情報処理学会論文誌 Vol. 49 No. 12 4039–4043 (Dec. 2008)

テクニカルノート

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

吉見毅彦[†1] 小谷克則[†2]
九津見毅[†3] 佐田いち子[†3]

本稿では，学習者による作文と母語話者による作文を訓練事例とした機械学習で構築した識別器を用いて学習者の作文を自動識別する手法について述べる．提案手法では，学習者と母語話者の作文の違いを表す特徴（素性）として，日本語作文とそれに対応する英文における単語どうしの対応に着目した．検証実験の結果，(1) 92.7%の識別精度で学習者と母語話者の作文の識別が可能であることと，(2) 学習者の作文を2段階で人手評価したとき，識別精度は下位群で有意に高く，上位群で有意に低いことが確認できた．これらの結果から，提案手法によって学習者の作文の評価を支援できることが示唆される．

Automatic Classification Using Word Alignment for Japanese Learners' Composition

Takehiko Yoshimi,[†1] Katsunori Kotani,[†2]
Takeshi Kutsumi[†3] and Ichiko Sata[†3]

This paper presents a method of automatically classifying sentences written by learners of Japanese as a second language. We use machine learning algorithms to construct classifiers that distinguish learners' sentences from native speakers' sentences. For this distinction, our method analyzes unnatural literal translation (word-for-word translation) in composition, and constructs a classifier based on features derived from word alignment, which presents the literal translation. In our experiment, we found that our method achieved a classification accuracy of 92.7%, and concluded that this method should assist the evaluation of Japanese learners' composition.

1. はじめに

第2言語教育においては，1人1人の学習者による作文を教員が正確に評価することが重要である．各学習者の作文を評価する教員の負担を軽減するため，学習者による作文（以下，学習者作文）を自動的に識別する試みがあり，そのような自動識別手法の1つとして，機械学習を利用した方法がこれまでに示されている[4]．この方法では，識別対象の学習者作文が流暢かつ適切であれば，その作文は母語話者による作文に近く，そうでなければ学習者によるものに近いと仮定する．この仮定の下で，母語話者による作文（以下，母語話者作文）と学習者作文を訓練事例とした機械学習によって識別器を構築する．そして，この識別器を用いて識別対象の学習者作文が流暢で適切な作文である（母語話者によるものに近い）かそうでない（学習者によるものに近い）かの判定を行い，学習者作文の評価を支援する．

このような方法で学習者作文の自動識別を行う場合，母語話者作文と学習者作文の違いを適切に表す特徴を機械学習で用いる素性として選ぶ必要がある．文献4)では，素性として，作文の構文解析に使用された構文解析規則，動詞と主語や目的語との間の共起関係，作文に現れる単純名詞句における主名詞と冠詞などが用いられている．また，文献1)では，N個（$N = 1, 2, 3$）の要素からなる単語列や品詞列などが用いられている．これに対して，本研究では，英語を母語とする日本語学習者による作文を識別対象とし，作文に現れる逐語訳（単語どうしの直訳）に着目した．2章で述べるように，母語話者作文での逐語訳と学習者作文での逐語訳の違いを表現するために，英文と母語話者作文の間，および英文と学習者作文の間で単語対応付けを行い，その結果を機械学習のための素性とする．

習熟度の低い学習者による作文に比べると，習熟度の高い学習者による作文は，母語話者作文との違いがより少ないため，母語話者作文との識別がより困難である．このため，習熟度の高い学習者による作文に対する識別器の識別精度は，習熟度の低い学習者による作文に対する識別精度よりも低いはずである．先行研究[1),3),4)]では，識別器の検証に用いられた実験データに対して一定の識別精度が得られることは示されているが，学習者の習熟度の向上にともなって識別器の識別精度が低下するかどうかは検証されていない．このため，学習

†1 龍谷大学
Ryukoku University
†2 関西外国語大学
Kansai Gaidai University
†3 シャープ株式会社
Sharp Corporation

者作文の自動識別手法として妥当であるかどうかは明らかではない[⋆1]．そこで，本稿では，学習者の習熟度の向上にともなって識別器の識別精度が低下するかどうかを検証し，提案手法の妥当性を明らかにする．

2. 着目した素性

学習者作文には様々な誤りが含まれている．学習者作文に含まれる誤りを分類する観点として，誤りの領域と誤りの原因がある．誤りの領域の観点からは，文字・表記の誤り，語彙・意味の誤り，文法の誤り，文体の誤りなどに分類できる．誤りの原因の観点からは，第2言語独特の用法の難しさによる誤り，母語の干渉による誤りなどに分類できる．母語の干渉による誤りは，母語での用法をそのまま第2言語にあてはめてしまったことによる誤りである．特に学習の初期段階においては，母語の言語習慣の干渉を受けて誤りを犯すことが多い[2)]．

