英語学習者発話コーパスを用いた習熟度判定

Thepchai Supnithi [†‡※] 内元 清貴[‡] 齋賀 豊美[†‡]
和泉 絵美[‡*] Virach Sornlertlamvanich[※] 井佐原 均[†‡*]

通信・放送機構[†]
通信総合研究所[‡]
神戸大学大学院[*]
Information Technology R&D Division, NECTEC, Thailand[※]

1. はじめに

国際化・情報化社会が進み、コミュニケーションの手段として英語を運用する能力が強く求められるようになってきた。英語教育分野の研究者との意見交換により、その能力の向上を支援する手段として、自然言語処理技術に対する期待が高まっていることも明らかになってきた。このような状況に鑑み、我々は自然言語処理技術を用い、英語学習を支援する環境を構築することを考えている。その一環として英語学習者の「話す」能力に重点を置き、英会話能力の習熟度を自動判定するシステムの開発を目指している。習熟度が自動判定できるようになれば、習熟度ごとの誤りの傾向を調べることによって、習熟度に応じた学習方法を提案できるようになると考えている。本研究では英語学習者発話コーパスを作成し、それを用いて習熟度判定を行なう。

本稿では次の2点を明らかにする。(1)コーパスの情報を用いて習熟度を判定することはどこまで可能か。(2)習熟度判定に有効な情報は何か。

2. 英語学習者発話コーパス

これまでに作成されてきた英語学習者コーパスは、書き言葉を対象とするものがほとんどであり、話し言葉の英語学習者コーパスは極めて少ない。また、現存するコーパスは小規模のものしかなく、学習段階別に分けられていない。本研究で行なっている大規模な英語学習者発話コーパスの作成は世界的に見ても初めての試みと言える。

2.1 概要

英語学習者発話コーパスは㈱アルクの実行する英語のスピーキング能力を判定するインタビュー形式のテスト(Standard Speaking Test :SST)のデータから作成したものである[1]。テストの受験者は日本人学習者である。SSTでは、各学習者に対し、およそ15分間、1対1方式のインタビューが行なわれる。英語学習者コーパスは、インタビューテストをテープに録音し、テープからインタビューにおけるすべての発話(発音の間違いは考慮しない)を書き起こすことによって作成される。コーパスに収録されるインタビューはおよそ1200件の予定である。学習者のコミュニケーションに関する情報、たとえば、繰り返しや言い直し、発話の重複、あいづちやポースなどには基本タグとして XML 形式のタグが挿入される。さらに、文法的な間違いに対しても同様に XML 形式でエラータグが付与されつつある。その他、本コーパスには学習者の習熟度に関する評価結果が9段階の数値情報として付与されている。

2.2 習熟度評価方法

一般にスピーキングに対する評価は非常に難しい。本コーパスには SST と同じ基準が適用されている。SST の場合は、2人または3人の評価官がインタビューのテープを聴いて、評価を行なう。評価官は専門の訓練を受けた方に限られている。各インタビューは、あらかじめ用意した評価基準に基づいて、言語機能、内容、発話の形、正確さから総合的に判断し、9 レベルに分けられる。正確さの基準は、語彙、流暢さ、文法的正確さ、発音、社会言語学上の適切さ、の5つの項目からなる[2]。

次節では本コーパスからこれらの条件に当てはまると考えられる情報を素性として利用した習熟度判定方法および、その実験結果について述べる。

3. 習熟度判定実験

3.1 実験に用いる素性

受験者の習熟度を判定するために、評価官は様々な観点から評価を行なっている。我々は、SSTの評価基準を考慮し、2 節で述べた英語学習者発話コーパスから利用可能な情報を次の 5 つの観点により分類した。ここで述べる個々の情報を以後、素性と呼ぶ。

語彙: 習熟度に応じて、使える語彙の幅が広がると考えられる。使用する頻度も習熟度と関係がありそうである。また、受験者の理解可能語彙と生成可能語彙を区別する必要があるかもしれない。そこで、受験者の発話に出現した語に関して、次のような観点で、素性を設定した。

(1)発話に出現した語(W)
(2)発話に2回以上出現した語(Wd)
(3)発話に出現した内容語(CW)
(4)発話に2回以上出現した内容語(CWd)
(5)(1)と発話に出現した連続する2語(単語 Bigram)(BW)
(6)(5)のうち、発話に2回以上出現したもの(BWd)
(7)発話に出現した受験者の生成可能語(受験者が発話した語から試験官が発言した語を除いた語)(Wb-a)
(8)レベル別語彙リスト「標準語彙水準 12000 リスト」

