英語論文からの表現集の自動生成

酒井 佑太† 小澤 俊介† 杉木 健二† 松原 茂樹‡
†名古屋大学大学院情報科学研究科 ‡名古屋大学情報基盤センター

1 はじめに

研究者が自身の研究を世界に主張するためには，英語で論文を作成することが必須である．英語ネイティブでない研究者にとって，英語論文の作成は大きな労力を伴う作業である．そのため研究者は，英語論文を作成する際，訳語や言い回し，書き方を調べるために，辞書や英語論文表現集，検索エンジンなどをよく利用する．特に，英語論文表現集は，掲載されている表現をそのまま利用できるため有用である．

しかし，現在市販されている表現集 [1] は，人手で作成されているため，表現の数が限られてしまう．また，用例の数も十分とは言えない．さらに，論文で用いられる表現は専門分野ごとに異なるため，表現集の対象分野以外の研究者にとって，利用価値は必ずしも高いとは言えない．

そこで本論文では，英語論文から論文作成に有用な表現を自動的に獲得する手法を提案する．この有用な表現を以下では**フレーズ**と呼ぶ．近年では，多くの学術論文が電子化され，Web 上で公開されている [2]．これらの論文から，頻出するフレーズを獲得することにより，既存の表現集の問題の解決を試みる．

獲得するフレーズは，イディオムや慣用句，連語も含む．また，フレーズの構成単位は基本句[1]とする．本手法では，論文に出現する単語列を基本句単位で抽出し，フレーズとしての適切さを判定する．フレーズとしての適切さの判定には，まず，論文中での頻度，長さ，連接する基本句の種類数の 3 つの統計情報を利用する．まだフレーズとして不適切な基本句列が含まれるため，文法に基づく除去ルールにより不適切な基本句列を取り除く．

フレーズ獲得実験をした．実験には，国際会議 ACL の 8 年分の論文を使用した．実験の結果，提案手法によるフレーズ獲得の精度は 76.2%に達し，本手法の利用可能性を確認した．また，フレーズの利用例として，フレーズ検索システムを作成し，獲得フレーズの有効性を示した．

2 フレーズの特徴

フレーズには，イディオムや慣用句，連語も含む．フレーズの例として，“In this paper, we propose 〜” や，“In addition to,” などが挙げられる．

[1]基本句とは入れ子を持たない最少の句のことである．例えば，“In this paper, we propose a new method.” という文を基本句列に変換すると，“[PP In] [NP this paper] , [NP we] [VP propose] [NP a new method] .” となる．ここで，[ ] で囲まれた部分が基本句である．PP は前置詞句，NP は名詞句，VP は動詞句を表している．

表現集に掲載するのに適切なフレーズの特徴を明確にするために，書籍に掲載されている表現を分析した．分析データには，表現集として大いに利用されている書籍 [1] を用いた．文献 [1] の全 1,119 表現を調査したところ，以下の特徴が見られた．

2.1 構成単位

表現集には意味的にまとまりのある単位で表現が記載されている．例として，“in the early part of the paper” や，“As a beginning, we will examine” が挙げられる．このことから，フレーズの構成単位を単語ではなく，句とすることが適切であると考えられる．書籍中の各表現が基本句から構成されているか否かを調査した．書籍中の全表現 1,119 のうち，96.7%にあたる 1,082 が基本句で構成おり，基本句をフレーズの構成単位とすることが妥当である．

2.2 句記号の存在

表現全体の 76.8%にあたる 859 表現には，省略表示が含まれていた．省略表示とは，“With the exception of 〜” の “〜” であり，ここでは名詞もしくは名詞句が入ることを示している．“〜” を**句記号**と呼ぶ．本研究では，句記号を含むフレーズも獲得対象とする．文献 [1] を参考に，付与を行う句記号の種類を，以下の 2 種類に定めた．

- <**NP**>：名詞句が挿入されることを示す
- <**CL**>：節が挿入されることを示す

2.3 統計的特徴

文献 [1] で現れる表現には以下の統計的な特徴が見られた．

- **論文に頻出する**
  書籍には，英語論文に頻繁に出現する表現が掲載されている．
- **短すぎない**
  表現の長さを調査したところ，表現に含まれる基本句数が 2 以下の表現は全体の 6.9%に過ぎなかった．
- **連接する基本句の種類が多い**
  フレーズは様々な文脈において利用されるため，フレーズには様々な種類の基本句が連接すると考えられる．基本句列 “in spite of” と，それより 1 基本句短い基本句列 “in spite” に連接する基本句の種

類数をみると，実験で用いた論文データでは，“in spite” は 36 回出現していた．36 回全ての出現において基本句 “of” が連接していた．一方，表現集に掲載されている基本句列 “in spite of” は 36 回出現しており，それぞれ異なる基本句が連接していた．このようにフレーズには，連接する基本句の種類が多いという傾向があるといえる．

