LK-007

未知文書からの母語話者／非母語話者に特徴的な表現の抽出

Obtaining patterns peculiar to native/non-native speakers from unknown texts

永田 亮† 掛川 淳一† 宮井 俊也‡ 淀 雅昭‡ 深田 剛継‡ 河合 敦夫‡
Ryo Nagata Jun-ichi Kakegawa Toshiya Miyai Masaaki Yodo Takatsugu Fukada Atsuo Kawai

1. はじめに

英文から，母語話者が使用しないような非母語話者特有の不自然な表現を自動的に見つけ出すことができれば，英文を書く際に大きな助けとなる．特に，権威の高い国際学会では，研究内容だけでなく，英語の質も高いものが求められるため，不自然な表現を見つけ出し修正することが重要となる．逆に，母語話者の書いた英文から，母語話者特有の自然な表現を発見できれば，非母語話者が英文を書く際の手引きとなる．更に，母語話者／非母語話者に特徴的な表現を網羅的に獲得することは，第二言語習得に関する知見の蓄積に繋がる．

このような背景から，母語話者／非母語話者に特徴的な表現（以後，単に特徴表現と省略）をコーパスから抽出する様々な手法[1, 7, 8, 9]が提案されているが，従来手法は次の 3 点の問題のうちいずれかを有する（ただし，手法[9]は，特徴単語を抽出する手法）．第一に，非常に限られた表現しか抽出できない．杉浦[7]は，母語話者／非母語話者コーパスから，単語 n-gram の出現頻順位表を作成し，両者の上位数十件を特徴表現とする手法を提案している．単語 n-gram とは，n 単語が隣接して生じる単語の共起関係のことである．例えば，n=2 の場合，“It is fine.” という文からは，“It is” と “is fine” が得られる．この手法では，順位表の中盤以降に存在する特徴表現は獲得できない．例えば，非母語話者コーパスでの順位が 100 位，母語話者コーパスでは 10000 位であるような n-gram は特徴表現である可能性が高いが抽出できない．この例からわかるように，特徴表現の抽出では，母語話者／非母語話者コーパスを独立に用いるのではなく，相補的に用いることが重要となる．

第二に，手法[7]や統計指標に基づく手法[9]では，個人の特徴から受ける影響が大きいという問題点がある．ある個人が特定の表現を繰り返し使用すると，その n-gram の頻度は高くなるが，母語話者／非母語話者の特徴表現ではなく，個人の特徴表現である可能性が高い．特徴表現の抽出では，その表現がどれぐらい一般的に母語話者／非母語話者の間で使用されるかという一般性も考慮しなければならない．

第三に，品詞列に基づく手法[1, 8]では，特徴表現の解釈が困難である．これらの手法では，特徴表現が品詞列の形で得られるため，特徴表現を解釈するためには，コーパス中の用例に戻り，分析を人手で行わなければならない．また，品詞列を得るために使用する品詞タガーは，通常，母語話者の英文用に設計されているため，非母語話者に特徴的な品詞列を誤って解析してしまう可能性があり[6]，特徴表現が正しく抽出できない恐れもある．

そこで，本論文では，これら 3 つの問題を解決した手法を提案する．提案手法は，n-gram を出現頻度順に並べた n-gram プロファイル[2, 3]から特徴表現を抽出する．青木ら[2]は品詞 n-gram プロファイルが母語話者／非母語話者の文書識別に有効であることを示していることから，n-gram プロファイルには，母語話者／非母語話者の特徴表現が多く含まれていると期待できる．ただし，本論文では，特徴表現の解釈が容易になるよう，品詞 n-gram ではなく，単語 n-gram プロファイルを利用する．更に，個人の特徴から受ける影響を少なくするため，特徴表現の候補となる n-gram が，未知文書中でどれぐらい頻繁に使用されるかを予測しながら特徴表現の抽出を行う．

2. 基本アイデア

既に述べたように，提案手法では単語 n-gram プロファイル（以下，単にプロファイルと省略）を特徴表現の抽出に利用する．プロファイルとは，n-gram を出現頻度の降順に並べたリストである．出現頻度は，学習データとして与えられた母語話者／非母語話者コーパスから求める．ここでは，一般の n-gram とは異なり，$1 \leqq n \leqq k$ を満たす全ての整数 n についての n-gram を対象とする．よって，プロファイルは様々な長さの n-gram を含むことになる．

