学術論文からの英語表現集の自動生成:

○松原茂樹[1)], 酒井佑太[1)], 小澤俊介[1)], 杉木健二[1)]
名古屋大学大学院情報科学研究科[1))]
〒464-8603 名古屋市千種区不老町
Tel: 052-789-4387 FAX: 052-789-4387
E-mail: matubara@nagoya-u.jp

Automatic extraction of English useful expressions from scientific papers:

MATSUBARA Shigeki[1)], SAKAI Yuta[1)], KOZAWA Shunsuke[1)], SUGIKI Kenji[1)]
Graduate School of Information Science, Nagoya University[1)]
Furo-cho, Chikusa-ku, Nagoya 464-8603 Japan
Phone: +81-52-789-4387 Fax: +81-52-789-4387
E-mail: matubara@nagoya-u.jp

【発表概要】

英語論文のライティング支援を目的に、英語表現集を自動生成する方法について述べる。学術機関リポジトリやオンライン論文集など、学術論文がインターネット上で流通しており、論文データを再利用するための環境が整ってきている。論文データから論文作成に有用な表現を獲得する手法を開発し、その性能を評価した。本手法により、英語表現を大規模に収集できる。対象となる論文データを選別することにより、分野に特化した表現集の生成も可能となる。自然言語処理分野の学術論文から表現を抽出し、英語表現検索システム SCOPE を構築した。SCOPE では、キーワード入力により論文に頻出する表現を検索できる。英語表現を使用する際には、実例を参照することにより、その表現の出現頻度や文脈を考慮することができる。

【キーワード】

情報抽出, 情報検索, ライティング支援, 学術コンテンツ, テキスト処理

1. はじめに

研究者が自らの研究成果を発信する上で、論文を英語で執筆することは重要である。しかし、英語ネイティブでない研究者にとって、英語論文の作成は必ずしも容易ではなく、作成にあたっては、対訳辞書や英文検索エンジン(例えば[1])、英語表現集などを駆使することになる。

この中でも英語表現集は、掲載されている表現の再利用性が高く、英文作成に有用な情報資源である。しかしながら、英語表現集を利用する場面では、

- 掲載されている英語表現の数が少なく、それを使用して英文を作成できることは必ずしも多くない。
- 各表現に付随する用例の数が少なく、実際にその表現を使用するのが適切かどうかの判断が難しい。

といった不都合が生じることが多い。

そこで本論文では、学術論文データから英語表現集を自動的に生成する方法について述べる。学術機関リポジトリやオンライン論文集など、大量の学術論文がインターネット上に流通しており、論

文データを再利用するための環境が整ってきている。本研究ではこのような状況を背景に、論文データから英語論文の作成に有用な表現を獲得する手法を開発し、その獲得性能を評価した。本手法により、英語表現を大規模に収集することができ、表現の獲得対象となる論文データを適切に選択することにより、分野に特化した表現集の生成も可能となる。

自然言語処理分野の学術論文から抽出した表現をもとに、英語表現検索システム SCOPE を構築した。SCOPE では、キーワード入力により論文に頻出する表現を検索できる。英語表現の使用に際しては、実例を参照することにより、その表現が出現する頻度や文脈を考慮できる。

2. 論文作成に有用な英語表現の特徴

本稿では、論文の作成に有用な表現を**フレーズ**と呼ぶ。英語表現集に掲載されている表現のことである。

フレーズは単語列で構成される。例えば、以下の単語列はフレーズである。

- in this paper, we proposes ...
- with the exception of ...
- if we assume that …

論文作成者は、このようなフレーズを活用することにより、英文作成のコストを軽減できる。例えば、"if we assume that …" というフレーズを使って、ある仮定から導かれる帰結を示すことができる。

本研究では、フレーズを論文データから自動獲得することを目指す。すなわち、"In this paper, we propose a novel randomized language model. "という論文中の文から、例えば"In this paper, we proposes ..."をフレーズとして取り出す。その際、問題となるのは、論文に出現するある単語列がフレーズであるか否かをどう判断するかという点である。

これを整理するために、フレーズとなる英語表現の特徴について検討した。具体的には、市販の英語表現集[2]を使用し、そこに掲載されている英語表現を参照した。以下では、本研究で考慮した6つの特徴について論じる。

まず、フレーズとなる英語表現には、

(A) 高頻度で出現する,

という特徴がある。そもそも、稀にしか出現しない表現は、それが使用される機会も少なく有用性が低い。

その一方で、たとえ頻出したとしても、"This is a ..." のように、論文に限らずあらゆるタイプの文章に出現する語彙からなる表現、あるいは、論文に特有であっても、"paper" のようにあまりに短い表現は、表現集に掲載すべき有用性を備えていない。すなわち、

(B) 論文に特有の語彙を含む,

(C) 短すぎない,

もまたフレーズの特徴とみなせる。

さらに、フレーズの構成単位として単語を採用すべきかどうか、という点を考慮する必要がある。例えば、

"in the early part of the paper",

はフレーズとみなせる一方で、

"in the early part of the",

は、フレーズとはいえない。"the" は、"the paper"という意味的なまとまりの構成素であり、そのようなまとまりとは無関係に、他の単語と連結した表現は、利用に供さないためである。すなわち、

