E-016

表層情報のみを利用した英語母語話者／非母語話者の文書判別

Distinguishing between texts written by native/non-native speakers based on superficial information

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

1. はじめに

母語話者の英文と非母語話者の英文を判別する技術は，良質なコーパスを構築する際に重要な役割を果たす．特に，Web ページをコーパスとして利用する場合，新聞記事とは異なり，英語の質は保証されないため，母語話者／非母語話者の判別は重要となる．すなわち，Web ページを一般的な英語コーパスとして利用する場合，非母語話者の英文はノイズとして除去する必要がある．逆に，非母語話者コーパスを構築する場合は，非母語話者の英文のみを選び出す必要がある．最近では，非母語話者コーパスや学習者コーパスは，非母語話者の特徴抽出[1, 5, 12]，第二言語習得に関する知見の獲得[11, 13]，誤り検出・訂正[3, 9]などに利用されており，非母語話者の英文に対する需要も高まってきている（非母語話者コーパス（学習者コーパス）の利用可能性については，文献[10]が詳しい）．

このような背景を受け，母語話者／非母語話者の文書を判別する手法が提案されている．藤井ら[6]は，母語話者と非母語話者における品詞 tri-gram の分布の差異を Skew Divergence[8]で定量化し，母語話者／非母語話者を判別する手法を提案している．藤井らは，この手法の判別精度が様々な手法（KL Divergence，SVM，ナイーブベイズ分類器に基づいた手法）より良いことを実験により示している．青木ら[2]は，品詞 n-gram プロファイル[4]と呼ばれる，品詞 n-gram を出現頻度順に並べたリストを利用した手法で，更に高い判別精度を実現している（判別精度 90.0%）．

しかしながら，これらの従来手法には，次の 3 点について改善の余地がある．第一に，これらの従来手法は，品詞列に基づくため，単語に含まれる，母語話者／非母語話者の特徴が欠落してしまう．例えば，“number of the” [1]という表現は，非母語話者に特徴的な表現であり，判別の重要な手がかりとなるが，品詞列に変換した場合，“名詞 前置詞 冠詞”となり，その特徴が失われてしまう．このような単語に含まれる特徴を判別の際に考慮できれば，判別精度の向上につながる．第二に，品詞タガーの解析誤りが挙げられる．通常，品詞タガーは，母語話者の英文用にデザインされている．そのため，非母語話者の書いた誤りを含む文や不自然な文を正しく解析できない可能性が高い．このような誤りや不自然さは，母語話者／非母語話者の判別のための重要な情報であり，正しく解析できることが好ましい．第三に，品詞解析にかかる時間も問題となる．特に，Webページを対象とした場合，判別文書数が非常に多く，品詞解析にかかる時間が大きな問題となる．

これらの問題点を解決するため，本稿では，表層情報のみに基づいて母語話者／非母語話者の文書を判別する手法を提案する．具体的には，単語 n-gram で，n-gram プロファイルを作成し，母語話者／非母語話者の判別を行う．単語 n-gram のみを利用するので，品詞解析の誤りや品詞解析にかかる時間の問題が解決されるのは明らかである．一方で，単語 n-gram を利用した場合，（1）n-gram の種類数が増加する，（2）母語話者／非母語話者の判別に寄与しないノイズが増加するという新たな問題が生じる．本稿では，これらの問題を解決し，単語 n-gram プロファイルに基づいて母語話者／非母語話者の判別を行う手法を提案する．

以下，2. で提案手法を詳細に説明する．3. で評価実験について述べる．4. で実験結果を考察する．

2．提案手法

2.1 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）を母語話者／非母語話者の判別に利用する．プロファイルとは，単語 n-gram を出現頻度の降順に並べたリストのことである．図 1 に非母語話者プロファイルの例を示す．図 1 では，プロファイルの 1 列目と 2 列目が，それぞれ，出現頻度と対応する単語 n-gram である．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 からわかるように，よく使用される n-gram ほど，プロファイルの上位に位置する．

プロファイルは，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，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 n−1 個入れて，文頭と文末の n-gram も抽出できるようにする．全ての単語は小文字に変換する．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．

図 2 に，判別処理の概要を示す．まず，判別対象として与えられた文書からプロファイルを作成する．作成の手順は，母語話者／非母語話者プロファイルの場合と同様である．作成されたプロファイル中の各 n-gram の順位は，書き手が非母語話者の場合，非母語話者プロファイルでの順位と類似していると予想できる．逆に，書き手が母語話者であれば，母語話者プロファイルに類似していると予想できる．そこで，各 n-gram の順位の差を判別に利用する．例えば，図 2 は，“number of the” という n-gram の順位の差（図 2 の d）を示している．“number of the” は非母語話

