英語論文に頻出する表現の獲得と分類

酒井 佑太† 杉木 健二† 松原 茂樹‡

†名古屋大学大学院情報科学研究科 ‡名古屋大学情報連携基盤センター

1 はじめに

学術論文では，文章の記述において論文特有の表現や言い回しを用いる必要がある．特に，母国語以外の言語を用いる場合，特有の表現や言い回しの知識が十分でない．これを克服するために，英語を対象とした表現集などの書籍を利用することは効果的である．

一方，電子化された論文を利用するシステムの開発が近年盛んになってきている．例えば，Google Scholar や CiteSeer[1] などの論文検索システムや，用例検索システム [2, 3] などが挙げられる．

本研究では，オンライン表現集の自動構築を試みる．具体的には，英語論文に頻出する表現を自動的に獲得し，論文の構成に即して表現を分類する．筆者の知る限り，論文から表現集を作成するといった試みは行われていない．

2 英語論文表現集の特徴

英語論文を対象とする表現集の書籍（[4, 5] など）では，以下の特徴がみられる．

- タイプ別に分類されている．
  例：「研究の目的を説明する表現」
- 論文の構成ごとに分類されている．
  例：「序論」「先行研究」「方法論」
- 表現ごとに用例がいくつか掲載されている．

上記の特徴は，表現集として必要な要件である．ただし，既存の表現集では，以下の点に問題がある．

- 表現の種類や数が限られている．
- 用例の数が少ない．
- 掲載表現が統計的に頻出するものであるとは限らない．
- 専門分野ごとの表現の違いに対応できていない．

本研究では，大量の論文を用いることにより，これらの問題を解決する．つまり，多くの論文に頻出する表現を獲得し，実際の論文を用例として利用する．さらに，分野ごとに本手法を適用することにより，その分野の表現集を作成できる．

Acquisition of Useful Expressions from English Papers and their Classification

† Yuta Sakai (Nagoya University)

sakai@el.itc.nagoya-u.ac.jp

† Kenji Sugiki (Nagoya University)

‡ Shigeki Matsubara (Nagoya University)

3 フレーズの獲得と分類

本研究では，連続した単語 N-gram 表現を獲得し，それをフレーズと呼ぶ．また，フレーズを論文の構成に基づいて分類する．

本手法の流れは以下のとおりである．(1) PDF 形式の論文を pdftotext[6] を用いてテキスト形式に変換し，文に分割する．(2) これらの文を N-gram の集合に変換し，そこから表現集として適切なフレーズを獲得する．(3) 獲得したフレーズを論文の構成に即して分類する．以下，(2), (3) の処理について述べる．

3.1 フレーズの獲得

本手法では，表現集として適切なフレーズを獲得するために，N-gram を利用する．依存構造解析を用いることも考えられるが，解析が誤ることもあり，特に，長いフレーズを正確に取得することが困難である．

N-gram を用いた場合，適切なフレーズでない文字列が大量に取得される．そのため，本研究では，以下のような不適切な表現，すなわち (1) ノイズを含む N-gram，(2) あるフレーズの部分文字列となるような不完全な N-gram，を除去する．(1) では，以下のものをノイズを含む N-gram とする．

- 図や表で現れる数字やアルファベットの羅列
- 括弧がフレーズ内で閉じていない
- 冠詞で終わっている

次に，(2) に関して，まず，フレーズの部分文字列の例を以下に説明する．フレーズ “The rest of this paper organized as follows.” に対して，それより短い N-gram “The rest of this papaer is organized as” “rest of this paper is organized as follows.” は，部分文字列となる．これらの部分文字列はフレーズとして完全ではない．

一方，N-gram“The rest of this paper” も同様に上記のフレーズの部分文字列となるが，不完全なフレーズではないため，この場合，除去するのは好ましくない．このような N-gram の場合，複数の上位の N-gram に含まれ，逆に，部分文字列となるような N-gram の場合，単一の上位の N-gram に多く含まれると考えられる．

以上から，本研究では，k-gram の頻度のうち，$(k+1)$-gram の部分文字列となる頻度の割合が閾値 α 以上の場合，その k-gram を除去する．

3.2 論文構成に基づくフレーズの分類

獲得したフレーズを論文に共通の章構成に分類する．論文の構成は，比較的固定されており，共通のものも多い．本研究では，すべての論文の構成を，「序論」「先行研究」「提案手法」「実験」「結論」の 5 つのクラスから成るとす

表 1: 実験結果

構成	フレーズ評価	分類評価			
クラス	正解率 (%)	(a)	(b)	(c)	(d)
序論	66.7 (20/30)	16	0	13	1
先行研究	63.3 (19/30)	21	0	8	1
提案手法	76.7 (23/30)	14	0	16	0
実験	73.3 (22/30)	23	0	7	0
結論	63.3 (19/30)	25	0	5	0
合計	68.7 (103/150)	99	0	49	2

