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

小林 雄一郎　大阪大学大学院
田中 省作　立命館大学
冨浦 洋一　九州大学大学院

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，論文内のメタ談話標識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，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することができた。

Pattern Recognition of English Scientific Papers Using Metadiscourse Markers

Yuichiro Kobayashi, University of Osaka
Shosaku Tanaka, Ritsumeikan University
Yoichi Tomiura, Kyusyu University

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 metadiscourse markers. With the accuracy of 81.79% 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，文章内のメタ談話標識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，どのようなメタ談話標識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

英文の質判定や自動採点の研究は，これまでも国内外で行われてきたが，その評価項目は，作文の総語数，異語数，平均文長，あるいは一部の文法項目などに限られていた。それに対して，本研究の特色の 1 つは，論文の質判定にパターン認識の技術を応用していることである。

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。

英語科学論文の自動分類に関して，[2] は，母語話者による論文と非母語話者による論文から品詞 *n*-gram（$n = 3 \sim 8$）を算出し，ベイズ識別と仮説検定に基づいて，両者を 92.5%の精度で分類している。自然言語処理の分野において，*n*-gram モデルは，機械翻訳，形態素解析，構文解析，文書分類などを含む多くの課題に対して有効であることが知られている。その一方で，*n*-gram モデルには，データスパースネスやゼロ頻度問題が不可避であり，多くの場合，何らかのスムージングを行わなければならない [3]。

また，[2] のような品詞 *n*-gram モデルは，テキストにおける構文情報や文法情報に注目する方法である。

それに対して，テキストの談話情報や語彙情報に注目した研究として，[4] が挙げられる。この研究では，[2] と同様に母語話者による論文と非母語話者による論文からメタ談話標識（3.2.節を参照）の頻度を算出し，ランダムフォレストを分類器として用いて，88.74%の精度で分類している。本研究は，[4] における素性と分類器の組み合わせを用いて，英語科学論文の質判定を行う。

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，[5] 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，[5] 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

図 1 は，この過程を視覚化したものである。そして，[6] 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