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

[1]母語話者では“number of 無冠詞”が一般的である．

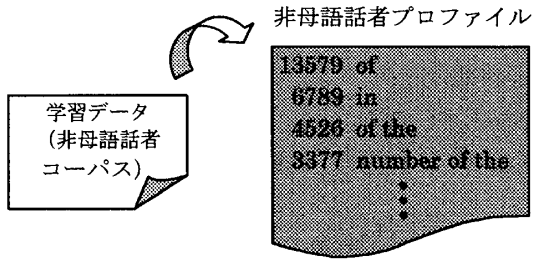

図 1 プロファイルの例

者に特徴的な表現であるので，図 2 では，判別対象プロファイルと非母語話者プロファイルでの順位の差を小さく示している．通常，判別対象の文書は，複数の n-gram を含むので，順位の差も複数得られる．そこで，全ての順位の差について，母語話者／非母語話者プロファイルごとに和をとり，最終的な判別の基準とする．順位の差の和が小さいほうのプロファイルの話者を判別結果とする．

2.2 基本アイデアの問題点と解決策

2.1 で述べた基本アイデアは，単語 n-gram に基づくため，品詞 n-gram では欠落してしまう情報も利用でき，判別精度の向上が期待できる．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，品詞に比べ単語は種類数が多く，プロファイルのサイズが非常に大きくなるという問題点が挙げられる．プロファイルの作成はソートを含むため，n-gram の種類数が増えると計算時間も長くなる．1. で述べたように，大量の文書を含む Web などを対象とした場合，できるだけ短い時間で判別が行えることが望ましい．第二に，プロファイルに含まれるノイズが挙げられる．n-gram の種類数が増えると，判別に寄与しないものも増える．このような n-gram は，他の判別に寄与する n-gram の順位に影響を与えノイズとなる．したがって，判別に寄与する n-gram だけを抽出してプロファイルを作成し，ノイズの影響を減らすことが重要である．

提案手法では，これらの問題を解決するために，前置詞を含む n-gram のみを対象にしてプロファイルを作成する．前置詞は，名詞や動詞などの内容語に比べ，種類数が少ないため，n-gram の種類数も少なくなる．同時に，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，前置詞の左右には，動詞句と名詞句が出現するため，動詞句や名詞句の差異もプロファイルに含めることができる．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，冠詞に関する特徴もプロファイルに含められる可能性がある．

以上を踏まえ，JLE コーパス[7]中での前置詞に関する誤り数を基準として，10 種類[2]の前置詞[3]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．例えば，

the number of the white houses

というフレーズからは，

	of	
number	of	
number	of	the
number	of	the white
the number	of	
the number	of	the
the number	of	the white

の 7 種類の n-gram が得られる．以後，特に断らない限り，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．

2.3 提案手法の定式化と処理の流れ

提案手法を定式化するため，次の記号を導入する．いま，C を，母語話者または非母語話者を表す変数とする．変数 C は，N（母語話者）と NN（非母語話者）を値に取るとする．また，判別対象文書から得られたプロファイル，母語話者／非母語話者プロファイルを，それぞれ P，P_N，P_{NN} 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，$r(in,P) = r(\text{in}, P_N) = r(\text{in}, P_{NN}) = 2$ 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，母語話者／非母語話者プロファイル中に存在しない x については，母語話者／非母語話者プロファイルの最大の順位のうち大きいほうの順位+1 を与えるとする．