2.4 構文的特徴

書籍中の表現には構文的な特徴が存在するた．例えば，“stem from” は書籍に掲載されているが，“stem in” や “stem with” は掲載されていない．これは，“stem from” が句動詞であり，この並びに特別な意味があるためであると考えられる．

また，書籍には，“In other words” といった様々な文書で利用できる表現だけでなく，“The purpose of this paper is to” や “The result of the experiment was that” といった論文特有の表現が掲載されていた．これより，フレーズ獲得において，論文特有であるか否かが手がかりの 1 つとなり得る．

3 フレーズ獲得手法

前章に示した特徴をもとに，フレーズを獲得する．まず，対象とする論文から基本句列を抽出する．このとき，基本句列内の名詞句を <NP> に置き換えたものも加える．次に，抽出した基本句列の中から統計的な特徴を満たす基本句列を獲得する．その後，文法情報をもとに構文的な特徴を満たさない基本句列を除去する．最後に，基本句列に <CL> を付与する．以下では，句記号の付与，統計的特徴の判定，及び，構文的特徴の判定について説明する．

3.1 句記号の付与

本手法では，獲得するフレーズとして 2.2 節に述べた 2 種類の句記号，すなわち <NP> と <CL> を含む基本句列も含める．

- **<NP> を含むフレーズの獲得**
  まず，候補として獲得した基本句列のうち，名詞句を含む基本句列に対して，名詞句を <NP> に置き換える．基本句列内に複数の名詞句が存在する場合，すべての組み合わせに対して <NP> に置き換える．例えば，基本句列中に 3 つの名詞句が存在した場合，7 通りの置き換えた基本句列を生成する．ただし，当てはまる名詞句が 1 種類のみ存在する場合，<NP> に置き換えない．

- **<CL> を含むフレーズの獲得**
  以下の 3.2 節，3.3 節の処理により得られた基本句列のうち，末尾基本句が従属節を伴う従属接続詞である基本句列の末尾に <CL> を付与する．

3.2 統計的特徴の判定

フレーズの候補である基本句列に対して，2.3 節に示した統計的な特徴を満たすかどうかを判定する．以下の，統計的な特徴を満たす基本句列のみを獲得する．

まず前処理として，以下の基本句列を除去する．

- 相対文書頻度が 1%未満
- 基本句の数が 2 未満

統計的特徴を判定するために，池野らの専門用語獲得の手法 [4] を利用する．すなわち，基本句がどの程度統計的な特徴を満たすかをスコア関数 $Lscore$，$Rscore$ を用いて計算する．池野らの文献 [4] を参考に，スコア関数を以下のように設定した：

$$Lscore(E) = \log(tf(E)) \times length(E) \times H_l(E)$$

$$Rscore(E) = \log(tf(E)) \times length(E) \times H_r(E)$$

ここで，E は基本句列を示す．また，$length(E)$ は E に含まれる基本句の数であり，$tf(E)$ は E の対象論文内における出現頻度である．$H_l(E)$ と $H_r(E)$ は，それぞれ右側と左側に連接する基本句の確率分布のエントロピーである．連接する基本句の種類が多く，それらの出現頻度が均一である場合，高い値となる．$H_l(E)$ と $H_r(E)$ は，以下の式により計算される．

$$H_l(E) = -\sum_i Pl_i(E) \log Pl_i(E)$$

$$H_r(E) = -\sum_i Pr_i(E) \log Pr_i(E)$$

ここで，Pl_i，Pr_i はぞれぞれある基本句 Xi が E の左，右に連接する確率であり，以下の式により計算される．

$$Pl_i(E) = P(X_iE|E) = \frac{P(X_iE)}{P(E)} \approx \frac{tf(X_iE)}{tf(E)}$$

$$Pr_i(E) = P(EX_i|E) = \frac{P(EX_i)}{P(E)} \approx \frac{tf(EX_i)}{tf(E)}$$

$Lscore$，$Rscore$ は，それぞれ第 1 項が長さ，第 2 項が出現頻度，第 3 項が連接する基本句の種類数に相当する．基本句列 E が，より長く，より頻出し，より連接する基本句の種類数が多いとき高い値となる．つまり，2.3 章に述べた統計的特徴を満たすほど，これらのスコア関数の値は高くなる．

ある基本句列 E が以下の 2 つの式を満たす場合に，E を獲得する．

$$Lscore(E) > Lscore(XE)$$

$$Rscore(E) > Rscore(EX)$$

ここで，X は 1 つの基本句であり，XE は E より左に 1 つ長い基本句列である．同様に，EX は E より右に 1 つ長い基本句列である．この 2 つの式により，ある基本句列 E が，自身よりもスコアの高い，左（前方）または右（後方）に長い基本句列 XE, EX に包含される場合，基本句列 E をフレーズでないと判定し，取り除く．