の各レベル(12段階)を単語クラスとした場合の、各単語クラスの出現頻度[1](Alc)

文法: 文法に関しては、次の 2 種類の観点が考えられる。「どのくらい幅広く文法を使いこなせるか」と「どのくらい文法の誤りを犯しているか」ということである。本稿では前者の観点からどのくらい多くの品詞を使えるかに焦点を当て、また、品詞(Ty)は文法の一部であると考え、各品詞の出現頻度を素性として利用する。後者については、現段階ではコーパスにエラータグを付与している途中であることから、今回は考慮しなかった。

流暢さ:流暢さは習熟度判定を行うために重要な情報であると考えられる。受験者がスムーズに会話できている場合には流暢さがあると考えられる。そのスムーズさを表わす指標という観点から、コーパスから利用可能な情報のうち、次の要素を素性として利用する。

(1) 言いよどみや繰り返しを表わす基本タグの数(F)
(2) 1 単語あたりの発話平均時間(Tpw)
(3) 単語の総数(Wc)
(4) 文の平均長(Wps)
(5) 文の数(S)
(6) インタビューに要した時間(Ti)

コミュニケーション能力: コミュニケーションが成立しているかどうかを判定するためには 2 者間の会話を考慮する必要がある。そこで、SST の試験官の発話および試験官とのやりとりに関する下記の情報も素性として利用することにした。

(1) 試験官の発話における各単語とその出現頻度(Wa)
(2) 会話中に起こる試験官と受験者のターンの数(Tu)

試験官はインタビューの過程で、受験者のレベルに合った文や句、単語を用いると考えられる。また、会話中に起こる試験官と受験者のターンの数は 2 者間の会話がスムーズ進んでいるかどうかのひとつの指標になると考えられる。

3.2 機械学習モデル

習熟度判定はクラス分類問題の一種と考えられる。そのため本稿では初期段階の実験として、テキスト分類でよく利用される Support Vector Machine と Maximum Entropy Model を用いた。

(1) Support Vector Machine (SVM)

Support Vector Machine[5,6]とは空間を超平面で分割することにより二つのクラスからなるデータを分類する二値分類器のことである。二つのクラスを正例と負例とすると、学習データにおける正例と負例のマージンを最大にする超平面を求め、それを用いて分類を行なう。通常は、学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする拡張(ソフトマージン)や、超平面の線形の部分を非線形とする拡張(カーネル関数の導入)などがなされたものが用いられる。これらの拡張によりクラスを判別することは、以下の識別関数の出力値が正か負によってクラスを判別することと等価である。

$$f(\mathrm{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathrm{x}_i, \mathrm{x}) + b\right) \quad (1)$$

$$b = -\frac{\max_{i,\, y_i=-1} b_i + \min_{i,\, y_i=1} b_i}{2}$$

$$b_i = \sum_{j=1}^{l} \alpha_j y_j K(\mathrm{x}_j, \mathrm{x}_i)$$

ここで $\mathbf{x}$ は識別したい事例の文脈(素性の集合)を、x_i と y_i $(i = 1,\cdots,l, y_i \in \{1,-1\})$ は学習データの文脈とクラスを意味する。また、関数 sgn(x)は、$x \geq 0$ のときに 1, $x < 0$ のときに-1 となる 2 値関数であり、各 α_i は式(3)と式(4)の制約のもと式(2)の $L(\alpha)$ を最大にするものである。

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathrm{x}_i, \mathrm{x}_j) \quad (2)$$

$$0 \leq \alpha_i \leq C (i = 1,\ldots,l) \quad (3)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \quad (4)$$

