機械学習と属性撰択によるスピーキングレベルの識別

廣川 佐千男
九州大学情報基盤研究開発センター
hirokawa@.cc.kyushu-u.ac.jp

Brendan Flanagan
九州大学大学院システム情報科学府 & JSPS
b.flanagan.885@s.kyushu-u.ac.jp

金子 恵美子
会津大学コンピュータ理工学部
kaneko@u-aizu.ac.jp

和泉 絵美
同志社大学全学共通教養教育センター
eizumi@mail.doshisha.ac.jp

1 はじめに

多くの機械可読データが公開され、言語学習支援に対する機械学習の応用についての期待が高まっている。本稿では、NICT で開発し公開されている日本人英語学習者の発話コーパス（SST と呼ばれるスピーキングテストを書き起こしたものであるため、本稿では SST データと呼ぶことにする）[1, 2, 3] を対象とし、そこで利用されている単語によりどれだけスピーキングレベルの識別が可能か、機械学習の手法である SVM(support vectore machine) による分析を試みた。SST データは 1280 名分のインタビュー形式のスピーキングテストの発話を書き起こしたもので、3 つの異なったタスクと、試験官との対話のデータが含まれる。各被験者については、専門家によりレベル 1 からレベル 9 までの評価値が割り当てられている。本稿では、SST レベルそのものではなく、国際的に利用されている Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) (Council of Europe, 2001) [4] で規定されているレベルでの識別を行った。SST レベルと CEFR,CEFR-J の対応付けは、投野らの調査結果を利用した (表 1) [5]。なお、SST4 については、CEFR レベルとして A1 あるいは A2 へに割当るという二通りの可能性があり、本稿では前者を CEFR1、後者を CEFR2 という記号で表すことにする。本稿では、A2 に割当てた場合 CEFR2 の評価を行った。A1 に割当てた場合 CEFR2 の解析は今後の課題とする。CEFR1、CEFR2 について各レベルのデータ数は表 2 の通りである。なお、SST9 については B2 に割り当てて分析を行った。

SST データは、1280 人の被験者を対象として 5 つのステージについて行われたインタビューを書きおこしたものである。本稿では、一人の被験者についての結果をまとめたものを 1 文書とし、1280 件の文書群についてのレベル推定問題として分析を行った。ただし、CEFR で Pre A1 となる SST レベル 1 は除外した。単語数は 9626 個である。今回の分析の主目的ではないが、各単語について 11 種類の品詞 (SUBST, VERB, PRON, ADJ, PREP, ADV, CONJ, INTERJ, ART, STOP, UNC) ならびに、ランカスター大学が開発した品詞タグセット CLAWS5,CLAWS7 [1] もデータとして抽出した。

CEFR	CEFR-J	SST
-	PreA1	1
A1	A1.1	2/3
	A1.2	3
	A1.3	4
A2	A2.1	4
	A2.2	5
B1	B1.1	6/7
	B1.2	8
B2	B2.1	9
	B2.2	9
C1	C1	9
C2	C2	9

表 1: CEFR-SST 対応表

[1] CLAWS5: http://ucrel.lancs.ac.uk/claws5tags.html, CLAWS7: http://ucrel.lancs.ac.uk/claws7tags.html

	件数	
レベル	CEFR1	CEFR2
A1	738	717
A2	236	257
B1	263	263
B2	40	40

表 2: データ件数

2 SVM と属性選択によるレベル識別

各被験者についての書き起こし文書における単語の出現数を使って、各被験者をベクトル化した。これに基づき検索エンジン GETA で検索システムを構築した。[2]

2.1 評価実験手順

二つのレベルの機械的識別性能評価のために、レベル X の文書群を正例、レベル Y の文書群を負例として機械学習の SVM を適用した。具体的には、svm_perf [7] を利用した。実験手順は大きく分けると、全ての単語を使って識別モデルを作る Step 1、単語の重要度を求める Step 2、重要単語で属性撰択したベクトル化で識別を行う Step 3 の 3 つの段階に分られる。Step 3 では、重要度の個数 N を N=1,2,...10,20,..,100 と増やして実験を繰り返した。推定性能の評価については、5 分割交差検定を適用した。