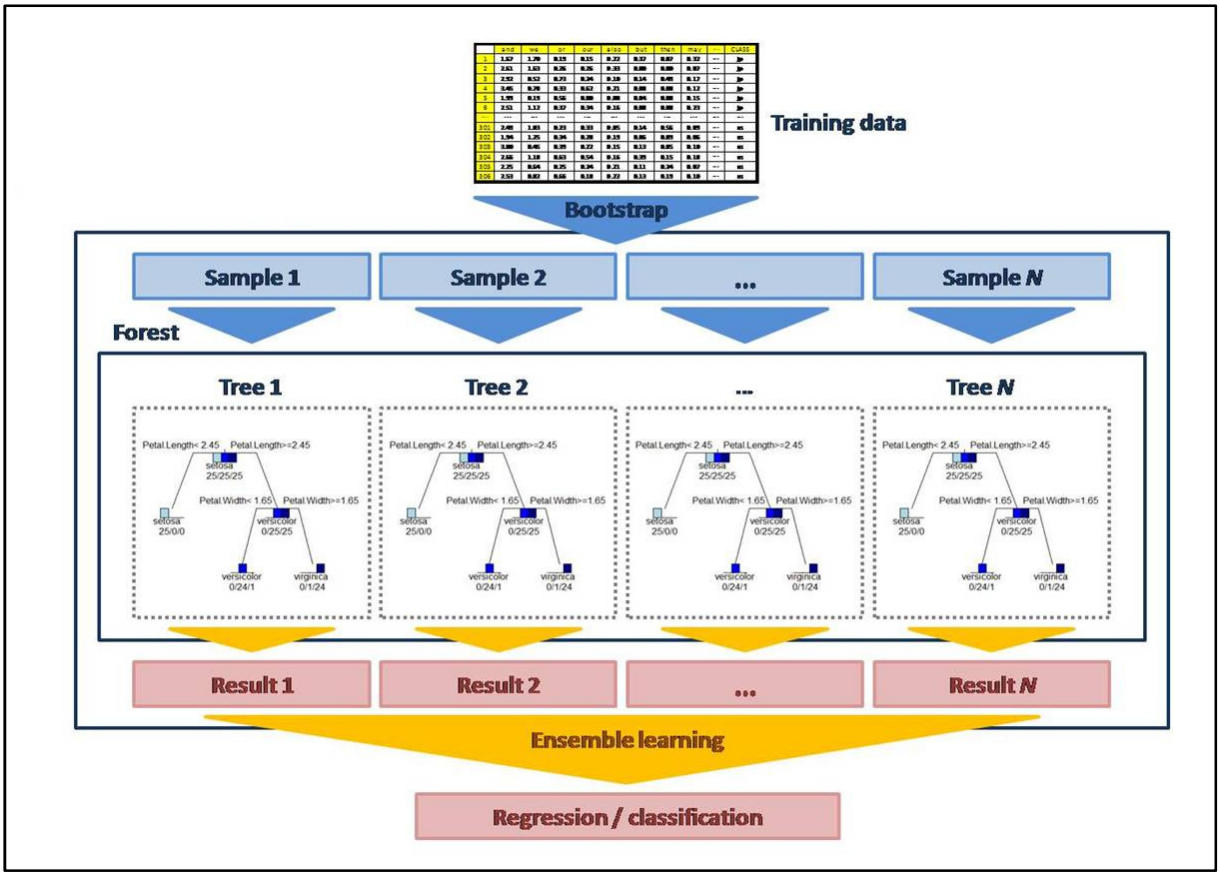

図 1: ランダムフォレストの手順

- 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）

3.2. 素性

本研究で実験に用いる素性は，メタ談話標識（metadiscourse markers, MDM）と呼ばれる談話表現である。メタ談話標識とは，「書き言葉，あるいは話し言葉のテキストにおける言語要素で，命題内容に何かを付け加えるものではなく，聞き手や読み手が与えられた情報を系統立て，解釈し，評価することを助けるためのもの」[7] と定義されている。

また，メタ談話標識の研究において，最もよく使われる枠組みは，約 400 種類の談話表現を網羅的に収録した [8] である。このリストは，[7] など先行研究をベースとして，表 1 のような 10 種類のカテゴリーに分類される約 400 種類の談話表現を網羅的に収録したものである（個々の表現のリストは，[8] を参照）。また，これは，コーパスに基づく統計的研究を想定して作成されたものであり，これまでにアカデミック・ライティングを始め，教科書，学位論文，ビジネスレターなど，様々な言語データの分析で成果を上げている。

表 1: メタ談話標識の意味カテゴリー

Category	*Function*	*Examples*
Interactive resources	***Help to guide reader through the text***	
Transitions (TRA)	Express semantic relation between main clauses	in addtion / but / thus / and
Frame markers (FRM)	Refer to discourse acts, sequences, or text stages	finally / to conclude / my purpose here is to
Endophoric markers (END)	Refer to information in other parts of the text	notes above / see Fig / in section 2
Evidentials (EVI)	Refer to source of information from other texts	according to X / (Y, 1990) / Z states
Code glosses (COD)	Help readers grasp functions of ideational material	namely / e.g. / such as / in other words
Interactional resources	***Involve the reader in the argument***	
Hedges (HED)	Without writer's full commitment to proposition	might / perhaps / possible / about
Boosters (BOO)	Emphasize force or writer's certainty in proposition	in fact / definitely / it is clear that
Attitude markers (ATM)	Express writer's attitude to proposition	unfortunately / I agree / surprisingly
Engagement markers (ENG)	Explicitly refer to or build relationship with reader	consider / note that / you can see that
Self-mentions (SEM)	Explicit reference to author(s)	I / we / my / our

4. 実験

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，[9] 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 2 および Appendix 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。

表 2: 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 語当たり 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 語当たり 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メタ談話標識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. メタ談話標識の抽出

