Vol.2012-IFAT-105 No.12
Vol.2012-NL-205 No.12
2012/1/21

N-gram を素性とするパターン認識を用いた英語科学論文の質判定

小林雄一郎[†]　田中省作[††]　冨浦洋一[†††]

本研究は，英文添削の専門家が表現上の質を評価した英語科学論文を対象に，論文内の単語 n-gram と品詞 n-gram の使用頻度を素性の候補とし，論文における表現上の質を統計的に推定することを試みる。ランダムフォレストを分類器として用いたパターン認識実験を行った結果，77.75%の精度で，十分な質を持った論文と不十分な質の論文を正しく分類することができた。

Pattern Recognition of English Scientific Papers Using N-grams

Yuichiro Kobayashi[†] Shosaku Tanaka[††] Yoichi Tomiura[†††]

The aim of the present study is to assess the quality of formal expressions in English scientific papers through random forests. The explanatory variables are the frequencies of word and POS n-grams. With the accuracy of 77.75% over the entire set of corpus texts, this study clarifies the difference between *good* and *poor* papers.

1. はじめに

文章における表現の質（以後，単に「質」と記す場合はこの表現上の質を表すこととする）を規定する要因は，言語学的に非常に興味深い問題である。また，そのような要因は，英語教育で学習者が書く英文章に対する質を評価する際にも重要な観点となる。

そこで本研究は，英文添削の専門家が表現上の質を評価した英語科学論文を対象に，文章内の単語 n-gram と品詞 n-gram の使用頻度を素性の候補とし，論文における表現上の質を統計的に推定することを試みる。その推定に採用した統計モデルは，構築後にどのような素性が判定に有効であるかを見直すことができるため，どのような n-gram が質と関わっているのかを知ることができる。また，このモデルは，完全に自動判定する場合だけでなく，複数の評価者の判定が一致しない場合に参考とすることもでき，言語教育分野にとっても有意義なものである。

2. パターン認識問題としての質判定

パターン認識は，音声認識，手書き文字認識，顔画像認識，X 線画像・CT 画像からの病気の診断，指紋・静脈・虹彩などによる本人識別などを含む様々な分野で用いられている。そして，これらは全て，対象の特徴を表す何らかの量を手がかり（素性）とし，対象の属性を表す識別子（クラス）を推定するという形で定式化される 1)。

言語データに対するパターン認識の適用例としては，検索キーワード（素性）から適切なウェブページであるか（クラス）を判定したり，テキスト中のキーワード（素性）からスパムメールやスパムブログ（クラス）を自動選別したりする技術が知られている。そこで，本研究では，その頻度に論文の質が如実に反映される言語項目を手がかり（素性）とし，論文の質（クラス）を推定する。

2) は，英文添削の専門家が表現上の質を評価した英語科学論文を対象に，論文内のメタ談話標識の使用頻度を素性の候補とし，ランダムフォレストを分類器として用いたパターン認識実験を行った結果，81.79%の精度で，十分な質を持った論文と不十分な質の論文を正しく分類した。そして，分類に寄与したメタ談話標識を吟味したところ，hedge (e.g. *could, likely, would, perhaps, may, appear, apparent, almost, largely*) が優れた書き手と稚拙な書き手を分ける言語表現であることが分かった。

† 大阪大学大学院言語文化研究科／日本学術振興会
Graduate School of Language and Culture, University of Osaka / Japan Society for the Promotion of Science
†† 立命館大学文学部
College of Letters, Ritsumeikan University
††† 九州大学システム情報科学研究院
Faculty of Information Science and Electrical Engineering, Kyushu University

Vol.2012-IFAT-105 No.12
Vol.2012-NL-205 No.12
2012/1/21

3. 分類手法

3.1 ランダムフォレスト

本研究で実験に用いる分類器は，3) によって提案されたランダムフォレストである。端的に言えば，ランダムフォレストとは，決定木のアンサンブル学習である。決定木は，非線形判別分析，非線形回帰分析の 1 つとして位置付けられ，素性の値を何らかの基準で分岐させ，判別・予測のモデルを構築する。分岐の過程は，木構造で図示することができ，IF-THEN のような簡単なルールで表すこともできる。また，アンサンブル学習は，必ずしも精度の高くない複数の分類器の結果を組み合わせ，精度を向上させるパターン認識の手法である。