る（以下，**構成クラス**と呼ぶ）．この構成クラスは，情報工学分野の論文を想定して決めた．各論文の章をこれら 5 つの構成クラスに分類する．以下，フレーズを構成クラスに分類する方法を説明する．

まず，論文を章に分割するため，各章のタイトルを特定する．各章のタイトルは共通のフォーマットが用いられているため，1 章のタイトル部分を正規表現により判定し，このパターンを後続の章に適用する．

次に，各章を構成クラスに分類する．多くの論文では，各構成クラスに対して共通の単語が含まれている．本研究では，文字列を手がかり表現として利用する．例えば,「実験」に対しては，“result” “experiment” “evaluation” “discussion” の 4 つを手がかり表現とする．

最後に，フレーズを構成クラスに分類する．獲得したフレーズに対して各クラスごとの頻度を求め，それに基づきフレーズの局所性を判定する．局所性とは，そのフレーズがどの程度そのクラスに偏って出現するかを示す．

すなわち，局所性（locatilty）を以下の式で計算する．

$$locality(p,c) = \frac{ndf_{p,c}}{\sum_{c_k \in C} ndf_{p,c_k}}$$

$$ndf_{p,c} = \frac{df_{p,c}}{N_c}$$

ここで，p はフレーズ，c はクラスを表す．C はすべてのクラスの集合を表す．$ndf(p,c)$ は，クラス c を含む論文数 N_c のうち，フレーズ p がクラス c に含まれる論文数 $df_{p,c}$ の割合である．フレーズの出現頻度ではなく，出現論文数を用いたのは，一人の著者が多用したフレーズが含まれるのを防ぐためである．また，$ndf(p,c)$ を用いるのは，各クラスに分類される章の数の差を緩和するためである．局所性の値が閾値 β 以上となるクラスに分類する．すべてのクラスにおいて，フレーズの局所性の値が閾値以下となり，どのクラスにも分類されない場合は，論文上のどの位置にも出現するフレーズとみなす．

4 評価実験

4.1 実験方法

提案手法の実現可能性を確認するために，獲得したフレーズの適切さ，及び，フレーズの分類の適切さについて評価した．各構成クラスごとに頻度上位 30 件のフレーズに対して評価した．フレーズの適切さは完全であるかどうかを基準として判定した．また，分類の適切さについては，(a) 局所性がありかつ正しい，(b) 誤っている，(c) 汎用的である，(d) 分からない，の 4 つの基準により判定した．実験データとして，国際会議 ACL の 2001 年から 2008 年までの論文 1,232 件を利用した．部分文字列の除去の閾値，フレーズの分類における局所性 ($locality$) の閾値は共に 0.5 に設定した．

表 2: 獲得した表現 (結論)

獲得フレーズ	出現数	フレーズ評価	分類評価
have shown that	213	○	(a)
We have presented	115	○	(a)
we plan to	114	○	(a)
We have shown that	78	○	(a)
We have also	76	×	(a)
In this paper, we have	64	×	(a)
We plan to	59	○	(a)
In the future, we	58	×	(a)
we intend to	57	○	(a)
we have shown that	46	○	(a)

4.2 実験結果

実験の結果を表 1 に示す．フレーズ評価では，68.7%が適切なフレーズと判定された．適切でないと判定されたフレーズは，上位の N-gram の部分文字列であった．これらのフレーズは，上位の N-gram において複数の表現の部分文字列となっていたため，除去されなかった．

分類評価では，ほとんどが正解，もしくは，汎用的な表現と判定され，誤った分類はなかった．汎用的な表現として判定されたフレーズは，局所性が閾値に近い 5 割〜6 割程度であり，そのクラスに出現しやすい傾向は示していた．

実験で獲得したフレーズの例を表 2 に示す．これは構成クラス「結論」で頻度が高かった上位 10 件である.「結論」でよく用いられるフレーズが多く獲得されていた．

5 まとめ

本稿では，英語論文に頻出する表現の獲得と論文構成への分類手法を提案した．実験の結果，本手法を用いた表現集の実現可能性を示した．今後の課題として，表現内に名詞句が含まれる表現の獲得とその分類があげられる．

参考文献

[1] S. Lawrence, C.C. Giles, and K. Bollacker: “Digital libraries and autonomous citation indexing”, Computer, vol.32, no.6, pp.67-71, 1999.

[2] 松原茂樹，加藤芳秀，江川誠二：英文作成支援ツールとしての英文検索システム ESCORT，情報管理，Vol. 51，No. 4，pp.251-259，2008.

[3] 難波英嗣，森下智史，相沢輝昭：論文データベースからのイディオム用例検索情報処理学会研究報告. 自然言語処理研究会報告，Vol.2005，No.117，pp.53-59，2005.

[4] 杉野敏子，伊藤文彦：英語論文の書式と使える表現集，ナツメ社，2008.

[5] 崎村耕二：英語論文によく使う表現，創元社，1991.

[6] Xpdf：http://www.foolabs.com/xpdf/