分類実験の前処理として，個々の論文に表れているメタ談話標識の頻度を算出し，“論文×メタ談話標識”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 3: 頻度行列（一部）

	and	but	so	…	us	G/P
1	20.74	0.67	1.34	…	1.34	G
2	26.26	1.22	1.22	…	0.61	G
3	19.14	0.63	0.00	…	0.21	G
…	…	…	…	…	…	…
779	28.05	1.54	2.06	…	0.26	P
780	22.38	2.03	0.23	…	0.00	P
781	13.12	8.11	3.82	…	3.10	P

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。図 2 は，木の数と誤判別率の関係である。横軸が木の数，縦軸が誤判別率をそれぞれ表し，3 種類の線は正例（G 論文），負例（P 論文），OOB（out of bag）の値を表している。因みに，OOB とは，ブートストラップサンプルの 3 分の 1 を評価用として外してモデルを作成し，外した 3 分の 1 を用いて評価を行った結果の誤判別率である。

図 2 を見ると，木の数が 100 以上であれば誤判別率は比較的安定するため，本実験における 200 という木の数が妥当であることが確認される。

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。図 3 は，G 論文と P 論文の分類において，寄与の大きい 30 表現（メタ談話標識）をまとめたものである。縦軸が重要な表現，横軸が重要度（ジニ係数）を表している。

これらの 30 種類の表現には，*on the other hand, also, therefore, but* (TRA), *subsequently, purpose* (FRM), *called, or, such as* (COD), *could, likely, would, perhaps, may, appear, apparent, almost, largely* (HED), *shown, realize, must, shows, evident, found* (BOO), *allow, determine, ensure, order, see* (ENG), *we* (SEM) が含まれている（これらの表現に関しては，次節を参照）。