表 1: 文法に基づく除去ルールの一部

文法パターン	例
名詞句 of <NP>	the threshold of <NP>
代名詞 動詞句	we extracted <NP>
<NP> be 動詞 名詞句	<NP> are words

表 2: 実験データの規模

論文数	1,232
文数	204,788
基本句数	2,683,773
単語数	5,516,612

3.3 構文的特徴の判定

フレーズは，2.4 節で述べた構文的特徴を満たす必要がある．しかし，基本句列は，様々な構文的特徴を持つため，画一的な判定は困難である．そのため，構文的特徴を満たさない基本句列の特徴をパターン化する．このパターンを適用して不適切な基本句列を除去する．

除去ルールを作成するために，対象データ中の基本句列から，ランダムに 809 個抽出し，フレーズとしての適切さを人手により判定した．これらのフレーズのうち，誤り率が高く，かつ出現数の多いものから，品詞か基本句の種類に基づく 25 パターンの除去ルールを作成した．作成した除去ルールの一部を表 1 に示す．

このパターンによりイディオムなども除去されてしまう．そのため，既存の辞書に掲載されているフレーズは除去しない．

3.4 関連研究

関連研究としてイディオムに着目した研究が挙げられる．イディオムは，本研究で獲得するフレーズの一部に該当する．これまでにも，イディオムに着目した研究がいくつか行われてきた．Liu はコーパス中に出現するイディオムの頻度を利用することにより，よく利用されるイディオムを獲得している [5]．しかし，獲得できるイディオムは既存の辞書に掲載されているものに限られ，新たな表現の獲得はできない

これに対し，イディオムを自動検出する研究が行われている [6, 7]．これらの研究では，複合名詞と動詞句を抽出し，それらがイディオムであるかを判定している．また，Widdows らは「A and/or B」というイディオムの自動獲得をしている [8]．しかし，獲得されるイディオムは非常に限定的な表現であるため，本研究の目的とするフレーズ獲得には利用ができない．

4 評価実験

提案手法におけるフレーズ獲得性能の有効性を確認するために，評価実験を実施した．

4.1 実験の概要

実験には，国際会議 ACL の 2001 年から 2008 年までの 1,232 論文を利用した．これらの論文から提案手法を用いてフレーズを獲得した．データ規模を表 2 に示す．

表 3: 実験結果

	精度 (%)	再現率 (%)	F 値
基本句列	24.9 (249/1000)	100.0 (249/249)	39.9
提案手法	51.8 (99/191)	39.8 (99/249)	45.0

論文の PDF 形式から TXT 形式への変換には pdftotext[2]を使用した．また，基本句のチャンキングには，JTextPro[3]を使用した．

評価は，対象データからランダムに選択した 1000 個の基本句列を利用し，英語論文の作成に精通した評価者 1 名によって，フレーズとしての適切性を判定した．この評価を正解セットとし，提案手法の精度，再現率，F 値を評価した．精度と再現率を以下の式により算出した．

$$\text{精度} = \frac{\text{獲得された正解フレーズ数}}{\text{獲得したフレーズの数}} \times 100$$

$$\text{再現率} = \frac{\text{獲得された正解フレーズ数}}{\text{全正解フレーズの数}} \times 100$$

4.2 実験結果

実験の結果を表 3 に示す．ランダムに選択した基本句列 1000 個のうち，フレーズとして適切なものは 249 個存在した．提案手法は精度 51.8%，再現率 39.8%を達成した．基本句列のランダム抽出よりも再現率は下がるものの，精度，F 値ともに向上した．これにより本手法の有効性が確認された．

4.3 考察

獲得したフレーズの誤り原因について調査した．誤りであったフレーズ 92 の 15.3%は，<NP> で始まり次に前置詞句と動名詞，若しくはカンマが続くものであった．例を以下に示す．

- <NP> of selecting <NP>
- <NP>, we discuss <NP>

1 つ目の例は “the problem of selecting source domain data” といった形で使われていた．よく使われる表現ではあるが，構成要素のほとんどが <NP> であるため，具体性が無く，どう使うのかが分からないフレーズであ

[2]xpdf: http://www.foolabs.com/xpdf/
[3]JTextPro: http://jtextpro.sourceforge.net/

