日本語学習者コーパス I-JAS を用いた母語識別

西島光洋 [1,2]　劉穎 [1]　中田和秀 [2]
[1] 清華大学人文学院中国語言文学系　[2] 東京工業大学工学院経営工学系
nishijima.m.ae@m.titech.ac.jp, yingliu@tsinghua.edu.cn,
nakata.k.ac@m.titech.ac.jp

1　はじめに

母語識別 (Native Language Identification, NLI) とは，ある文書が母語話者以外によって書かれたことが分かっている前提のもとで，その文書の書き手の母語を識別するタスクである．NLI は第 2 言語習得 (Second Language Acquisition, SLA) や法言語学の分野で応用がされ得る．たとえば，第 2 言語 (L2) 学習者が書いた文書に対する自動誤り訂正研究において，書き手の母語を考慮することで訂正の質が向上することが指摘されており [1]，NLI により書き手の母語を自動的に認識することができれば，より質の高いフィードバックを学習者に与えられるようになる．また [2] では，NLI における機械学習モデルを解釈することで，既存の SLA 研究で得られた知見に新たな証拠を与えたり，新たな仮説を提唱したりできるとしている．そして NLI は，文書の書き手の属性を推定するタスクである著者プロファイリングの一種としても捉えることができ，犯罪捜査における活用も期待される [3]．日本語学習者の数は年々増加傾向にあり [4]，さらに 2019 年度から特定技能制度が開始されたことで，日本語を L2 とする人は今後も増加していくと考えられる．したがって，これまで L2 英語を中心にして行われてきた NLI 研究を L2 日本語に拡張することは重要である．

本稿は，L2 としての日本語で書かれた文書（以下 L2 日本語文書）に対して初めて NLI を行ったものである．また，既存の NLI 研究で有効とされていた特徴に加えて，日本語特有の特徴として文字種を新たに提案する．これらの特徴を用いて L2 日本語文書の NLI を行った結果を報告する．さらに，[2] の手法を参考に，線形サポートベクトルマシン (SVM) の重みベクトルを用いて，中国語母語話者の過剰使用表現について分析する．

2　関連研究

NLI 研究の先駆けは，Koppel らによる研究 [5] である．彼らは，機能語，スペルミスなどの誤用，文字 Ngram と品詞 (POS) Ngram を特徴として，5 つの言語をそれぞれ母語とする人によって書かれた文書に対して線形 SVM で分類を行なった結果，最高で 80.2%の正解率 (accuracy) が得られたとしている．これを契機として，どのような特徴や分類器を用いることで NLI の識別性能を高められるかという研究が進むことになる．2013 年には，11 の言語をそれぞれ母語とする人によって書かれた文書に対する分類正解率を競う競技会が行われた．その競技会の報告論文 [6] では，単語，POS や文字の Ngram が特徴として多く採用され，SVM が分類器として多く採用されたことが報告されている．

以上の研究はすべて L2 英語を対象にした NLI であった．そこで，英語に対して有効性が確認された NLI の手法が他の言語でも有効なのかという観点から，英語以外の言語を対象とした NLI 研究が 2011 年以降盛んに行われてきた [7, 8, 9, 10, 11, 12, 13]．

一方で，日本語に目を向けてみると，与えられた日本語文書が母語話者によって書かれたものかどうかを機械学習手法を用いて識別する研究 [14] はあったとしても，日本語を対象にした NLI 研究は管見の限り存在しない．

3　手法

本節では，本稿で用いたコーパスとその処理方法，そして特徴ベクトルを作成する際に着目した文体特徴と文書を分類するときに用いた分類器を述べる．

3.1　コーパスとその処理

本稿では，日本語学習者コーパスとして『多言語母語の日本語学習者横断コーパス (I-JAS)』[15] を使用

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

した．I-JAS には，中国語（中），英語（英），フランス語（仏），ドイツ語（独），ハンガリー語（洪），インドネシア語（尼），韓国語（韓），ロシア語（露），スペイン語（西），タイ語（泰），トルコ語（土），ベトナム語（越）の計 12 言語をそれぞれ母語とする日本語学習者によって産出されたデータが収録されている．今回は I-JAS に含まれるデータのうち，2 種類のコマ割り漫画を見てそのストーリーをそれぞれ描写するストーリーライティング課題で産出された文書を用い，2 種類の漫画からそれぞれ産出された 2 つの文書を各学習者ごとに結合させたものを 1 文書として扱った．中国語，英語，韓国語以外の 9 言語の学習者データは母語ごとにそれぞれ 50 件あるので，それらをすべて用いた．一方で，中国語，英語，韓国語母語話者のデータは収録数がそれぞれ 50 件を超えていたため，なるべく前述の 9 言語の学習者全体の日本語習熟度分布と一致するようにそれぞれ 50 件選択した[1)]．つまり，各母語話者ごとに 50 文書，計 600 文書を用意した．これらの文書に対して，日本語自然言語処理ライブラリ GiNZA[16] を用いて形態素解析と依存構造解析を行なった．