(D) 意味的まとまりの列である,

こともまたフレーズの特徴といえる。

その他の特徴として、単語列の一部を汎化した表現の存在が挙げられる。これは、"With the exception of ..."の"..."に相当する部分であり、この例では名詞句が入る。このような汎化は、フレーズとしての有用性を高めることになる。よって、

(E) 省略表示を含む,

もフレーズの特徴である。ただし、表現における"..."の部分にはあらゆる単語列が挿入されうるわけでなく、名詞句などの

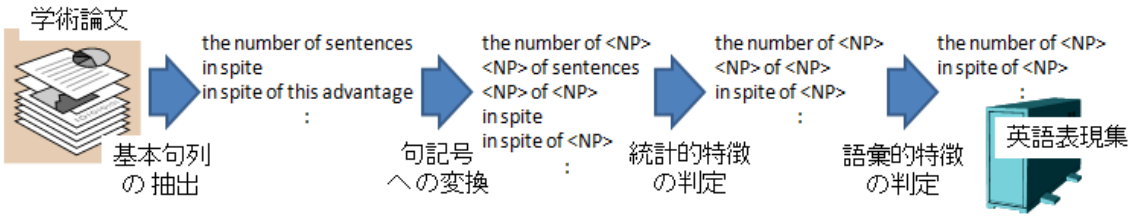

図 1. フレーズ獲得の流れ

In/ this paper/,/ we/ propose/ a new method/ for/ word translation disambiguation/ using/ a bootstrapping technique/ we/ have developed/./

図 2. 基本句への分割例

文法役割によってタイプ分類されるのが一般的であり、そのような役割を表す記号は**句記号**と呼ばれる。

残された視点として、フレーズの両端に位置する単語に備わる性質がある。本研究では、フレーズが接続する表現の多様性という視点に着目する。すなわち、**(F) 様々な種類の表現と連接する,** というものである。例えば、"in spite of" はフレーズといえるが、"in spite"はそうとはいえない。"in spite"の右に連接する表現の多様性は小さく(ほとんどが "of")、“in spite of”では大きい。

本研究では、これら(A)～(F)の特徴を考慮し、フレーズ獲得手法を開発する。

3. フレーズの自動獲得と評価

3.1 フレーズ獲得の流れ

フレーズ獲得の流れを図 1 に示す。以下に、図 1 の各処理について説明する。

3.1.1 基本句列の抽出

特徴(D)の観点から、フレーズの構成単位として基本句を採用する。**基本句**とは、入れ子をもたない最少の句のことである。この処理では、論文データ中のすべての基本句列を抽出する。図 2 に単語列の基本句への分割を示す。

3.1.2 句記号の付与

特徴(E)に従い、句記号"<NP> (名詞句)"と"<CL>(節)"を含む基本句列を以下の通り生成し、フレーズ候補とする。

- <NP>を含むフレーズの生成
  基本句列に名詞句が含まれているとき、それを<NP>に置き換えたものも基本句列の候補とする。
- <CL>を含むフレーズの生成
  基本句列の末尾の基本句が補文標識(that, which, so that など)であれば、末尾に<CL>を付ける。

3.1.3 統計的特徴の判定

フレーズ候補である基本句列に対して、(A),(C),(F)の特徴を満たすかどうかを判定する。池野らの手法[3] を参考に、スコア関数 L_s, R_s を以下の通り設定した。

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

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

ここで、E は基本句列を示す。$tf(E)$ は E の全論文での出現頻度、$length(E)$ は E に含まれる基本句数である。$H_l(E)$ 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り、以下で計算する。これは、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。

$$H_l(E) = -\sum_i P_{li}(E) \log P_{li}(E)$$

$$H_r(E) = -\sum_i P_{ri}(E) \log P_{ri}(E)$$

ここで、$P_{li}(E)$, $P_{ri}(E)$ は、ある基本句 X_i が E の左、右にそれぞれ連接する確率であり、以下の式により計算される。

$$P_{li}(E) = P(X_i E \mid E) \approx \frac{tf(X_i E)}{tf(E)}$$

$$P_{ri}(E) = P(E X_i \mid E) \approx \frac{tf(E X_i)}{tf(E)}$$

Ls, Rsは、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し、それぞれ特徴(C), (A), (F)に対応する。

ある基本句列Eが、Eより左に 1 つ長い基本句列XE、右に長いEXとの間で以下の両式を満たすとき、Eをフレーズとして獲得する。

$$Ls(E) > Ls(XE),\quad Rs(E) > Rs(EX)$$

すなわち、ある基本句列が、自身よりもスコアの高い、左または右に長い基本句列に包含される場合、その基本句列はフレーズでないとみなし、候補から除く。

3.1.4 語彙的特徴の判定

フレーズに適さない基本句列をパターン化し、これを適用して不要な基本句列を除去する。品詞または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。なお、適切な基本句列が除去されることを避けるために、既存の辞書に掲載されている基本句列、及び、論文特有の名詞や動詞を含む基本句列は除去しない(特徴(B)を反映)。

