統計的語彙情報に基づく日本人学習者の英語習熟度の分析

鍔木　元*1,4　安田　圭志*2　山本　博史*3　匂坂　芳典*1,2,4

*1 早稲田大学 GITI　*2 NICT　*3 近畿大学 理工学部　*4 早稲田大学 ことばの科学研究所

hjm-tsubaki@asagi.waseda.jp, keiji.yasuda@nict.go.jp,

yama@info.kindai.ac.jp, yoshinori.sagisaka@gmail.com

１．はじめに

器械による音声言語能力の自動評定は、評定の省力化や評定者の能力に左右されない客観的評価の可能性、有用性が期待できる。人間の翻訳、作文に関する言語能力の自動測定についてはこれまで、翻訳システムの自動評価手法を適用した、人間による翻訳の評価[1]、英語コミュニケーション能力の自動測定[2]等の研究がなされてきている。また、語彙のリスト及び難易度を用いた、人間の自由英作文の評価[3]、英語コミュニケーション能力試験における正解率の分析[4]も進められている。我々は日英翻訳を対象として、統計的自動翻訳の統計的情報量(単語 N グラム及び翻訳確率)を用いた、人間の翻訳文の評価に関する研究[5]を進めてきた。

本研究では、従来の統計的自動翻訳で用いる統計的情報の相補的情報として、翻訳文評価尺度の語彙そのものが持つ統計的性質の英語習熟度評価への利用可能性を検討した。

２．統計的語彙情報による英語習熟度推定

先に行った統計的自動翻訳の統計的情報量を用いた日本人英語文の客観評価では、翻訳する元の日本語単語に対応する英語訳単語の適切さを示す翻訳確率の有用性が示された[5]。 この反面、もう一つの統計量である単語 N グラムを用いた英語文としての特徴量は寄与しないことが示された。英語語彙と学習者の能力には何らかの関係があるように思われるが、この結果に示されるように、日英翻訳文１文といった少量の語彙から客観評価に有用な特徴量を導くことは難しいかもしれないが、一定量の日英翻訳文データ中の語彙の使用分布は学習者能力の差異を求められるのではないかと考えた。このため、訳出英語文中の使用単語分布を特徴とした評価を試みた。

特徴量としては、日本人英語学習者の日英翻訳文を対象として、英単語の基本語彙リストを用いて抽出した統計的語彙情報を考えた。学習者の英語習熟度推定手法にはニューラルネットワークを用い、その統計的語彙情報から推定される学習者の TOEIC スコアと、実際のスコアとの相関関係を分析した。

表1 JACET8000 基本語彙リスト

単語	出現頻度順位
the	1
and	2
≧	≧
midway	7999
alight	8000

英語習熟度を評価する基本語彙リストとして、「大学英語教育学会基本語リスト（以下、JACET8000）」を用いた。JACET8000 は、大学英語教育学会が British National Corpus 内の語彙とそれらの出現頻度を基準として、日本の英語教育の現状を反映した言語資料(中学、高校の検定教科書等)及び、アメリカ英語を比較的多く含めた言語資料(アメリカの新聞、雑誌、映画のスクリプト等)を用いて補正を施し選定した、8000 語の基本語彙リストである。本リストの語彙には、上記コーパスに従った、一般的な出現頻度順位が付与されている(表1)。JACET8000 に基づき、学習者の日英翻訳文から基本語彙毎の出現総数を抽出し、本研究の統計的語彙情報として用いる。

統計的語彙情報に基づいた学習者の TOEIC スコアの推定には、3層のフィードフォワード型のニューラルネットワークを使用する。統計的語彙情報が入力情報として入力される入力層、変換関数にシグモイド関数を使用した中間層(今回はユニット数を3及び4とした)及び、学習者の TOEIC スコアが出力される出力層の3層構造である。統計的語彙情報を学習者毎に集計したデータ(表2)を入力情報として使用する。入力情報は、以下のように3つに分割する。

(a) ニューラルネットワークの予測モデル作成のための学習データ

(b) 予測モデルを適応させるための開発データ

(c) オープンテストのための評価データ

前述のニューラルネットワークと入力情報を用い、次の手順で学習者の TOEIC スコアを推定する。

(1) 学習データ入力値及び、出力値としての学習者の TOEIC スコア実測値を与え、ニューラルネットワークの学習を行う。

(2) 学習したニューラルネットワークに対し、開発データ入力値及び、出力値としての学習者の TOEIC スコア実測値を用い、出力値が実測値に適合するよう、学習回数等のパラメータを調整する。

(3) 適合後のニューラルネットワークに対し、評価データの入力値によって、学習者の TOEIC スコアを推定する。

3. 英語習熟度推定実験

JACET8000 に基づいて日本人英語学習者の日英翻訳文データから抽出した統計的語彙情報から、ニューラルネットワークを用いて学習者の TOEIC スコアを推定し、スコアの実測値との相関関係を分析した。

3.1 分析データと入力情報

日英翻訳文データには、ATR で収録された旅行会話基本表現データ(以下、BTEC)を用いた。BTEC 日本語会話文から日本語文 394 文を選定し、それらに対して日本人英語学習者が日英翻訳及び発話したものを書き下し、翻訳文データを作成した。翻訳文データは、以下のような構成になっている。