Step 1	全ての属性を使ったベクトル化で SVM を適用しモデルを構築した。
Step 2	一つの属性 wi だけから成る仮想的文書についてモデルを適用した推定値をその属性の重要度 weight(wi) とした。
Step 3	Step 2 で得られる属性重要度の降順で上位の属性を撰択し、それらの属性だけでベクトル化して SVM を適用し、推定性能が最適な属性数 N を求めた。

図 1: 実験手順

なお、Step 2 で得られる各属性のスコアは SVM の分離超平面からの距離を表わしている。運用能力が上位の被験者の特徴に関連する属性の weight は正、下位の被験者の特徴属性の weight は負となる。

2.2 属性選択尺度

全ての属性を使った場合の識別性能は、例えば、A1 と A2 については、Precision, Recall, F-measure, Accuracy がそれぞれ、0.8923, 0.8117, 0.8491, 0.7830 であった。これは十分高い推定性能だが、どの項目が識別に有効か分らない。そこで本論文では、文献 [6] の属性撰択による機械学習を適用した。

属性 w について Step 2 で得られた単語重要度 weight(w) を使って、表 3 の 6 通りの評価尺度について Step 3 を実行した。df(w) は属性 w の出現頻度、つまり、属性 w を満す被験者数、abs は絶対値を表す。

記号	尺度
w.o	weight(wi)
d.o	weight(wi)*df(wi)
l.o	weight(wi)*log(df(wi))
w.a	abs(weight(wi))
d.o	abs(weight(wi)*df(wi))
l.a	abs(weight(wi)*log(df(wi)))

表 3: 属性撰択尺度

絶対値を使わない尺度の場合、属性撰択としては、weight が正の属性上位を N 個、weight が負の属性上位を N 個、合計 2*N 個の属性で対象被験者をベクトル化した。絶対値を使う場合、絶対値の降順で上位 2*N 個の属性を使った。

3 レベル識別性能

この章に、下の図を入れて説明する予定です。

図 2 は、上位 2N 個の単語を使った場合の識別性能 (Accuracy) をプロットした図である。A2*B1,B1*B2 では、曲線はなだらかに増加しており使う属性数が増えるに従って識別性能が徐々に高くなっていることが分る。これは、逆にいえば、少数の単語では識別ができないことを意味する。一方、A1*B2、A2*B2、A2*B2 では N=10 のあたりで、急激に識別性能が高くなっている。これは、少数の単語で十分高い識別が可能ということを意味している。

[2] http://geta.ex.nii.ac.jp

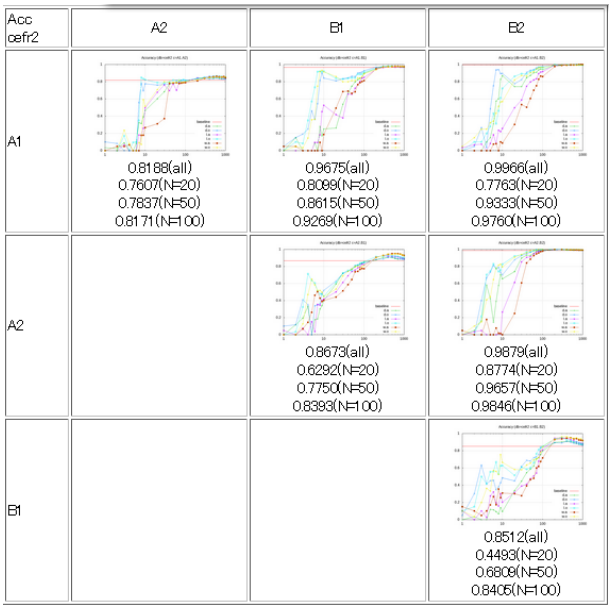

図 2: 属性選択による識別性能

3.1 ベースラインでの識別性能

全ての単語を使った場合の識別性能を表 4 に示す。隣接するレベルについては性能が低くなっていることが分る。

Acc	A2	B1	B2
A1	0.8188	0.9675	0.9966
A2		0.8673	0.9879
B1			0.8512

表 4: 全単語での識別性能

3.2 属性選択による識別性能

図 1 は、A1 を正例、A2 を負例として、上位 2N 個の単語で属性撰択した場合の、識別性能 (Accuracy) の変化を示したものである。N=200 以上では、全ての単語を使うより僅かではあるが性能が良くなっている。二つの尺度 l.o、d.o については N=9 でも全ての単語を使った場合と変らない識別性能となっており、少数の単語で、識別できるという結果を示している。