そして，学習したモデルを評価用データセットに適用した結果，81.79%の精度で G 論文と P 論文を正しく分類することができた（表 4）。

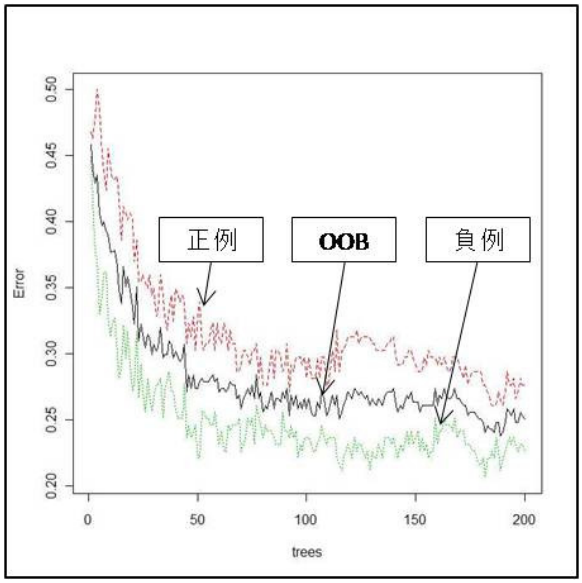

図 2: 木の数と誤判別率の関係

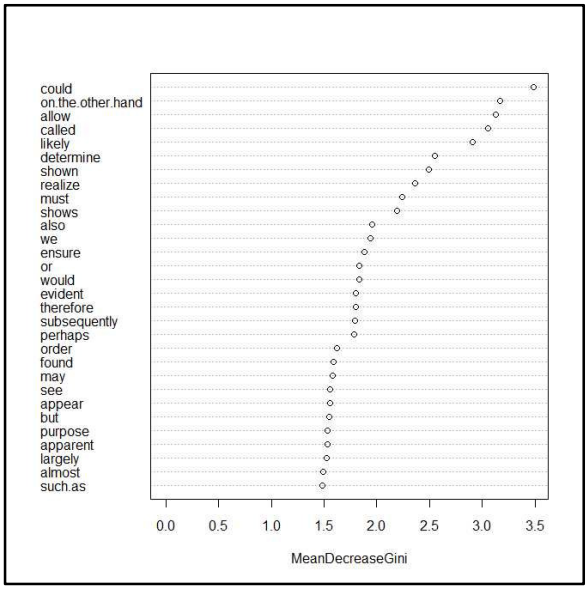

図 3: 素性の重要度

表 4: 分類結果の混同行列

	Good	Poor
Good	154	38
Poor	33	165

4.4. 分類に寄与した素性の分析

本節では，G 論文と P 論文の分類に寄与した素性について，詳しく見ていく。表 5 は，図 3 で挙げられている 30 種類の素性に関して，G 論文と P 論文のどちらに多く生起しているのかを見るために，それぞれの 100 万語あたりの相対頻度 (Freq.) と差異係数 (DC) をまとめたものである。なお，差異係数の算出は以下の式に基づき，その結果は-1 から+1 までの値となる。

$$\text{difference coefficient} = \frac{\text{Freq. of G - Freq. of P}}{\text{Freq. of G + Freq. of P}}$$

表 5 において，差異係数が正の値の表現は G 論文に高い頻度で生起することを表し，負の値の表現は P 論文に高い頻度で生起することを表している。

表 5: 重要な素性の相対頻度と差異係数

	MDM	Freq. (PMW)		DC
		Good	Poor	
TRA	on the other hand	129.96	339.25	-0.45
	also	1792.34	2163.05	-0.09
	therefore	472.63	848.38	-0.28
	but	2124.94	1448.45	0.19
FRM	subsequently	120.83	16.36	0.76
	purpose	120.20	193.93	-0.23
COD	called	297.99	554.36	-0.30
	or	3763.41	2932.99	0.12
	such as	847.71	1067.81	-0.11
HED	could	814.67	483.62	0.25
	likely	368.47	153.03	0.41
	would	1377.93	717.97	0.31
	perhaps	149.78	26.47	0.70
	may	1553.82	1105.34	0.17
	appear	207.05	107.79	0.32
	apparent	107.93	37.05	0.49
	almost	182.19	245.90	-0.15
	largely	69.23	33.68	0.35
BOO	shown	778.49	1248.26	-0.23
	realize	25.80	92.87	-0.57
	must	765.58	478.81	0.23
	shows	578.04	1162.61	-0.34
	evident	39.65	13.47	0.49
	found	537.45	495.17	0.04
ENG	allow	213.34	129.45	0.24
	determine	268.73	171.79	0.22
	ensure	74.89	47.64	0.22
	order	657.02	926.33	-0.17
	see	825.06	689.58	0.09
SEM	we	5534.99	7252.35	-0.13

ただ，表 5 を見る際に注意しなければならないのは，差異係数は，実験データにおける G 論文全体と P 論文全体の頻度から求められたものであり，個々の論文における頻度から求められたものではないということである。従って，単純に差異係数が大きいからといって，分類実験に寄与しているとは限らない。ここでは，前述のように，前節の分類実験において重要であることが確認された 30 種類の素性に関して，G 論文と P 論文のどちらに多く生起しているのかを見るために差異係数を示している。

表 5 を見ると，これらの 30 表現のうち，21 表現が論文の読み手を議論に巻き込んでいく interactional resources (HED, BOO, ENG, SEM) であることが分かる。とりわけ，30 表現の中に HED が 9 種類も含まれていることは特筆に値する。

メタ談話標識の研究において，HED は，書き手の習熟度を如実に表す項目の 1 つであることが知られている [10]。言うまでもなく，学術的な議論を展開するにあたって，書き手の懐疑 (doubt) や確信 (certainty) を適切に表現することは非常に重要なことである。特に，HED は，自らの主張 (claim) を弱めることで，逆に議論 (argument) そのものを強くする働きを持っており [11]，それを巧みに使いこなせるかどうかが優れた書き手と稚拙な書き手を分ける“rhetorical gap” [12] である。

図 4 は，表 5 に含まれている 9 種類の HED に関して，G 論文と P 論文における相対頻度 (100 万語あたり) を視覚化したものである。横軸が HED，縦軸が相対頻度を表している。また，図中の黒いバーと灰色のバーは，それぞれ G 論文における頻度と P 論文における頻度を表している。