本研究では，母語話者作文と学習者作文の違いを適切に表現するための特徴として，母語の干渉による誤りに着目した．母語の干渉は，文字・表記の誤り，語彙・意味の誤り，文法の誤り，文体の誤りなどの原因となるが，本研究では，母語話者作文と学習者作文の識別のための言語的特徴を自然言語処理技術によって作文から自動的に抽出することを前提としているため，現状の技術でもほぼ適切に扱える語彙の誤り（の一部）を対象とする．

学習者が第2言語の適切な表現を習得していない場合，母語の影響を受けて逐語訳を行うことがある．たとえば，次の英文（E1）は学習者作文（L1）に対する学習者本人による英語訳である[⋆2]．これに対して，日本語として自然で適切な文は（N1）のようになるだろう．

（E1） The students in these classes had very high motivation and their talents were very strong.

（L1） このクラスの学生達の動機がとても高くて能力が強かったです.

（N1） このクラスの学生は大変やる気があり才能にも恵まれていた.

母語話者作文（N1）と学習者作文（L1）を比べると，両者の主な違いとして “high motivation” の翻訳と “talents were strong” の翻訳があげられる．母語話者作文（N1）では，前者は「やる気がある」，後者は「才能に恵まれていた」と訳され，自然な翻訳になっている．一方，学習者作文（L1）では，そのまま「高い動機」，「能力が強かった」という逐語訳になっている．このような違いが主な原因で，学習者作文（L1）が母語話者作文（N1）に比べて不自然になっていると考えられる[⋆3]．

表 1　英文（E1）と母語話者作文（N1）の単語対応付け結果と英文（E1）と学習者作文（L1）の単語対応付け結果
Table 1　Word alignments of (N1) and (L1) with (E1).

母語話者作文（N1）	学習者作文（L1）
align(student, 学生)	align(student, 学生)
align(in, の)	align(in, の)
align(these, この)	align(these, この)
align(class, クラス)	align(class, クラス)
align(very, 大変)	align(very, とても)
align(motivation, やる気)	align(high, 高い)
align(talent, 才能)	align(motivation, 動機)
align(be, いる)	align(talent, 能力)
align(., .)	align(be, です)
non-align(the)	align(strong, 強い)
non-align(have)	align(., .)
non-align(high)	non-align(the)
non-align(and)	non-align(have)
non-align(their)	non-align(and)
non-align(very)	non-align(their)
non-align(strong)	non-align(very)
non-align(は)	non-align(達)
non-align(が)	non-align(の)
non-align(ある)	non-align(が)
non-align(に)	non-align(て)
non-align(も)	non-align(が)
non-align(恵まれる)	non-align(た)
non-align(て)	
non-align(た)	

母語話者作文と学習者作文の間には以上のような違いがあるため，単語どうしの対応付けを英文と母語話者作文の間で行った場合と，英文と学習者作文の間で行った場合を比べると，後者のほうが単語対応が付きやすいと予想される．英文（E1）と母語話者作文（N1）に対して単語対応付けを行うと**表 1** の左側のようになり，英文（E1）と学習者作文（L1）の間で行うと表1の右側のようになる．表1において，align(X, Y) は X と Y が対応付けら

⋆1 学習者作文ではなく機械翻訳システムによる翻訳を対象とした実験[5)] では，機械翻訳システムによる翻訳を翻訳品質が上位のものと下位のものに分けたとき，下位群での識別精度より上位群での識別精度のほうが低いことが示されている．

⋆2 学習者作文（L1）は，3 章で述べる実験で用いた作文データベースに実際に含まれている文である．

⋆3 逐語訳が常に不自然さの原因になるわけではない．

れた単語の対であることを表し，non-align(Z) は Z が対応付けられずに残った単語であることを表す．以下，align(X, Y) を対応単語対と呼び，non-align(Z) を未対応単語と呼ぶ．表 1 を見ると，母語話者作文（N1）では “high” も “strong” も単語対応が付いていないのに対して，学習者作文（L1）ではそれぞれ「高い」，「強い」と単語対応が付いていることが分かる．これらのことから，母語話者作文と学習者作文の違い（の一部）を対応単語対と未対応単語によって表現できると考えられる．

本稿では機械学習にサポートベクタマシン [6] を用いる．サポートベクタマシンによる機械学習では，母語話者作文（N1）と学習者作文（L1）をそれぞれ，表 1 に示したような対応単語対と未対応単語を成分とする素性ベクトルで表す．具体的には，母語話者作文と学習者作文における対応単語対と未対応単語を素性名番号（整数）に変換し，その素性値を 1 とする．

