LE_004

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
Documents Discrimination between Native English Documents and Nonnative Ones Based on Language Identification Technique

青木 さやか† 冨浦 洋一‡ 行野 顕正† 谷川 龍司†
Aoki Sayaka Tomiura Yoichi Yukino Kensei Tanigawa Ryuji

1. はじめに

英語非母語話者が流暢な英文を書くためには，単語や文法の知識だけでなく，自然な単語の組み合わせ (コロケーション) が必要である．また，母語話者の書く英文には品詞や単語の自然な遷移 (リズム) があり，これから大きく外れる英文は不自然に感じる．従来はこのような不自然な英文を避けるためには，実際に母語話者にチェックしてもらうしかなかった．しかし，周りに母語話者がいないケースも多い．従って，これらを自動的に指摘するシステムがあれば有用である．

このようなシステムの開発には，母語話者の書いた程度に質の良い英文書 (N) と，非母語話者の書いた様々な間違いや不自然さを含んだ英文書 (NN) の2種類の文書が大量に必要となる．そこで本稿では，少数の学習データから，高精度でN/NNを自動的に判別する手法を提案する．このようなシステムを用いることにより，インターネット上などにある大量の英文書を，その母語話者らしさに応じて分類し，上記2種類のコーパス§を構築することが可能になる．

提案手法では，言語識別分野で提案されている特徴設定手法・相違度尺度をN/NN判別に応用する．提案手法では，英文書を品詞の列に変換し，その部分品詞列を判別対象文書やクラスの特徴として扱い，その相違度を測ることによりN/NN判別を行う．部分品詞列としては様々な長さが考えられるが，本手法では言語識別で提案されている知見を利用し，長い品詞列を使用する．長い部分品詞列はN/NNのどちらかにしか存在しないものが多く，N/NNそれぞれの特徴を表すと考えられるので，従来手法で使われてきた短い部分品詞列よりも有用であると考えられる．判別対象文書–クラス間の相違度を測る尺度として，KL-Divergenceに基づく相違度尺度と品詞列出現順位に基づく相違度尺度を適用する．

長い品詞列の有効性を確かめるために，使用する品詞列の長さを1～7に変えて判別実験を行った．また，従来手法との比較を行った．その結果，長い品詞列がN/NN判別に有効であり，提案手法が従来手法より高い精度で判別できることが判った．

以降，2節では，提案手法である長い品詞列を用いたN/NN判別手法と，N/NN判別の従来手法について説明する．3節では，実験により，長い品詞列の有効性と，提案手法の従来手法に対する優位性を示す．4節では，実験結果に対する考察を行う．

†九州大学大学院システム情報科学府
‡九州大学大学院システム情報科学研究院
§電子化されたテキストの集積．言語データベースとして用いられる．

She likes a cake or something like that.
↓ TreeTaggerにより品詞列化
PP VBZ DT NN CC NN IN DT SENT
↓ 部分品詞列を抽出
1-品詞列: PP, VBZ, DT, NN, CC, NN, IN, DT, SENT
2-品詞列: *-PP, PP-VBZ, VBZ-DT, DT-NN, NN-CC, CC-NN, NN-IN, IN-DT, DT-SENT
3-品詞列: *-*-PP, *-PP-VBZ, PP-VBZ-DT, VBZ-DT-NN, DT-NN-CC, NN-CC-NN, CC-NN-IN, NN-IN-DT, IN-DT-SENT
⋮
(ただし，*は文頭記号，-は品詞を結合するための記号を表す)

図 1: 品詞列の抽出手順

2. 品詞列を文書特徴とする母語話者文書・非母語話者文書判別手法

2.1 文書の特徴量

本稿で提案する手法では，英文書を品詞の列として扱い，その部分品詞列の相対出現頻度分布に文書の母語話者らしさないし，非母語話者らしさ) が表れると考える．英文書から，その部分品詞列を抽出する手順を図1に示す．英文書は，文書中の単語を品詞に変換することで，品詞の列として捉えることができる．その品詞列から，連続したn個の品詞列を結合し，抽出したものを，n-品詞列と呼ぶ．

