Vol.2011-CH-90 No.6
2011/5/21

# 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

小林雄一郎[†]　田中省作[††]　冨浦洋一[†††]

近年，非母語話者が書く英語科学論文（以下，論文）と母語話者の論文の分類を通して，両者の様々な言語的差異を抽出することが試みられている。本研究では，まず，論文中の談話表現に注目し，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する。その分類精度は 88.74%で，類似研究の中でも高いものであった。そして，構築された分類器の素性を分析することで，母語話者の論文と非母語話者の論文にそれぞれ特徴的な談話表現を抽出する。

# Classification and Assessment of English Scientific Papers Using Random Forests

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

The aim of the present study is to assess English scientific papers through random forests. The explanatory variables are the frequencies of metadiscourse markers. With the accuracy of 88.74% over the entire set of corpus texts, this study clarifies the difference between papers by native speakers and those by non-native speakers.

## 1. はじめに

科学論文やビジネスといった特定分野の英語（English for Specific Purposes, ESP）では、基本的な語彙や表現が一般的な英語（English for General Purpose, EGP）と大きく異なることが知られており，英語教育の分野でも盛んに研究されている。その中でも，自然科学に関する学術文書の作成において，ESP の重要性が高い。その理由としては，1) が指摘しているように，研究論文をはじめ，自然科学分野での英語使用の割合が極めて多いことが挙げられる。

従来の日本の英語教育では語彙や文法は重要視されてきたため，研究者による科学論文であればスペリングや文法に関する稚拙な誤りは比較的少ない。しかしながら，英語による科学論文の書き方に不慣れな日本人が書いた国際会議などの論文と熟達者が書いた論文との間に何かしらの大きな質的隔たりがあることも事実である。では，その隔たりとは何か。それらを部分的にでも明らかにしていくことは，日本の専門英語教育において非常に重要な課題の 1 つである。

このような課題を解決するべく，本研究では，対照中間言語分析（Contrastive Interlanguage Analysis, CIA）という方法論を用いる。これは，異なる言語を比較するという伝統的な意味での対照分析（Contrastive Analysis, CA）ではなく，2) が述べているように，「比較可能な状況において，同一言語の非母語話者と母語話者がそれぞれどのように振舞うかを比較・対照する」ものである。また，3) が定義しているように，中間言語とは，目標言語の完全な習得にはまだ達していない非母語話者の言語を指す。4) によれば，主として，CIA には 2 種類の比較研究が存在する。

- NL vs. IL　母語話者（Native Language, NL）の言語と中間言語（Interlanguage, IL）の比較
- IL vs. IL　異なる中間言語の比較

NL と IL の比較は，非母語話者の言語的特徴を明らかにしようとするものである。この分析モデルは，5) が述べているように，次のような研究を可能にする。

- 母語話者と比較した場合，非母語話者が有意に過剰使用，または過少使用しがちな目標言語の言語的特徴にはどのようなものがあるか

† 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

- 非母語話者の目標言語における振舞いには，母語からの影響（母語転移）がどの程度あるか
- 非母語話者が目標言語で十分に表現できない場合は，「回避ストラテジー」を使うが，その言語領域にはどのようなものがあるか
- 非母語話者が母語話者的に運用したり，あるいは非母語話者的に運用したりする言語領域にはどのようなものがあるか

本研究は，このような CIA の考え方に基づいて，母語話者と非母語話者によって書かれた英語科学論文を比較するものである。具体的には，メタ談話標識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する。そして，構築された分類器の素性を分析することで，非母語話者の論文に特徴的な談話表現を抽出する。