(2) Maximum Entropy (ME) Model

このモデルでは、素性 $f_j (1 \leq j \leq k)$ の集合をとするとき、式(5)を制約とし、式(6)で表される目的関数つまりエントロピーを最大にするような確率分布を求め、その確率分布にしたがって求まる各クラスの確率のうち、最も大きい確率値を持つクラスを最適なクラスとする[3,4]。

$$\sum_{a \in A, b \in B} p(a,b) g_j(a,b) = \sum_{a \in A, b \in B} \tilde{p}(a,b) g_j(a,b) \quad (5)$$

$$for\ \forall f_j\ (1 \leq j \leq k)$$

$$H(p) = -\sum_{a \in A, b \in B} p(a,b) \log(p(a,b)) \quad (6)$$

ただし、A,B はそれぞれクラスと文脈の集合を意味し、$g_j(a,b)$は文脈 b に素性 f_j があってかつクラスが a の場合 1 となりそれ以外で 0 となるとなる 2 値関数である。また $\tilde{p}(a,b)$ は、既知データでの (a,b) の出現の割合を意味する。

3.3 習熟度判定の実験設定

SVM についての実行環境としては Chang ら[8]が開発した LIBSVM を利用した。

カーネル関数は $K(x,y) = (ax*y+b)^d$ であり、$a = 1/k$, $b = 0$, $d = 1, 2, 3$とした。ここで、kは入力データの長さを表す。

SVM は 2 値分類器であるため、本稿では LIBSVM で用意されている one-against-one という手法を利用する。これは 2 値分類器のSVMを多値分類器に拡張する手法である。

[1] 「標準語彙水準 12000 リスト」は日本人の英語学習者にとって有用であると思われる英語語彙 1 万 2000 語を基礎から上級まで 12 のレベルに区分した段階別学習語彙リストである。

表 1. 全素性を含む素性セットと最適な素性セットの精度比較(SVM)

素性セット	d = 1	d = 2	d = 3
[Wb-a Wa Ty Alc F S Tu Ti Tpw Wps Wc]	63.78%	57.00%	56.20%
[Wb-a Ty Alc F S Ti Tpw Wps]	65.57%	56.91%	57.09%

ME で扱う素性は基本的に、存在するか、しないかの 2 値のものであるため、素性の値を量子化する必要がある。実験では、語彙に関する素性を単語の出現頻度に応じて以下のように量子化した。

-(A)各単語を、その単語が出現したか、しなかったかの 2 種類に分割。

-(B)各単語を、出現頻度により、0、1 から 5、6 から 10、11 以上の 4 種類に分割。

-(C)各単語を、出現頻度により、0、1、2 から 5、6 から 10、11 以上の 5 種類に分割。

-(D)各単語を、出現頻度により、0 から 10 までのそれぞれと、11 以上の 12 種類に分割。

語彙に関する素性のみを用いた実験では、上記の(B)が最も良い結果になったため、ME の実験では(B)の量子化の方法を採用した。

3.4 最適な素性セットの選択

習熟度判定に最適な素性セットは、10 分割のクロスバリデーションによる判定精度が最大となるものをトップダウンに探索することにより決定する。まず、3.3 節で述べたすべての素性を用いて(これを root とする)、SVM と ME それぞれのモデルに対する精度を求める。次に root から、素性を一種類取り除いた素性セットを作成し、同様の実験を行なう。一種類取り除いた素性セットのうち、精度が root より高くかつ最大となる素性セットをより良い素性セットとして選択する。この削除と選択を繰り返すことにより、最終的に最高の精度が得られる素性セットが決まる。

4. 実験および評価

本節では、SVM と ME それぞれに対する最適な素性セットを決定し、全素性セットを用いた場合との精度を比較するとともに、機械学習モデル間に共通して有効であった素性と、共通して有効でなかった素性、個別に有効であった素性を明らかにする。また、語彙に関する素性と精度との関係についても述べる。

4.1 SVM を用いた実験の結果

結果を表 1 に示す。最適な素性セットは[Wb-a Ty Alc F S Ti Tpw Wps]であった。ここで、[ ]内の各記号は、3.1 節にあげた各素性の集合を意味する。つまり、受験者の生成語彙、品詞、レベル別単語クラス、言いよどみや繰り返しの数、文の数、発話時間、1 単語あたりの発話平均時間、文の平均長に関する素性を用いた場合が最適であった。表1にカーネル関数の多項式の指数 d がそれぞれ 1、2、3 の時の精度をあげる。すべての素性を含む素性セットと、探索の結果得られた最適な素性セットそれぞれを用いたときの精度をあげた。精度が最も良いのは d = 1 の時であった。

表 4. 全素性を含む素性セットと最適な素性セットの精度比較(ME)

素性セット	正解率
[BWd Wa Ty Alc F S Tu Ti Tpw Wps Wc]	60.39%
[BWd Wa Ty F S Tu Ti Wps Wc]	61.11%

