AUTOMATING THE SCORING OF ELICITED IMITATION TESTS

Deryle Lonsdale and Carl Christensen

Department of Linguistics & English Language
Brigham Young University
Provo, UT 84602

ABSTRACT

This paper explores the role of machine learning in automating the scoring for one kind of spoken language test: elicited imitation (EI). After sketching the background and rationale for EI testing, we give a brief overview of EI test results that we have collected. To date, the administration and scoring of these tests have been done sequentially and the scoring latency has not been critically important; our goal now is to automate the test. We show how this implies the need for an adaptive capability at run time, and motivate the need for machine learning in the creation of this kind of test. We discuss our sizable store of data from prior EI test administrations. Then we show various experiments that illustrate how this prior information is useful in predicting student performance. We present simulations designed to foreshadow how well the system will be able to adapt on-the-fly to student responses. Finally, we draw conclusions and mention possible future work.

1. INTRODUCTION

Acquiring a second language is a challenging task for the learner. Assessment of a learner's proficiency in a second language is also difficult, particularly where oral (i.e. spoken) language is concerned. Several oral proficiency testing techniques have been developed, each with its own merits and issues. In this paper we focus on only one particular technique, elicited imitation (EI). In this section we sketch our prior work in EI testing of English. Our goal is to automate the testing procedure we have implemented to date, and this requires machine learning techniques, as described below.

1.1. EI testing

Testing for learners' oral proficiency involves assessing a myriad of factors. These factors quantify mastery of a range of language skills: pronunciation, prosody, vocabulary, morphology, syntax, and semantics, as well as pragmatic skills such as social interaction, conversational ability, cultural understanding, and so forth. Over time a few standardized tests have been developed to perform these assessments, for example the Oral Proficiency Interview (OPI). The OPI is an oral interview lasting 20-30 minutes with free-form responses. Administering and grading the OPI requires skilled personnel, is fairly costly and time-consuming, and is largely holistic and somewhat subjective in nature. To obtain an accurate holistic score, oral proficiency testing must elicit sufficient data which results in longer tests and a sparsity of relevant features.

Consequently, alternative oral testing methods have been proposed and developed. One of these involves elicited imitation [1, 2]. A student being tested is given sentences of varying complexity one by one, and she must repeat them as exactly as possible. The response is recorded and later graded. One commonly accepted grading scheme allocates a perfect score of 4 to a correctly repeated sentence; this score is decremented by 1 for each error, down to a score of 0 (which means there were more than 4 errors). Errors are usually scored at the syllable level; if a syllable is missing or incorrect, it counts as 1 error. Part of the attraction of EI testing is that it can be administered by computer.

The theory behind EI testing is that a detectable threshold exists for the student's linguistic abilities; if sentences are too complex linguistically, the student will not be able to repeat back the sentence [3]. To be sure, the sentences must be carefully engineered for length to ensure that the student is not just parroting the stimulus. When test sentences are carefully engineered with features contributing to linguistic complexity, they can be combined into an EI test that correlates well with more traditional oral testing methods. Though there are some critics of the approach, there is also a growing consensus that EI does gives linguistic insight into oral language ability. We will not pursue the debate on the merits of this test or others in this paper.

Researchers have also investigated how various linguistic and demographic features influence the performance of students on EI test items. In particular, lexical choice, syntactic elements, and naturalness of the stimuli have all received attention. Many methods of oral proficiency rating also focus on the rate of speech [4] or some other type of fluency measure like a quality of pronunciation score [5].

Adaptive testing is desirable for tasks that can be entirely automated, because the questions can be calibrated to the test subject's abilities [6, 7, 8].

1.2. The data

We have developed over 300 EI items for English using language resources such as vocabulary lists, treebanks, and large-scale corpora [9]. From these items we have developed several EI tests that we have administered to over 1200 subjects, all adult learners of English in the U.S. at varying levels of proficiency. The result is a database archive of over 175,000 EI responses with their associated linguistic features[1]. We have also hand-scored these items at the syllable level (i.e. well over a million syllables) to assign each item a score from 0 to 4 as described above.

Some of these items were scored by more than one person, though we have shown that there is a high degree of agreement between scorers, obviating the need for massive redundancy in scoring. In addition to hand-scoring these items, we have also developed a system that has performed automatic speech recognition (ASR) scoring for these test items [10]. Even though the test subjects speak with foreign accents, the task is very tractable since the target sentence is known *a priori*, so the response needs only undergo a process called forced alignment to recognize whether all of the words were present. We have achieved, for English, correlation of greater than 0.9 for ASR-scored items with human-scored EI test items. We have also described elsewhere our ASR scoring pipeline and how well it correlates with the OPI [11]. Table 1 shows the distribution of the ASR-scored items. Note that over half of the responses were scored at '0'.