## 2. 関連研究

まず，母語話者と非母語話者による英語科学論文を計量的に比較した研究として，6) と 7) を挙げることができる。

6) は，品詞 *n*-gram 分布に基づく論文分類モデルを用いて，非母語話者の論文に存在する「非母語話者性」（統語的には誤りではないが，不自然さに強く関連する品詞レベルの要因）を抽出している。その結果，非母語話者の論文では，名詞による名詞の修飾・重出，現在分詞による名詞の修飾，関係節（先行詞主格）による後置修飾，*to* を除く名詞の後置修飾，文頭の前置詞，文頭の接続詞・副詞（連結語），受動態を過剰使用し，形容詞の限定用法，過去分詞による名詞の後置修飾，*to* 句による後置修飾，形容詞の重出，文頭の名詞句，主語・述語間の副詞，副詞節前文（受動態）の分詞化，副詞節後文の分詞化，前置詞句における名詞句の省略を過少使用すると報告している。

また，7) 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*=3～8）を抽出し，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。

そして，英語科学論文をデータとして分析したものではないが，母語話者と非母語話者による英文における談話表現を統計的に比較した研究として，8) と 9) を挙げることができる。これらは，談話分析の観点からメタ談話標識を素性とし，判別分析と回帰木を用いて，母語話者と非母語話者の大学生による英作文を 92%の精度で分類している。さらに，分類に寄与した素性を分析した結果，書き手の可視性（self-mentions），心的態度（hedges, boosters），テキスト構造に関わる談話表現（frame markers）の使用法に両者の違いが見られたと報告している。

## 3. 実験の手順

### 3.1 実験データ

本研究で用いるデータは，母語話者（アメリカ人，us）と非母語話者（日本人，jp）が書いた information technology（コンピュータ・通信・情報処理などの情報学分野）と industrial technology（機械・化学・電気などの工学分野）に関する英語科学論文である。表 1 は，実験データの論文数（*n*），総語数（tokens），異語数（types）をまとめたものである。

表 1　実験データの概要

| | information technology | | industrial technology | | overall |
|---|---|---|---|---|---|
| | jp | us | jp | us | |
| *n* | 151 | 155 | 179 | 179 | 664 |
| tokens | 745597 | 1212999 | 915208 | 1547475 | 4421279 |
| types | 23265 | 33239 | 30243 | 45148 | 75957 |

このデータは，教育機関の web サイトで個人的に公開されているものを，日本（jp）とアメリカ（us）のドメイン別に収集したものである。jp ドメインであるからといって著者が必ずしも日本語を母語としているとは限らず，us ドメインについても同様である。そこで，このように収集した論文に対して，第一著者が jp ドメインについては日本人らしい名前であること，us ドメインについては少なくとも日本人らしい名前ではないことを人手で確認し，フィルタリングしたのが表 1 の論文データである。[1]

さらに，このデータには英語を母語とした英文添削の専門家によって各論文の表現上の質的評価やコメントなどの情報が付与されている。表現上の質的評価とは，もちろん内容（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」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」によって規定されている。その詳細に関しては，10) を参照されたい。

なお，本研究では information technology と industrial technology の分野ごとに上記 jp／us の分類を行う際に質情報は活用していない。従って，jp ドメインではあるが十分な質を有するもの，us ドメインであるが不十分なものも含まれている可能性がある。

[1] jp ドメインで日本人らしい名前であっても，英語を母語とした著者である可能性，あるいはバイリンガルである可能性もある。また，us ドメインについても同様である。従って，完全に母語を特定したものではないが，特に jp ドメインについては，かなりの割合が日本語を母語とした著者の論文となっていることが期待される。

### 3.2 実験手法

本研究で用いる分類手法は，2001 年に 11) によって提案された，ランダムフォレストである。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。

- 与えられたデータセットから，B 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

その長所として，12) は，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群の個体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
- 群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（個体の類似度の計算など）

また，言語分析にランダムフォレストを用いた例としては，日本語の著者推定をした 13)，日本語の政治的談話を分析した 14)，女性シンガーソングライターの歌詞を分類した 15) などが知られている。なお，具体的な計算に関しては，統計処理環境 R の randomForest パッケージを使用する。

### 3.3 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」16) 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 17) である。このリストは，16) など先行研究をベースとして，表 2 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，17) を参照されたい）。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。

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，母語話者／非母語話者という群情報を判定する分類実験を行う。