3.2 文体特徴

各文書の特徴ベクトルを作成する際に着目した文体特徴とその説明ないし例を示す．括弧内の英文字は，以下で用いる略称である．なお，以下で用いる「1-Ngram」とは，ある言語要素の 1gram から Ngram までのすべてを指す．

(1) 形態素の基本形 (Lemma) 1-Ngram：句読点や記号なども含める．［例］、 二人 (2gram)
(2) 文字 (CHAR) 1-Ngram：句読点や記号なども含める．［例］ ！」と (3gram)
(3) 助詞と助動詞の基本形 (FW) 1-Ngram：英語の機能語に対応する特徴．品詞情報の第 1 階層が「助詞」または「助動詞」である形態素の基本形．［例］に られる (2gram)
(4) 品詞 (POS) 1-Ngram：最深階層までの品詞情報．［例］名詞-固有名詞-人名-名 (1gram)
(5) 依存関係のラベル (DEP)：Universal Dependencies のラベル．［例］nmod
(6) 依存関係のラベルと 2 つの Lemma の 3 つ組 (DEPL)：直接の依存関係がある 2 つの Lemma と両者間の依存関係のラベル．［例］case(犬, は)
(7) 依存関係のラベルと 2 つの POS の 3 つ組 (DEPP)：特徴 (6) での Lemma をその POS で置き換えたもの．［例］nsubj(動詞-一般, 代名詞)
(8) 文字種 (CType)：漢字，ひらがな，カタカナそれぞれの頻度．アルファベットやアラビア数字，記号などはカウントしない．

特徴 (1) から特徴 (7) までは NLI 研究ですでに広く使用されている特徴である [17]．本稿では，特徴 (1) から特徴 (4) までの N として 2 から 6 までの 5 通りを考慮する．それらに加えて本稿では，日本語文書において特有な，NLI の特徴として (8) の文字種を新たに提案する．文字種を特徴として用いる理論的根拠は，漢字圏の日本語学習者は漢語を多く使用する傾向にあることが SLA 研究においてしばしば指摘されていることによる [18]．今回は集計のしやすさから，漢語，和語，外来語の代わりとして，漢字，ひらがな，カタカナの頻度をそれぞれ集計した．以上の特徴らの出現頻度ベクトルを $l2$ ノルムで正規化したものを，各文書の特徴ベクトルとした．

3.3 分類器

これまでの NLI 研究では分類器として SVM がよく選択される傾向にある [6]．しかし，SVM よりもロジスティック回帰 (LR) の方が正解率が高いという報告 [19] や，NLI と類似のタスクである著者推定や著者プロファイリングの研究のうち日本語文書を対象にしたものでは，SVM よりもランダムフォレスト (RF) の方が有効であるといった報告 [20, 21] もある．そのため本稿では，解釈のしやすさも考慮して，線形 SVM，LR，RF 計 3 種類の分類器の性能を比較することにした．SVM と LR は 10×10 の入れ子式交差検証を行い，内部の交差検証ではグリッドサーチによりハイパーパラメータ C を決定した．また RF は，決定木の本数を 1000 本に固定して，10 交差検証を行った．そして，10 交差検証で得られる 10 個の正解率の平均値を最終的な正解率とした．

4 実験

4.1 各特徴および 3 種類の分類器の比較

本項では，3.2 項で挙げた特徴にそれぞれ着目して特徴ベクトルを作成し，線形 SVM，LR，RF をそれぞ

1) ただし I-JAS の特性上，韓国語母語話者の習熟度が文書選択後も依然として他の母語話者よりも平均的に高いなど，各母語話者ごとの習熟度が厳密に統制されていないという問題点はある．しかし本研究では，機械学習で必要とされるデータ数の確保のために，習熟度を厳密には統制しなかった．なお，I-JAS における各母語話者ごとの習熟度分布の詳細は [15] を参照せよ．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　各特徴と各分類器による分類正解率 (%)