以下に，ランダムフォレストのアルゴリズムを簡潔に示す（詳細については，3) を参照）。

- 与えられたデータセットから，N 組のブートストラップサンプルを作成
- 各々のブートストラップサンプルデータを用いて，未剪定の最大の決定・回帰木を生成（但し，分岐のノードは，ランダムサンプリングされた素性のうち最善のものを使用）
- 全ての結果を統合し（回帰問題では平均，分類問題では多数決），新しい予測・分類器を構築

そして，4) は，ランダムフォレストの長所として，以下を挙げている。

- 精度が高い
- 大きいデータに効率的に作用し，何百・何千の素性を扱うことができる
- 分類に用いる素性の重要度を推定する
- 欠損値の推測，多くの欠損値を持つデータの正確さの維持に有効である
- 分類問題における各クラスのケース数がアンバランスであるデータにおいてもエラーのバランスが保たれる
- 分類と素性の関係に関する情報を計算する
- クラス間の近似の度合が計算できる
- 外的基準がないデータにも適用できる（ケースの類似度の計算など）

なお，具体的な計算に関しては，統計解析環境 R の randomForest パッケージを使用する。

3.2 素性

本研究で実験に用いる素性は，単語 n-gram と品詞 n-gram である。具体的には，単語 2-gram，単語 3-gram，単語 4-gram，品詞 2-gram，品詞 3-gram，品詞 4-gram の頻度上位 100 種類ずつを抽出する。その際，単語は表記形のままとし，品詞は TreeTagger によって自動付与された Penn Treebank tagset 5) を用いる。そして，それらを素性として，以下の 12 種類の探索的実験を行う。

- 単語 2-gram による分類実験（素性の数は 100）
- 単語 3-gram による分類実験（素性の数は 100）
- 単語 4-gram による分類実験（素性の数は 100）
- 単語 2~4-gram による分類実験（素性の数は 300）
- 品詞 2-gram による分類実験（素性の数は 100）
- 品詞 3-gram による分類実験（素性の数は 100）
- 品詞 4-gram による分類実験（素性の数は 100）
- 品詞 2~4-gram による分類実験（素性の数は 300）
- 単語 2-gram と品詞 2-gram による分類実験（素性の数は 200）
- 単語 3-gram と品詞 3-gram による分類実験（素性の数は 200）
- 単語 4-gram と品詞 4-gram による分類実験（素性の数は 200）
- 単語 2~4-gram と品詞 2~4-gram による分類実験（素性の数は 600）

4. 結果と考察

4.1 実験データ

本研究の実験データは，Web 上で公開されている英語科学論文を収集したものである（具体的な収集方法に関しては，6) を参照）。また，それぞれの論文には，英語を母語とする複数の英文添削の専門家によって，各論文の表現上の質評価やコメントなどの情報が付与されている。表現上の質評価とは，内容（新規性や論理性など）に関する評価ではなく，科学論文としての表現に関する評価を指し，「英文章中の表現の誤りの種類（軽微な誤り／非母語話者特有の誤り）と回数」（観点 A）と，「各分野で高い評価を得ている学術雑誌にそのまま掲載できるものかどうか」（観点 B）によって規定されている（表 1 を参照）。なお，「軽微な誤り」とは，科学論文に通じた母語話者（NS）でも犯すようなミススペリングや編集ミスといったものである。「非母語話者（NNS）特有の誤り」とは，NS は決して犯さない文法的誤りや不自然なコロケーション，科学論文としては不自然な表現（まわりくどい表現，古風な表現，カジュアルな表現）などである。

Vol.2012-IFAT-105 No.12
Vol.2012-NL-205 No.12
2012/1/21

表 1　科学論文における表現の質の区分

Lv.	誤りの種類と回数（観点 A）	学術雑誌への掲載（観点 B）
L5	十分に良質で，修正の必要はない	そのままで掲載可
L4	軽微な誤りが 250 語あたり 2 箇所以下，なおかつ NNS 特有の誤りは皆無である	
L3	軽微な誤りと NNS 特有の誤りがいずれも 250 語あたり 2 箇所以下，または NNS 特有の誤りが 250 語あたり 3〜4 箇所ある	そのままで掲載可，または軽微な修正の上で掲載可
L2	NNS 特有の誤りが 250 語あたり 8 箇所以下である	掲載不可
L1	NNS 特有の誤りが 250 語あたり 8 箇所より多い	