論文に特有な単語の獲得には、新聞などの一般的な文書と学術論文との間の出現頻度の差を利用する。本手法では、ある単語wが以下の条件を満たすとき、wを論文特有な単語とする。

- 論文での相対文書頻度がα%以上
- 対象論文での相対出現頻度が、一般文書での相対頻度のk倍以上

3.2 評価実験

フレーズ獲得実験を実施した。フレーズ獲得対象として、国際会議 ACL2001～2008 の 1,232 論文(204,788 文、5,516,612 単語)を利用した。

3.1.4 の辞書には、英辞郎[4]を使用した。論文特有の単語を獲得するための比較文書として、Wall Street Journal を使用した。閾値αを 1、kは名詞で 4、動詞で 2 と定め、論文特有語として名詞 1119 単語、動詞 226 単語を獲得した。

表 3. 実験結果

	精度(%)	再現率(%)	F 値
基本句列	16.20	100.00	27.88
提案手法	57.53	51.85	54.55

表 4. 自動獲得されたフレーズ（一部）

<NP> is set to <NP>
<NP> is shown in Figure <digit>.
As a result,
adding <NP> to <NP>
<NP> divided by the total number of <NP>
<NP> is consistent with <NP>
the results obtained with <NP>
extracting <NP> from <NP>

PDF 形式の論文からの文抽出には pdftotext[5]を使用し、基本句チャンキングには、JTextPro[6]を使用した。

評価は、ランダムに選択した 500 個の基本句列を利用し、英文作成に精通した評価者 1 名が、フレーズとしての適切性を判定した。この結果を正解セットとし、提案手法の精度、再現率、及び、その調和平均である F 値を測定し、評価した。

結果を表 3 に示す。ランダムに選択した基本句列 500 個のうち、フレーズとして適切なものは 81 個存在した。基本句列を単純に用いるよりも高い達成しており、本手法の有効性を確認した。表 4 に獲得フレーズの一部を示す。

4. フレーズ検索システム SCOPE

英語フレーズ検索システム SCOPE (System for Consulting Phrasal Expressions)を、Perl で実装した。

4.1 SCOPE の構成

SCOPE は、英語フレーズとその用例文からなる英語表現集、及び、検索インタフェースから構成される。

英語表現集を自動生成するために、自然言語処理に関する国際会議 ACL 2001～2008, COLING 2000～2008

図 5. SCOPE の検索画面

図 6. クエリに対する検索結果

図 7. 選択したフレーズの用例文

の論文集のすべての学術論文を使用した。提案手法を用いて 7,769 のフレーズを獲得し、データベースに格納した。

一方、検索インタフェースでは、ユーザのキーワード入力に対して、該当するフレーズを出力する。SCOPE の検索画面を図 5 に示す。SCOPE の特徴として以下の 2 点が挙げられる。

1. キーワード入力に加え、フレーズが出現する論文部分を指定できる。部分として、「序論」「関連研究」「提案手法」「評価実験」「結論」を定めた。
2. 検索されたフレーズの頻度と例文を参照できる。例文は、フレーズの使われ方を知る上で有用である。これは、フレーズが実データから獲得されたために実現できた機能である。

4.2 SCOPE の利用例

実験結果について英語で言及する場合には、「結果」というキーワードを手がかりに、"result"をクエリとして入力すればよい。参照したいフレーズは、論文の「実験」に出現していると考え、検索対象クラスとして「評価実験」の指定もできる。

このときの検索結果を図 6 に示す。なお、フレーズ中の"<digit>"は数字が挿入されることを示している。ユーザは図 6 の検索結果を参照し、"Table <digit> shows the results for <NP>"や"we present the results of <NP>"などのフレーズを用いて、実験の結果を記すことができる。フレーズを選択すれば、図 7 に示すような用例文を参照でき、そのフレーズを使うことの適切さ及び使用法を確認できる。

5. おわりに

本稿では、学術論文から英語表現集を自動生成するためのフレーズ獲得手法について述べた。また、英語フレーズ検索システム SCOPE について記した。SCOPE は、以下で公開されている。

http://scope.itc.nagoya-u.ac.jp/

参考文献

[1] 松原, 加藤, 江川, 英文作成支援ツールとしての用例文検索システム ESCORT, 情報管理, Vol.51, No.4, pp.251-259, 2008.

[2] 崎村, 英語論文によく使う表現, 創元社, 1991.

[3] 池野, 濱口, 山本, 井佐原, Web 文書集合からの専門用語獲得, 情処論, Vol.47, No.6, pp.1717-1727, 2006.

[4] 英辞郎 第 4 版, アルク, 2008.

[5] http://www.foolabs.com/xpdf/

[6] http://jtextpro.sourceforge.net/