表 2 は語彙に関する素性のみを変えた場合の各素性セットの精度を表わす。語彙に関する素性のみを利用する場合は、単語 Bigram を用いた場合が最も精度が良かった。しかし、他の素性と組み合わせた場合には、語彙に関する素性として、全単語を用いるより、受験者の生成語彙に関する素性のみを用いた方が良い結果が得られている。試験官は受験者が発話した単語が生成語彙か理解語彙かを判断の基準にしていると考えられる。同様の性質を素性として用いた場合に最適な素性セットとなった点は興味深い。

4.2 ME を用いた実験の結果

結果を表 4 に示す。最適な素性セットは[BWd Wa Ty F S Tu Ti Wps Wc]であった。つまり、受験者と試験官の発話に現れた単語、品詞、言いよどみや繰り返しの数、文の数、ターンの数、発話時間、文の平均長、総単語数に関する素性を用いた場合が最適であった。

表 3 は語彙に関する素性のみを変えた場合の各素性セットの精度を表わす。語彙に関する素性のみを利用した場合、他の素性と組み合わせた場合ともに、単語 Bigram を用いた場合に最も精度が良かった。

表 5 は Baseline、ME、そして SVM を用いた場合の精度を表わす。Baseline とは常にコーパスにおける最多のレベルを出力するモデルである。最多のレベルは 4 であった。ME と SVM は Baseline に比べて、20%以上良かった。

表 5. 機械学習モデルの精度比較

システム	正解率
Baseline	40.18%
ME	61.06%
SVM	65.57%

5 考察

本稿の目的のひとつは、コーパスから得られる情報を用いて英語学習者の発話の習熟度をどこまで判定できるか

表 2. 語彙に関する素性間の精度比較(SVM)

素性セット	W	Wd	Wb-a	CW	CWd	BW	BWd
語彙に関する素性のみ	58.93%	58.25%	51.70%	54.11%	53.97%	59.59%	**61.02%**
[Ty Alc F S Ti Tpw Wps] (最適)	64.41%	64.59%	**65.57%**	63.51%	63.43%	60.04%	60.04%
[Wa Ty Alc F S Tu Ti Tpw Wps Wc] (すべて)	63.16%	63.36%	**63.78%**	63.51%	63.60%	61.20%	61.37%

表 3. 語彙に関する素性間の精度比較(ME)

	W	Wd	Wb-a	CW	CWd	BW	BWd
語彙に関する素性のみ	56.11%	55.22%	50.41%	50.85%	47.28%	**57.00%**	55.40%
[BWd Wa Ty F S Tu Ti Wps Wc] (最適)	58.34%	58.70%	58.25%	58.16%	58.70%	58.07%	**61.06%**
[BWd Wa Ty Alc F S Tu Ti Tpw Wps Wc] (すべて)	60.39%	60.39%	59.68%	59.95%	58.61%	60.21%	**60.48%**

表 6. 各レベルのデータ数と最適素性セットを用いたときの精度

レベル	データ数（ファイル）	精度(0)	精度(±1)
1	3	0.00%	100.00%
2	29	65.52%	100.00%
3	198	68.69%	99.50%
4	453	83.23%	99.78%
5	226	44.69%	98.67%
6	127	47.24%	92.13%
7	52	26.92%	82.69%
8	24	12.50%	70.83%
9	9	0.00%	11.11%
All	1121	65.57%	96.52%

を明らかにすることであった。機械学習モデルである SVM と ME を用いて実験をした結果、SVM を用いた場合の精度は 65.57%であった。この結果は一般的なテキスト分類のタスクと比較すると、若干低い。また、Mayfield ら[7]は機械学習の手法を用いて Native と Non-Native の発話を分類する実験を行ない、高い精度が得られることを報告している。しかし、習熟度判定のタスクは、SSTの試験官の間でも初期段階の判定結果が割れることも多いことから、テキスト分類や Native と Non-Native の区別に比べて分類先のカテゴリの境界が曖昧であると考えられる。ちなみに、レベルの判定が一段階違っても正解と認めるという緩い基準では、表 6 で示したように精度が 96%を超えるという結果であった。また、SSTの試験官にシステムの結果を見てもらうと、かなり良いという評価であった。これらのことを考慮し、コーパスを用いた習熟度判定は実際の判定結果とかなり近いところまで可能であると言える。