本研究では，表 1 の L4〜5 にあたる論文を「質の高い論文」（G 論文）とし，L1〜2 にあたる論文を「稚拙な論文」（P 論文）とする。実験データにおける論文の総数は 781 本（総語数は 5256051 語）で，そのうち，専門家が G 論文であると判定したものが 384 本（総語数は 3177966 語），P 論文であると判定したものが 397 本（総語数は 2078085 語）含まれている。

本研究では，個々の論文におけるそれぞれの n-gram の相対頻度を素性の候補とし，G 論文／P 論文というクラス情報を判定する分類実験を行う。

4.2 N-gram の抽出

分類実験の前処理として，個々の論文に表れている n-gram の頻度を抽出し，表 2 のような論文×n-gram の形で表わされる頻度行列を作成する。その際，個々の論文の語数が異なるため，観測頻度は相対頻度（本研究では，1 万語あたり）に変換する。

表 2　頻度行列の一部（単語 2-gram）

	of the	in the	to the	and the	on the	･･･	CLASS
1	155.19	53.35	75.17	38.80	24.25	･･･	good
2	99.75	63.37	31.69	23.47	9.39	･･･	good
･･･	･･･	･･･	･･･	･･･	･･･	･･･	･･･
781	40.02	28.90	17.79	6.67	8.89	･･･	poor

4.3 分類実験

まず，実験データに含まれる 781 本の論文を 2 分割し，半数の 391 本（G 論文 192 本，P 論文 199 本）を学習用のデータセットとし，残りの 390 本（G 論文 192 本，P 論文 198 本）を評価用のデータセットとする。

分類実験にあたって，ランダムサンプリングする素性の数は，素性の数の正の平方根（ランダムフォレストの考案者が推奨する数）を取り，ランダムフォレストに含まれる木の数は 200 とする。

次に，学習用データセットから構築したモデルを見ていく。表 3〜5 は，12 種類の分類実験において寄与の大きい 30 表現（n-gram）をそれぞれまとめたものである。因みに，G 論文に高い頻度で生起する n-gram を黒字，P 論文に高い頻度で生起する n-gram を赤字で表している。

分類に寄与した単語 n-gram（表 3）を見ると，英語科学論文において比較的一般的な表現が目立っている。その中で，G 論文には *can be used to*，*due to*，*as well as*，*the* [*context / end / presence / rest / size*] *of* などが高い頻度で生起し，P 論文には *on the other hand*，*in this paper*，*for example*，*and so on* などが高い頻度で生起している。

また，分類に寄与した品詞 n-gram（表 4）を見ると，名詞句に関わる表現が多く見られる（e.g. DT-NN-IN）。そして，G 論文には副詞や形容詞の重出（e.g. RB-RB, JJ-JJ），*to* を使った表現（e.g. TO-DT-JJ-NN）などが高い頻度で生起し，P 論文には受動態（e.g. VBZ-VVN），等位接続（e.g. CC-DT-NN）などが高い頻度で生起している。

そして，単語 n-gram と品詞 n-gram の両方を使った場合（表 5）では，文頭の *on the other hand*（例文 1）と，受動態（例文 2）が目立つ。

(1) **On the other hand**, there is a famous theory about human needs called “needs hierarchy theory” in clinical psychology (Maslow, 1943, 1954, 1967). （P 論文）
(2) An outline of this system **is (VBZ) shown (VVN) in (IN)** Figure 1. （P 論文）

