母語話者・非母語話者言語判別システムによる学習者言語の評価法

小谷 克則[†/‡]　吉見 毅彦[††/‡]　九津見 毅[‡‡]
佐田 いち子[‡‡]　井佐原 均[‡]

[†]関西外国語大学　[‡]情報通信研究機構　[††]龍谷大学　[‡‡]シャープ株式会社

１．はじめに

第二言語学習者の作文の評価法の一つとして，学習者の作文にみられる言語的特徴を判別の要因として分析し，学習者の作文を母語話者言語か学習者言語かに判別する手法が提案されている（Lee 2007）[1]．先行研究の提案手法では，学習者の作文の言語的な特徴を分析し，母語話者言語か学習者言語のいずれかに判別する．この判別結果に基づいて，学習者の作文を評価する．また，判別の手がかりとして，母語話者言語と学習者言語の間にみられる使用語彙や構造の違いといった言語的特徴を利用することから，評価対象となる学習者の作文の言語的特徴を調査することも可能となる．

このように言語判別システムを用いて第二言語学習者の作文を評価することは，コンピュータ支援言語教育への援用といった教育的見地，および第二言語習得の発達段階調査といった研究的見地からも有益であると考えられる．本稿は母語話者・非母語話者言語判別システムにおける判別素性として，第二言語学習者の母語による影響に着目し，第二言語の作文においてを学習者の母語からの影響が大きい言語（学習者言語）か小さい（あるいは皆無の）言語（母語話者言語）かという尺度に基づいて判別を行うシステムを提案し，その検証実験結果を報告する．

２．言語判別システム

2.1　動機

第二言語習得の問題の一つとして，学習者が第二言語を使用する際に，学習者の母語にみられる言語的特徴により引き起こされると考えられる誤りが挙げられる．この問題は，特に学習者の母語と第二言語において，言語的差異が大きい場合により顕著になる．

日本語と英語のように使用する文字形式に始まり統語構造や談話構造まで幅広く異なる言語間においての第二言語習得を考えた場合，母語の言語特徴に起因する誤りが第二言語の運用時に散見される．例えば，日本語と英語における文法的な違いとして，一致（Agreement）に関する情報の表出の違いがある．日本語では単数や複数といった一致に関する情報は名詞句や動詞句において義務的に顕在化される必要がないのに対し，英語では義務的に顕在化される．そのため，日本語を母語とする英語学習者が一致に関する情報を顕在化させずに誤った英語文“He have a pen.”を出力することがある．また，日本語と英語は文主語の生成に関してもふるまいが異なる．英語では時制を伴う節において文主語が義務的に顕在化されるのに対し，日本語では文主語の省略が頻繁に容認される．そのため，英語を母語とする日本語学習者が日本語文において不必要に主語を生成することがある[2]．

このように第二言語において母語の介入による誤りは，初級学習者に顕著にみられると思われる．また，中級以上の学習者であっても，第二言語における未修得語彙や構造を補完する際に母語の影響による誤りが生じると考えられる．したがって，学習者の作文を母語の介入の程度に基づいて評価すると，学習者の習熟度を測定できるだけでなく，学習者の未修得語彙などの検出にも役立つと考えられる．そこで，学習者の作文を母語からの影響度により判別するシステムを構築することにした．

2.2　判別素性

第二言語学習者の作文において母語の介入を示す言語形式として，学習者の母語からの直訳と考えられる形式に着目した．学習者の作文中の直訳表現は，第二言語の適切な言語表現が未修得の場合，母語からの機械的な置換作業による翻訳の結果もたらされると考えられる．その結果，第二言語として自然な表現が生成される場合，問題はない．しかし，不自然と判断される場合，不自然な直訳は第二言語の学習の結果ではないと考えられるから，学習者の母語からの影響による誤りでとして位置づけられる．

もし学習者が第二言語のみで思考を言語表現化できれば，語彙使用や文法といった言語形式に表出する母語からの影響はないはずである．しかし，第二言語の習得段階にある学習者にとっては第二言語のみの運用が困難であるため，母語による支援が必要となる．そのため，母語により生成された言語表現形式を第二言語による表現形式へと翻訳する必要が生じる．初級学習者であればあるほど，第二言語の語彙知識や文法知識が少ないため，母語からの直訳への依存が高くなると考えられる．したがって，第二言語による作文において直訳度の高い箇所の有無は学習者の習熟度を示すだけでなく，母語による影響を如実に映し出していると考えられる．