3. 実験と考察

3.1 実 験 設 定

提案手法の検証実験に必要な言語資源は，英文，学習者作文，母語話者作文である．今回の実験では，国立国語研究所で作成された「日本語学習者による日本語作文と，その母語訳との対訳データベース」[⋆1] を用いた．このデータベースには，日本語学習者が執筆した日本語作文と，作文執筆者本人が日本語作文を訳した英文が含まれている．このデータベースから，英語が母語であると考えられる日本語学習者のデータを 689 件抽出した．これらの英文を日本語母語話者 1 名が日本語に翻訳して，英文，学習者作文，母語話者作文の 3 つ組を作成した．

これらの英文と学習者作文の間，および英文と母語話者作文の間で単語対応付けをそれぞれ行った．単語対応付けにはシャープ（株）で開発されたソフトウェア [7),⋆2] を用いた．このソフトウェアでは，まず日本語作文を形態素に分割し，日本語の形態素と英語の単語を対応付ける．

英文と学習者作文の間の単語対応付け結果と英文と母語話者作文の間の単語対応付け結果とを合わせた 1,378 件を事例集合とした．この事例集合を 5 分割し，交差検定を行った．

サポートベクタマシンによる機械学習には TinySVM [⋆3] を利用した．サポートベクタマシンでは，どのようなカーネル関数を用いるかや，ソフトマージンのパラメータ（訓練事例の誤分類に対するペナルティ）C をどのように設定するかなどによって識別器の識別精度に影響が出る．カーネル関数は，自然言語処理研究の分野で用いられることが比較的多い d 次多項式（$d = 1$，2，3，4）とした．ソフトマージンのパラメータ C は，0.00001，0.0001，0.001，0.01，0.1，1 と変化させた．カーネル関数とソフトマージンのパラメータ以外は，TinySVM で標準設定されている値を使用した．

表 2　使用した素性ごとの識別精度
Table 2　Feature dependency of classification accuracy.

使用素性	対応単語対	未対応単語	両方
識別精度	78.7%	85.3%	92.7%

学習者作文 689 文を日本語母語話者 1 名が 100 点満点で採点し，50 点を境界として上位群と下位群に分けた．上位群に分類された文は 478 文（69.4%）であり，下位群に分類された文は 211 文（30.6%）であった．

3.2 素性の種類の識別精度への影響

識別器の構築の際に使用する素性の種類によって提案手法の識別精度がどのように変化するかを検証した．このために，素性として対応単語対 align(X, Y) だけを使用して構築した識別器，未対応単語 non-align(Z) だけを使用して構築した識別器，対応単語対と未対応単語の両方を使用して構築した識別器についてそれぞれ識別精度を求めた．カーネル関数として 4 次多項式を用い，ソフトマージンのパラメータ C を 0.00001 としたときの識別器による試験事例の学習者作文に対する識別精度を**表 2** に示す．数値は 5 分割交差検定の平均値である．

表 2 より，未対応単語だけを用いた場合の識別精度が対応単語対だけを用いた場合の識別精度よりも高くなっていることが分かる．このことから，母語話者作文と学習者作文を区別する特徴としては，未対応単語のほうが対応単語対よりも有効であるといえる [⋆4]．また，対応単語対と未対応単語の両方を用いることによって最も高い識別精度が得られていることが分かる．

3.3 学習者作文の人手評価結果と識別精度の関係

学習者作文に対する人手評価の向上にともなって識別器の識別精度が低下するかどうかを検証した．カーネル関数として 4 次多項式を用い，ソフトマージンのパラメータ C を

⋆1 http://www2.kokken.go.jp/eag/
⋆2 http://www.slc.atr.jp/IWSLT2006/proceedings/EC_15_SLE_slides.pdf
⋆3 http://chasen.org/~taku/software/TinySVM/
⋆4 対応単語対よりも未対応単語を用いたほうが識別精度が高くなる理由についての詳細な分析は今後の課題である．

表 3　対応単語対のみによる識別器の識別精度と人手評価
Table 3　Human assessment and accuracy of classifier based on aligned pairs.

	正	誤	識別精度
下位	172	39	81.5%
上位	370	108	77.4%

表 4　未対応単語のみによる識別器の識別精度と人手評価
Table 4　Human assessment and accuracy of classifier based on non-aligned words.

	正	誤	識別精度
下位	194	17	91.9%
上位	394	84	82.4%

表 5　対応単語対と未対応単語による識別器の識別精度と人手評価
Table 5　Human assessment and accuracy of classifier based on aligned pairs and non-aligned words.