相対出現頻度とは，品詞列の総出現頻度に対する出現頻度の割合である．n-品詞列 x の相対出現頻度 $p(x)$ は x の出現頻度を $f(x)$，n-品詞列のありうる全ての組み合わせからなる集合を X とすると式 (1) となる．

$$p(x) = \frac{f(x)}{\sum_{x \in X} f(x)}. \quad (1)$$

提案する手法では，品詞列の中でも，長い品詞列における相対出現頻度分布の差に着目する．長い品詞列は，短い品詞列に比べて特定のクラスのみに出現する傾向が強いため，その相対出現頻度分布は，それぞれのクラスの特徴を良く表していると考えられる．

N/NN間におけるn-品詞列の相対出現頻度分布の差を利用して，未知の文書がNかNNかを判別するためには，既知のN/NN集合からそれぞれ推定した各クラスの相対出現頻度分布と，判別対象文書の相対出現頻度分布との相違度を比較し，判別対象文書とより類似した分布を持つクラスと決定すればよい．分布の相違度を測る尺度として，本稿では「KL-Divergenceに基づく相違度尺度」と「品詞列出現頻度順位に基づく相違度尺度」

を提案する．以降，2.2 で KL-Divergence を相違度尺度に用いた手法を，2.3 で品詞列出現順位の差を相違度尺度に用いた手法をそれぞれ説明する．

2.2 KL-Divergence に基づく相違度比較

一つ目の手法では，n-品詞列の相対出現頻度分布を N クラス，NN クラス，判別対象文書それぞれに対して算出し，その分布間の相違度を KL-Divergence で測ることにより，N/NN 判別を行う．

KL-Divergence とは，確率分布間の相違度をはかる尺度であり，クラス C と判別対象文書 d の相違度は，$p_C(x)$ を x のクラス C における相対出現頻度，$p_d(x)$ を x の判別対象文書 d における相対出現頻度，X を判別対象文書に出現した n-品詞列の集合とすると式 (2) で表される．

$$D_{KL}(d \parallel C) = \sum_{x \in X} p_d(x) \log \frac{p_d(x)}{p_C(x)}. \quad (2)$$

式 (2) では，$p_d(x) > 0$ かつ $p_C(x) = 0$ となる x が一つでも存在すると，$D_{KL}(d \parallel C)$ の値は ∞ となる．これはゼロ頻度問題と呼ばれ，$p_C(x)$ の推定精度が低い場合，大きな問題である．本稿では，加算法を使用し，ゼロ頻度問題に対処する．加算項を δ とし，加算法を施した相対頻度 $p(x)$ の求め方を式 (3) に示す．

$$p(x) = \frac{f(x) + \delta}{\sum_{x \in X} f(x) + |X|\delta}. \quad (3)$$

KL-Divergence に基づく N/NN 判別法のアルゴリズムを以下に示す．

1. あらかじめ，文書クラス $C(C \in N, NN)$ それぞれに属する学習データから式 (3) によりそれぞれの文書クラスに対する n-品詞列の相対頻度分布 $p_C(x)$ を求めておく．
2. 判別対象文書 d に対する n-品詞列の相対頻度分布 $p_d(x)$ を求める．
3. 式 (2) に示す評価関数を用いて $p_C(x)$,$p_d(x)$ 間の相違度を求める．
4. 判別対象文書 d の発生源クラス $\hat{C}$ は式 (4) より決定される．

$$\hat{C} = \arg\min_C D_{KL}(d \parallel C). \quad (4)$$

2.3 品詞列出現順位に基づく相違度比較

二つ目の手法では，品詞列の出現順位表を求め，順位表間の相違度を測る事により N/NN 判定を行う．出現順位表とは k-品詞列 ($k = 1 \sim n$) の出現頻度を計数し，出現頻度の降順に並べた表である．クラス C の順位表を $Prof(C)$，クラス C での品詞列 x の順位を $r_C(x)$，文書 d の順位表を $Prof(d)$，文書 d での品詞列 x の順位を $r_d(x)$ とする．