このとき，n-gram の順位の差を，

$$d(x,P,P_C) = \left| r(x,P) - r(x,P_C) \right| \qquad (1)$$

で定義する．したがって，判別の基準となる順位の差の和は，

$$s(P,P_C) = \sum_{x \in X} d(x,P,P_C) \qquad (2)$$

で表される．ただし，X は，P 中の n-gram を要素とする集合を表すとする．

式(2)で表される順位の差の和が小さくなるほうの話者を判別結果とする．すなわち，

$$\hat{C} = \arg\min_{C} s(P,P_C) \qquad (3)$$

となる $\hat{C}$ を判別結果とする．

以上をまとめると，提案手法の判別の流れは，次の 3 ステップからなる：

A) 判別対象をプロファイルに変換
B) 式(2)を用いて順位の差の和を計算
C) 式(3)から得られる判別結果を出力

3. 評価実験

3.1 実験条件と実験手順

母語話者／非母語話者を判別する実験を行った．実験条件は，青木ら[2]の実験を参考にして決定した．

[2] 選択した前置詞：about, at, by, for, from, in, of, on, to, with

[3] これらの単語は，文脈に応じて，副詞など他の品詞で用いられることもあるが，提案手法は表層情報のみを利用するので，便宜上全て前置詞として扱う．

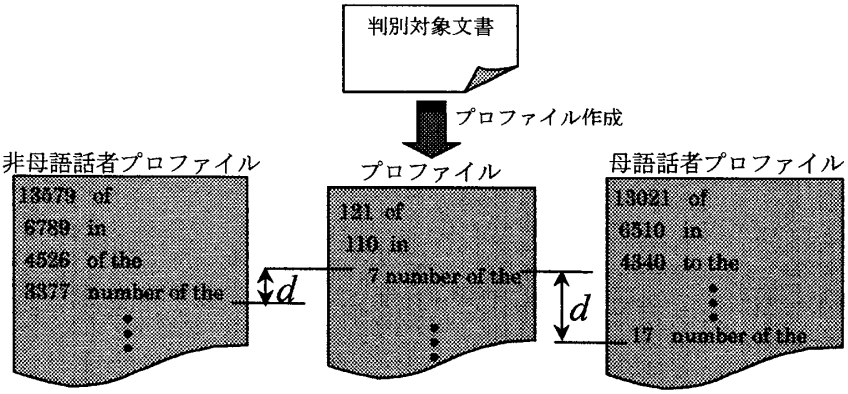

図 2　判別処理の概要

判別対象の文書として，青木ら[2]でも対象とされている科学技術論文を使用した．われわれの実験では，材料化学分野の論文（Journal of Non-Crystalline Solids [4]）を選択した．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，計 240 編選び出した．各論文の最後に記載されている謝辞は予め削除した．評価尺度は，正しく判別できた論文数を判別数で除した判別精度とした．判別精度は，6-fold cross validationにより算出した．学習データとテストデータとの比は 8:2 とした．

比較対象として，現状で一番精度が良いとされる青木らの手法[2]を選択した．この手法で必要となる品詞情報は，TreeTagger[5]を用いて取得した．また，提案手法で利用する前置詞を含むn-gramの効果を見るために，全てのn-gramを対象にしてプロファイルを作成する手法も比較対象とした．

3.2　実験結果

表 1 に，実験結果を示す．表中の“単語 n-gram”とは，全ての n-gram を対象にしてプロファイルを作成する手法を表す．また，“平均精度”とは，6-fold cross validation の結果得られた各判別精度の平均を表す．

表 1 から，提案手法が最も判別精度が良いことがわかる．実際，提案手法と青木らの手法[2]では，有意水準 5%で有意差が見られた（paired *t*-test）．また，単語 n-gram に基づく手法は，提案手法には劣るものの，青木らの手法[2]より判別精度が良い結果となった．

表 1　実験結果

手法	平均精度
提案手法	0.904
青木らの手法[2]	0.829
単語 n-gram	0.875

4. 考察

評価実験の結果，提案手法は，従来手法より判別精度が良いことが確認できた．実験結果を分析したところ，提案手法のプロファイルには，母語話者／非母語話者に特徴的なn-gramが多数含まれおり，判別に寄与していることが判明した．例えば，1. でも例として取り上げた“number of the”も判別に寄与していることが確認できた．“number of the”の母語話者／非母語話者プロファイルでの順位は，それぞれ，135721 位[6]／470 位であり，その差は，135251 と大きい．更に，順位の差が大きいだけでなく，“number of the”を含む論文は，母語話者の論文で 1 論文，非母語話者の論文で 21 論文と，非母語話者の論文での出現数が圧倒的に多い．言い換えると，ある特定の個人のみが頻繁に使用する表現でなく，非母語話者に共通した特徴的なn-gramであるといえる．このようなn-gramは，“with each other”，“while that of”，“and so on 文末”など，多数確認できた．これらのn-gramは誤りではないが，母語話者があまり使用しない，どちらかと言えば不自然な表現である．