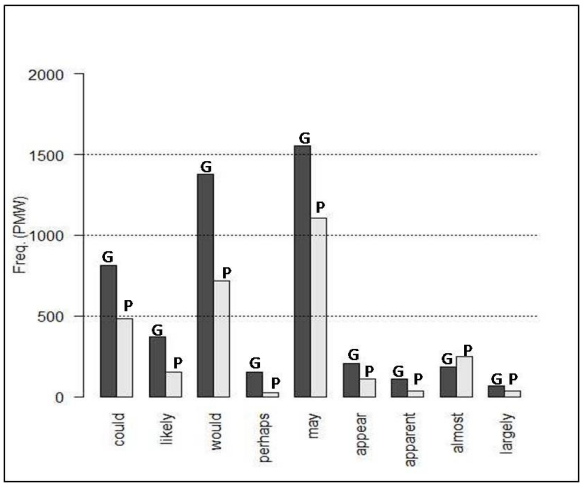

図 4: HED の相対頻度

この図を見ると，*almost* を除く 8 種類の HED は，G 論文に高い頻度で生起している。このことからも，HED が質の高い科学論文に不可欠な表現であることが分かる。

HED 以外の意味カテゴリーでは，ENG が G 論文に特徴的で，TRA と SEM が P 論文に特徴的である。その中で，TRA は，稚拙な書き手が過剰使用する言語項目であることがすでに知られている [13]。また，SEM の過剰使用に関しては，本研究の実験データにおける P 論文の書き手に日本人が多く含まれていることと関係がある。[14] が報告しているように，他の様々な言語を背景に持つ書き手（母語話者も含む）と比べて，日本語を背景とする書き手は，*I* や *we* などの SEM を極めて高い頻度で用いることが知られている。

4.5. 誤分類された論文の分析

ランダムフォレストで誤って分類された論文に関して，表現上の質情報や添削者のコメントといったメタ情報を参照したところ，誤分類された論文の大半は，複数の添削者の間で英文中の誤りの数が一致しない論文や，論文として出版するに相応しい形式かどうかに関する評価が分かれている論文であった。この点については，今後詳しく調査する。

4.6. 分類器の精度比較

近年，パターン認識の分野で数多くの分類器が提案されている。そこで，本研究で用いたランダムフォレストの有効性を確認するため，線形判別分析（LDA），ナイーブベイズ（NB），決定木（CART），*k* 最隣法（KNN），ニューラルネットワーク（NNET），学習ベクトル量子化（LVQ），サポートベクターマシン（SVM），バギング（BAG），ブースティング（BOO），ランダムフォレスト（RF）の 10 種類の分類器の精度比較を行った。

図 5 は，その結果を視覚化したものである。横軸が分類器，縦軸が分類精度（百分率）を表している。

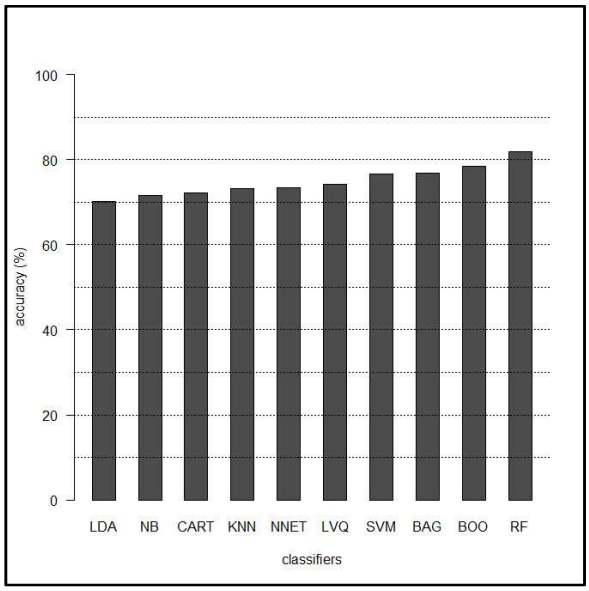

図 5: 分類器の精度比較

この図を見ると，アンサンブル学習を用いた分類器（BAG，BOO，RF）の精度が高く，その中でランダムフォレストによる分類精度が最も高いことが分かる。勿論，分類精度は用いたテキストや素性に依存するために絶対的なものではないが，少なくとも本研究の分類実験においてはランダムフォレストが最も優れていることが分かった。