・日本語文1文あたり、21 名の日本人学習者が日英翻訳文を生成した(計 8274 文)。
・挨拶などの決まり文句の翻訳は、翻訳文データに含まれない。

また、21 名の日本人学習者の TOEIC スコアは 300 点台から 900 点台に分布しており、各 100 点台に3名ずつ存在する。これらの TOEIC スコアを実測値、ニューラルネットワークの出力値とした。

JACET8000 に基づいて日英翻訳文データから基本語彙毎の出現総数を抽出した。今回は、全学習者を通して出現数が1以上の基本語彙毎の出現総数を統計的語彙情報として使用した。対象の基本語彙は 1209 語で、それらの語彙毎の出現総数を学習者毎に集計し、ニューラルネットワークへの入力情報を作成した。

ここで統計的語彙情報特性として、JACET8000 の出現頻度順位順に従った場合の統計的語彙情報特性、日英翻訳文データで学習者が高頻度で使用した基本語彙順に従った場合の統計的語彙情報特性の2つについて検討した。

表2 統計的語彙情報の差異

特性	JACET8000 出現順位順		日英翻訳データ内の語彙出現順	
順位	単語	出現総数	単語	出現総数
1	the	1620	I	2786
2	and	316	be	2285
3	to	1390	you	1633
≈			≈	
200	each	4	need	33
201	course	7	open	33
≈			≈	
300	stay	8	desk	20
301	wait	112	soul	20

学習者の英語習熟度と対応するのは、英語コーパス基づく高頻度出現語彙、いわゆる重要英単語の使用分布なのか、それとも日英翻訳タスクの内容を反映するために高頻度で使われた語彙の使用分布なのかを調べるためである。日英翻訳タスクの内容によって、両者の語彙の使用分布は異なる。今回使用した日英翻訳文データに基づく統計的語彙情報にも、両者の差異が観察される(表2)。

どちらの統計的語彙情報特性が英語習熟度に影響を与えるのかを調べるために、次の2つの条件で並び替えた入力情報を用いる。

JACET 頻度順:
JACET8000 の出現頻度順位順に並び替えた入力情報

当該タスクに基づいた頻度順:
日英翻訳文データで学習者が高頻度で使用した基本語彙順に並び替えた入力情報

また、統計的語彙情報量と英語習熟度推定との関係を調べるために、英語習熟度推定に使用する情報量の度合いについて以下の2つの特徴量を考えた。

高頻度語彙に限定した特徴量:
語彙個々の出現数が多くしかも学習者間でばらつきが顕著な情報量を有すると思われる基本語彙に限定した、すなわち、JACET8000 の出現頻度順位の上位からN語まで、もしくは出現総数の降順に上位からN語までに限定したデータが、英語習熟度に影響を与える。

全体的な語彙頻度に基づく特徴量:
基本語彙の全ての出現数、すなわち使用された全基本語彙の全体としての出現分布が、英語習熟度に影響を与える。

表3 各パターンの学習者 TOEIC スコアの推定値と実測値の相関

統計的語彙情報特性	中間層ユニット数	上位 200 までの語彙毎の情報量	5語彙毎にまとめた全語彙情報量	10 語彙毎にまとめた全語彙情報量
JACET8000 の出現頻度順位順	3	0．45	0．54	0．64
	4	0．50	0．64	0．59
学習者使用語彙の高頻度順	3	0．79	0．56	0．74
	4	0．77	0．77	0．60

高頻度語彙に限定した特徴量は、高頻度語彙それぞれの統計量をしようしているので、それらに対する精度は高いが、データの個数が少ないため信頼度に欠けるおそれがある。また、全体的な語彙頻度に基づく特徴量は全体の分布特性を反映しているが、単語固有の特性は捉えきれていないため精度が十分取れていないきらいがある。今回は、高頻度語彙に限定した特徴量については上位からの語数を 200 語としたデータを使用した。また、全体的な語彙頻度に基づく特徴量については 1209 語の全出現数データを使うべきであるが、ニューラルネットワーク処理の際の負荷と要する時間を考慮し、今回は一定の語彙単位(5語彙毎と 10 語彙毎)にまとめたデータを用いた。上記の2つの条件と組み合わせ、計6種類の入力情報で英語習熟度推定実験を行った。

3.2 ニューラルネットワークによる英語習熟度推定

21 名の日本人学習者の入力情報を7名ずつ学習、開発及び評価データに分割した。各 100 点台に3名ずつ存在ので、各データについて TOEIC スコアの 300 点から 900 点の範囲で各 100 点台に1名ずつ均等に配置した。

学習データと学習者の TOEIC スコアの実測値で予測モデルを作成した。そのモデルに対して、開発データと学習者の TOEIC スコアの実測値を用いて、予測モデルの出力値が実測値に適合するよう、学習回数等のパラメータを調整した。調整後のパラメータを用い、学習データ及び開発データ、学習者の TOEIC スコアの実測値で予測モデルを作成し、評価データで学習者の TOEIC スコアを推定し、実測値との相関を求めた。結果を表4に示す。また、最も高い相関値 0.79 を示したパターンの TOEIC スコア推定値と実測値との対応を図1に示す。

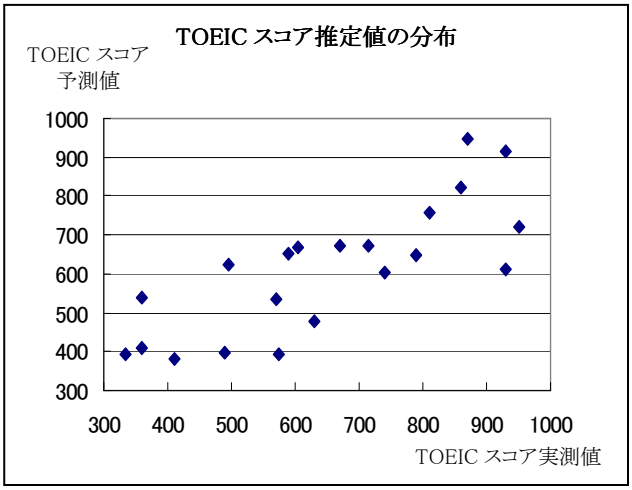

図1 TOEIC スコアの推定値と予測値の対応

4．英語習熟度推定実験結果の考察

表4より、統計的語彙情報特性と英語習熟度推定について、日英翻訳タスクにおける学習者の高頻度使用語彙順を反映した特性のほうが、JACET8000 の出現頻度順位順を反映した特性よりも、TOEIC スコアの実測値に対して高い相関を示す傾向にある。このことから、日英翻訳の中で学習者が実際に使用した語彙に基づく統計的語彙情報が、TOEIC スコアの実測値と良い対応関係を示すことが明らかになった。

また、統計的語彙情報量と英語習熟度推定について、高頻度単語の統計的語彙情報量を使用したほうが、全体的な統計的語彙情報量(5語彙毎と 10 語彙毎にまとめた情報量)を用いるよりも、TOEIC スコアの実測値と対応関係において、高い相関を示した。データ量が少なくとも、個々のデータが学習者の基本語彙分布を色濃く反映する、情報量の高いデータを用いることで、統計的語彙情報が TOEIC スコアの実測値と対応関係を示すことが明らかになった。

日英翻訳データから抽出される統計的語彙情報は、学習者の語彙の使用特性を反映する統計的情報量である。英語習熟度推定実験の結果より、この統計的語彙情報と英語習熟度との対応関係が示された。しかも、タスク依存性が高く、日英翻訳タスクにおける高頻出語彙から抽出した情報を使用する場合及び、信頼度の高いデータを使用する場合が英語習熟度とより強い対応関係を示すという特徴を有することがわかった。このような特徴量が、日英翻訳における英語習熟度の統計的評価に有用であることが判明した。

５．まとめ

日英翻訳文データから抽出した統計的語彙情報の英語習熟度評価への利用可能性を検討した。JACET8000 に基づいて抽出した基本語彙毎の出現総数分布を統計的語彙情報として、ニューラルネットワークを用い、学習者の英語習熟度(TOEIC スコア)を推定した。評価データによる推定において、学習者が実際に使用した基本語彙の統計的語彙情報特性に基づいた統計的語彙情報が、0.8 に近い相関を得た。推定値と実測値の間には対応関係が見られ、英語習熟度評価への適用可能性が示された。

今回検討した統計的語彙情報は、これまでの研究で扱ってきた統計的言語翻訳の翻訳確率とともに、人間の翻訳文評価のための統計的尺度になり得る。前者は使用語彙分布の特性を測ることで翻訳における使用語彙の妥当性を、後者は訳語選択の特性を測ることで翻訳の妥当性を反映する。正解翻訳との一致度を基準とする翻訳文の評価手法とは異なり、人間の言語運用能力の側面から翻訳文を評価する手法を今後、検討していきたい。

謝辞

本研究にて使用しました JACET8000 に関する情報を提供して頂きました早稲田大学 中野美知子教授、東京音楽大学 大和田和治准教授に感謝致します。

参考文献

[1] 山本誠一，菅谷史昭，安田圭志，隅田英一郎，“音声翻訳技術開発の経験に基づく外国語能力評価法の提案”，電子情報通信学会技術報告書, pp30-31, 2003

[2] 安田圭志，隅田英一郎，山本誠一，柳田益造，前川喜久雄，菅谷史昭“英語コミュニケーション能力の自動測定技術の提案”，情報処理学会研究報告 pp65-70, 2003

[3] 水本 篤“自由英作文における語彙の統計指標と評定者の総合的評価の関係”，統計数理研究所共同研究リポート 215 pp15-28, 2008

[4] 中條清美他“語彙力と実用コミュニケーション能力の関係”，Language Education & Technology 第 39 号 pp105-115, 2002

[5] 鍔木 元,安田圭志,山本博史,匂坂芳典 ″統計的翻訳評価尺度に基づく日英翻訳文の訳質分析″，言語処理学会 第 14 回年次大会論文集 pp1117-1119, 2008