一方，母語話者に特徴的なn-gramとして，前置詞の使い分けに関するものが多数見られた．例えば，“data”に対して，母語話者は，“data for”，“data of”，“data from”，“data in”など，多様な前置詞を使い分けていた．非母語話者では，“data of”が主流であり，“data from”を含む論文は 1 つも存在しなった（母語話者では 39 論文，順位の差 13512.5）[7]．同様な例として，“ability to”，“ability of”，“ability in”（“ability to”は，母語話者 18 論文，非母語話者 0 論文），“was performed on”，“was performed to”，“was performed in”（“was performed on”は，母語話者 12 論文，非母語話者 0 論文）などが挙げられる．

提案手法と異なり，従来手法である青木らの手法[2]や藤井らの手法[6]では，品詞列に基づくため，これらの差異を捉えることはできない．例えば，“data from”と“data of”を品詞列にすると，両方とも“名詞 前置詞”となり，母語話者の論文にも非母語話者の論文にも共通して頻出する n-gram となってしまう．このような品詞変換の際に生じる情報の欠落より，提案手法より判別精度が低くなったと分析できる．

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

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[6] ここでの順位は，学習データとテストデータを合わせた母語話者の論文 120 編全てから作成したプロファイル中での順位である．以下の順位も同様である．

[7] プロファイル中に同順位の n-gram がある場合，平均順位とするため，順位の差が少数となることがある．

全ての単語 n-gram を利用した手法については，青木らの手法より判定精度が良い．これは，全ての単語 n-gram を利用した手法では，上述の母語話者／非母語話者に特徴的な n-gram もプロファイルに含むためである．しかしながら，2.2 で述べたように，全ての単語 n-gram を対象としてプロファイルを作成すると，判別に寄与しない n-gram が増加する．判別に寄与しない n-gram がノイズとなり，提案手法よりは判別精度が低くなったと考えられる．

判別精度以外にも，提案手法は次の 2 つの利点がある．一点目として，判別に要する時間が短いことが挙げられる．前置詞を含む n-gram のみを対象としてプロファイルを作成するため，青木らの手法より n-gram の種類数が少なくなり，判別に要する時間が短くできる．3．の評価実験では，提案手法における n-gram の種類数は，平均して青木らの手法の約 40%に削減できた．二点目として，解析ツールを必要としないことが挙げられる．このことは，品詞解析誤りからの影響を全く受けないことを意味する．また，非母語話者専用の品詞解析ツールを開発する必要もない，更に，品詞解析を要せず，判別処理に要する時間も短くなる．

5. おわりに

本稿では，母語話者／非母語話者の文書を判別する手法を提案した．提案手法は，前置詞を含む n-gram に基づいてプロファイルを作成し，母語話者／非母語話者の判別を行う．実験の結果，提案手法は，従来手法よりも判別精度が良いことを確認した（平均判別精度 0.904）．判別精度以外にも，提案手法には，（a）判別処理にかかる時間が少ない，（b）品詞解析ツールなどの解析ツールを必要としない，の 2 つの利点があることを確認した．

今後の課題として，判別対象の拡張が挙げられる．本稿の実験では，母語話者／非母語話者を，それぞれアメリカ人／日本人とした（従来手法[2, 6]でも同様である）．しかしながら，Web などを対象とした場合，母語話者，非母語話者とも，それ以外の書き手による文書も判別対象に含まれることになる．書き手の数が増えると判別は難しくなり，判別精度が低下する．今後は，多数の書き手が含まれる文書集合を対象として，提案手法を評価する予定である．

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] C. Brockett, W.B. Dolan, and M. Gamon. 2006. Correcting ESL Errors Using Phrasal SMT Techniques. *In Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of ACL*, pp. 249-256.

[4] 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.

[5] D. Cock and S. Granger. 1998. An automated approach to the phrasicon of EFL learners. In *Learner English on computer*, pp. 67-79.

[6] 藤井, 冨浦, 田中. 2005. Skew Divergenceに基づく文書の母語話者性の推定. 自然言語書処理, pp. 79-96.

[7] 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.

[8] L. Lee. 1999. Measures of distributional similarity. In *Proc. 37th Annual Meeting of the ACL*, pp. 25-32.

[9] R. Nagata, A. Kawai, M. Koichiro, and N. Isu, A Feedback-Augmented Method for Detecting Errors in the Writing of Learners of English. *In Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of ACL* 1999. Measures of distributional similarity. pp. 241-248.

[10] 齋賀, 和泉, O. Chung, 井佐原. 2002. 日本人英語学習者コーパスの作成とその利用可能性. 言語処理学会第8回年次大会, pp. 164-167.

[11] 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.

[12] 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.

[13] Y. Tono. 2000. A corpus-based analysis of interlanguage development. In *PALC.*, pp.323-340.