機関リポジトリを活用した英語学術表現リストの階層的構築

田中 省作
立命館大学 文学部
sho@lt.ritsumei.ac.jp

冨浦 洋一
九州大学大学院
システム情報科学研究院
tom@inf.kyushu-u.ac.jp

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，分野に依拠した英語学術表現リスト（以後，適宜，表現リストと記す）の作成は重要な課題の一つである [2]．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，そのような英語学術表現リストの言語資源として活用する．多くの場合，IR には，組織構造が反映されたかたちで著作物が蓄積されており，当該機関が扱う研究分野と組織構造に依拠した，粒度の細かい英語学術表現リストの作成が可能となる．

2　言語資源としての機関リポジトリ

2.1　利点

IR は当該機関の関係者がかかわった著作物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・テーマに関するものに集中する．そのような著作物に基づいた表現リストは，当該機関の関係者に関連が深いものが列挙されている可能性が高く，英語論文執筆等の際に大きな助けとなる．また，IR の著作物だけではなく，そこで参照されているような文献を集積することで，当該機関の取り組んでいるテーマの周縁的な言語資料の構築も期待できる．

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[1]は，“community” という概念によって資料を束ねている[2]．このような組織情報を参照しつつ，著作物を活用することで，組織構造を反映した表現リストの作成が容易となる．

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，整備が比較的進んでいるといわれる九州大学の IR でさえ，英語著作物は 2012 年 7 月時点で 5,838 点である[3]．

著作権など他機関との兼ね合いで必然的に登録対象から外されるものもあるが，本来，IR に登録されていても良いような未登録の著作物が，CiNii のような外部データベースでは蓄積・公開されているようなこともある．そこで，本研究プロジェクトの一環として，IR 立ち上げ支援・データ補完を目的に，ハーベスティングなどによる外部データベースとの連携も検討している．本稿では，まず，現状の IR を素直に利用した事例を報告する．

[1] http://www.dspace.org
[2] 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．
[3] 異なり数である．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．

3　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．具体的には，次節で述べるようなスコアなどで優先順位付けされた表現の集合である．[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，抽出した表現リストを最終的に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] の抽出法を簡易化し，英語学術表現リストの作成を試みる．

また，IR から得られる組織の階層性を強く意識する．たとえば，「A 大学 B 学部 C 学科」の場合，「A 大学の表現リスト」「B 学部の表現リスト」「C 学科の表現リスト」の 3 つの表現リストを整備する．ここで，「A 大学の表現リスト」とは，A 大学の IR にある著作物全体から生成されるような表現リストで，上位の表現は比較的どの学部・学科でも使われるようなものであることが予想される．「B 学部の表現リスト」は，IR 内の B 学部の著作物全体から生成されるような表現リストで，「C 学科の表現リスト」も同様である．「A 大学→ B 学部→ C 学科」の方向性は，「A 大学内における EGAP（一般学術目的の英語）から ESAP（特定学術目的の英語）」[2] におおむね対応する．さらに，この階層の最上部に，EGP の粗い近似として日本の中高英語を置く．それに対応する著作物は，中高の英語教科書や参考書などである．つまり，表現リストを「日本の中高英語→『A 大学→ B 学部→ C 学科』」といった具合に階層的に整備する．

このような表現リストの間で，組織の階層関係を考慮し，次のような調整を考える．上部組織の表現リストである一定より上位に列挙される表現は，それよりも下部組織の表現リストでは含めないよう，重複処理を行う．このようにすることで，一つ一つの表現リストがコンパクトで，表現リストの意味付けもより明確化される．その結果，識者がそれらを編纂する際も判断が下しやすく，最終的な表現リストも使いやすくなると考えられる．