	正	誤	識別精度
下位	203	8	96.2%
上位	436	42	91.2%

0.00001 とした．検証は，識別器の構築の際に素性として対応単語対だけを使用した場合，未対応単語だけを使用した場合，対応単語対と未対応単語の両方を使用した場合について行った．各識別器について，上位群・下位群別に，学習者作文であると正しく識別された文数と母語話者作文であると誤って識別された文数を集計した．集計結果を**表 3**，**表 4**，**表 5** にそれぞれ示す．

各識別器についての集計結果に対して χ^2 検定を行った．その結果，対応単語対だけを用いた識別器の場合，母語話者作文であると誤って識別された学習者作文の件数の偏りは有意ではなかった（$\chi^2(1) = 1.47$，$p = 0.225$）．未対応単語だけを用いた識別器の場合，母語話者作文であると誤識別された学習者作文の件数は，下位群で有意に少なく，上位群で有意に多かった（$\chi^2(1) = 10.60$，$p = 0.001$）．対応単語対と未対応単語の両方を用いた識別器の場合も母語話者作文であると誤識別された学習者作文の件数の偏りは有意であった（$\chi^2(1) = 5.43$，$p = 0.020$）．これらのことから，未対応単語だけを用いて構築した識別器の識別精度と対応単語対と未対応単語の両方を用いて構築した識別器の識別精度は，学習者の習熟度が高くなると低下するといえる．したがって，これら 2 つの識別器は学習者作文を適切に識別できる妥当な自動識別手法であることが示唆される．

4. おわりに

本稿では，母語話者作文と学習者作文を訓練事例とした機械学習によって構築した識別器を用いて学習者作文を自動識別する手法について述べた．提案手法では，母語話者作文と学習者作文の違いを表す言語的特徴として，英文と母語話者作文の間，および英文と学習者作文の間で単語対応付けを行った結果を利用した．検証実験の結果，(1) 92.7%の高い識別精度で母語話者作文と学習者作文の識別が可能であることと，(2) 学習者作文を 2 段階で人手評価したとき，評価の高いほうが識別精度が有意に低いことが確認できた．これらの結果は，学習者作文の識別器としての提案手法の妥当性を示している．

参考文献

1) Baroni, M. and Bernardini, S.: A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text, *Literary and Linguistic Computing*, Vol.21, No.3, pp.259–274 (2006).
2) Ellis, R.: *Second Language Acquisition*, Oxford University Press (1997). 牧野高吉（訳）：第 2 言語習得のメカニズム，筑摩書房 (2003).
3) Kotani, K., Yoshimi, T., Kutsumi, T., Sata, I. and Isahara, H.: Classification of Language Learners' Sentences into Native-like or Non-native-like Language based on Word Alignment Distribution, *Proc. International Technology, Education and Development Conference* (2008).
4) Lee, J., Zhou, M. and Liu, X.: Detection of Non-native Sentences using Machine-translated Training Data, *Proc. Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies*, pp.93–96 (2007).
5) Sun, G., Liu, X., Cong, G., Zhou, M., Xiong, Z., Lee, J. and Lin, C.: Detecting Erroneous Sentences using Automatically Mined Sequential Patterns, *Proc. 45th Annual Meeting of the Association for Computational Linguistics*, pp.81–88 (2007).
6) Vapnik, V.N.: *Statistical Learning Theory*, Wiley (1998).
7) Whitelock, P. and Poznanski, V.: The SLE Example-Based Translation System, *Proc. International Workshop on Spoken Language Translation*, pp.111–115 (2006).

(平成 20 年 7 月 5 日受付)
(平成 20 年 9 月 10 日採録)

吉見　毅彦（正会員）

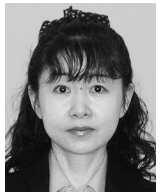
1987 年電気通信大学大学院計算機科学専攻修士課程修了．1999 年神戸大学大学院自然科学研究科博士課程修了．（財）計量計画研究所（非常勤），シャープ（株）を経て，2003 年より龍谷大学理工学部勤務．2004 年より 2008 年まで（独）情報通信研究機構専攻研究員を兼任．

小谷　克則

2004 年関西外国語大学外国語学研究科博士課程修了（英語学博士）．2002 年より 2008 年まで（独）情報通信研究機構特別研究員．2004 年より関西外国語大学外国語学部講師．

九津見　毅

1965 年生まれ．1990 年大阪大学大学院工学研究科修士課程修了（精密工学―計算機制御）．同年シャープ株式会社に入社．以来，機械翻訳システムの研究・開発に従事．



佐田いち子

1984 年北九州大学文学部英文学科卒業．同年シャープ（株）に入社．現在，同社情報通信事業本部要素技術開発センター第一開発部副参事．1985 年より機械翻訳システムの研究開発に従事．