図 1 に，提案手法の基本アイデアを示す．図 1 の母語話者／非母語話者プロファイルが，母語話者／非母語話者コーパスそれぞれから得られたプロファイルである．1 列目は出現頻度を表し，2 列目が対応する n-gram である．このとき，非母語話者プロファイル中での順位が高く母語話者プロファイル中での順位が低い n-gram は，非母語話者のみがよく使用する表現であり，特徴表現の可能性が高い．逆の場合は，母語話者の特徴表現である可能性が高い．すなわち，順位の差（図 1 の d）が大きいほど，特徴表現である可能性が高い．したがって，順位の差で n-gram を降順にソートすることで，特徴表現候補のリストが得られる．

しかしながら，この基本アイデアには，大きな問題が 2 つある．第一に，1. で述べた個人の特徴からの影響がある．すなわち，ある個人が，特定の n-gram を繰り返し使用すると，出現頻度は高くなるが，母語話者／非母語話者の特徴でなく個人の特徴である可能性が高い．そのため，候補となる n-gram が，どれぐらい一般的に母語話者／非母語話者の間で使用されるかという一般性の評価が重要と

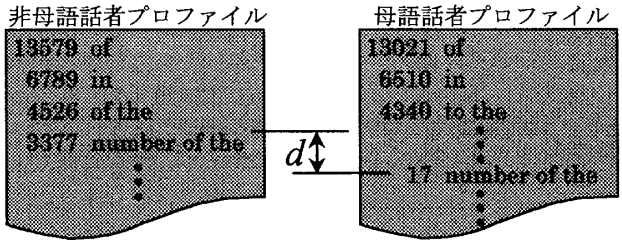

図 1　提案手法の基本アイデア

† 兵庫教育大学, Hyogo University of Teacher Education
‡ 三重大学, Mie University

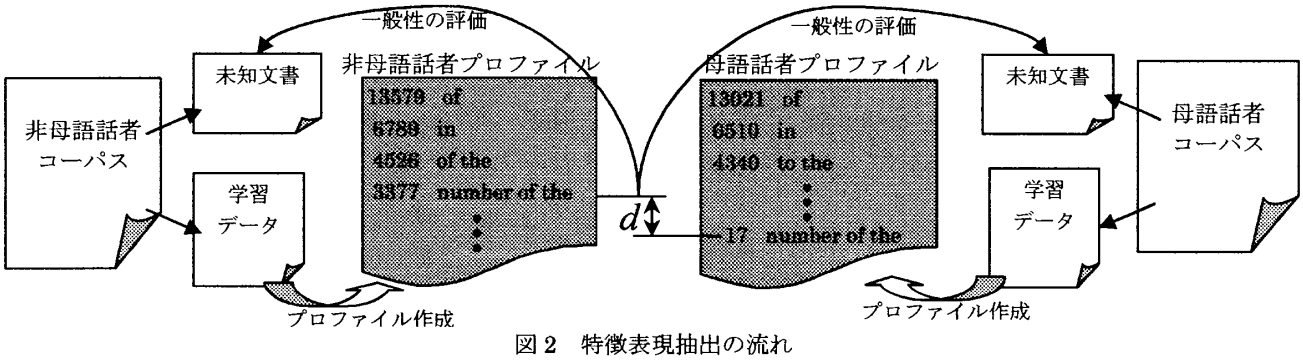

図 2　特徴表現抽出の流れ

なる．第二に，品詞 n-gram でなく単語 n-gram を利用する場合，候補となる n-gram の種類数が大きく増加するという問題がある．このことは，プロファイルに特徴表現でない n-gram も多数含まれ，ノイズとなることを意味する．また，プロファイルの作成には，ソートが含まれるため計算時間も問題となる．したがって，何らかの方法で候補を限定して，効率良く特徴表現を抽出することも重要となる．

第一の問題点の解決策として，提案手法では，プロファイルから得られた候補 n-gram が未知文書にどれぐらいよく出現するかを予測することで一般性の評価を行う．未知文書に頻出する候補 n-gram は，真に特徴表現である可能性が高い．実際に，未知文書を予め用意することは不可能であるので，未知文書は仮想的に作成する．図 2 に示すように，母語話者／非母語話者コーパスから，一部の文書を取り出し（この文書を未知文書と呼ぶ），残りの文書を学習データとしてプロファイルを作成する．図 2 に示されるように，未知文書は学習データには含まれないので，仮想的に未知とみなすことができる．プロファイル中で順位の差 d が大きい，且つ，未知文書に頻出する n-gram を特徴表現として抽出する．提案手法では，この手順を異なる学習データと未知文書との組み合わせで複数回行うことで，信頼性の高い抽出を実現する．

第二の問題は，前置詞を含むn-gramだけに候補を限定することで解決する．前置詞は，非母語話者にとって用法の難しい単語であり，母語話者／非母語話者の特徴が出やすいと予想される．例えば，“on/in the street” の微妙な意味の差を使い分けることは非母語話者にとって困難であり，どちらか偏って使用する可能性が高い．実際，Artsら[1]は，母語話者と非母語話者では，前置詞を含むn-gramの使用に差異があることを報告している．また，日本語表現「AのB」に対する “B of A; e.g., king of England”，“B for A; e.g., key for success”，“B to A; e.g., key to success” の例のように，複数の選択肢があることも前置詞の用法を難しくする．更に，前置詞の両端には，動詞句と名詞句が出現するため，動詞句や名詞句に関する特徴表現の抽出も期待できる．そのほか，前置詞は，冠詞の用法にも影響を及ぼすことから，冠詞に関する特徴表現も抽出できる可能性がある．そこで，本論文では，JLEコーパス[5]中での誤り数を基準として，30 種類の前置詞[1]を選び出した．これらの前置詞を中心とした左右 0〜2 単語の組み合わせから成るn-gramで，プロファイルを作成する．以後，n-gramとは，この前置詞を中心としたn-gramを指すことにする．

3. 提案手法

3.1 プロファイルの作成

プロファイルは，入力として与えられた母語話者／非母語話者コーパス，それぞれから作成する．以下，非母語話者プロファイルの作成方法を説明するが，母語話者プロファイルについても同じ手順で作成する．

まず，非母語話者コーパスを文に分割する．次に，分割した文から，n-gram を抽出する．このとき，仮想的な文頭・文末記号を適宜入れて，文頭と文末の n-gram も抽出できるようにする．例えば，前置詞の左 2 単語まで考慮する n-gram では，文頭記号（BOS）を 2 つ入れて抽出を行う（例えば，“BOS BOS in”）．全ての単語は小文字に変換する．数字とカンマは，それぞれ “NUM” と “COM” に置き換える．最後に，各 n-gram の出現頻度を計算し，出現頻度の降順で n-gram をソートする．その結果得られるリストが，非母語話者プロファイルとなる．

2. で述べたように，提案手法では，コーパスを異なる複数の組合せ（以後，L 組とする）で学習データと未知文書を用意する．したがって，プロファイルも L 種類作成されることになる．

3.2 特徴表現の抽出

提案手法を定式化するため，次の記号を導入する．いま，母語話者，非母語話者を，それぞれ N, NN で表すとする．また，j 番目（$1 \leqq j \leqq L$）の学習データから作成された母語話者／非母語話者プロファイルを，それぞれ P_{jN}，P_{jNN} で表す．また，プロファイル中の n-gram を x で表し，その順位を $r(x,P)$ で表す．例えば，図 2 では，$r(\text{in},P_N)=r(\text{in},P_{NN})=2$ である．ただし，同順の n-gram がある場合は，r は平均順位とする．また，プロファイル中に存在しない x については，両プロファイルの最大の順位で大きいほうの順位+1 を与える．このとき，n-gram の順位の差を，

$$d_j(x) = r(x,P_{jN}) - r(x,P_{jNN}) \qquad (1)$$

で定義する．プロファイルは L 種類あるので，順位の差も L 種類得られる．そこで，実際の抽出では，各順位の差の平均を利用する．すなわち，

$$\overline{d}(x) = \frac{1}{L}\sum_{j=1}^{L}\{r(x,P_{jN}) - r(x,P_{jNN})\} \qquad (2)$$

に基づいて特徴表現を抽出する．

[1] 選択した前置詞：about, above, across, against, along, among, as, at, beneath, between, by, down, during, for, from, in, into, of, off, on, onto, out, over, to, under, up, upon, with, within, without.

式(2)を用いて，非母語話者の特徴表現を抽出するためのスコアを，

$$s(x, NN) = \overline{d}(x)\sum_{j=1}^{L} f_j(x, NN) \quad (3)$$

で定義する．ただし，$f_j(x, NN)$ は，j 番目の未知文書中で，プロファイル P_{jNN} の x を含む文書の数とする．したがって，$\sum_{j=1}^{L} f_j(x, NN)$ は，x が出現する未知文書の数を表し，一般性の指標と捉えられる．式(3)全体では，順位の差（の平均）が大きい，且つ，未知文書に頻出する特徴表現候補に高い値を与えることになる．また，どれだけ順位の差が大きくとも，未知文書に出現しない n-gram にはスコア 0 を与えることになる．

最後に，式(3)のスコアに基づいて特徴表現の抽出を行う．まず，非母語話者プロファイル中の全ての x に対して式(3)のスコアを計算する．次に，スコアの降順で x をソートする．最後に，ソートされた x の上位 M 件を非母語話者の特徴表現として抽出する．同様に，NN を N に読み替えることで，母語話者の特徴表現も抽出できる．

4. 実験と考察

4.1 実験条件と実験手順

特徴表現をコーパスから抽出する実験を行った．コーパスとして，材料化学分野の論文（Journal of Non-Crystalline Solids[2]）を利用した．第一著書の所属が，アメリカ／日本である論文を母語話者／非母語話者とし，それぞれ 120 編選び出した． 120 編中，100 編を学習データ，20 編を未知文書とする組合せを 6 種類作成し，抽出を行った．両コーパスとも，謝辞は削除した．

評価は定性的に行った．なぜなら，対象とする母語話者／非母語話者コーパスに含まれている全ての特徴表現を予め特定しておくことは難しく，抽出精度などを定量的に評価することは困難であるからである．提案手法で抽出された上位 10 件及び 91～100 件の特徴表現を定性的に評価した．また，比較対象として，単語 n-gram の出現頻順位表中の上位数十件を特徴表現とする手法[7]を用いた．本実験では，tri-gram を対象にして，頻度上位 10 件を特徴表現とした．また，特徴単語を抽出する手法[9]も比較対象とした．手法[9]は，特徴単語を抽出する手法であるが，高い平均精度で特徴単語が抽出できると報告されているので対象とした．手法[9]で提案されている尺度のうち，単独での精度が最も良く，且つ，複数の尺度を組み合わせた方法とも精度にあまり差がない CSM を利用した（CSM の精度 0.369，複合尺度 0.390）．手法[7]と同様に，tri-gram を対象として特徴表現を抽出した．

4.2 実験結果と考察

表 1 に，提案手法で抽出された特徴表現の上位 10 件及び 91～100 位を示す．表中の「分類」とは，特徴表現の種類を表す．なお，表中で，スコアが同一かつ部分文字列である特徴表現は，長いほうの特徴表現に集約した（ただし，“BOS BOS”のみ“BOS”に集約した）．

表 1 から，提案手法では，前置詞の選択に関する特徴表現が多く抽出されていることがわかる．例えば，非母語話者の特徴表現 1 位の“mechanism *of* the”は，一見特徴表現に見えないが，プロファイル中の“mechanism”を調べると，非母語話者では，“of”が圧倒的に多い一方で，母語話者では“of”だけでなく“in”や“for”なども使い分けていることがわかる．すなわち，非母語話者では，“mechanism *of*”を偏って使用していることになる．逆に，母語話者の特徴表現 1 位の“ability *to*”では，母語話者だけが“to”を使用し，非母語話者では“in”と“of”が主流である（同様な例：“expressed *by*”（非母語話者），“data *from*”（母語話者））．

提案手法は，冠詞・限定詞の選択に関する特徴表現も抽出している．例えば，“number of *the*”（非母語話者）は，“number of 無冠詞”とするのが一般的である（非母語話者では 21 論文で，母語話者では 1 論文でのみ出現）．同様な特徴表現として，“*a* part of”が抽出された．“part of”の指定する部分が曖昧で特定されない場合，“無冠詞 part of”となることが一般的である[4]が，このような冠詞用法は非母語話者にとって難しい．そのため，“*a* part of”を多用し特徴表現として抽出されたと分析できる（母語話者では“無冠詞 part of”が大部分を占める）．

そのほか，提案手法で，不自然な表現“with each other”（非母語話者），イディオム“allow for”（母語話者）など多様な特徴表現が抽出できている．このような特徴表現は，順位 100 位付近でも抽出されている（例：前置詞の選択“curves *in*”（非母語話者 100 位），希少表現“clustering of the”（母語話者 97 位））．

提案手法との比較のために，表 2 に，手法[7]で抽出された特徴表現を示す．表 2 では，7 割の表現が母語話者と非母語話者とで共通しており，殆どが特徴表現でない．例えば，表 2 中の“NUM NUM NUM”（このパターンの大部分が文献番号の列挙）は，両コーパスとも頻出するため，リストの上位に位置するが，特徴表現とはいえない．この原因として，手法[7]では，母語話者／非母語話者コーパスを単独で利用し，両者の比較を行わないことが挙げられる．特徴表現を抽出するためには，提案手法のように，両者の比較を行うことが重要である．

表 3 に，CSM[9]で抽出された特徴表現を示す．表 3 では，“than that of”や“COM which is”など特徴表現が抽出されている．しかしながら，100 位付近では，“favored structures and”や“the hole burning”など個人，又はトピックに依存した表現であると思われるものが含まれる結果となった．例えば，“the hole burning”では，母語話者／非母語話者コーパスでの頻度は，15／0 回であるが，母語話者の 120 論文のうち 2 論文でしか出現していない．このような表現の混入を防ぐためには，提案手法のように，候補となる表現の一般性を評価する必要がある．また，頻度差が大きいため，“NUM NUM NUM”を非母語話者の特徴表現として抽出しているが，実際には，両話者に共通する表現である．

以上の結果から，特徴表現の抽出では，(a) 母語話者／非母語話者コーパスを比較する，(b) 一般性を考慮する，の 2 点が重要であるといえる．この 2 点を満たすため，提案手法では，前置詞の選択や冠詞・限定詞の選択に関する特徴表現，不自然な表現，イディオムなど多様な特徴表現の抽出に成功した．また，品詞列に基づく手法[1, 8]と比べても，(1) 解析ツールに依存しない，(2) 特徴表現の

[2] http://www.elsevier.com/wps/find/journaldescription.cws_home/505709/description#description

表 1　提案手法で抽出された特徴表現

非母語話者（NN）				母語話者（N）		
順位	分類	特徴表現	コメント	分類	特徴表現	コメント
1	PP	mechanism *of* the	N は{in, for}も使用	PP	ability *to*	NN は{in, of}が主流
2	O	to clarify	N/NN での出現 0/33 回	PS	for a given	NN は受動態の given が多数
3	DT	number of *the*	通常 number of 無冠詞	PP	data from	NN は of が多数，from は 0
4	O	to clarify the	N/NN での出現 0/27 回	SP	*differences* between	NN での使用なし
5	UN	with each other	N/NN での出現 0 /18 回	R	to minimize	N/NN での出現 18/1 回
6	O	the *relation* between the	N では “correlation” が多数	?	be found in	
7	UN	while that of	N/NN での出現 1/15 回	I	allow for	「考慮に入れる」の意
8	PP	the mechanism *of* the	N は{in, for}も使用	R	as measured by	N/NN での出現 22/0 回
9	O	ascribed to the	N/NN での出現 6/27 回	R	indicative of	N/NN での出現 18/1 回
10	PP	expressed *by*	N は “in terms of” も使用	PP	was performed on	NN は{at, for, in, by}が主流
91	?	works as		PP	fluctuations *in*	NN では of が主流
92	SP	*dimension* of	N では複数形	PP	well as *in*	NN では 0 回
93	PP	evaluated from the	N では{for, by, at}も	?	*presented in* fig NUM	NN では"shown in"が主流
94	UN	BOS from these results	N/NN での出現 1/10 回	?	is important to note	
95	?	by irradiation		R	over all	N/NN での出現 9/0 回
96	DT	*a* part of	N は “無冠詞 part of” が主流	DT	accuracy of the	N では"accuracy of 無冠詞"
97	O	*BOS* contrary to	N では文頭での使用 1 回	R	clustering of the	NN では cluster のみ出現
98	PP	were calculated by	N では “for” も	R	at this point	NN での使用なし
99	O	*COM* in the present	N は文頭でのみ使用	PS	has resulted in	NN では現在完了はなし
100	PP	curves *in*	N では{for, with}が主流	PS	*attempts* to	N は名詞としても使用

