4G-01

日本人英語学習者の短い発話を自動採点するシステムの実現可能性の検討

近藤　悠介†

早稲田大学グローバルエデュケーションセンター†

1. はじめに

受検者が与えられた短い発話を求められる談話完成タスク(Discourse Completion Task: DCT)を利用して日本人英語学習者の発話を自動的に採点するシステムを構築することが本研究の目的である。自動採点の方法は、音声認識機で得られた音声データの特徴量と音声データに付与された評価値との関係を検証し、何らかの予測方法を選択し、新たな受検者の音声データの特徴量から評価を予測するという方法を用いる。本稿では、談話完成タスクを利用して収集したデータを用いて、発話自動採点システムの実現可能性を報告する。

2. 背景

外国語としての英語を評価・測定する際にコンピュータが用いられるようになり、言語テストに大きな影響を与えた。多肢選択式の問題を項目反応理論や潜在ランク理論に基づいて分析するコンピュータ適応型テストはその顕著な例である。しかしながら、客観的に評価することが難しいと考えられる発話能力の評価・測定においては、コンピュータを利用したテストが普及しているとは言い難く、開発過程とも言える。発話自動採点システムはいくつかの問題を抱えており、現在実用化されている、Pearson VUE の Versant™ と Educational Testing Service の SpeechRater™ は、これらに対し異なる方法を選んでいる。

発話自動採点システムが抱えるひとつめの問題は、受検者の発話能力を評価するために使用されるタスク（写真を描写する、文章を読み上げるなど）および評価の観点が音声認識の技術に制限されることである(Xi, 2010. p. 294)。一般に音声認識技術において発話の自由度が上がるほどその認識精度は下がる。発話される単語やその順番が限定されていれば高い認識精度が得られるが、自然な発話などでは、高い精度を期待することはできない。外国語の音声認識では、さらにその精度は下がる。Versant™ は、視覚提示された文を読み上げたり、提示された語の反対の意味の語を言うタスクを採用し、発話の自由度を制限している。Versant™ はテストの構成概念を Facility in L2 とし、Facility を「日常的な事柄について話されている言葉を理解する能力および分かりやすい言葉を使って母語話者同士の会話と同じペースで適切に話すことができる能力」と定義している(Bernstein, Van Moere, & Cheng, 2010, p. 358)。Versant™ は測定しようとする能力を間接的に測定していると解釈できる。一方で、SpeechRater™ は、自然に発話を測定の対象としているが、単語認識率は受検者によって大きく異なり、認識率の平均は約 50%と報告されている。正確に認識される単語が約 50%で、そこから算出される特徴量を用いて評価を予測しているということになる。予測の精度とともに特徴量の信頼性という観点からも議論の余地がある。

発話能力の評価の目的は、会話や発表をする能力を評価するといことである。一方で、発話能力の評価に使用されるタスクは、学習者に「どのようなことができるようになれば良いのか」を示すものであるため、タスクの選定には十分な注意が必要である。

3. 本研究の目的

このような背景を鑑み、本研究では実際に運営されている英語教育プログラムで使用されている教科書をもとにタスクを作成した。このタスクにおける評価を発話の特徴量から予測するシステムを構築することが本研究の最終的な目的である。

4. 方法

4.1 タスク

本研究で作成したタスクは、早稲田大学グローバルエデュケーションセンターが提供する英語科目 Tutorial English で使用されている教科書に基づく。Tutorial English の各課ではテーマが設定されている。たとえば、Asking about personal

Feasibility demonstration of automated scoring system for short speech of Japanese learners of English

†Yusuke Kondo, Global Education Center, Waseda University

plan というテーマの課では以下のような表現が学習の対象となり、実際にこのような表現を適切に使用する練習を行っている。

- Where are you up to this weekend?
- Do you have any plan for the week end?
- Are you busy this week end?
- I have to work all weekend.
- I'm visiting my grandparents in Osaka.
- I have a biology test on Friday.

4.2 データの概要

本研究で分析の対象とした DCT は以下である。

You want to end your conversation. What would you (A) say in the conversation below?
A: (　　).
B: See you.

この DCT を利用して、24 人の日本人大学生から音声データを収集し、学習用データとした。この音声データの書き起こしを中等教育、高等教育において英語教育の経験がある 3 人の評定者が 3 段で表現の適切さという観点から評価した。評価基準はヨーロッパ言語共通参照枠の Overall Spoken Interaction (Council of Europe, 2001, p. 74)を参考に C2 および C1、B2 および B1、A2 および A1 をそれぞれひとつのカテゴリとみなした。この評価の信頼性の指標として、評定者の一致度の指標であるフライスの κ を使用した。この評価におけるフライスの κ は.30 であった。3 人の評価が一致しない場合には 2 人の評定者が付与した評価を最終的な評価とした。

評価用データとして 36 人の日本人大学生から音声データを収集し、前述の評定者のうち 2 人の評定者が評価した。

Hidden Markov Model Toolkit を用いて隠れマルコフモデルに基づいた音声認識機を作成した。音響モデルの作成には 101 人のアジア人英語学習者の読み上げ音声および本研究の分析対象である DCT と同様のタスクを用いて収集した日本人英語学習者の 10606 発話を学習データとして用いた。言語モデルは先述の 24 人の日本人大学生の音声データの書き起こしから 2-gram モデルを作成した。学習用データの単語認識率は.66、単語認識精度は.63 であった。

5. 評価の予測

表現の適切さという観点から 3 段階で付与された評価を予測するために、tf-idf を指標として用いた。学習用データで出現した全単語の tf-idf を算出し、評価用データで出現した単語に tf-idf を付与した。この指標を用いて 3 段階のどのレベルに属するかを予測した。表 1 に例を示す。表 1 では、1 番左の列に受検者が発話した単語が示され、その横にそれぞれの単語の各レベルでの tf-idf が示されている。この tf-idf の平均値を各レベルで算出し、平均値が最も大きいレベルをその発話のレベルとした。

表 1: tf-idf に基づく予測方法の例

単語	レベル 1	レベル 2	レベル 3
see	0.2	0.5	0.3
you	0.3	0.2	0.3
soon	0.1	0.2	0.6
平均	0.2	0.3	0.4

音声認識機に得られた書き起こしを用いてこの方法で評価を予測したところ、評定者 2 人の評価との一致度（フライスの κ）は.27 であり、人間による書き起こしを用いた場合は.32 であった。

6. おわりに

本研究では音声認識機より得られた書き起こしを利用して、発話の表現の適切さという観点の評価を予測した。音声認識機を用いた書き起こしを利用した場合、一致度は.27 であり評価の一致度としては十分ではないが、学習用データにおける評定者の評価の一致度が.30 であり、大きく一致度が下がってはいない。このことから自動採点システムを評定者の 1 人として扱う可能性は十分にある。また、評価用データにおいて人間による書き起こしを使用した場合では、わずかに一致度があがるため、認識率の向上も必要である。

参考文献

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355–377.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: CUP.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291–300.