	SVM	LR	RF
Lemma 1-2gram	**68.0**	65.2	63.0
1-3gram	**68.5**	66.7	60.0
1-4gram	**68.7**	67.2	59.8
1-5gram	**69.2**	66.3	61.0
1-6gram	**68.7**	66.2	59.7
CHAR 1-2gram	**68.7**	68.2	59.7
1-3gram	**70.5**	67.8	60.8
1-4gram	**71.5**	67.5	62.5
1-5gram	**70.8**	67.7	61.0
1-6gram	**71.0**	68.2	60.2
FW 1-2gram	**36.0**	35.5	34.5
1-3gram	**39.5**	36.7	35.2
1-4gram	**40.5**	38.8	36.2
1-5gram	**40.7**	38.3	35.8
1-6gram	**40.5**	37.8	35.7
POS 1-2gram	44.5	**45.7**	45.0
1-3gram	**48.0**	46.3	47.3
1-4gram	**49.2**	48.2	49.0
1-5gram	**50.5**	48.8	47.5
1-6gram	49.2	**49.8**	46.0
DEP	30.2	30.2	**31.2**
DEPL	**63.0**	61.7	53.3
DEPP	**45.7**	42.8	44.2
CType	14.2	**14.8**	11.2
ベースライン	8.3 (=1/12)		

れ用いて分類した場合の分類正解率を調査する．表 1 にその結果を記し，各特徴ごとに 1 番正解率が良かった部分を太文字で表示している．

表 1 より，各特徴の有効性については，CHAR 1-Ngram が最も有効で，Lemma 1-Ngram，DEPL，POS 1-Ngram，DEPP，FW 1-Ngram，DEP，CType の特徴が順に続いていることが分かる．今回新しく提案した文字種は，単独ではほとんど識別能力が無かった．また，分類器間で分類正解率を比較してみると，線形 SVM の分類正解率が最も高く，次いで LR，最後に RF というおおむねの傾向を読み取ることができる．そこで，次項以降の実験では，1 番正解率の高かった線形 SVM のみを分類器として用いることにする．

4.2　特徴を組み合わせたときの分類正解率

本項では，3.2 項の特徴を組み合わせて特徴ベクトルを作成したときに，分類正解率がどう変化するのか

表 2　特徴を組み合わせたときの分類正解率 (%)

順位	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	正解率
1		○			○			○	72.0
1		○				○	○	○	72.0
1	○	○				○	○	○	72.0
4		○			○	○	○	○	71.8
4	○	○			○	○	○	○	71.8
6		○				○		○	71.7
6	○	○					○	○	71.7
ベースライン：(2) のみ									71.5

表 3　文字種がある場合と無い場合での正解率 (%) の比較

順位	文字種あり	文字種なし	差
1	72.0	70.8	1.2
1	72.0	71.5	0.5
1	72.0	71.5	0.5
4	71.8	70.5	1.3
4	71.8	71.3	0.5
6	71.7	71.3	0.3
6	71.7	70.8	0.8

を検証する．本項の実験では，4.1 項の実験結果を鑑みて，Lemma 1-5gram，CHAR 1-4gram，FW 1-5gram，POS 1-5gram，DEP，DEPL，DEPP，CType の計 8 種類の特徴を組み合わせることを考える．すべての特徴を使わない場合を除いて，これら 8 種類の特徴をそれぞれ使うか使わないかのすべての場合（$2^8 - 1 = 255$ 通り）で特徴ベクトルを作成し，分類正解率を測定した．表 2 がその結果であるが，紙幅の関係上，4.1 項の実験で正解率が最も高かった CHAR 1-4gram をベースラインとして，CHAR 1-4gram を特徴として用いたときの正解率 71.5%よりも高かった 7 つの組み合わせのみを掲載している．

表 2 から，最高で 72.0%の正解率を達成できたことが分かる．また，CHAR 1-4gram と CType がベースラインよりも正解率が高かった 7 つの組み合わせすべてに出現しており，次いで DEPL と DEPP が 5 回，Lemma 1-5gram と DEP が 3 回出現し，FW 1-5gram と POS 1-4gram は 1 回も組み合わせの中に出現しなかったことも分かる．注目すべきなのは，単独ではほとんど識別能力が無かった文字種 (CType) が表 2 ではすべての組み合わせに出現しており，したがって文字種は他の特徴と組み合わせたときに分類正解率を高める力があると示唆される点である．実際に，文字種以外の特徴を表 2 の通りに固定して文字種がある場合と無い場合とで正解率を比較したところ，表 3 に示すように，最大で 1.3%正解率が上昇していることが分かる．