3.2　手順

IR に含まれる英語著作物を事前に組織階層別に分け，それぞれで次のように英語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，
   $$[_{\mathrm{NC}}\ \mathrm{DT\ paper\_NN}]\ \mathrm{show\_VV}\ [_{\mathrm{LC}}\ \mathrm{that\_IN}\ ...]$$
   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 句であることを表している．なお，文構造を成していないもの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，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，それぞれで n-gram を生成する．さきほどの例で $n = 3$ の場合，“@ 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．また，n も 2 ∼ 10 といったように動かし，累積的に計数する．

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．

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

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．

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

$\in \{L, R\}$ で，$P_L(y \mid \boldsymbol{x})$ は y が n-gram $\boldsymbol{x}$ に前接する確率，$P_R(y \mid \boldsymbol{x})$ は後接する確率で，次のように与える．

$$P_L(y \mid \boldsymbol{x}) \approx f(y\boldsymbol{x})/f(\boldsymbol{x})$$
$$P_R(y \mid \boldsymbol{x}) \approx f(\boldsymbol{x}y)/f(\boldsymbol{x})$$

4. **フィルタリング**
自組織よりも上部組織の表現リスト s で上位 β_s% までに列挙されている $\boldsymbol{x}$，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，抽出対象となっている $\boldsymbol{x}$ に完全に包含されるような $\boldsymbol{x}'$ も削除する．

4 実験

4.1 データと条件

2012 年 7 月時点の九州大学機関リポジトリ QIR に含まれる英語著作物 5,838 点のうち，形態素数が $2,000 \sim 10,000$ のもの 4,462 点を対象とした．これらを学部・研究科レベルに相当する 27 部局に細分化し，九州大学全体に加え，それぞれの部局の表現リストを作成する．中高英語の著作物には，平成 14 年度版検定済中高英語教科書（中学 7 シリーズ，高校 28 シリーズ）の本文部分を採用した．$n = 3 \sim 7$ で，最低頻度を 5，$\beta_{\text{中高英語}} = 10, \beta_{\text{九州大学}} = 0.1$ とした．なお，NC,LC の同定は，TreeTagger [4] のチャンキング結果と品詞情報を勘案し，行った．

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

$\boldsymbol{x}$	score($\boldsymbol{x}$)	$f(\boldsymbol{x})$
⟨NC⟩ don't	47025	762
there be ⟨NC⟩	33635	729
when ⟨NC⟩ be	22378	356
DT lot of ⟨NC⟩	14478	207
I don't	13953	327
go to ⟨NC⟩	8176	266
⟨LC⟩ see ⟨NC⟩	8066	163
say ⟨LC⟩ ⟨NC⟩	7880	217
think ⟨NC⟩ be	6203	171
look at ⟨NC⟩	5925	193
tell ⟨NC⟩ ⟨LC⟩	4681	123
more than ⟨NC⟩	4513	132
give ⟨NC⟩ ⟨NC⟩	4217	165
out of ⟨NC⟩	3840	148
take ⟨NC⟩ to	3777	96

表 1: 中高英語教科書の抽出結果 （上位 15 組）

4.2 抽出例

中高英語教科書では 680 組の表現が得られた．その上位 20 組の表現を表 1 に示す（品詞情報は省略する）．九州大学全体では 10,755 組の表現が得られた（表 2）．学部・研究科については，紙面の都合上，情報科学系の独立研究科であるシステム情報科学研究院と農学部・研究院の結果の一部を示す．システム情報科学研究院は，229 点の著作物から 1,192 組の表現が得られた（表 3）．農学部・研究院は，808 点の著作物から 3,651 組の表現が得られた（表 4）．

九州大学全体の表現リストでは，比較的論文の執筆マニュアルなどで挙げられるような定型表現がみられた．特に，理系学部・研究院の著作物数が多く，その影響が少なくないと考えられる．2 つの学部・研究院の表現リストは，具体的な語の表現パタンが多くを占める．また，スコアにエントロピーを直に掛け込んだこともあり，複数の定型的な記号列が混合しているものが散見され，⟨NC⟩ に付随する前置詞などが含まれるようなものが少なくない．

5 まとめ

本稿では，言語資源としての IR の活用，英語学術表現リストの階層的な構築に対する基本アイディア，実験結果の一部を示した．2.2 節で述べ

$\boldsymbol{x}$	score($\boldsymbol{x}$)	$f(\boldsymbol{x})$
DT number of	806805	6226
based on	566340	6281
there be ⟨NC⟩	505452	6497
when ⟨NC⟩ be	426625	4120
where ⟨NC⟩ be	400443	4371
in order ⟨LC⟩	349408	3620
such as	348494	4569
due ⟨LC⟩ ⟨NC⟩	272857	3603
according to ⟨NC⟩	252775	3345
where ⟨NC⟩ be ⟨NC⟩	246711	2916
by using ⟨NC⟩	239129	2762
shown in ⟨NC⟩	226665	5559
as follow :	200795	2106
⟨NC⟩ be shown in ⟨NC⟩	195657	2390
with respect to ⟨NC⟩	168394	1697
on DT other hand	121365	2508
used in ⟨NC⟩	118555	2189
denote ⟨NC⟩ of	111933	1318
consist of ⟨NC⟩	103511	1837
using of ⟨NC⟩	101584	1304

表 2: 九州大学全体の抽出結果（上位 20 組）

$\boldsymbol{x}$	score($\boldsymbol{x}$)	$f(\boldsymbol{x})$
reduce ⟨NC⟩ of	5600	141
we assume ⟨LC⟩ ⟨NC⟩	4205	108
represent ⟨NC⟩ of	2401	83
gure CD show ⟨NC⟩	2215	100
mean ⟨LC⟩ ⟨NC⟩	1996	102
correspond to ⟨NC⟩	1850	82
which ⟨NC⟩ consist of ⟨NC⟩	1807	58
it be necessary to	1758	59
focus on ⟨NC⟩	1616	88
at least	1562	53
proceeding of ⟨NC⟩ on	1551	63
if there be ⟨NC⟩	1462	53
for low power	1409	54
apply ⟨NC⟩ ⟨LC⟩	1391	62
evaluate ⟨NC⟩ of	1345	66
make ⟨NC⟩ of	1308	62
solve DT problem	1289	69
for ⟨NC⟩ using ⟨NC⟩	1272	51
consider ⟨NC⟩ of ⟨NC⟩	1263	53
depend on ⟨NC⟩ of ⟨NC⟩	1192	48

表 3: システム情報科学研究院の抽出結果（上位 20 組）

$\boldsymbol{x}$	score($\boldsymbol{x}$)	$f(\boldsymbol{x})$
found in ⟨NC⟩	10255	347
most of ⟨NC⟩	9576	226
observed in ⟨NC⟩	7846	262
derived from ⟨NC⟩	7529	232
collected from ⟨NC⟩	7504	284
according to ⟨NC⟩ of ⟨NC⟩	6782	141
covered with ⟨NC⟩	6709	247
similar to ⟨NC⟩ of	6628	123
followed by ⟨NC⟩	6585	243
report ⟨LC⟩ ⟨NC⟩	6542	223
with DT pair of ⟨NC⟩	6354	106
longer than ⟨NC⟩	6245	260
change in ⟨NC⟩	5408	170
affected by ⟨NC⟩	5364	204
examine ⟨NC⟩ of	5361	139
by using ⟨NC⟩	5324	108
divided into CD	5016	139
more than ⟨NC⟩ of ⟨NC⟩	5007	91
Euler's ⟨NC⟩	4987	150
seem to be	4947	150

表 4: 農学部・研究院の抽出結果（上位 20 組）

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).

[2] 田地野彰, 水光雅則: 大学英語教育への提言 -カリキュラム開発へのシステムアプローチ-, これからの大学英語教育（竹蓋幸生, 水光雅則編）, 岩波書店, pp.1-46 (2005).