PP：前置詞の選択，DT：冠詞・限定詞の選択，UN：不自然な表現，PS：品詞異なり，I：イディオム，O：overuse，R：希少表現，SP：単複の選択，?：分類不能，BOS：文頭記号，COM：カンマ，NUM：数字，N：母語話者，NN：非母語話者

解釈が容易であるという 2 つの特徴を有する．更に，前置詞を含む n-gram という制約を除き，且つ，単語 n-gram を文字単位の n-gram に変更することで，提案手法は言語依存しない特徴表現抽出手法となる．

表 2 手法[7]で抽出された特徴表現

順位	非母語話者（NN）	母語話者（N）
1	NUM NUM NUM	NUM NUM NUM
2	NUM NUM and	NUM NUM and
3	NUM and NUM	NUM and NUM
4	fig NUM shows	in fig NUM
5	in fig NUM	shown in fig
6	shown in fig	due to the
7	due to the	fig NUM shows
8	on the other	the formation of
9	the other hand	a function of
10	other hand COM	from NUM to

表 3 CSM[9]で抽出された特徴表現

順位	非母語話者（NN）	母語話者（N）
1	on the other	as well as
2	the other hand	the presence of
3	other hand COM	the use of
4	NUM NUM NUM	fig NUM and
5	fig NUM shows	NUM and NUM
6	NUM shows the	between NUM and
7	than that of	was used to
8	COM which is	a function of
9	as shown in	COM as well
10	COM and the	to determine the

5. おわりに

本論文では，母語話者／非母語話者コーパスから特徴表現を抽出する手法を提案した．実験の結果，母語話者／非母語話者コーパスを比較し，且つ，未知文書での出現を考慮するという提案手法の考え方は，特徴表現の抽出に効果的であることが確認できた．その結果，冠詞・限定詞の選択に関する特徴表現，不自然な表現，イディオムなど多様な特徴表現の抽出に成功した．なお，提案手法を実装したツールを http://www.ai.info.mie-u.ac.jp/~nagata/tools/にて公開している．

謝辞

本研究の一部は文部科学省科学研究費補助金・若手研究（B）（課題番号：19700637）により実施した．

参考文献

[1] J. Aarts and S. Granger. 1998. Tag sequences in learner corpora. In *Learner English on computer*, pp. 132-141.

[2] 青木, 冨浦, 行野, 谷川. 2006. 言語識別技術を応用した英語における母語話者文書・非母語話者文書の判別. *FIT2006*, pp. 85-88.

[3] W.B. Cavnar and J.M. Trenkle. 1994. N-gram-based text categorization. In *Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-175.

[4] 樋口. 2003. 現代英語冠詞事典, 大修館書店.

[5] 和泉, 内元, 井佐原. 2004. 日本人1200人の英語スピーキングコーパス, アルク.

[6] S. Granger, E. Dagneaux, F. Meunier. 2002. *International Corpus of Learner English*. Presses Universitaires de Louvain.

[7] 杉浦. 2000. 第二言語習得研究のための英語学習者コーパスの構築とその利用. 科学研究費補助金研究成果報告書.

[8] 田中, 藤井, 冨浦, 徳見. 2006. NS/NNS論文分類モデルに基づく日本人英語科学論文の特徴抽出. *英語コーパス研究*, 13:pp.75-87.

[9] 内山, 中條, 山本, 井佐原. 2004. 英語教育のための分野特徴単語の選定尺度の比較. *自然言語処理*. 11(3): pp.163-197.