5. おわりに

本研究では，英語科学論文における談話表現に注目し，そ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 81.79%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，文法情報や構文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。

註

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] 冨浦洋一・青木さやか・柴田雅博・行野顕正 (2009). 「仮説検定に基づく英文書の母語話者性の判別」『自然言語処理』 16 (1), pp. 25-46.

[3] 言語処理学会 (編) (2009). 『言語処理学事典』 東京: 共立出版.

[4] 小林雄一郎・田中省作・冨浦洋一 (2011). 「ランダムフォレストを用いた英語科学論文の分類と評価」『情報処理学会研究報告』 2011-CH-90, pp. 53-68.

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

[6] 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.

[7] Crismore, A., Markkanen, R., & Steffensen, M. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish students. *Written Communication,* 10, pp. 37-71.

[8] Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. New York: Continuum.

[9] 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.

[10] 小林雄一郎 (2010). 「多変量アプローチによる英語学習者のレトリック分析」 田畑智司(編)『統計学的アプローチによるテクスト分析』 (統計数理研究所共同研究リポート245) (pp. 1-22)

[11] Meyer, P. G. (1997). Hedging strategies in written academic discourse: Strengthening the argument by weakening the claim. In R. Markkanen & H. Shroder (Eds.), *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts* (pp. 21-41). Berlin: Walter de Gruyter.

[12] Hyland, K. (1995). The author in the text: Hedging scientific writing. *Hong Kong Papers in Linguistics and Language Teaching*, 18, 33-42.

[13] 小林雄一郎 (2010). 「多変量アプローチで見る英語学習者の接続表現」 田畑智司(編)『電子化言語資料分析研究 2009-2010』 (言語文化共同研究プロジェクト2009) (pp. 3-27)

[14] 小林雄一郎 (2010). 「多変量解析による世界の英語学習者の談話分析」 『人文科学とコンピュータシンポジウム論文集―人文工学の可能性』(pp. 41-48)

Appendix: Assessment Guidelines

Items to input

- Level 1, 2, 3, 4 or 5
- Quality (adequate/inadequate) in all levels
  adequate : Assessor judges the manuscript as having *adequate* English quality to be published in higher-level international journals such as *IEEE transactions* and *ACM transactions.*
  inadequate : Assessor judges the manuscript as requiring proofreading before publishing in such higher-level international journals.
- The issue of "eloquence" in style and tone of writing is addressed in levels 4 and 5 by the "quality" radio buttons, in which adequate/inadequate is determined in large part by the paper's level of eloquence.
- Readability (good/poor) in levels 1 and 2.
- **Number of native-speaker errors in levels 1, 2 and 3.**

Assessor Level Check Guidelines

Level 5:

Basically no need for proofreading changes as defined by the Error Guidelines (no non-native speaker errors).

*Possible exceptions:

- A very small number of typos.
- An editor might make some punctuation changes, but these are few and not major (e.g., not errors that result in run-on sentences or possible misunderstandings).
- Zero or an average of less than 1 error per 250 words: typos or "native-speaker careless errors" in grammar.
- Possibly some minor stylistic "errors" that are not truly errors as defined in our guidelines, but more preference-based.

Level 4:

-1 to 2 (per 250 words) errors or typos (e.g. "the the") that a native speaker might make, and no non-native-speaker errors.

Level 3:

-1 to 2 non-native-speaker errors (per 250 words).

Level 2:

-3 to 6 non-native-speaker errors (per 250 words).

Level 1:

- 7 or more non-native-speaker errors (per 250 words).