因みに，表 3〜5 のいずれの場合においても，P 論文に高い頻度で現れる表現が上位に並んでいることは注目に値する。

Vol.2012-IFAT-105 No.12
Vol.2012-NL-205 No.12
2012/1/21

表 3　単語 n-gram

	2-gram	3-gram	4-gram	2~4-gram
1	of these	are shown in	the other hand ,	On the other
2	In this	For example ,	, as well as	the following
3	the other	, there are	On the other hand	In this
4	the following	) , which	In this paper ,	the other hand
5	used to	it is not	(e. g. ,	used to
6	based on	, we have	(i. e. ,	of these
7	due to	, in the	this paper , we	On the other hand
8	, as	, for example	In this section ,	other hand ,
9	, or	In this paper	and so on .	, or
10	which is	, where the	is one of the	, as well as
11	and the	a number of	in the context of	In this paper ,
12	(e.	there is a	as a function of	the other hand ,
13	as well	that is ,	the end of the	based on
14	shows the	the fact that	the size of the	shows the
15	to a	in terms of	is based on the	, as well
16	e. g.	with respect to	, i. e. ,	the other
17	, but	, which is	However , it is	this paper ,
18	In the	i. e. ,	, which is a	, as
19	It is	, i. e.	, in order to	and the
20	, the	e. g. ,	In addition , the	due to
21	) .	, respectively .	the center of the	, and a
22	, we	et al. ,	, however , the	(e.
23	such as	, however ,	, it is not	as well as
24	of a	shown in Figure	can be used to	In this paper
25	in a	of the system	in the form of	(i. e.
26	it is	such as the	shown in Fig .	, the
27	at the	the number of	in the presence of	(e. g.
28	of an	be used to	) , and the	as follows .
29	as the	On the other	this section , we	, and
30	shown in	in which the	the rest of the	which is

表 4　品詞 n-gram

	2-gram	3-gram	4-gram	2~4-gram
1	VBZ VVN	VBZ VVN IN	NN VBZ VVN IN	VBZ VVN
2	NN SENT	CC DT NN	DT NN VBZ VVN	VBZ VVN IN
3	CC DT	DT NP NN	DT NN NN VBZ	NN SENT
4	NN VBZ	NN VBZ VVN	VBZ VVN IN DT	CC DT NN
5	RB RB	NN NN VBZ	DT JJ NN IN	DT NP NN
6	RB VVN	NNS IN DT	IN DT NNS IN	NN VBZ VVN
7	NNS IN	IN DT NNS	NN NN VBZ VVN	NN VBZ VVN IN
8	JJ TO	DT NN SENT	DT NN IN NN	CC DT
9	JJ JJ	DT JJ JJ	NN IN DT JJ	DT NN VBZ VVN
10	VBP VVN	JJ NN SENT	DT NN NN IN	RB VVN
11	DT NNS	NN NN SENT	IN DT NP NN	DT JJ NN IN
12	RB JJ	VBP VVN IN	JJ NN NN SENT	VVN IN NP
13	NN NNS	VVN IN NP	IN DT NP NP	NNS IN DT
14	IN NP	DT NN NN	DT JJ NN SENT	IN DT NNS IN
15	JJ NNS	IN NNS IN	DT JJ NN VBZ	NN NN VBZ
16	NN CC	DT NNS IN	NNS IN DT JJ	NN VBZ
17	DT NN	JJ NN NNS	NNS IN DT NN	JJ NNS IN
18	PP MD	JJ NNS IN	JJ NNS IN DT	DT NN NN VBZ
19	NN IN	NN NN IN	JJ NN IN NN	DT NN IN NN
20	RB VV	JJ NN VBZ	NN IN NN SENT	IN DT NN
21	VBD VVN	DT NP NP	IN DT NN SENT	RB RB
22	NNS SENT	DT NN VBZ	JJ NNS IN JJ	JJ TO
23	NP CD	IN DT NP	TO DT JJ NN	DT NN
24	NNS VVP	JJ JJ NNS	NN IN NNS IN	JJ NNS IN DT
25	IN VVG	DT NN IN	DT JJ NNS IN	IN DT NP NN
26	IN NN	DT NN CC	JJ NN IN NNS	NP NP
27	NP IN	JJ NN IN	IN JJ NNS IN	JJ NN NN SENT
28	DT JJ	NNS VBP VVN	NN NN IN DT	VBZ VVN IN DT
29	NN NN	DT NN VVZ	JJ NN IN JJ	JJ NN IN NN
30	NP NP	NN NN CC	NNS VBP VVN IN	TO DT JJ NN