本稿のもうひとつの目的は、習熟度判定に有効な情報が何かを明らかにすることであった。実験から、SVM と ME に共通する素性は、語彙、品詞、言いよどみや繰り返しの数、インタビューに要した時間、文の平均長に関するものであることが分かった。これらは、3 節で示したような語彙、文法、流暢さを表わす情報として習熟度判定に有効な素性であると言える。語彙に関する素性の中で比較すると、習熟度判定に有効な素性は学習モデルによって異なる。表 2 や表 3 のように語彙に関する素性のみを利用する場合はできるだけ多くの情報を利用した方が良いということが分かった。しかし、これは他の素性と組み合わせた場合の結果とは必ずしも一致しない。この結果は、本稿で行なったように最適な素性セットを探索する必要があることを示している。4.1 節でも述べたように、SVM では、語彙の素性として、全単語ではなく、その部分集合である学習者の生成語彙に関する素性を用いた場合が最適な素性セットであった。一方、ME では、単語 Bigram まで用いた場合が最適な素性セットであった。これは、ME で用いられる素性は量子化をする必要があるために多少情報が欠落し、その結果、最適素性セットとして必要となる語彙関係の素性としてSVMよりも多くの情報を持つものが選ばれたのではないかと考えられる。文法、流暢さを表わす素性はどの学習モデルにおいても習熟度判定に有効であったが、コミュニケーション能力に関する素性の有効性については学習モデルによって異なる。例えば、SVM では有効でなかったタ
ーンの数に関する素性が ME では有効であった。

最後に、今回の実験に用いなかった 81 件のデータに対し、最適素性セットとすべての素性を含む素性セットの精度を比較してみたところ、それぞれ、64.20%、62.59%であり、最適素性セットの方が精度が良いことが分かった。このことから、今回の実験で得られた最適素性セットは閉じたデータに対する最適素性セットではなかったことが分かる。

6 おわりに

本稿では「コーパスの情報を用いて習熟度を判定することはどこまで可能か」「習熟度判定に有効な情報は何か」の 2 点を明らかにした。まず、機械学習の手法を用いた習熟度判定の実験により、コーパスを用いた習熟度判定は実際の判定結果とかなり近いところまで可能であることが分かった。次に、10 分割のクロスバリデーションによって習熟度判定に最適な素性セットを探索することにより、習熟度判定には、語彙、文法（品詞）、流暢さ（言いよどみや繰り返しの数、インタビューに要した時間、文の平均長）に関する情報が有効であることが分かった。

今後の課題として、まず、現在進行中のエラータグの付与結果を利用することが考えられる。また、エラータグの自動抽出に関する研究も進めており、その結果の利用も考えている。次に、表 6 に示されるように、最低と最高のレベルについてはまだデータの数が少なく十分な精度が得られていないため、全レベルに対する精度を向上されるためには、コーパスのデータを増やす必要がある、ということがあげられる。さらに、現段階ではコーパスに記述されていない情報、たとえば、発音やイントネーション、アクセントなどの音声情報をコーパスに情報として記述し、利用できるようにしていくことも今後の課題のひとつである。

参考文献

[1] 通信・放送機構、(2002) 平成 13 年先端技術移転加速型研究開発プロジェクト適合型コミュニケーション技術の研究開発研究報告書

[2] 田辺洋二、(2002)スタンダード・スピーキング・テスト(SST)に関わる考察と評価，早稲田大学オーラルコミュニケーション研究所

[3] Jaynes, E. T.(1957). “Information Theory and Statistical Mechanics” Physical Review, 106, 620-630

[4] Jaynes E. T. (1979). “Where do we Stand on Maximum Entropy?.” In Levine, R.D and Tribus, M.(Eds.), The Maximum Entropy Formalism, p15. M.I.T Press.

[5] Cortes C.,and Vapnik, V.(1995). Support vector networks. Machine Learning 20:273-297

[6] Joachims, T. (1998). Text Categorization with support vector machines. In Proc. of European Conference on Machine Learning (ECML)

[7] Laura Mayfield, Tomokiyo and Rosie Jones, (2001) You’re Not From ‘Round Here, Are you? Naïve Bayes Detection of Non-native Utterance Text, 2nd Meeting of NAACL

[8] Chin-Chung Chang, Chih-Jen Lin. (2002) LIBSVM: a Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf