Vol.2014-NL-219 No.21
2014/12/17

上層句情報の利用と学習データの選別による母語推定の精度向上

田中正浩[†1]　王瀾[†1]　山名早人[†2†3]

本稿では，英語文章を対象とした母語推定（Native Language Identification），すなわち英語文章執筆者の母語推定における推定精度向上のための手法として，i)上層句情報の利用， ii)学習データからの外れ値削除手法を提案する．2005年から始められた母語推定タスクは著者の属性推定タスクの一つであり，第二言語習得など多くの応用先があるとされる．2013 年には NLI Shared Task 2013 が行われるなど，近年盛んに研究されている．本提案手法を NLI Shared Task 2013 と同一のデータセットを用いて評価を行った結果，2014 年の最高精度である 85.3%を超える 85.6%の精度を達成し，提案手法の有効性を確認した．

Improved Native Language Identification with Upper Phrase Information and Training Data Selection

MASAHIRO TANAKA[†1]　LAN WANG[†1]
HAYATO YAMANA[†2†3]

Native Language Identification, the task of identifying the native language (L1) of a writer based solely on a sample of his/her writing in non-native language (L2), is one of the authorship attribution problem. In this paper, we propose i) “upper phrase information” as a new feature, ii) discarding essay data which seem to be outliers from the training dataset. NLI is able to applicable to many other NLP tasks such as Second Language Acquisition. From 2005, many researchers have approached this task in different ways. Combining all the proposed techniques and existing methods, our system archives 85.6% accuracy on the NLI Shared Task 2014 data. To the best of our knowledge, this is a state-of-the-art accuracy in the NLI tasks.

1. はじめに

著者の属性推定のタスクは，広告推薦など多くの分野に応用できることから盛んに研究されている．著者属性推定タスクには，著者の性別推定 [1]や Native Language Identification（母語推定．以下，NLI）等がある．本稿では，これらの内 NLI を扱う．

NLI では，ある著者が非母語（以下，L2）として書いた文章を解析し，著者の母語（以下，L1）を推測する．NLI に関する研究は，広告推薦など著者の属性推定の一般的な応用分野のみならず，第二言語習得（Second Language Acquisition，以下 SLA）学習においても有用である．また，NLI はフィッシングサイト判別にも応用できる．フィッシングとは，銀行などを騙る偽 Web サイトへユーザを誘導し，パスワード等を盗む行為を示し，近年その被害額は増加の一途をたどっている． 2013 年度のフィッシングサイトによる被害金額は，2012 年度比で 29 倍[2]となっており，その対策が重要視されている．フィッシングサイトでは不自然な言語が使われていることがあることが報告[3]されており，NLI はそのようなフィッシングサイト判別のための一手法としての利用が期待できる．

NLI では，特に英語を L2 とする著者が書いた英語の文章を対象にすることが多く，2013 年には，TOEFL のテストで書かれたエッセイから作成された TOEFL11 データセットを用いた NLI Shared Task 2013 が開催されるなど，近年盛んに研究されている．NLI Shared Task 2013[4]での最高精度は 83.6%である．TOEFL11 データセットはその後，NLI の研究によく用いられるようになった．2014 年には Ionescu ら[5]が String Kernel を特徴量として用い，TOEFL11 データセットと NLI Shared Task 2013 と同一の条件で 85.3%の精度を達成した．

従来の NLI 研究では，表層格の直上の品詞の情報や，Context Free Grammar の生成規則が使われてきたが，直上より上層の句情報と表層格を直接組み合わせた特徴量や，学習データのエッセイに関する考察は十分に行われてこなかった．直上より上の句情報を用いることで，例えば文頭に使われやすい表現などを特徴量として利用することができる．また，学習する際に外れ値となってしまうようなデータを予め選別することにより，全体の精度向上に寄与すると考えた．そこで本稿では，TOEFL11 データセットを用いて NLI Shared Task 2013 と同一条件の元で，表層格の直上より上層の句情報を特徴量として利用し，また学習データの選別によって NLI の精度向上を目指す．