N=20	A2	B1	B2
A1	0.7607	0.8099	0.7763
A2		0.6292	0.8774
B1			0.4493

表 5: N=20 での識別性能

N=50	A2	B1	B2
A1	0.7837	0.8615	0.9333
A2		0.7750	0.9657
B1			0.6809

表 6: N=50 での識別性能

4 レベル A1 の特徴語

表 8 は、レベル A1 と他のレベルを比較したときに得られた上位 10 位までの特徴語集合を示す。「jp」は日本語の単語、「anonym.」は固有名詞を表す。どのレベルとの比較についても、A1 には名詞が多くみうけられ、cat, theater, boy, zoo, lion, monkey などは、SST のタスクの絵を見て話すタスクで使われる絵に現れるものの影響と思われる。一方、他のレベルには動詞、副詞、形容詞が多い。しかし、一般的に品詞を表す記号、例えば VERB や ADJ などは、特徴語としては現れなかった。つまり、単純な品詞の種別ではレベルの識別ができないことが確認できた。実際、分析データとしては品詞情報も含めていたが、結果として、上位に現れているのは、A2 と B1 の比較において、A1 の特徴になった c7=RGQ (程度を表す副詞) と、B1 の特徴になった C7=RRR(副詞比較級)、C7=DA(代名詞として使う形容詞) の 3 つしかなかった。また、表 8 から、比較対象によって A1 の特徴語集合は異っていることが分る。たった 20 個程度で識別ができるのは予想外だった。表 8 に含まれる単語の解釈は今後の課題である。

N=100	A2	B1	B2
A1	0.8171	0.9269	0.9760
A2		0.8393	0.9846
B1			0.8405

表 7: N=100 での識別性能

A1 レベルの特徴	比較対象レベルの特徴
look, please, jp, first work, just, picture what, friend, cat	(A2) home, find, when now, will, ask, other eat, think, if
ten, c7=RGQ, story, speak, theater, boy our, bring, anonym., favorite	(B1) really, also, ask call, actually, different c7=RRR, your, stay c7=DA
theater, pardon, cold color, zoo, lion, monkey shinjuku, recently, tv	(B2) an, into, drive brother, anything, club fun, once, teacher explain

表 8: A1 レベルの特徴と他レベルの特徴

5 まとめと今後の課題

日本人英語発話コーパスを対象に、機械学習の手法である SVM を適用し、CEFR レベルの識別を試みた。属性撰択により 20 個前後の単語で、全ての単語を使う場合と同様に 90%以上の識別に成功した。隣接レベルの識別は 10 た。A1 と B1 の識別ならびに B1 と B2 の比較については、少数の単語では高い識別が得られないことが分った。A1 レベルの特徴語としては、日本語や固有名詞の他、単純な名詞が多かった。

本稿では SST と CEFR の対応については、SST4 を割り当てを A2 として行ったが、A1 に割り当てた場合との違いは今後の課題である。また、CEFR よりもより細かくレベル分けされている CEFR-J でのレベル判別も今後の課題である。

謝辞

本研究は科研究 24242017 および 15H02778 による。

参考文献

[1] 和泉絵美, 内元清貴, 井佐原均, 日本人 1200 人の英語スピーキングコーパス, 東京: アルク, 2004

[2] Izumi, E., Uchimoto, K., Isahara, H., The NICT JLE Corpus: Exploiting the language learner's speech database for research and education. International Journal of the Computer, the Internet and Management, 12(2), pp.119-125, 2004

[3] Izumi, E., Uchimoto, K., Isahara, H., The Overview of the SST Speech Corpus of Japanese Learner English and Evaluation through the Experiment on Automatic Detection of Learners' Errors 4th International Conference on Language Resources and Evaluation, pp. 1435-1438, 2004

[4] Council of Europe, Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press, 2001

[5] 投野由紀夫（編）, 英語到達指標 CEFR-J ガイドブック, 大修館書店, 2013

[6] Sakai, T., Hirokawa, S., Feature Words that Classify Problem Sentence in Scientific Article, Proc. iiWAS2012, pp.360-367, 2012

[7] Joachims, T., Training Linear SVMs in Linear Time, Proc. ACM Conference on Knowledge Discovery and Data Mining (KDD), pp.217-226, 2006