Vol.2012-IFAT-105 No.12
Vol.2012-NL-205 No.12
2012/1/21

表 5　単語 n-gram＋品詞 n-gram

	2-gram	3-gram	4-gram	2~4-gram
1	VBZ VVN	VBZ VVN IN	the other hand ,	other hand ,
2	the following	On the other	On the other hand	the other hand ,
3	of these	the other hand	In this paper ,	On the other hand
4	NN SENT	other hand ,	NN VBZ VVN IN	the other hand
5	the other	NN VBZ VVN	, as well as	VBZ VVN IN
6	In this	, and a	VBZ VVN IN DT	NN SENT
7	due to	(i. e.	NN IN DT JJ	of these
8	NN VBZ	NN NN VBZ	DT JJ NN IN	On the other
9	based on	this paper ,	DT NN VBZ VVN	In this paper ,
10	, as	, as well	DT NN NN VBZ	based on
11	used to	In this paper	DT NN NN IN	VBZ VVN
12	, or	(e. g.	JJ NNS IN DT	the other
13	RB RB	NNS IN DT	IN DT NNS IN	the following
14	VBP VVN	CC DT NN	TO DT JJ NN	, as well as
15	shows the	JJ NNS IN	NN NN VBZ VVN	used to
16	and the	, we can	DT JJ NN VBZ	VBZ VVN IN DT
17	e. g.	DT NN SENT	IN DT NN SENT	, as well
18	CC DT	JJ NN NNS	JJ NN NN SENT	(i. e.
19	(e.	DT NP NN	JJ NNS IN JJ	(e. g.
20	RB VVN	is shown in	DT NN IN NN	In this paper
21	to a	DT NN NN	NNS IN DT JJ	as well as
22	which is	based on the	IN DT NP NN	DT NN
23	of a	as shown in	JJ NN IN NN	DT NP NN
24	NN NNS	DT JJ JJ	NN IN NNS IN	NN VBZ VVN
25	as well	as well as	IN DT JJ JJ	, or
26	NNS IN	as follows .	JJ NN IN NNS	CC DT NN
27	DT NN	paper , we	NN NNS IN DT	In this
28	of an	NN NN SENT	DT JJ NN SENT	, as
29	In the	NNS IN NNS	NN NN IN DT	for a
30	at the	DT NN VVZ	DT NN IN NP	CC DT

そして，表 6 は，学習したモデルを評価用データセットに適用した結果をまとめたものである。

表 6　分類精度

	2-gram	3-gram	4-gram	2~4-gram
単語	75.38	76.14	74.19	76.42
品詞	74.06	74.30	73.23	74.87
単語＋品詞	76.07	76.21	75.81	**77.75**

12 種類の実験のうちで最も分類精度が高かったのは，単語 2~4-gram と品詞 2~4-gram の全てを素性に使った実験で，77.75%であった。また，同じ n の数であれば単語＋品詞，単語，品詞の順に精度が高く，n の数に関しては，2~4，3，2，4 の順に精度が高いという傾向が見られた（図 1）。