第二言語学習者の作文において母語の影響により不自然さが引き起こされる例として，先述の一致や文主語に関す

[1] 言語判別システムの先行研究として，機械翻訳と人手による翻訳との判別システム(Corston-Oliver 2001, Kulesza 2004)などが挙げられる．

[2] 英語と日本語における照応表現に関する違いは Tsujimura (2007)などを参照．

る現象がある．さらに，例（1）にみられる限定詞の扱いも挙げられる．英語名詞句は限定詞を伴って出現することが多く，（1a）のように限定詞 some により事物の存在も示すことが可能である．一方，日本語の場合，（1b）のように存在を示す英語限定詞 some に対応する名詞修飾表現「いくらか」を用いて存在の意味を示すよりも，存在を示す動詞を用いた（1c）の方が自然な表現である．

（1） a. Some students came.
b. ?いくらかの学生が来た．
c. 来た学生もいた．

筆者らは学習者の作文において直訳により生成された方言を検出するために，原文と翻訳文における単語対応付け情報を用いることにした．単語対応付け情報とは，原文と翻訳文の単語を辞書的情報に基づいて機械的に対応付けを行った結果を指す．直訳度の高い文とそうでない文では，単語対応付けの分布が異なる．例えば，（2a）の英語文を直訳すると（2b）となる．（2b）の直訳でなく文脈に応じた自然な日本文へと翻訳すると（2c）となる．（2b）は文法的には適切であるが，（2a）と同じ文脈で用いることができるかどうかと言った観点からは不自然と判断される．一方，（2c）は文法性と容認度のどちらの観点からも適切である．

（2） a. Today the sun shining.
b. ?今日，太陽は輝いている．
c. 今日は晴天だ．

原文（2a）と翻訳文（2b）の間，また，（2a）と（2c）の間，それぞれにおいて単語対応付けツールによって対応付けを行った．その結果は表 1 のとおりである．表 1 において align(A, B)とは対応付けられた単語ペアを示し，対応付けられなかった日本語単語と英語単語はそれぞれ non-align_jpn(C)と non-align_eng(D)によって示される．単語対応付け情報が適切に直訳度を反映しているのであれば，対応付けられた単語ペアと非対応単語との比率において（2a-b）と（2a-c）は異なると考えられる．実際に，対応付けの比率において，（2a-b）の方が（2a-c）よりも高いという結果（66%と 33%）が得られた[3]．

表 1：対応付けの結果

（2a-b）	（2a-c）
align(today, 今日)	align(today, 今日は)
align(is, ている)	align(is, だ)
align(sun, 太陽)	non-align_jpn(晴天)
align(shining, 輝い)	non-align_eng(the)
non-align_jpn(は)	non-align_eng(sun)
non-align_eng(the)	non-align_eng(shining)

このような対応付けと直訳の関連から，筆者らは単語対応付け情報を判別素性として利用することにした．そして，母語話者言語と学習者言語のそれぞれにおける対応付けの分布を分析し，母語話者言語に顕著にみられる直訳を自然な表現の結果生じる適切な直訳とし，学習者言語に顕著に見られる直訳を不自然な表現を引き起こす不適切な直訳として位置づける．

2.3 言語判別システムの構築

筆者らは言語判別システムを構築する手法としてサポートベクターマシンと呼ばれる機械学習法を利用した．サポートベクターマシンソフトウェアには，TinySVM を利用した．カーネル関数には一次の多項式を利用し，その他の設定はデフォルト値を利用した．学習素性には，2.1 節でみた学習者言語と母語話者言語それぞれと原言語との単語対応付け情報を用いた．

言語判別システムの構築には，母語話者言語と学習者言語が訓練データとして必要である．本稿では英語を母語とする日本語学習者の評価を目的とするため，必要な訓練データは，英語と日本語の対応データである．そして，英語データは母語話者により作成された母語話者言語データが必要であり，日本語データには母語話者言語データと学習者言語データの両方が必要となる．