表 2　メタ談話標識の意味カテゴリー

| ***Category*** | ***Function*** |
|---|---|
| ***Interactive resources*** | ***Help to guide reader through the text*** |
| Transitions (TRA) | Express semantic relation between main clauses |
| Frame markers (FRM) | Refer to discourse acts, sequences, or text stages |
| Endophoric markers (END) | Refer to information in other parts of the text |
| Evidentials (EVI) | Refer to source of information from other texts |
| Code glosses (COD) | Help readers grasp functions of ideational material |
| ***Interactional resources*** | ***Involve the reader in the argument*** |
| Hedges (HED) | Without writer's full commitment to proposition |
| Boosters (BOO) | Emphasize force or writer's certainty in proposition |
| Attitude markers (ATM) | Express writer's attitude to proposition |
| Engagement markers (ENG) | Explicitly refer to or build relationship with reader |
| Self-mentions (SEM) | Explicit reference to author(s) |

## 4. 結果と考察

### 4.1 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抽出し，表 3 のような論文×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 1 万語あたりの相対頻度に変換する。

表 3　頻度行列の一部（informational technology）

| | and | we | or | our | also | but | then | may | … | CLASS |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.67 | 1.70 | 0.19 | 0.15 | 0.22 | 0.37 | 0.07 | 0.32 | … | jp |
| 2 | 2.61 | 1.63 | 0.26 | 0.26 | 0.33 | 0.00 | 0.00 | 0.07 | … | jp |
| 3 | 2.92 | 0.52 | 0.73 | 0.24 | 0.10 | 0.14 | 0.49 | 0.17 | … | jp |
| … | … | … | … | … | … | … | … | … | … | … |
| 304 | 2.66 | 1.18 | 0.63 | 0.54 | 0.16 | 0.39 | 0.15 | 0.18 | … | us |
| 305 | 2.25 | 0.64 | 0.25 | 0.24 | 0.21 | 0.11 | 0.24 | 0.07 | … | us |
| 306 | 2.53 | 0.82 | 0.66 | 0.18 | 0.22 | 0.13 | 0.19 | 0.10 | … | us |

### 4.2 分類実験 1: Information Technology

まず，information technology（INF）に関する 306 本（jp 151 本，us 155 本）の論文を 2 分割し，半数の 153 本を学習用のデータセットとし，残りの 153 本を評価用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 500 とする。図 1 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，負例，OOB（out of bag）の値を表している。この図を見ると，木の数が 300〜400 に達するところで誤判別率は比較的安定するため，本実験における 500 という木の数は妥当であることが確認される。

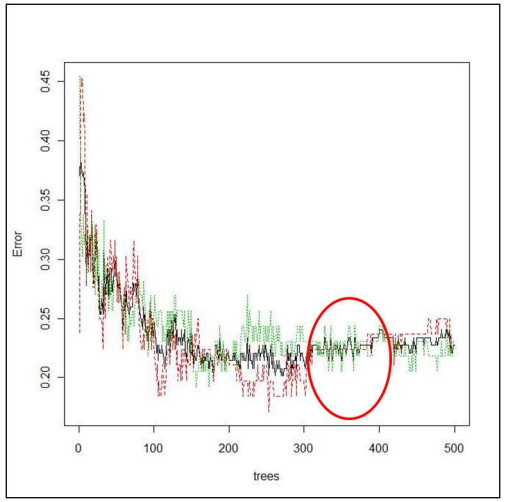


図 1　木の数と誤判別率の関係（INF）

次に，学習用データセットから構築したモデルを見てみよう。図 2 は，母語話者による論文と非母語話者による論文の分類において，寄与の大きい表現（メタ談話標識）の上位 30 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 種類の表現の中で，母語話者の方が高い頻度で用いていたものは *or, allow, increase, could, overall, still, use, determine, must, appropriate, back to, likely, largely, further, while, perhaps, may, notice, indeed, again, important, would, first, even, about, set* の 26 表現であり，非母語話者の方が高い頻度で用いていたものは *on the other hand, shows, although, moreover* の 4 表現であった。図 3 は，非母語話者による *on the other hand* の使用例（一部）である。

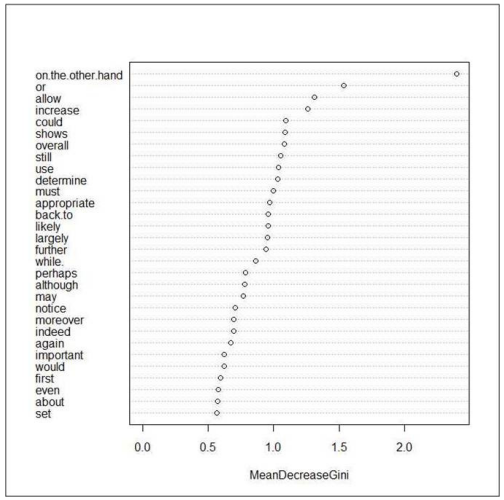


図 2　素性の重要度（INF）

| | | | |
|---|---|---|---|
| 1 | have a large spread around the true position. If, | on the other hand | , the intensity greatly varies around that point, |
| 2 | vices and environments. For internal uncertainty, | on the other hand | , there is no way of increasing the accuracy excep |
| 3 | degrades when parameter manipulations were large. | On the other hand | , sophisticated methods based on iterative procedu |
| 4 | space. Breslauer [Bre98] reduced it to O(n) time. | On the other hand | , our algorithm constructs a CDAWG for a (normal) |
| 5 | (S ) for S = {aaab$, aac$, aa$, abc$, bab$, ba$} | On the other hand | , for a set S = {w1 $1 , . . . , wk $k } where $i |
| 6 | reas failed to obtain a mapping near the optimum. | On the other hand | , the stochastic MDS network and the fast SA can u |
| 7 | nit directly through a one-to-one correspondence. | On the other hand | , the SOM suers some distortion in the output arra |
| 8 | agnanti-Orlin [2]. ) For traditional maximum ows, | on the other hand | , Edmonds-Karp [11] suggested two polynomialtime i |
| 9 | bottom plate and the lower part of the end plate. | On the other hand | , as is seen in Figure 4(b), the ux goes out from |
| 10 | on the coach side are sinkers of the magnetic ux. | On the other hand | , as is seen in Figure 4(b), the magnetic ux goes |
| 11 | oximation of the robust counterpart is crude. But | On the other hand | , from the engineering point of view, one might we |
| 12 | they have larger delay than other input patterns. | On the other hand | , the input of "hurrah", which has no common part |
| 13 | ations will be run on a local or remote computer. | On the other hand | , simulation methods can be developed without rega |
| 14 | he ExperimentalElement object is the application. | On the other hand | , a simulation using OpenSees only needs to obtain |
| 15 | n would have to be modied. The proposed approach, | on the other hand | , introduces typed articial identiers or surrogate |
| 16 | tion to control versioning. The present approach, | on the other hand | , treats versioning as a more explicit programming |
| 17 | e updated destructively. In the present approach, | on the other hand | , the order of consideration is reversed. Basicall |
| 18 | , may generate multiple copies of a single array. | On the other hand | , the database versioning simply locks the state a |
| 19 | nment where data of evidence changes dynamically. | On the other hand | , as parallel distributed systems become important |
| 20 | hey proceed to the upper part. The cold material, | on the other hand | , being warmed by the gas as they go down. In plan |

図 3　非母語話者による *on the other hand* の使用例（一部）（INF）

Vol.2011-CH-90 No.6
2011/5/21

R の randomForest パッケージは，MDSplot と呼ばれる多次元距離尺度法による視覚化機能を提供している。図 4 は，その機能を用いて，個体間の類似度を計算した結果を視覚化したものである。図中の△は母語話者による論文，○は非母語話者による論文をそれぞれ表している。また，図中で近い位置に布置された個体は類似した性質を持ち，遠い位置に布置された個体は異なる性質を持っている。図 4 を見ると，図中の右に母語話者（us）の論文，左に非母語話者（jp）の論文が布置されており，母語話者の論文と非母語話者の論文がうまく分離されている。

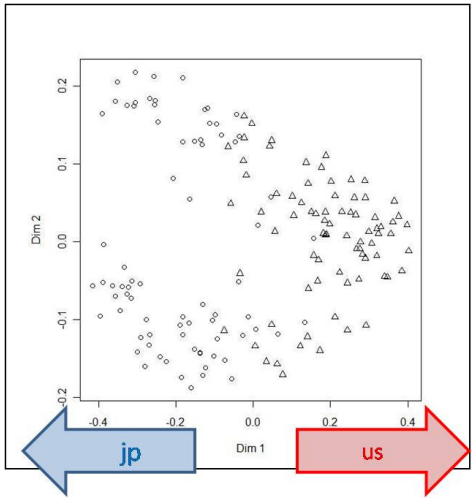


図 4　多次元距離尺度法による視覚化（INF）

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，88.74%という高い精度で母語話者による論文と非母語話者による論文を分類することができた。

### 4.3 分類実験 2: Industrial Technology

続いて，industrial technology（IND）に関する 358 本（jp 179 本，us 179 本）の論文を 2 分割し，半数の 179 本を学習用のデータセットとし，残りの 179 本を評価用データセットとする。

INF の場合と同様，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を取り，ランダムフォレストに含まれる木の数は 500 とする。図 5 は，木の数と誤判別率の関係である。この図を見ると，木の数が 300〜400 に達するところで誤判別率は比較的安定するため，本実験における 500 という木の数は妥当であることが確認される。

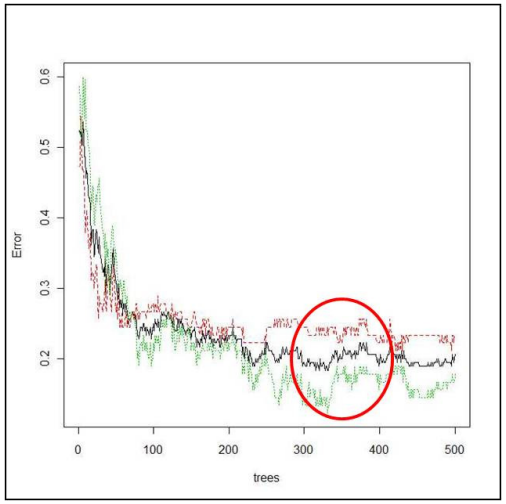


図 5　木の数と誤判別率の関係（IND）

次に，学習用データセットから構築したモデルを見てみよう。図 6 は，母語話者による論文と非母語話者による論文の分類において，寄与の大きい表現の上位 30 をまとめたものである。これらの 30 種類の表現の中で，母語話者の方が高い頻度で用いていたものは *allow, increase, result in, overall, determine, may, state, evident, while, or, largely, further* の 12 表現であり，非母語話者の方が高い頻度で用いていたものは *we, therefore, such as, almost, moreover, shows, called, because, our, table, in other words, order, according to, third, on the other hand, important, however, first* の 18 表現であった。図 7 は，非母語話者による *we* の使用例（一部）である。

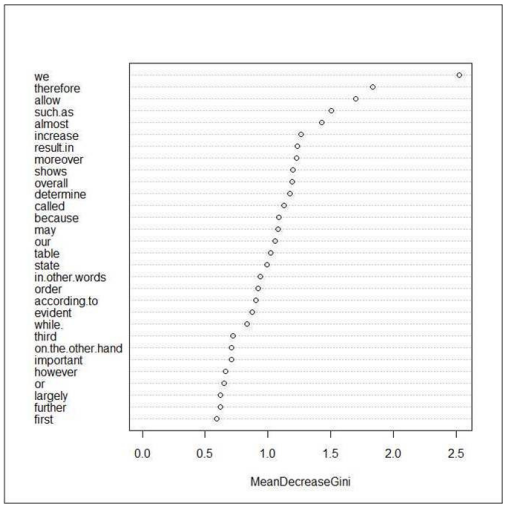


図 6　素性の重要度（IND）

| | | | |
|---|---|---|---|
| 1 | trees are expected to provide valid information. | We | focus on the evolution of the developmental patte |
| 2 | vity in vertebrate smooth muscle (Horowitz et al. | We | reconstructed a tree of the EF-hand superfamily ( |
| 3 | tries found. Number of compared amino acid sites. | We | used only sites which do not contain any gaps. Th |
| 4 | in (Cheney, Riley, and Mooseker 1993; Cope et al. | We | used class II myosin heavy chain to reconstruct a |
| 5 | yosin heavy chain to reconstruct a gene tree, and | we | used other classes to root this subfamily. Actin |
| 6 | tomyosin ATPase reaction in solution (Sugi 1993). | We | reconstructed a phylogenetic tree of the whole ac |
| 7 | es (Molkentin and Olson 1996; Yun and Wold 1996). | We | used the MASH family as an outgroup to root the M |
| 8 | base release 52.00. To identify each gene family, | we | chose a sequence which obviously belongs to the f |
| 9 | abase. The query sequences are listed in table 2. | We | conducted multiple alignments of each protein fam |
| 10 | 994) and removed sites at which any gaps existed. | We | used the neighbor-joining method (Saitou and Nei |
| 11 | ne was not available from the current literature, | we | did not map the gene to any tissue class. For any |
| 12 | class. For any case that was not described here, | we | omitted any correspondence bet |
| 13 | ailable data was not sufcient for our objectives, | we | provided exceptions: 1. When genes that shared th |
| 14 | tissue classes because of evolutionary interest. | We | thus reconstructed phylogenetic trees having tiss |
| 15 | Base (Ashburner and Drysdale 1994) was also used. | We | have prepared a |
| 16 | mpossible to know each correspondence among them. | We | thus developed an algorithm to superimpose multip |
| 17 | tree are different from those of the three genes | we | discusses before. There are two gene duplications |
| 18 | uced by Moncrief, Kretsinger, and Goodman (1990). | We | therefore used the maximum-likelihood method as f |
| 19 | odman's (1990) tree. Therefore, the topology that | we | consider to be reasonable can be represented in t |
| 20 | is inconsistency may be caused by short branches. | We | added 12 new invertebrate sequences to Moncrief, |

図 7　非母語話者による *we* の使用例（一部）（IND）

また，図 8 は，計量的多次元距離尺度法を用いて，個体間の類似度を計算した結果を視覚化したものである。図中の△は母語話者による論文，○は非母語話者による論文をそれぞれ表している。この図では，図中の右に母語話者（us）の論文，左に非母語話者（jp）の論文が布置されている。

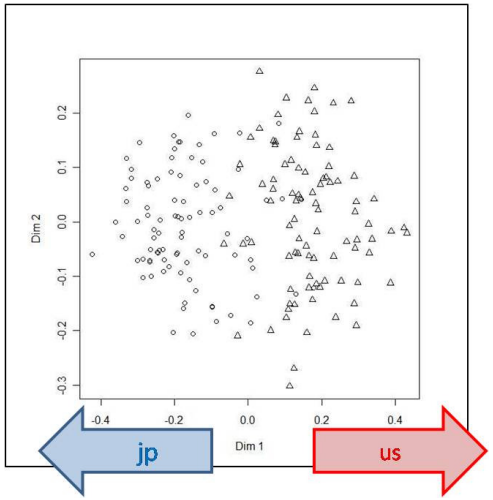


図 8　多次元距離尺度法による視覚化（IND）

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，81.56%の精度で母語話者による論文と非母語話者による論文を分類することができた。

### 4.4 母語話者と非母語話者にそれぞれ特徴的な談話表現

前掲の図 2 と図 6 において，母語話者による論文と非母語話者による論文の分類において寄与の大きい表現を示した。そして，その中の 14 種類の表現は，INF と IND という 2 つの分野で（上位 30 以内に）共通して現れていた。以下は，その 14 表現を，「INF と IND の両方で母語話者（us）が多く使っている表現」，「INF と IND の両方で非母語話者（jp）が多く使っている表現」，「一方で母語話者が多く，他方では非母語話者が多く使っている表現」の 3 タイプに分けて，まとめたものである。

- INF と IND の両方で母語話者（us）が多く使っている表現
  *or, allow, increase, overall, determine, largely, further, while, may*

- INF と IND の両方で非母語話者（jp）が多く使っている表現
  *on the other hand, shows, moreover*
- 一方で母語話者が多く，他方では非母語話者が多く使っている表現
  *important, first*

まず,「INF と IND の両方で母語話者（us）が多く使っている表現」を見ると，*allow*, *increase*, *determine* という 3 つの動詞が含まれている。18) などが報告しているように，習熟度の低い書き手の英語は名詞中心であり，習熟度の高い書き手の英語は動詞中心である。母語話者の論文に動詞が特徴的であることは，以上のような書き手の習熟度と関係している。また，19) は，母語話者が非母語話者と比べて，*largely* のような*-ly* 副詞や *while* のような従属接続詞を統計的に多く使用することを示している。さらに，8) や 9) などが指摘しているように，母語話者が非母語話者よりも *may* のような緩衝表現（hedges）を高い頻度で用いることが知られている。

次に,「INF と IND の両方で非母語話者（jp）が多く使っている表現」を見ると，*on the other hand* と *moreover* という接続表現が含まれている。20) が報告しているように，母語話者の英語と比べて，日本人英語学習者は接続表現を過剰使用する傾向がある。また，*shows* という動詞が含まれているのは，英語科学論文において図表や数式でデータを示す（*shows*）ことが不可欠であることと関係している。8) で言及されているように，母語話者の論文では，同一表現の繰り返しを避け，様々な動詞が巧みに使い分けられている。その一方，非母語話者の論文では，そのような使い分けが相対的に少ないため，*show* や *think* のような基本動詞が相対的に多く使われることになる。

このように，今回の実験では，言語学的に有意味な表現が重要表現として抽出されている。

## 5. 分類器の精度比較

近年，機械学習の分野で数多くの分類手法が提案されている。本節では，データマイニングや自然言語処理の分野で用いられることの多いナイーブベイズ（NB），回帰木（CART），ニューラルネットワーク（NNET），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 7 つの分類器の精度比較を行う。計算には，統計解析環境 R の e1071 パッケージ（NB），mvpart パッケージ（CART），nnet パッケージ（NNET），kernlab パッケージ（SVM），adabag パッケージ（BAG と BOO），randomForest パッケージ（RF）を使用する。

図 9 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。また，図中の黒いバーと灰色のバーは，それぞれ INF の分類実験の結果と IND の分類実験の結果を表している。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。

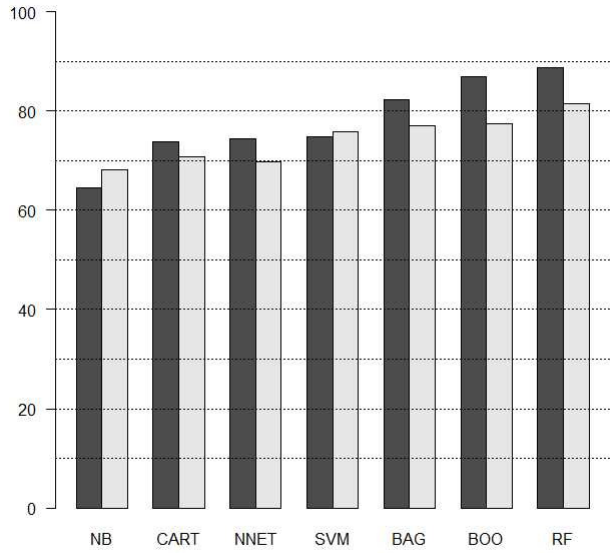


図 9　分類器の精度比較

## 6. 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 88.74%で，類似研究の中でも高いものであった。そして，構築された分類器の素性を分析し，母語話者

の論文と非母語話者の論文に特徴的な談話表現を抽出した。

今後の課題としては，以下の 3 つが挙げられる。まず，分類の結果として抽出された素性について，より深い言語学的・言語教育学的考察を行うことである。第 2 に，分類器による分類結果と人間による評価を比較することである。本研究の実験データには，今回は使用しなかったが，英文添削の専門家による評価やコメントなどのメタ情報が付与されている。今後は，分類器で誤分類された論文を中心に，自動分類の結果と専門家の評価の関係を吟味したい。第 3 に，7) で構築された分類器と本研究で構築した分類器を組み合わせることである。前者は，英文における品詞や構文の特徴に基づく分類器であり, 後者は, 英文における語彙や談話の特徴に基づく分類器である。これらを組み合わせることで分類精度が向上し，より多角的に論文を評価することが可能になるかも知れない。

**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) 小山由紀江・水本篤 (2010).「単語連鎖にみる科学技術分野と他分野の英語表現比較」『ESP コーパスからの特徴表現の抽出』(pp. 1-11). 東京: 統計数理研究所.
2) Pery-Woodley, M. M. (1990). Contrastive discourses: Contrastive analysis and a discourse approach to writing. *Language Teaching,* 23, pp. 143-151.
3) Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics,* 10, pp. 209-231.
4) Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.), *Learner English on computer* (pp. 3-18). London: Longman.
5) Leech, G. (1998). Preface. In Granger, S. (ed.), *Learner English on computer* (pp. xiv-xx). London: Longman.
6) 田中省作・藤井宏・冨浦洋一・徳見道夫 (2006). 「NS/NNS 論文分類モデルに基づく日本人英語科学作文の特徴抽出」 『英語コーパス研究』13, pp. 75-87.
7) 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.
8) 小林雄一郎 (2009).「NS/NNS テキスト分類モデルに基づく日本人英作文の特徴抽出」 『人文科学とコンピュータシンポジウム論文集—デジタル・ヒューマニティーズの可能性』(pp. 261-268). 東京: 情報処理学会.
9) 小林雄一郎 (2010). 「回帰木を用いた NS/NNS テキスト分類」 『言語処理学会第 16 回年次大会発表論文集』 (pp. 318-321). 東京: 言語処理学会.
10) 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』13, pp. 61-71.（印刷中）
11) Breiman, L. (2001). Random forests. *Machine Learning,* 24, pp. 123-140.
12) 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.
13) 金明哲・村上征勝 (2007). 「ランダムフォレスト法による文章の書き手の推定」 『統計数理』55(2), pp. 255-268.
14) Suzuki, T. (2009). Extracting speaker-specific functional expressions from political speeches using random forests in order to investigate speakers' political styles. *Journal of the American Society for Information Science and Technology,* 60(8), pp. 1596-1606.
15) 細谷舞・鈴木崇史 (2010). 「女性シンガーソングライターの歌詞の探索的分析」 『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. 195-202). 東京: 情報処理学会.
16) 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.
17) Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing.* New York: Continuum.
18) 小林雄一郎 (2007). 「The NICT JLE Corpus における発達指標の研究—コレスポンデンス分析によるタグ頻度解析」 『言語処理学会第 13 回年次大会発表論文集』 (pp. 486-489). 東京: 言語処理学会.
19) Granger, S., & Rayson, P. (1998). Automatic profiling of learner texts. In Granger, S. (ed.), *Learner English on computer* (pp. 119-131). London: Longman.
20) Narita, M., & Sugiura, M. (2006). The use of adverbial connectors in argumentative essays by Japanese EFL college students. *English Corpus Studies,* 13, pp. 23-42.