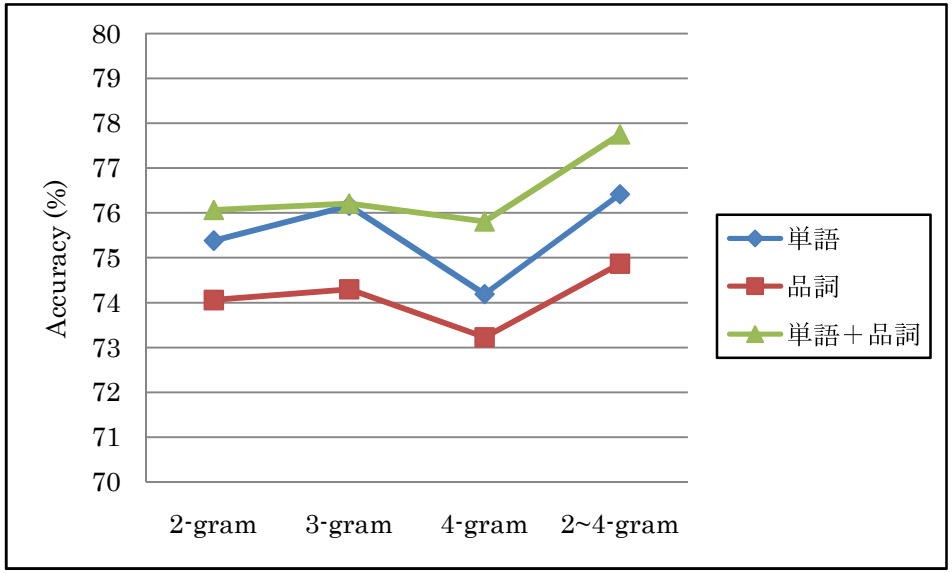

図 1　素性による分類精度の比較

Vol.2012-IFAT-105 No.12
Vol.2012-NL-205 No.12
2012/1/21

5. おわりに

本研究では，英語科学論文における n-gram の頻度を素性の候補とするランダムフォレストに基づく分類器を構築した。その分類精度は 77.75%であった。

今後の方向性としては，言語学や言語教育の知見に基づく分類器の改良，談話情報や誤り情報に基づく分類器との統合，2 クラス分類モデルから多クラス分類モデルへの拡張などが考えられる。

謝辞 本研究の一部は，科学研究費補助金（基盤研究(B)）「Web 上からの母語話者/非母語話者英語論文コーパスの作成・公開とその利用」（代表: 冨浦洋一）（2008～2011 年度），科学研究費補助金（特別研究員奨励費）「テキストマイニングを用いた学習者作文における談話標識の研究」（代表: 小林雄一郎）（2010～2011 年度）によって行われたものである。

参考文献

1) 金森敬文・竹之内高志・村田昇 (2009). 『パターン認識』（R で学ぶデータサイエンス 5）東京: 共立出版.

2) 小林雄一郎・田中省作・冨浦洋一 (2011). 「メタ談話標識を素性とするパターン認識を用いた英語科学論文の質判定」 『人文科学とコンピュータシンポジウム論文集—「デジタル・アーカイブ」再考』(pp. 51-58) 東京: 情報処理学会.

3) Breiman, L. (2001). Random forests. *Machine Learning,* 24, pp. 123-140.

4) 金明哲 (2007).『R によるデータサイエンス—データ解析の基礎から最新手法まで』 東京: 森北出版.

5) Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics,* 19, pp. 313-330.

6) 田中省作・柴田雅博・冨浦洋一 (2011). 「Web を源とした質情報付き英語科学論文コーパスの構築法」 『英語コーパス研究』18, pp. 61-71.

付録: TreeTagger tagset

CC	Coordinate conjunction	TO	*to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb *be*, base form
EX	Existential *there*	VBD	Verb *be*, past tense
FW	Foreign word	VBG	Verb *be*, gerund or present particle
IN	Preposition or subordinate conjunction	VBN	Verb *be*, past particle
JJ	Adjective	VBP	Verb *be*, non-3rd person singular present
JJR	Adjective, comparative	VBZ	Verb *be*, 3rd person singular present
JJS	Adjective, superlative	VH	Verb *have*, base form
LS	List item marker	VHD	Verb *have*, past tense
MD	Modal	VHG	Verb *have*, gerund or present particle
NN	Noun, singular or mass	VHN	Verb *have*, past particle
NNS	Noun, plural	VHP	Verb *have*, non-3rd person singular present
NP	Proper noun, singular	VHZ	Verb *have*, 3rd person singular present
NPS	Proper noun, plural	VV	Verb others, base form
PDT	Predeterminer	VVD	Verb others, past tense
POS	Possessive ending	VVG	Verb others, gerund or present particle
PP	Personal pronoun	VVN	Verb others, past particle
PP$	Possessive pronoun	VVP	Verb others, non-3rd person singular present
RB	Adverb	VVZ	Verb others, 3rd person singular present
RBR	Adverb, comparative	WDT	Wh-determiner
RBS	Adverb, superlative	WP	Wh-pronoun
RP	Particle	WP$	Possessive wh-pronoun
SYM	Symbol	WRB	Wh-adverb