日本語スピーキングテストにおける文生成問題の自動採点の検討*

☆大久保梨思子，山畑勇人，山田武志，△今井新悟，△石塚賢吉 (筑波大)
篠崎隆宏 (千葉大)，西村竜一 (和歌山大)，牧野昭二，北脇信彦 (筑波大)

1　はじめに

日本語学習者の日本語能力をインターネット上で測定するアダプティブテストとして，J-CAT(Japanese Computerized Adaptive Test)[1] が国内外で広く用いられている．現在のところ，J-CAT には聴解，語彙，文法，読解の能力を測定するセクションはあるものの，発話能力の評価は行っていない．そこで我々は，J-CAT における自動採点形式のスピーキングテストとして SCAT(Speaking section of J-CAT) の開発を進めている．SCAT には，文読み上げ，選択肢読み上げ，空所補充，文生成，自由発話の 5 つのタスクが設定されている．この順に解答の自由度が高くなり，自動採点も難しくなる．

本稿では，設問に対して受験者がワンセンテンスで解答を行う文生成問題を対象とした自動採点手法を検討する．文生成問題は解答に多少の自由度があるため，発話音声だけでなく発話内容も採点の対象となる．従来の自動採点手法の多く (例えば [2]) は発音のみに注目しており，発話内容も評価する自動採点手法はほとんど提案されていない．そこで，我々は新たに発話音声と発話内容の両方を評価する自動採点手法を提案する．

文生成問題における日本語教師による採点は，0〜4 点の 5 段階絶対尺度を用いた総合的な印象評価のみによって行われている．以下，これを総合点と呼ぶ．この評価の際には発音のような個別要因を意識していると考えられる．そこで，まず総合点に影響を及ぼすと考えられる要因を主観評価実験によって明確にする．次にこの結果に基づいて，各要因の主観値から総合点を推定するモデルを構築する．最後に各要因を推定するための特徴量を提案し，本手法の有効性を検証する．

2　提案手法

提案手法の処理フローを Fig. 1 に示す．提案手法では，まず解答音声から特徴量を抽出し，それを用いて各要因の主観値を推定する．次に，各要因の推定値を用いて総合点を推定する．なお，前者の推定対象となる要因は事前に決定する．後者の推定に用いる総合点推定モデルは，人間の主観評価値に基づいて決定する．

提案手法の特徴は，各要因の推定値から総合点を推定することにある．これは特徴量から直接総合点を推定するよりは容易であると考えられる．また，受験者に対し総合的な評価結果だけでなく，発話能力を改善するためのアドバイスを示すような応用も可能である．

3　文生成問題の採点に影響を及ぼす要因

3.1　要因の設定

文生成問題の採点に影響を及ぼす要因として，文献 [3]，[4] を参考に発話音声に関する要因群と発話内容に関する要因群を設定した．発話音声に関する要因群は，発音 (X_1)，イントネーション (X_2)，アクセント (X_3)，流暢さ (X_4)，ラウドネス (X_5)，発話内容に関する要因群は，聴解力 (X_6)，表現力 (X_7)，文法力 (X_8)，語彙力 (X_9) である．次節では以上の要因について調査する．

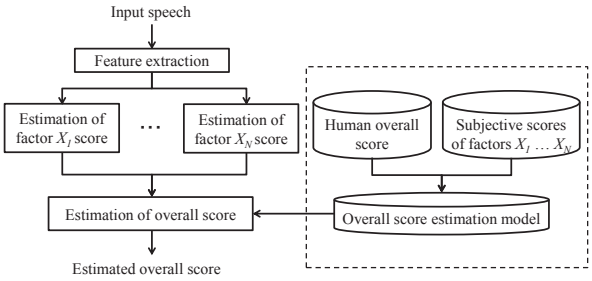

Fig. 1　Overview of the proposed method

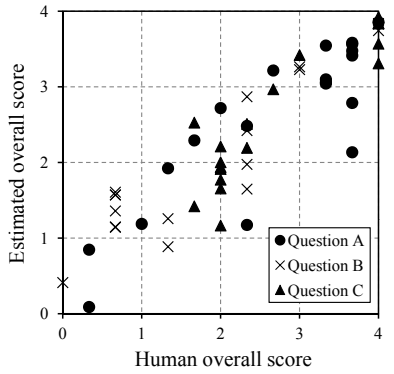

Fig. 2　Relationship between the subjective overall score and the overall score estimated from the subjective scores of each factor