本稿は以下の構成を取る．まず 2 節で NLI における関連研究についてまとめる．3 節で本稿において使用する

†1 早稲田大学大学院　基幹理工学研究科
†2 早稲田大学　理工学術院
†3 国立情報学研究所

TOEFL11 データセットを説明した後，4 節で提案手法について述べる．5 節で実験と結果について述べた後，6 節で本稿をまとめる．

2. 関連研究

NLI では機械学習が用いられており，特徴ベクトルとして何を採用するかによって大きく 2 つに分類できる．具体的には, 個々のデータから n-gram などを用い特徴ベクトルを独立に生成する手法と，String Kernel などを用いてデータ間の相互関係から特徴ベクトルを生成する手法がある．

2.1 n-gram などの方法で特徴ベクトルを生成する手法

Tomokiyo ら[6]は発話を文章として書き起こしたものを対象として，用いられた言語が話者にとって L1 か L2 かの 2 値分類を行った. 用いたデータセットは Tomokiyo らが作成したもので, 45 人の話者が話した内容を書き起こしたものである．話者の L1 は中国語，日本語，英語の 3 言語であり，96%の精度を達成した．

Koppel ら[7]は機能語や品詞 n-gram，文字 n-gram，英語のスペルミスを特徴量として Support Vector Machine （以下, SVM）で学習した．5 つの言語（Czech, French, Bulgarian, Russian, Spanish）をそれぞれ L1 とする著者らが書いた 258 個ずつの文章をデータセットとして用いて分類し，結果として 80.2%の精度を達成した．

これまで NLI 研究に用いられてきたデータセット（～2012 年）は，International Corpus of Learner English（以下，ICLE）である．しかし，2013 年に Brooke ら[8]は，「ICLE にはトピックのバイアスが存在し，NLI の研究には向かない」という事実を示した．そこで Tetreault ら[9]が新しく TOEFL11 データセットという NLI の研究に向いたデータセットを作成し，以降の研究では TOEFL11 データセットがよく用いられている．

2013 年には，NLI の初めてのコンペティションとして，NLI Shared Task 2013[3]が行われた．TOEFL11 データセットを用いて，29 チームで NLI の精度が競われた．最高精度は 83.6%である．2014 年には，Bykh ら[10]が Context Free Grammar （以下，CFG）を利用して，表層格と品詞の置き換えに関する最適化を行い, TOEFL11 データセットを用いて NLI Shared Task 2013 と同一の条件で精度 84.8%を得た．

2.2 データ間の相互関係を特徴ベクトルとする手法

Ionescu ら[5]は String Kernel と呼ばれる手法を用いて抽出した特徴量を学習に利用し，精度として現在で最高精度の 85.3%を得た．String Kernel とは，ある 2 つの文字列間の類似度を測る手法の一つであり，学習段階ではトレーニングデータ内のエッセイ同士で相互に String Kernel を用いて類似度を求めベクトル化したものを特徴ベクトルとして学習する．テストではテストデータ中のエッセイとトレーニングデータ内のエッセイ同士で同様に String Kernel を用いて類似度を求め，結果として得られたベクトルを特徴ベクトルとして，L1 の判別を行っている．

3. データセット

本節では，本稿での実験に用いた TOEFL11 データセットの詳細について説明する．

TOEFL11 データセットは，表 1 に示す 11 言語を L1 とする著者らが書いた L2 のエッセイから成る．ここで，L1 とはエッセイ著者の母語であり, L2 とは著者にとっての非母語である．TOEFL11 データセットでは，L2 は全て英語である．

表 1 TOEFL11 データセットの言語

A. Arabic,
B. Chinese
C. French
D. German
E. Hindi
F. Italian
G. Japanese
H. Korean
I. Spanish
J. Telugu
K. Turkish

また，エッセイは Training（以下，Train）, Development（以下，Dev）, Test と分けられている．内訳は表 2 のとおりである．

表 2 TOEFL11 データセットの言語ごとのエッセイ数

	エッセイ数
Train	900
Dev	100
Test	100

表 2 の数値はそれぞれの言語毎のエッセイ数である．つまり, TOEFL11 データセット全体のエッセイ数は 12,100 エッセイである．NLI Shared Task 2013 では，予め L1 ラベルが付随した Train データと Dev データがまず配布された．NLI Shared Task 2013 の参加者は Train データを用いてモデルを構築し，Dev データを用いて構築したモデルの評価を行いモデルの改善を行った．その後に L1 ラベルの付随しない Test データが配布され，参加者は各々が構築したモデルを用いて Test データにあるエッセイの L1 を判別する．

4. 提案手法

提案手法は，関連研究において説明した「個々のデータ

から独立に特徴ベクトルを生成する手法」をベースとしている．Ionescu らの手法[5]では，データ間の相互関係を特徴として抽出しているため，言語に依存しない NLI が可能となる．しかし，ある L1 にとっての L2 で使いがちな単語や表現などの特徴量を用いることができない．これは， 特徴ベクトルとしてエッセイ間の類似度を用いていることに起因する．エッセイ中の表現自体を特徴ベクトルの要素として採用していないため，L1 に依存する表現などを抽出することができず，SLA などに応用することができないと考えた．関連研究との大きな差異は 2 つある．

- 学習に不適当なエッセイの削除
- 上層の句情報の特徴量としての利用

従来の NLI 研究では，表層格の直上の品詞の情報や，Context Free Grammar の生成規則が使われてきたが，直上より上層の句情報と表層格を直接組み合わせた特徴量や，学習データのエッセイに関する考察は十分に行われてこなかった．直上より上の句情報を用いることで，例えば文頭に使われやすい表現などを特徴量として利用することができる．また，学習する際に外れ値となってしまうようなデータを予め選別することにより，全体の精度向上に寄与すると考えた．

4.1 学習に不適当なエッセイの削除

図 1 に，TOEFL11 データセット中のそれぞれのエッセイに含まれる単語数の分布を示す．

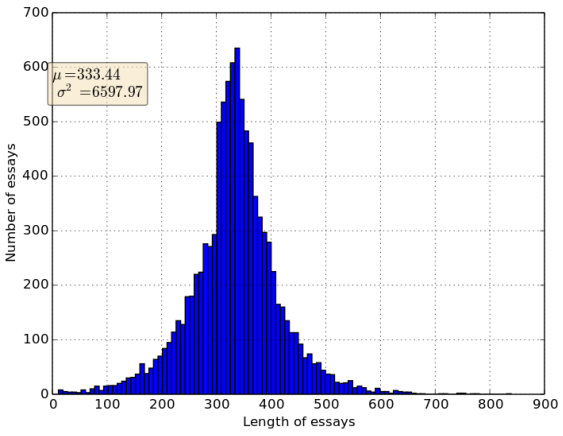

図 1 TOEFL11 データセットの単語数の分布

ここで，単語数のエッセイあたりの平均は 333.44,であり，分散は 6597.97 である．図 1 からわかるように，それぞれのエッセイに含まれる単語数には偏りがあり，一つのエッセイに含まれる単語数が 10 語に満たないものや 800 語を超えるものがある．既存の研究では単語数の大きく異なるエッセイでも同様に学習データとして扱っていたが，極端に単語数が多いエッセイ，あるいは極端に単語数が少ないエッセイは学習する際に不適当であると考えた．何故なら，極端に単語数が多いまたは少ないエッセイから特徴量を抽出すると外れ値になると考えられるためである．極端に単語数の少ないエッセイの例を表 3 に示す．

表 3 TOEFL11 データセット中の短いエッセイの例

ファイル名	エッセイ
1474522.txt	I agree.