る．フレーズ中の <NP> の割合も考慮した構文ルールを作成する必要があると言える．

2つ目の例は，“In this paper, we discuss <NP>” や，“In the following, we discuss <NP>” といった形で使われていた．左に1基本句長い基本句列 “this paper, we discuss <NP>” や “the following, we discuss <NP>” には，ほとんど基本句 “In” しか連接しないため，左に連接する基本句の種類数，頻度の観点において，統計的特徴を満たしている．しかし，文法的にまとまっていないためフレーズとして不適切である．文法に基づく除去ルールを増強することで対応できると考えられる．

5 フレーズ検索システム SCOPE

英語論文作成を支援するために，本手法を用いて獲得したフレーズを利用し，フレーズ検索システム SCOPE(System for Consulting Phrasal Expressions) を作成した．検索例を図1に示す．本システムでは，頻出するフレーズを提示し，かつ，実際の論文からの大量の用例を提示することが可能である．また，我々が提案した，論文の章構成に基づくフレーズの分類手法 [9] を用いて，「序論」や「評価実験」といった特定の章に頻出するフレーズの提示も可能である．

本システムの特徴は以下の通りである．

- 入力した英単語を含むフレーズの検索ができる
- 特定の章に頻出するフレーズの検索ができる
- フレーズの利用頻度，用例が確認できる

利用例として，ユーザが実験の結果を記述する場合を考える．ユーザは「結果」という単語を手掛かりに，「評価実験」の章に頻出する “result” を含むフレーズを検索したとする．その検索結果を図1に示す．ユーザは図1の結果から，実験でよく使われているフレーズとして，“Table <digit> shows the results of <NP>” や “we present the results of <NP>” などのフレーズを利用できることがわかる．<digit> は数字が入ることを示している．さらに，出現頻度などの統計情報から，どの程度論文に利用されているかを確認したり，実際に用例を確認したりすることにより，ユーザにとって適切な表現を確認できる．

SCOPE は下記のサイトで実験的に運用されている．
http://scope.itc.nagoya-u.ac.jp/

6 おわりに

本論文では，表現集自動作成を目的とし，英語論文からの表現獲得手法を提案した．本手法では，論文中の基本句列をフレーズの候補とし，統計的特徴と，構文的特徴を用いて，フレーズとして適切な基本句列を獲得した．フレーズの獲得実験では，精度 76.2%を達成し，提案手法の利用可能性を確認した．また，獲得したフレーズの利用例として，フレーズ検索システム SCOPE を作成し，獲得したフレーズの利用可能性を確認した．

フレーズ	出現回数	出現論文数	LRスコア	長さ	スコア
The results are shown in Table <digit>.	17	16	5.89	5	4.56
Table <digit> shows the results of <NP>	40	30	6.50	4	5.11
we present the results of <NP>	12	12	3.35	4	3.44
<NP> is the result of <NP>	22	22	4.07	3	3.40
<NP> report results for <NP>	16	15	3.22	3	3.05
<NP> shows the result of <NP>	18	12	3.60	3	3.18
<NP> shows the results for <NP>	19	16	3.72	3	3.23
<NP> shows the results of <NP>	60	48	5.33	3	4.50
<NP> summarizes the results of <NP>	13	12	3.14	3	2.82
Figure <digit> shows the results	13	12	3.77	3	2.82
Table <digit> shows the results	93	73	7.09	3	4.98

図 1: フレーズ検索システム SCOPE の “result” に対する検索結果

参考文献

[1] 崎村耕二：英語論文によく使う表現，創元社，1991.

[2] S. Lawrence, C.C. Giles, and K. Bollacker: Digital Libraries and Autonomous Citation Indexing, *Computer*, Vol.32, no.6, pp. 67-71, 1999.

[3] 国語辞典，集英社，1993.

[4] 池野篤史，浜口佳孝，山本英子：Web 文書集合からの専門用語獲得，情報処理学会論文誌，Vol.47, No.6，pp. 1717-1727，2006.

[5] D. Liu: The Most Frequently Used Spoken American English Idioms: A Corpus Analysis and Its Implications, *TESOL Quarterly*, Vol. 38, No. 4, pp. 671-700, 2003.

[6] D. Lin: Automatic Identification of Non-compositinal Phrases, In *Proc. of the 37th Annual Meeting of Association for Computational Linguistics*, pp. 317-324, 1999.

[7] P. Cook, A. Fazly, and S. Stevenson: Pulling their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context, In *Proc. of the LREC Workshop Towards a Shared Task for Multiword Expressions*, pp. 19-22, 2008.

[8] D Widdows and B. Dorow: Automatic Extraction of Idioms using Graph Analysis and Asymmetric Lexicosyntactic Patterns, In *Proc. of the ACL2005 Workshop on Deep Lexical Acquisition*, pp. 48-56, 2005.

[9] 酒井佑太，杉木健二，松原茂樹：英語論文に頻出する表現の獲得と分類，情報処理学会第 71 回全国大会講演論文集 (2)，pp. 275-276，2009.