母語話者言語による英日対応データとして，筆者らは英語新聞記事とその日本語対応文からなるロイター英日対訳コーパス（Utiyama2003）を選定した．このコーパスにおける言語データは英語，日本語ともに新聞記事として通用することから母語話者言語として適格であると判断した．このコーパスから，判別システムの訓練データとして 129,000 文を抽出した．

学習者言語による英日対応データとして，日本語学習者の日本語作文とその英語対応文からなる英日対訳コーパス（国立国語研究所 2001）を利用する．このコーパスにおける日本語データは日本語学習者により作成されたエッセイであるため，学習者言語による日本語データである．また，英語データは学習者自身が日本語で作成したエッセイに対して自らの母語（あるいは母語と同程度ともくされる言語）である英語により付記した対訳である．したがって，英語データは母語話者言語データとして適格である．この英日対訳コーパスから抽出した英日対訳文 689 文を学習者言語データとした．

機械学習を行う際，一般的に学習結果の信頼性は訓練データの規模に依拠すると考えられる．言語判別システムを 689 文の学習者言語データから構築することは可能であるが，母語話者言語の量と比較して十分とはいえない．そこで，大規模な学習者言語データとして，機械翻訳システムにより生成された機械翻訳言語データを代替データとして利用することにした．

機械翻訳システムにより生成される言語には学習者の言語と類似する点が多くあり，言語判別システム構築に必要な学習者言語データの代替データとして有効であることが確認されている（Lee 2007）．機械翻訳言語と学習者言語の類似点として，いずれも生成する文に何らかの誤りが含まれている点が挙げられる．機械翻訳システムの誤りには，原言語と翻訳言語の置換作業による翻訳過程で生じる誤りがある．この種の誤りは，学習者言語における母語の介入による誤りと類似していると考えられる．（両言語の類似

[3]機械翻訳文における直訳度と単語対応付けとの関連は今村(2002)でも指摘されているが，学習者言語の場合にも当てはまるかどうかは今後の課題とした．

性を単語対応付けの分布の観点からの調査結果からも支持される．詳細は次節で述べる．）

機械翻訳言語の学習者言語の代替データとしての有効性を示す先行研究の調査結果などに加え，データの収拾の容易さが機械翻訳言語データを積極的に使用する理由である．機械翻訳言語データは，機械翻訳システムと原文さえ用意できれば言語データの入手は容易である．これらの理由から，筆者らは学習者言語データの代替データとして機械翻訳言語データを利用することにした．

言語判別システム構築の訓練データにおける学習者言語データは，母語話者言語データとして利用した英日対訳コーパス（Utiyama2003）から抽出した．英日対訳コーパス（Utiyama2003）から抽出した英語原文データ 129,000 文を対象に，日本国内で市販されている機械翻訳システムにて翻訳処理し，翻訳言語データを作成した．利用した機械翻訳システムは MT-A と MT-B の二種類であった．その結果，訓練データとしての学習者言語データは MT-A 学習者言語データと MT-B 学習者言語データの二種類が得られた．

上記の母語話者言語データと学習者言語データに，訓練データとして必要な単語対応付け情報を付記した．単語対応付け情報は，筆者らが実験用に開発した単語対応付けツールを利用した．訓練データに付記された対応付け情報の件数は，表 2 に示すように，英日対訳母語話者言語データが 568,259 件，英日対訳学習者言語データ MT-A が 518,894 件，英日対訳学習者言語データ MT-B が 537,460 件となった．最終的に言語判別システムの構築と検証のために準備した言語データは以下の通りである．訓練データには英日対訳母語話者言語データ（129,000 文）と英日対訳学習者言語データ MT-A と MT-B（それぞれ 129,000 文），そして，試験データには英日対訳学習者言語データ（689 文）を使用した．

2.4 学習者言語と機械翻訳言語の類似性

機械翻訳言語が学習者言語の代替言語として妥当であれば，両言語の間に様々な類似性が確認されるべきである．そこで，本稿で判別素性として利用する単語対応付け情報の観点から両言語の類似性を確認する．

2.2 節でみたように学習者言語は母語話者言語と比較して，単語対応付けにおいて対応付けられる単語ペアの出現率が高いと考えられる．もし，機械翻訳言語データが学習者言語の代替データとして有効であれば，機械翻訳言語データにおいても対応付けられる単語ペアの出現率が母語話者と比較して高いと予測される．そこで，機械翻訳言語データと英日対訳母語話者言語データ（Utiyama2003）において単語対応の分布が異なるかどうかを調査した．英日対訳機械翻訳言語データとして，二種類の機械翻訳システム（MT-A，MT-B）により得られた言語データを利用した．

機械翻訳言語データと母語話者言語データにおける単語対応付け情報を比較した結果，表 2 に示されるように，機械翻訳言語において対応付けられた単語ペアの出現率が高いことがわかった．

そこで，機械翻訳言語データと母語話者言語データにおける対応付けの分布の違いに対して，分散分析を行った．その結果，単語対応付けの分布において両言語の間に有意差が確認できた．さらに，母語話者言語データを基準としてテューキー法により多重比較を行い，母語話者言語と機械翻訳言語の間に有意差があることを確認した（$F(3, 56)=616.10, p<0.0001$）．

これらの結果により，機械翻訳言語が単語対応付けの観点から母語話者言語と異なることが明らかになった．さらに，この結果が機械翻訳言語と学習者言語との類似性を示唆するものとであると筆者らは考えた．

表 2：単語対応の分布

	N（件）	対応ペア（%）	非対応単語（%）	対応率（%）
MT-A	518,894	37.1	62.9	59.0
MT-B	537,460	36.4	63.6	57.3
母語話者言語	568,259	24.1	75.9	31.7

3. 検証実験

3.1 目的

この実験の目的は，学習者の作文を母語による影響に基づき，母語話者言語か学習者言語かを判別する本稿の提案システムの検証である．言語判別システムの検証として，まず，提案システムの判別精度を単独で調べた．次に，判別精度が学習者の習熟度に応じて変化するかどうかを調べた．

言語判別システムは，学習者の作文にみられる言語的特徴から，母語話者か学習者言語かを判別する．したがって，ある学習者が母語話者と全く同程度の言語を出力した場合，言語判別システムの判別精度は低くなるはずである．一方，学習者が母語話者と全く異なる言語を出力した場合，言語判別システムの判別精度は高くなるはずである．このように言語判別システムの精度が学習者の習熟度に依存していると考えられることから，言語判別システムの検証として，学習者の習熟度と独立して評価するだけでなく，学習者の習熟度に応じて判別精度が変化するかどうかも調べることにした．

3.2 実験結果と考察

先行研究（Lee 2007）の調査結果や 2.4 節での機械翻訳言語と母語話者言語との比較結果から，機械翻訳言語が学習者言語の代替データとして有効であると考え，機械翻訳言語を訓練データとして言語判別システムを構築した．この言語判別システムに対して，英日対訳学習者言語データ（日本語学習者による言語データ）を試験データとして，判別精度を調べた．また，訓練データとして使用する機械翻訳システムの違いにより，判別精度が異なるかどうかも確認するために，二種類の機械翻訳システム（MT-A と MT-B）により言語判別システムを構築した．さらに，それぞれの言語判別システムにおいて，判別素性による影響を調べるために，二種類の判別素性を利用して，（1）対応・非対応情報に基づき構築された判別システムと（2）非対応情報だけに基づき構築された判別システムとを構築した．本実験では訓練データや判別素性の違いから 4 種類の判別システム（MT-A1，MT-A2，MT-B1，MT-B2）を用いて検証を行った．それぞれの判別システムにおける判別精度は表 4 に示されるとおりである．

学習素性に対応・非対応情報を利用した判別システム（1）の場合，判別精度は約 73%であった．学習素性に非

対応情報のみを利用した判別システム（2）の場合，判別精度は約 80%であった．この結果から，学習者の母語による影響に基づく言語判別システムは有効であると筆者らは考えた．