表 3 のようなエッセイは，データセット中の単語数の平均が 333.44 であることを考えると極端に短く，特徴量を抽出しても外れ値になり得る．そこで我々は，外れ値となるようなエッセイを学習データから除くという手法を提案する．具体的には，エッセイの長さが正規分布に近似できると仮定した上で， L1 毎にエッセイの長さの分布を取り，平均と標準偏差を L1 毎に計算する．計算した値を用いて，式(1)に示す条件に合致したエッセイのみを学習に用いる．

$$-n_1 \times \sigma \leq lenEssay \leq n_2 \times \sigma \qquad (1)$$

ここで$lenEssay$はエッセイ中の単語数を表す． $n_1, n_2 \in \{1,2,3,4\}$とし，σは言語ごとのエッセイ中の単語数の標準偏差とする．

4.2 上層句情報の利用

4.2.1 Bykh らの手法

句情報を利用した NLI 研究において，Bykh ら[2]は CFG を特徴量として利用する際に，以下の 3 つを特徴量として利用している．

- CFGRph: 句情報のみの生成規則
  - 規則の例: ROOT→N VP, VP→V S, …
- CFGRlex: 表層格と，直上の句情報の利用
  - 規則の例: N→John, V→hit, D→the, …
- CFGRph∧lex: CFGRph と CFGRlex を同時に利用

Bykh らは[2]において，CFGRph∧lex を使用した場合が最も高い精度が得られたと報告している．具体的には，CFGRph∧lex と，単語 n-gram，品詞 n-gram，及び見出し語化した単語 n-gram を併用し，最高精度 84.82%を達成している．

図 2 に示す構文解析木を例にとって説明する．Bykh らが使用した CFGRph は，図 3 のようになる．また，CFGRlex を図 4 に示す．

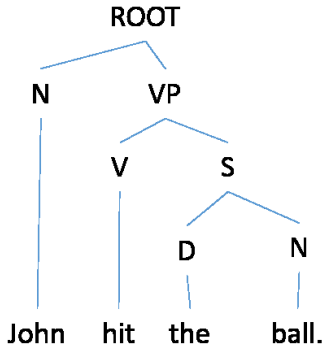

図 2 構文解析木の例

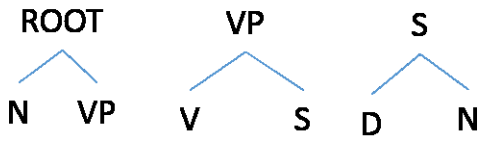

図 3 CFGRph の例

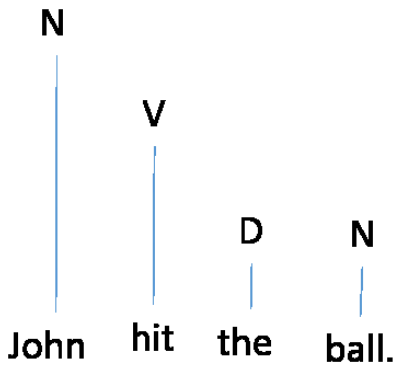

図 4 CFGRlex の例

4.2.2 単純化した句情報の利用

Bykh らの手法では，得ることが困難な特徴として「文書の書き出し」などがある．例えば，図 4 において「John」という固有名詞が「ROOT」として用いられているという特徴を利用できないことなどが挙げられる．これは，CFG の生成規則では一つ上位の概念しか定義されないためである．そこで，我々は図 5 のように分解したもの（上層句情報）を特徴量として利用することを考えた．

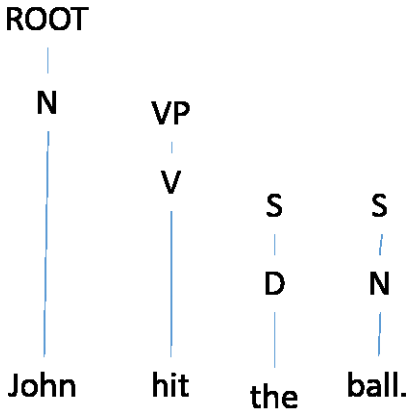

図 5 単純化した句情報の例

具体的には，構文解析木を構築した後，表層格の 2 つ上までの句情報を利用する．構文解析器には Stanford Parser を利用した．

このような手法をとることで，固有名詞が ROOT として使われているということが特徴量として利用できる．

5. 実験と結果考察

実験には，3 節で示した TOEFL11 データセットを使用する．NLI Shared Task 2013 の設定では，Train データ，または Train データと Dev データを両方用いたデータを学習用として用いて，Test データをテストとして用いる．本稿では Train データと Dev データを両方用いたデータセットを学習データとした．

5.1 学習に不適当なエッセイの削除

表 4 に式（1）に示した n1 と n2 をそれぞれ変化させた場合の精度の変化を示す．なお，ここで使用する特徴量は単語{1,2}-gram であり，Gebre らの手法と同様に TF-IDF で重み付けを行っている．使用する機械学習は線形カーネルを用いた SVM であり，実装としては liblinear を用いた．

表 4 学習データを減らした際の精度の変化

		n1			
		1	2	3	4
n2	1	0.826	0.83	0.829	0.827
	2	0.824	0.823	0.829	0.829
	3	0.822	0.821	**0.834**	0.833
	4	0.825	0.826	0.826	0.831

結果として，$n_1 = 3, n_2 = 3$としたときが最も精度が良かった．このことから，学習データから学習に不適当なデータを取り除いた場合のほうが精度向上につながることが確かめられた．

5.2 単純化した句情報の利用

上層句情報の利用により精度向上ができることを確認するために，上層句情報を利用した場合と利用しなかった場合とで精度の比較を行った．なお，この実験の際には 5.1 項で示した学習データに不適当なエッセイの削除で最も精度の高かった式(1)において$n_1 = n_2 = 3$として学習データの削除を行った．結果を表 5 に示す．

表 5 上層句情報の有無での精度の変化

	精度
上層句情報を用いた場合	85.6%
上層句情報を用いなかった場合	84.8%

句情報以外に用いた特徴量を表 6 にまとめる．これらは NLI の研究で広く使われているものであり，新規性はない．

表 6 最終的なシステムに使用した特徴量

種類	n （複数の n-gram を併用）
単語 n-gram	1,2,3,4
品詞 n-gram	1,2,3
文字 n-gram	2,3,4

表 7 最終的なシステムの混同行列

		予測されたラベル										
		Arabic	Chinese	French	German	Hindi	Italian	Japanese	Korean	Spanish	Telugu	Turkish
実際のラベル	Arabic	84	0	1	2	3	1	1	2	5	1	0
	Chinese	2	86	0	1	1	0	6	2	0	0	2
	French	2	0	86	4	2	2	1	1	1	0	1
	German	0	0	2	94	1	1	9	9	1	0	1
	Hindi	1	1	0	0	86	0	0	1	2	8	1
	Italian	1	1	2	1	0	91	1	0	3	0	0
	Japanese	3	2	0	2	0	1	88	4	0	0	0
	Korean	0	6	1	0	0	1	8	82	1	0	1
	Spanish	2	1	5	1	1	5	1	1	82	0	1
	Telugu	1	0	0	0	21	0	0	0	0	78	0
	Turkish	2	4	1	1	0	2	3	0	2	0	85

5.3 最終的なシステムの結果と考察

最終的なシステムでは，学習に不適当なエッセイを削減した上で，上層句情報の他に単語 n-gram や品詞 n-gram を併用した．なお，これらの特徴量は我々の実験で最も高い精度を出した特徴量の組み合わせである．また，それぞれの特徴量は Gebre らと同様に TF-IDF で重み付けを行った．また，機械学習には 5.1 項と同様に線形 SVM を用いた．表 7 に 85.6%を達成した際の混同行列を示す．

提案手法と同様に「n-gram などを用いた特徴ベクトルを作成する手法」を用いた Bykh らの手法と比べると精度は 0.8%ほどの向上である．また，Ionescu らの String Kernel を用いた手法と比べると精度向上は 0.3%であるが，4 節で述べたように Ionescu らの手法と異なり，提案手法では L1 に特異な特徴を重み付けられるという利点がある．

5.4 誤判定の例と考察

本項では，提案手法のシステムで誤判定したエッセイの例を示す．

表 8 誤判定したエッセイの例

ファイル名	エッセイ
1001053.txt	I AGREE
1024129.txt	I agree that it is more important （中略） Second, when we understand ideas and concepts, we will remember easily, For example, please think about **Chinese characters**. We learned （以下略）

表 8 の 1 つ目の例では，エッセイの長さが短すぎるために，十分な特徴が取れないので，判定を間違えていると判断できる．

表 8 の 2 つ目の例では，エッセイ中に”Chinese characters”という表現が存在する．”Chinese character”という表現は中国語と判定されるような重みが強いため，L1 を中国語と判定したと推測される．

また，表 7 を見ると，”Telugu”というラベルがついているエッセイに対して”Hindi”と誤判定しているケースが多い．”Telugu”とはインドの一部で用いられている言語であり，”Hindi”と文法上良く似ている．そのために誤判定しているケースが増えたと考えられる．こうした文法上似ている言語を L1 とする著者が書いた L2 のエッセイを判別できる手法を考えることが，今後重要であると考えられる．

6. まとめと今後の課題

本稿では，単純化した句情報と特徴量選定の最適化による母語推定の精度向上について示した．結果として，2014 年時点での最高精度である 85.3%を超える 85.6%を達成した．しかし，精度向上の度合いはわずかであり，更なる精度向上が望まれる．今後は，新たな特徴量の選定を行い，更なる精度向上を行いたい．また，TF-IDF とは異なる重み付けを検討している．

参考文献

1) Moshe Koppel, Shlomo Argamon and Anat Rachel Shimoni, “Automatically Categorizing Written Texts by Author Gender”, LLC, pp. 401-412 (2002).
2) フィッシングレポート 2014, “https://www.antiphishing.jp/report/pdf/phishing_report_2014.pdf”, 2014-11-07 アクセス.
3) J-cast テレビウォッチ, “http://www.j-cast.com/tv/2013/11/19189320.html”, 2014-11-05 アクセス.
4) Joel Tetreault, Daniel Blanchard and Aoife Cahill, “A Report on the First Native Language Identification Shared Task”, BEA Workshop, pp. 48-57 (2013).
5) Radu Tudor Ionescu, Marius Popescu and Aoife Cahill, “Can characters reveal your native language? A language-independent

approach to native language identification", EMNLP pp. 1363-1373 (2014).
6) Laura Mayfield Tomokiyo and Rosie Jones, "You're Not From 'Round Here, Are You? Naïve Bayes Detection of Non-native Utterance Text", NAACL, pp. 1-8, (2001).
7) Moshe Koppel, Jonathan Schler and Kfir Zigdon, "Determining an Author's Native Language by Mining a Text for Errors", KDD, pp. 624-628 (2005).
8) Julian Brooke and Graeme Hirst, "Robust, Lexicalized Native Language Identification", COLING, pp. 391-408 (2012).
9) Daniel Blachard, Joel Tetreault, Derrick Higgins, Aoife Cahill and Martin Chodorow, "TOEFL11: A Corpus of Non-Native English", Technical report, Educational Testing Service Research Report, No. RR-13-24 (2013).
10) Serhiy Bykh and Detmar Meurers, "Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization", COLING, pp. 1962-1973 (2014).
11) Binyam Gebrekidan Gebre, Maros Zampieri, Peter Wittenburg and Tom Heskes, "Improving Native Language Identification with TF-IDF Weighting", BEA Workshop, pp. 216-223, (2013)