3.2　主観評価実験

被験者は大学院生 5 名であり，防音室内でヘッドホンにより音声サンプルを受聴し，上述した全ての要因についてそれぞれ評価した．音声サンプルは 3 つの設問に対して留学生 20 名が発話した計 60 個である．各要因の評価尺度は，一般の日本人が日常会話で使用する標準的な日本語を基準とする，0〜4 点 (非常に悪い〜非常に良い) の 5 段階である．各要因の主観値は 5 名の評価点の平均である．また総合点は日本語教師 3 名の平均である．

3.3　実験結果と考察

教師による総合点を目的変数，各要因の主観値を説明変数とする重回帰分析，及びステップワイズ法により，発音 (X_1)，聴解力 (X_6)，表現力 (X_7) が選択され，式 (1) が導出された．

$$\text{総合点} = 0.43X_1 + 0.59X_6 + 0.40X_7 - 1.52 \quad (1)$$

これは，Fig. 1 における総合点推定モデルにあたる．選択された説明変数は，それぞれ音声生成能力，リスニング能力，文章構成能力に相当すると言える．なお，設問の区別はせず，すべての設問を用いて 1 つのモデルを構築した．これは設問に依存せず，どのような設問に対しても使用できる共通のモデルを構築するためである．

教師による総合点と，式 (1) に各要因の主観値を代入して推定した総合点の関係を Fig. 2 に示す．相関係数は 0.92，RMSE は 0.50 であった．この推定精度は，各要因を正確に推定できた場合の上限値にあたる．

*A study of automatic scoring method for sentence generation task in SCAT Japanese speaking test. by N. Okubo, Y. Yamahata, T. Yamada, S. Imai, K. Ishizuka (Univ. of Tsukuba), T. Shinozaki (Chiba Univ.), R. Nisimura (Wakayama Univ.), S. Makino, N. Kitawaki (Univ. of Tsukuba)

4　提案手法の有効性の評価

4.1　各要因を推定するための特徴量

発音 (X_1) を推定する特徴量として，文献 [5] と類似の特徴量を用いた．具体的には，ディクテーションの認識結果に対するアライメントと，連続音素認識の認識結果に対するアライメントを行い，両者の対応を比較することで算出される以下の特徴量である．

x_{1a}：発話区間において，両者の音素が一致しているフレームの割合

x_{1b}：発話区間において，両者の音素が一致しておらず，かつ両者の音響尤度の差が閾値 ($\theta = 2.25$) 以上のフレームの割合

x_{1a} は一定レベルの発音をしている部分，x_{1b} は発音が特に悪い部分に着目している．文献 [5] は文読み上げ問題を扱っているので，読み上げ対象の文へのアライメントを行っている．しかし文生成問題では発話内容が未知なので，ディクテーションの認識結果へのアライメントを行うことにした．ただしディクテーションの認識結果には誤認識が含まれ得るので，認識結果の信頼度から算出される x_{1c} を導入する．

x_{1c}：単語の時間長で重み付けた，ディクテーションによる認識結果の単語信頼度の平均

単語信頼度とは，競合する仮説候補との尤度差を 0 から 1 の値で表したものであり，1 に近いほど信頼度が高い [7].

次に，聴解力 (X_6) と表現力 (X_7) を推定するための特徴量について述べる．文生成問題においては自動採点を行うことを見越して，解答に含まれるべきキーワードが設定されている．例えば，日付を問うような設問だった場合，その日付がキーワードとなる．これを重要キーワードとする．重要キーワードが発話されているか否かは，採点結果に多大な影響を及ぼす．また，3.2 節の音声サンプルを分析した結果，文末語を正しく発話できている場合に，表現力が高く評価される傾向があった．そこで，想定される複数の文末語を文末キーワードとする．これらのキーワードをディクテーションとキーワードスポッティングにより抽出する．2 種類の抽出を行うのは抽出漏れを防ぐためである．以上のキーワード抽出により算出される以下の特徴量を提案する．

x_{6a}, x_{7a}：解答に含まれるべきキーワード数に対する，抽出されたキーワード数の割合

ここで，重要キーワードは文末キーワードの 2 倍の重みでカウントしている．x_{6a}, x_{7a} は質問の意図を理解して，適切な解答をしているかに着目している．

またキーワードは単語レベルの特徴量であり，文章全体の適切性を必ずしも反映していない．そこで，ディクテーションの認識結果と複数の模範解答との文字単位での編集距離を特徴量として併用する．

x_{6b}, x_{7b}：ディクテーションの認識結果と各模範解答から算出した編集距離のうちの最小値 (文章の長さで正規化)

この値が小さいほど模範解答に近いとみなせる．

4.2　提案手法の有効性の評価

まず各要因の推定モデルを決定する．音声認識器は Julius[6]，音響モデルは日本語話し言葉コーパス (CSJ) で学習したトライフォンモデルを留学生の解答音声で適応したものである．またディクテーションにおける言語モデルは，留学生の解答音声の書き起こし，ウェブテキスト，新聞記事の融合モデルである．キーワードスポッティングでは，連続音素認識をガーベージとした．以上によりに算出した特徴量の値と，各要因の主観値の重回帰により，式 (2)(3)(4) を得た．

Table 1　Estimation accuracy of the scores of each factor

Factor	Correlation coefficient	RMSE
X_1	0.73	0.43
X_6	0.74	0.79
X_7	0.76	0.57

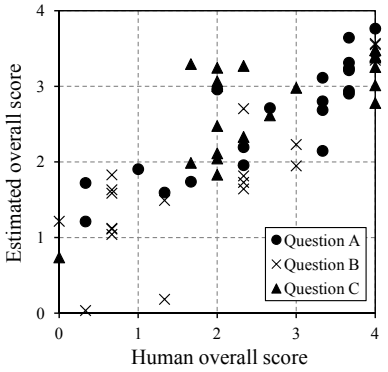

Fig. 3　Relationship between the subjective overall score and the overall score estimated from the estimated scores of each factor

$$X_1 = 1.16x_{1a} - 3.11x_{1b} + 1.67x_{1c} + 1.37 \quad (2)$$

$$X_6 = 2.08x_{6a} - 0.72x_{6b} + 1.77 \quad (3)$$

$$X_7 = 0.77x_{7a} - 1.53x_{7b} + 2.62 \quad (4)$$

聴解力 (X_6)，表現力 (X_7) の推定には同じ特徴量を用いているが，重みが異なっているのが分かる．

3.2 節の音声サンプルから算出した特徴量を，式 (2), (3), (4) にそれぞれ代入し，各要因の主観値を推定した．推定精度を Table 1 に示す．

次に，各要因の推定値を式 (1) に代入して総合点を推定した．教師による総合点と，推定した総合点の関係を Fig. 3 に示す．相関係数は 0.82，RMSE は 0.72 であった．Fig. 3 から，設問 C は特にばらついている．設問 C は比較的長い文章での解答が求められる．よって語順等が影響して編集距離に基づく特徴量が上手く機能しなかった可能性がある．

5　おわりに

本稿では SCAT における文生成問題を対象として，総合点に影響を及ぼす要因を考慮した自動採点手法を提案し，その有効性を評価した．今後は各モデルの性能向上，及び従来手法との比較を行う．

謝辞　本研究をご支援いただいた J-CAT メンバーに深く感謝する．本研究は科研費 (22242041) の助成を受けた．

参考文献

[1] J-CAT, http://www.j-cat.org/.
[2] M. Suzuki, *et al.*, “Integration of multilayer regression with structure-based pronunciation assessment,” Proc. INTERSPEECH2012, pp. 586–589, 2010.
[3] 藤代 他, “ブレンド型授業による英語の音読力と自由発話力に及ぼす効果,” 日本教育工学会論文誌 32(4), pp. 395–404, 2009.
[4] “Versant English test,” http://www.versanttest.co.uk/pdf/ValidationReport.pdf
[5] 山畑 他, “日本語スピーキングテストにおける文章読み上げ問題の自動採点の検討,” 音講論, Sep. 2012.
[6] A. Lee, *et al.*, “A efficient two pass search algorithm using word trellis index,” Proc. ICSLP1998, pp. 1831–1834, 1998.
[7] A. Lee, *et al.*, “Real-time word confidence scoring using local posterior probabilities on tree trellis search,” Proc. ICASSP2004, vol. 1, pp. 793–796, 2004