順位表間の相違度を測る尺度として，順位の差の総和を利用する．クラス C と判別対象文書 d の品詞列順位表に基づく相違度を式で表すと，式 (5) となる．

$$D_{TC}(d \parallel C) = \sum_{x \in Prof(d)} |r_d(x) - r_C(x)|. \quad (5)$$

ただし，学習データ中に出現しない品詞列 x のクラス C における順位は，式 (6) で与える．

$$r_C(x) = \max_{x' \in Prof(C)} r_C(x') + 1. \quad (6)$$

品詞列出現順位を用いた N/NN 判別法のアルゴリズムを以下に示す．

1. あらかじめ，文書クラス $C(C \in N, NN)$ に属する学習データから，それぞれの文書クラスの品詞列順位表 $Prof(C)$ を求めておく．
2. 判別対象文書 d から品詞列順位表 $Prof(d)$ を求める．
3. 式 (5) に示す評価関数を用いて $Prof(C)$，$Prof(d)$ 間の相違度を求める．
4. 判別対象文書 d の発生源クラス $\hat{C}$ は式 (7) より決定される．

$$\hat{C} = \arg\min_C D_{TC}(d \parallel C). \quad (7)$$

2.4 言語識別手法との関連

提案した二つの相違度尺度は，言語識別において提案された手法である．KL-Divergence に基づく手法は Sibun[1] らが，出現順位に基づく手法は Cavnar[2] が提案した¶．これらのでは手法は，それぞれ高い精度が確認されており，N/NN 判別においても効果的であることが期待できる．

なお，言語識別では，文書の言語特徴を表すものとして，バイト列の相対出現頻度分布が用いられている．バイト列の相対出現頻度分布は，文書をバイトの列‖として捉え，その部分バイト列を計数することで抽出される．

バイト列を特徴として用いると，文書中に出現した機能語や接辞などの情報が抽出されるため，言語識別では有用であった．しかし，本稿の目的である N/NN 判別では，両クラスにおいて機能語や接辞の出現傾向に大きな差はないと考えられる（両者とも英語であるため）．そのため，本稿では，バイト列ではなく，より大きな単位である，品詞の列をクラス特徴として用いている．

また，同じく言語識別の分野において行野らは，長いバイト列が類似言語間において判別に有効に働くことを示している [3]．一般に，言語が類似するほど，言語間におけるバイト列の相対出現頻度分布の差は小さくなる．

¶言語識別の分野では，Sibun らは，長さ2の byte 列を使用することにより，Cavnar は，400 位以下を切り捨てることにより．短く高頻度な byte 列を活用して判別をしている．

‖英語やドイツ語など，欧州の言語圏においては，一般的に文字列と同一と考えて構わない．中国語や日本語など，文字を表すのに複数のバイトを用いる言語圏（ないし，符号化手法）では，1 文字よりも細かい単位で，文書特徴を捉えていることになる．

文献 [3] では，長いバイト列ならば，類似言語間でも十分に大きな出現傾向の差が得られることに着目しており，短いバイト列を言語特徴とするその他の手法に比べ，高い精度が得られたことを報告している．

本稿の目的である N/NN 判別は，品詞列を特徴として見たとき，類似した言語間と同様の傾向があると思われる．例えば，N においても NN においても，基本的な文法事項に関しては正しく書かれていると推測されるため，多くの品詞列は似たような相対出現頻度になると考えられる．一方，複雑な構文などでは，徐々に両者の差が広がり，クラスの特徴が明確になると期待できる．提案手法では，このような類似性から，長い品詞列をクラス特徴として導入している．

2.5 従来手法

藤井らは， 品詞 tri-gram を文書やクラスの特徴として扱い，特徴間の相違度を Skew-Divergence に基づいて求めることにより．N/NN の判別を行っている [4]．

この手法は，十分に学習データがある場合，各クラスモデルからの判別対象文書の発生確率を比較することに等しい．藤井らの手法は，言語モデルとして品詞 n-gram モデルを用いるため，Sibun らの提案した KL-Divergence を用いる手法以上に，確率分布を正しく推定することが難しい．特に，n が大きくなると条件部の組み合わせ数も爆発的に増加し，各条件部に対応する学習データ数が小さくなり，各条件部に対する確率分布を求めることが難しくなる．そのため，本稿では，藤井らの手法において長い品詞 n-gram を用いることをせず，比較対象としてのみ扱った．