表 4：判別精度

	N（文）	判別精度（%）
MT-A1	689	74.2
MT-A2	689	80.4
MT-B1	689	73.4
MT-B2	689	80.0

次に，判別システムが学習者の習熟度を適切に反映して，その精度が習熟度に応じて変化するかどうかを調べた．本実験で試験データとして利用した英日対訳学習者言語データ（日本語学習者コーパス）は，学習者の学習歴に応じて四分割（一年未満，二年未満，三年未満，四年以上）できる．そこで，学習歴毎に学習者言語データを分割し，各データ群において判別精度を調べることにした．実際には，テストデータにおける文量を考慮して 689 文を二群（二年未満 486 文と二年以上 223 文）に分けて，判別精度の変化を調べた．その結果は表 5 のとおりである．

表 5：学習歴による低下率とその判別精度

	低下率（%）	2 年未満（%）	2 年以上（%）
MT-A1	1.5	72.6	71.5
MT-A2	3.5	80.2	77.4
MT-B1	4.1	73.0	70.0
MT-B2	8.9	80.6	73.4

判別システムの精度は，学習者歴が上がることにより低下している．また，判別精度は，学習者言語データや判別素性の違いに関係なく低下している．この結果から，判別システムが適切に学習者の習熟度を反映していると筆者らは考えた．

4．まとめと今後の課題

本稿は，英語を母語とする日本語学習者の作文の評価法として，学習者の作文にみられる母語による影響に基いて，母語話者言語か学習者言語かを判別するシステムを提案した．また，提案システムを学習者言語の判別精度と学習者の習熟度に応じた判別精度の変化といった観点から検証をおこなった．その結果，筆者らは提案システムの妥当性を確認した．

この提案システムは，学習者の作文において母語による影響が強く，かつ不自然な文を学習者言語として判別する．この母語による影響が強く不自然な表現は学習者の未修得の語彙や文法により生じると考えられる．そのため，提案システムをコンピュータ言語教育支援システムに組み込むことで，作文能力の育成を支援できると考えられる．また，第二言語習得段階において母語による影響に関する研究調査を目的とした利用も考えられる．

本稿において残された課題として，提案システムによる判別結果の定性的な分析が挙げられる．本稿では，学習者言語や判別素性の違いから四種類の判別システムを構築した．これらの判別システムにおいて，表 4 に示される結果から MT-A1 と MTB-1 において，どのような判別結果であったかを調べた．判別システム MT-A1 が判別を誤った文は 689 文中 178 文であったのに対し判別システム MT-B1 が誤った文は 183 文であった．どちらのシステムにおいても判別が誤られた文は 170 文とその大半を占めることがわかった．今後，これらの文を言語学的観点から調べ，どのような特徴がみられるかを明らかにする．

また，学習者言語データと機械翻訳言語の類似性に関しても課題が残された．本稿の実験で用いた訓練データとしての学習者言語データは新聞コーパスから抽出されたものであった．試験データとしての学習者言語データは日本語学習者の日常の話題に関するエッセイから抽出したものであった．どちらの言語データもテキストの特性としては説明文であるなどの共通点がある．その一方で，対象となる話題になどに大きな違いが見られる．例えば，新聞記事コーパスの場合，政治，経済，社会問題などが話題として取り上げられるのに対し，学習者コーパスの場合，「喫煙」などの日常的な話題に限定されている．このような話題の違いから，機械翻訳言語と学習者言語の差異が大きくなってしまった可能性も考えられる．このように機械翻訳言語データの学習者言語データとしての代替としての使用に関しても調査の余地がある．

参考文献

Corston-Oliver, S., M. Gamon and C. Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL01). 148-155.

今村賢治，隅田英一郎. 2002. 直訳性に着目した対訳コーパスフィルタリング. 第 1 回情報科学技術フォーラム（FIT2002），Vol.2 pp.185-186.

国立国語研究所 2001. 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.

Kulesza, A. and S. M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI04). 75-84.

Lee, J., M. Zhou, and X. Liu. 2007. Detection of non-native sentences using machine-translated training data. In Proceedings of the 2007 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT07). 93-96.

Tsujimura, N. 2007. An Introduction to Japanese Linguistics (2nd edition), Oxford, Blackwell.

Utiyama, M. and H. Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL). 72-79.