文字種特徴を他の特徴と組み合わせたときに正解率

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

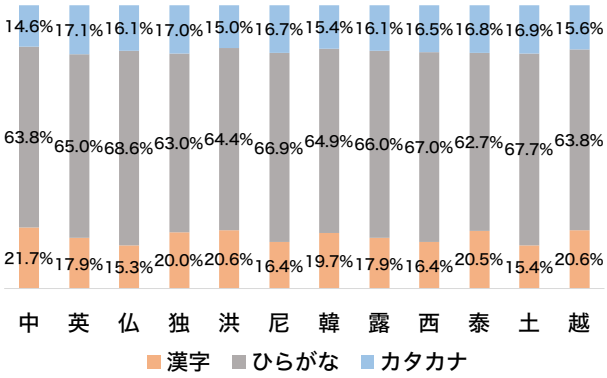

図 1　各母語話者ごとの各文字種の相対頻度

表 4　中国語母語話者が過剰使用する Lemma 1-5gram

順位	特徴	順位	特徴	順位	特徴
1	彼	5	ら は	9	とき
2	ら	6	犬 に	10	、
3	彼 ら	7	彼 ら は	11	彼 は
4	られる	8	、 彼	12	その 時 、

が上昇した現象に対して，特徴の多様性の観点からの説明が 1 つ考えられよう．Malmasi と Cahill は [22] で，異なる性質の特徴を組み合わせることで正解率が上昇することを指摘している．今回用いた特徴の性質を考えてみると，文字種以外の特徴は言語要素らの高々 1 文内でのミクロな関係を捉える特徴であるのに対して，文字種特徴は文書全体にわたるマクロな情報を捉える特徴であり，両者は相補的な関係を成していると言える．また，図 1 は今回用いた全 600 文書における各母語話者ごとの各文字種の相対頻度を示したものであるが，各文字種の相対頻度は各母語話者ごとに若干偏りがあることが分かる．そのため，文字種の情報だけでは母語を識別するための十分な情報とは言えないため単独ではほとんど識別能力が無かったが，文字種特徴で捉えたマクロな情報が文字 1-4gram などの特徴で捉えたミクロな情報を補い，結果として正解率が上昇したと推察される．

4.3　中国語母語話者の過剰使用表現

Malmasi と Dras は [2] で，線形 SVM の重みベクトルに着目することで各母語話者の過剰使用ないし過少使用の現象を捉えられ，そこから既存の SLA 研究で得られた知見に新たな証拠を与えたり，新たな仮説を提唱したりできるとしている．紙幅の関係上，本稿では中国語母語話者の過剰使用のみに焦点を絞ってその様子を捉えることにする．

表 4 は，Lemma 1-5gram を特徴としたとき，10 交差検証で生じた 10 個の SVM から中国語母語話者に対応する 10 本の重みベクトルを取りその平均ベクトルを計算し，平均ベクトルの要素値が大きい上位 12 個の部分に対応する具体的な特徴を記したものである．まず表 4 のうちのおよそ半分が，3 人称代名詞「彼」の過剰使用に関連するものである．石川は [23] で日本語母語話者と比較したときの中国語母語話者による 3 人称代名詞の過剰使用を指摘しているが，他の母語話者と比較したときにも中国語母語話者による 3 人称代名詞の過剰使用が示唆される結果となった．また，時間表現「とき」「その時、」も他の母語話者と比較したときに中国語母語話者が過剰使用している表現として挙げられる．その要因として，望月ら [24] で挙げられている“当 (dang) ＋［文］＋时 (shi)”（～とき）の他に，“这时 (zheshi)”“此时 (cishi)”（この時／その時）[2] という中国語における表現も影響しているのではないかと推測される．ただし，「時」という表現自体は中国語を母語とする学習者に限らず，学習者全体で過剰使用される表現であるという見解もあるし [26, 27]，加えて韓国語母語話者やタイ語母語話者に対しても表 4 と同様の表を描いたところ，20 位付近に「その 時」という特徴が存在していたため，この点については更なる調査が必要であると考えられる．

5　おわりに

本稿では，日本語学習者コーパス I-JAS を用いて初めて L2 日本語文書に対する NLI を行い，また日本語特有の特徴として新たに文字種を提案した．その結果，文字種を他の特徴と組み合わせて用いることで分類正解率の向上が確認され，最高で 72.0%の正解率で 12 の言語をそれぞれ母語とする人によって書かれた文書を分類することができた．また，SVM の重みベクトルを用いて中国語母語話者の過剰使用表現について分析を行った．今回は解釈のしやすさを考慮して分類器を選択したが，今後の課題の 1 つとして，アンサンブル学習 [28] などの手法を用いて分類正解率をより高めることが挙げられる．

参考文献

[1]Alla Rozovskaya and Dan Roth. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computa-*

2)　この表現からの影響について [24] では明示的に議論されてはいないものの，中国語母語話者による中国語での産出でこの表現が使用されているのが，[24] における例文 (17)b で観察される．また，「とき」「その時」という表現が中国語母語話者に過剰使用される要因については，[25] も参照のこと．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

tional Linguistics: Human Language Technologies, pp. 924–933, 2011.

[2]Shervin Malmasi and Mark Dras. Language transfer hypotheses with linear SVM weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1385–1390, 2014.

[3]財津亘. 犯罪捜査のためのテキストマイニング—文章の指紋を探り，サイバー犯罪に挑む計量的文体分析の手法—. 共立出版, 2019. 金明哲（監修）.

[4]文化庁国語課. 令和元年度国内の日本語教育の概要, 2019. https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/nihongokyoiku_jittai/r01/pdf/92394101_01.pdf（2020 年 12 月 25 日閲覧）.

[5]Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 624–628, 2005.

[6]Joel Tetreault, Daniel Blanchard, and Aoife Cahill. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 48–57, 2013.

[7]Felix Golcher and Marc Reznicek. Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus. In *Proceedings of Quantitative Investigations in Theoretical Linguistics 4*, pp. 29–34, 2011.

[8]Shervin Malmasi and Mark Dras. Chinese native language identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 95–99, 2014.

[9]Shervin Malmasi and Mark Dras. Arabic native language identification. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing*, pp. 180–186, 2014.

[10]Shervin Malmasi and Mark Dras. Finnish native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, pp. 139–144, 2014.

[11]Shervin Malmasi, Mark Dras, and Irina Temnikova. Norwegian native language identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 404–412, 2015.

[12]Iria del Río Gayo, Marcos Zampieri, and Shervin Malmasi. A Portuguese native language identification dataset. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 291–296, 2018.

[13]Nikita Remnev. Native language identification for Russian. In *2019 International Conference on Data Mining Workshops*, pp. 1138–1144, 2019.

[14]吉見毅彦, 小谷克則, 九津見毅, 佐田いち子. 単語対応付けに基づく日本語学習者による作文の自動識別. 情報処理学会論文誌, Vol. 49, No. 12, pp. 4039–4043, 2008.

[15]迫田久美子, 石川慎一郎, 李在鎬（編）. 日本語学習者コーパス I-JAS 入門—研究・教育にどう使うか—. くろしお出版, 2020.

[16]松田寛. GiNZA - Universal Dependencies による実用的日本語解析. 自然言語処理, Vol. 27, No. 3, pp. 695–701, 2020.

[17]Shervin Malmasi. *Native language identification: explorations and applications*. PhD thesis, Macquarie University, 2016.

[18]迫田久美子. 改訂版 日本語教育に生かす第二言語習得研究. アルク, 2020.

[19]Sze-Meng Jojo Wong and Mark Dras. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1600–1610, 2011.

[20]金明哲, 村上征勝. ランダムフォレスト法による文章の書き手の同定. 統計数理, Vol. 55, No. 2, pp. 255–268, 2007.

[21]財津亘, 金明哲. ランダムフォレストによる著者の性別推定—犯罪者プロファイリング実現に向けた検討—. 情報知識学会誌, Vol. 27, No. 3, pp. 261–274, 2017.

[22]Shervin Malmasi and Aoife Cahill. Measuring feature diversity in native language identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 49–55, 2015.

[23]石川慎一郎. 多様な外国語学習者の言語使用特性—中国人英語／日本語学習者の過剰・過小使用語彙—. 第 2 回学習者コーパスワークショップ予稿集, pp. 57–68, 2017.

[24]望月圭子, 申亜敏, 小柳昇. 日本語・英語・中国語双方向学習者コーパスにみられるテンス・アスペクトの習得. 日本語・日本学研究, No. 10, pp. 137–152, 2020.

[25]王蕊. 日本語上級レベル学習者の接続表現の使用状況に関する調査—中国語母語話者のストーリーテリングテストを中心に—. ポリグロシア, Vol. 17, pp. 117–128, 2009.

[26]小西円. 日本語学習者と母語話者の産出語彙の相違—I-JAS の異なるタスクを用いた比較—. 国立国語研究所論集, Vol. 13, pp. 79–106, 2017.

[27]小口悠紀子. 談話における出来事の生起と意外性をいかに表すか—中級学習者と日本語母語話者の語りの比較. 日本語／日本語教育研究, Vol. 8, pp. 215–230, 2017.

[28]Shervin Malmasi and Mark Dras. Native language identification with classifier stacking and ensembles. *Computational Linguistics*, Vol. 44, No. 3, pp. 403–446, 2018.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).