3. 実験

3.1 実験方法

長い品詞列を用いるという提案手法の有効性を確認するため，3 種類の実験を行った．実験 1 では，KL-Divergence を用いた手法における適切な加算項を調べるため，加算項 δ の探索を行った．実験 2 では，長い品詞列が判別に有効であることを示すため，提案手法において様々な長さの品詞列を文書特徴に用いた時の判別精度を比較した．実験 3 では，提案手法が従来手法よりも高い判別性能を発揮できることを示すため，藤井らの手法と判別精度を比較した．藤井らの手法は文献 [4] で，その他の基本的な分類手法を用いた場合に比べ，有意に高い判別精度が得られたことが報告されている．

判別実験の学習データには，電気情報系の国際会議で発表された英語科学技術論文を用いた．母語話者データとして，海外の国際会議で発表され，かつ著者の所属機関が USA の英語論文 217 本を，非母語話者データとして，日本などアジアで開催された国際会議で発表され，かつ著者が日本人の英語論文 316 本を用意した．これらのデータは N と NN で学習データサイズの差をなくすために，それぞれの文書集合に含まれる総品詞数が同じになるようにしてある．判別データには，校正専門家がネイティブチェックを行った論文 60 本 (N:NN=25:35) を用意した．提案手法では品詞列出現パターンの違いを見て判別するため，全ての文書データを品詞列化する必要がある．この処理には TreeTagger**を用いた．品詞セットには TreeTagger における品詞セット (Penn Tree Bank [5] の品詞タグセットを拡張したもの) をそのまま用いた．

n\ 加算項	0.01	0.1	0.5	1
1	68.3%	68.3%	68.3%	68.3%
2	70.0%	70.9%	71.7%	71.7%
3	76.7%	80.0%	80.0%	80.0%
4	80.0%	83.3%	83.3%	83.3%
5	88.3%	86.7%	86.7%	85.0%
6	**88.3%**	88.3%	88.3%	88.3%
7	85.0%	85.0%	86.7%	86.7%

表 1: 加算項を変えた場合の正解率：n は品詞列の長さ

n\ 相違度尺度	KL-Divergence	品詞列出現順位
1	68.3%	66.7%
2	70.0%	78.3%
3	76.7%	81.7%
4	80.0%	83.3%
5	88.3%	88.3%
6	**88.3%**	88.3%
7	85.0%	**90.0%**

表 2: 品詞列の長さごとの正解率：n は品詞列の長さ

品詞列を抽出する際，文をまたいで品詞列を求めず，1 文ごとに品詞列を抽出した．また，文頭表現を抽出するため，文頭に $n-1$ 個の文頭記号が存在するものとして，n-品詞列を抽出した．これにより，文書に含まれる品詞数と同数の n-品詞列が抽出できる．

実験の評価値として正解率を用いた．正解率を「判別を行ったすべての文書のうち，そのクラス (N/NN) を正しく判別できた文書の割合」と定義した．さらに，各クラスに対する相違度が同値で判別不可能となった文書については判別失敗とみなした．

すべての文書を，実験データ中により多く含まれている NN であると判別すると，58.3%の精度が得られる．この正解率を本実験のベースラインとする．

3.2 実験 1:KL-Divergence における加算項の設定

実験 1 では，KL-Divergence を用いた手法において，複数の加算項 δ を試すことで，実験的に加算項 δ を探索した．加算項を 0.01，0.1，0.5，1 に変えて判別実験を行った．その実験結果を表 1 に示す．実験結果より，加算項に 0.01 を用いるのが適当であることがわかった．

3.3 実験 2:長い品詞列の有効性

実験 2 では，長い品詞列が判別に有効であることを示すため，提案手法において様々な長さの品詞列の相対頻度出現分布を文書特徴に用いたときの判別精度を比較した．品詞列の長さを 1~7 に変えて二つの提案手法により N/NN 判別を行った．なお，KL-Divergence を用いた手法での加算項は 0.01 を用いた．その実験結果を表 2 に示す．どちらの手法においても，長い品詞列を用いると

**http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Method	Accuracy
ベースライン	58.3%
Skew-Divergence	81.7%
KL-Divergence	88.3%
品詞列出現順位	90.0%

表 3: 従来手法との比較

Method	Accuracy
ベースライン	58.3%
Skew-Divergence	81.7%
KL-Divergence	76.7%
品詞列出現順位	81.7%

表 4: 長さ 3 の品詞列を使用した場合の正解率

精度が良くなっている．実験結果より，長さ 5~7 程度の長い品詞列も識別に有効であることが判った．

3.4 実験 3:従来手法との比較

実験 3 では，提案手法が従来手法よりも高い判別性能を発揮できることを示すため，藤井らの手法と判別精度を比較した．提案手法と従来手法の判別精度を比較したものを表 3 に示す．実験データには，各手法とも 3.1 で示したものを使用した．

KL-Divergence に基づく手法では，加算項 0.01 でスムージングを施した 6-品詞列を，品詞列出現順位に基づく手法では，7-品詞列を，それぞれ文書特徴として用いた．Skew-Divergence では，2.5 で述べたように，品詞 tri-gram を文書特徴に用いている．

品詞列出現順位を用いた手法が，一番精度がよく，KL-Divergence を用いた手法においても，従来手法よりも精度がよかった．実験結果より，提案手法が従来手法よりも高い精度で N/NN の判別を行えることがわかった．

4. 考察

長さ 3 の品詞列における正解率の比較を表 4 に示す．文献 [4] でも述べられているとおり，3-品詞列における判別精度を比較すると，補間法の優れている従来手法の方が高い精度で判別できる．しかし，表 3 からわかるように，長い品詞列を用いて判別を行うと，あまり良いとは言えない補間法を使用した KL-Divergence でも，十分な精度で判別が行える．

短い品詞列では，個々の品詞列の出現はある程度正しく見積もられる．ところが，品詞列の分布のばらつきが小さく，N/NN で共通して出現する品詞列が多くなる．共通部分が多いと，各クラスの特徴を十分に表すことができない．

一方，長い品詞列では，学習データ内での個々の品詞列の出現は，真の出現に沿っているとは限らない．しかし，品詞列の分布のばらつきが大きく，N/NN それぞれ特有の品詞列が出現しやすいため，各クラスの特徴をよく表すことができる．

N/NN 判別は，個々の品詞列に対して相違度を測り，その和を取っていると解釈できる．個々の品詞列の出現が信頼出来なくても，品詞列全体の集合がクラスに発生する確率については，ある程度，信頼のおけるものとなり，クラス判別に有効に働くと考えられる．

長い品詞列を使用することにより，判別精度が向上したと考えられる．

5. おわりに

本稿では，長い品詞列を活用した N/NN 判別手法を提案した．提案手法では，品詞列の頻度分布を文書の特徴として用いた．特に，長い品詞列はどちらかにしか存在しないものが多く，N/NN それぞれの特徴を表すと考えられるので，長い品詞列を特徴として活用した．また，特徴間の相違度を測る尺度として，言語識別で提案されていた，KL-Divergence に基づく相違度尺度と品詞列出現順位に基づく相違度尺度を適用した手法を利用した．

提案手法の有効性を確認するために，N/NN 判別実験を行った．実験の結果から，長い品詞列が N/NN 判別に有効なことや，従来手法より高い精度で判別ができることが示された．

参考文献

[1] Penelope Sibun and Jeffrey C. Reynar. Language identification: Examining the issues. In *5th Symposium on Document Analysis and Information Retrieval*, pp. 125–135, Las Vegas, Nevada, U.S.A., 1996.

[2] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Symposium On Document Analysis and Information Retrieval*, pp. 161–176, University of Nevada, Las Vegas., 1994.

[3] 行野顕正, 田中省作, 冨浦洋一, 松本英樹. 低頻度 byte 列を活用した言語識別. 情報処理学会論文誌, Vol. 47, No. 4, pp. 1287–1294, 2006.

[4] 藤井宏, 田中省作, 冨浦洋一. Skew divergence に基づく文書の母語話者性の推定. 自然言語処理, Vol. 12, No. 4, pp. 79–96, 2005.

[5] Beatrice Santorini. Part-of-speech tagging guidelines for the penn treebank project. Techinical Report MS-CIS-90-47, 1990.