$ DT paper_NN] show_VV $[_{\mathrm{LC}}$ that_IN ...]
   となる．ここで，x_p は原形が x で品詞が p の語，$[_Y\ \boldsymbol{y}\ ]$ は語列 $\boldsymbol{y}$ が Y 句であることを表す．なお，文構造を成していないものについては，分析対象から除く．

2. **句構造を考慮し n-gram を生成**: 文の前後に文頭・文末を表す特殊記号 @ を付加し，n-gram を生成する．その際，NC,LC をまたぐ場合には，別途それらの語列を一旦 ‘⟨NC⟩’,‘⟨LC⟩’ という 1 記号に置換した列も別途考え，その組み合わせ全ての n-gram を生成する．さきほどの例で $n = 3$ の場合であれば，“@ DT paper_NN”, “DT paper_NN show_VV”, “paper_NN show_VV that_IN” に加え，“@ ⟨NC⟩ show_VV”, “⟨NC⟩ show_VV ⟨LC⟩”, “⟨NC⟩ show_VV that_IN” なども生成されることになる．また，n も $2 \sim 10$ といったように動かし，計数する．

3. **スコアリング**: 生成された各 n-gram $\boldsymbol{x}$ に対して，次のようにスコアを与える．

$$\mathrm{score}(\boldsymbol{x}) = f(\boldsymbol{x}) \cdot \ell(\boldsymbol{x}) \cdot \mathcal{H}_L(\boldsymbol{x}) \cdot \mathcal{H}_R(\boldsymbol{x})$$

   ここで，$f(\boldsymbol{x})$ と $\ell(\boldsymbol{x})$ はそれぞれ $\boldsymbol{x}$ の頻度と語数である．$\mathcal{H}_L, \mathcal{H}_R$ は前後に連接する語のエントロピーで，次のように与える．

$$\mathcal{H}_\alpha(\boldsymbol{x}) = -\sum_y P_\alpha(y \mid \boldsymbol{x}) \log P_\alpha(y \mid \boldsymbol{x})$$

$\boldsymbol{x}$	score($\boldsymbol{x}$)
DT number of ⟨NC⟩	2,680
based on	1,244
when ⟨NC⟩ be	1,198
in order ⟨LC⟩	1,182
by using ⟨NC⟩	938
there be ⟨NC⟩	840
according to ⟨NC⟩	809
where ⟨NC⟩ be	734
such as ⟨NC⟩	674
as follow :	617

表 1: 九大システム情報科学研究院の抽出結果

   $\alpha \in \{L, R\}$ で，$P_L(y \mid \boldsymbol{x})$ は y が n-gram $\boldsymbol{x}$ に前接する割合，$P_R(y \mid \boldsymbol{x})$ は後接する割合である．

4. **フィルタリング**: 頻度が小さい $\boldsymbol{x}$，末尾が DT で終わるような表現としては不自然な $\boldsymbol{x}$，内容語を含まない $\boldsymbol{x}$ などは対象外とする．さらに，$\mathrm{score}(\boldsymbol{x}) > \mathrm{score}(\boldsymbol{x}')$ で，$\boldsymbol{x}$ に完全に包含されるような $\boldsymbol{x}'$ も削除する．

4 実験

九州大学の機関リポジトリ QIR（2012 年 7 月時点）に含まれる英語著作物を学部・研究科に対応する 27 部局に細分化し，形態素数が $2,000 \sim 10,000$ のものを対象に $n = 3 \sim 7$，最低頻度を 5 として上記手法を適用した．なお，NC,LC の同定は，TreeTagger のチャンキング結果と品詞情報を勘案し，行った．たとえば，大学院情報系研究科であるシステム情報科学研究院の場合，229 編の著作物から 1,192 組の表現が得た．上位 10 組を表 1 に示す（品詞情報は省略）．

5 まとめ

本稿では，学術表現リスト作成への IR 活用の基本アイディアと，作成手法，実験結果の一部を示した．2.2 節で述べた問題への対応，リストの定量的な評価などが今後の課題である．

参考文献

[1] 松原茂樹, 酒井祐太, 小澤俊介, 杉木健二: 学術論文からの英語表現集の自動生成, 第 7 回情報プロフェッショナルシンポジウム, pp.41-44 (2010).