# items	Score
68946	0
11917	1
14283	2
16685	3
22921	4

Table 1. Distribution of ASR-scored items

1.3. Adaptive testing

We expect to be able to implement the EI test with automatic grading and in an adaptive testing framework.

Our current EI test administration can be run standalone or over the web, and makes use of a combination of open-source technologies—Java, Red5, and the Sphinx4 ASR engine. The system begins by registering the subject and asking pertinent questions about her language background [12]. The subject is then presented with a diagnostics page to ensure their audio equipment can interact with the online application correctly. After verifying their equipment works the application then proceeds to give them a sample test item to familiarize them with the testing procedures. The test is administered as follows; an item is played for the subject to hear, and their response is recorded. Sentences are chosen from a predetermined store of items, either randomly or sequentially (a parameter chosen by the test designer). This process continues until all questions have been asked. User responses are stored as sound files but are not scored until after the test is completed.

Of course, the adaptive system we envision will require some modification to the current processing scenario. Once responses are received the system will need to evaluate them immediately in order to be maximally reactive to the student's proficiency level. Ideally this will involve some component of machine learning. The adaptive test algorithm, then, is as follows:

- Administer at randon two of the easiest items.
- Score the two items using ASR and machine learning.
- If the subject does poorly (i.e. 4-score of 0 or 1 for this item), drop down a level if possible. If the subject does average (i.e. 4-score of 2 for this item), stay at this level. If the subject does well (i.e. 4-score of 3 or 4 for this item), increase difficulty a level.
- Iterate until some threshold is reached.

At least three thresholds are possible:

- Some predetermined number of test questions has been asked.
- Some stability is reached by the student at a given level.
- The test becomes exhausted because all items at a required level have been used.

Figure 1 sketches the adaptive algorithm.

2. EXPERIMENTS AND RESULTS

In this section we describe several experiments meant to explore and document progress towards automating the testing procedure.

2.1. Using collateral features

The first test we ran was to assess how well the system could guess the score of an individual EI test item based on the relevant features for that item: personal information on the student who provided the response, and linguistic features for the sentence itself. Predictions were made using TiMBL [13], a k-nn (i.e. nearest neighbor) system with the Overlap metric used for the calculations. The system was trained on all of the items and then tested under the leave_one_out paradigm.

Table 2 shows the results for the baseline condition, guessing the exact 4-score, and guessing within 1 point of

[1]Some students haven taken several different EI tests.

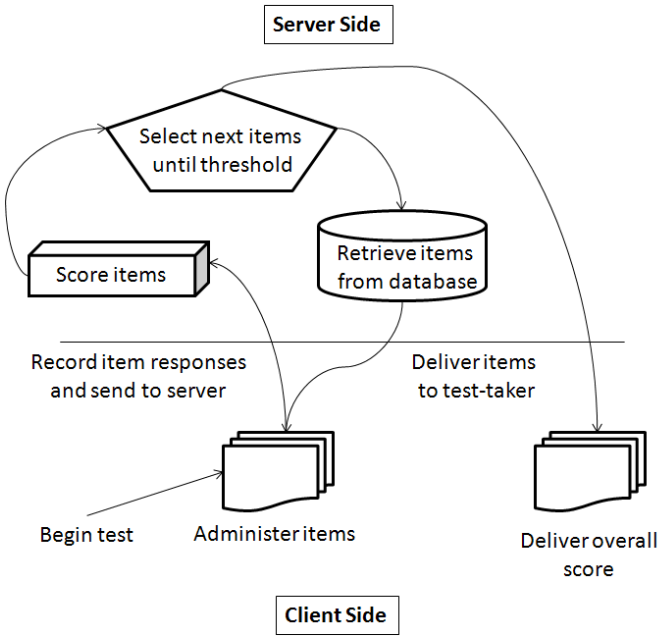

Fig. 1. Adaptive system for EI testing.

Test	Operation	Result
Baseline	Guess 0	0.51
Exact	Guess exact 4-score value	0.62
Within-1	Guess 4-score value $\pm$ 1	0.80

Table 2. Predicting item score based on features.

the target 4-score. Within-1 scoring is usually allowed when humans grade items; as long as their assigned rating is within one point of the other grader's, they are deemed in agreement. The contribution of the various features for this evaluation is noteworthy; see Table 3 for a summary of the features' significance.

2.2. Simulating the test

In the second machine learning experiment, we ran simulated test administrations. This involved choosing a scored item at random and, based on the item's features that are—of course—known *a priori*, guessing the exact 4-score for how well the student would perform on that item. Training data consulted for this guess include all scored items in the entire EI history database, excluding those for this particular student. Then that item is added to the training data, another of the student's test items is selected at random, and a guess is made to predict the score for that item based on its features. That item is also added to the training data, another item is drawn at random and predicted, and so on. At each stage the guess is thus predicated on the various student information features, the linguistic features for the sentence, and the full database archive of scored EI items for all the other students, plus the previously predicted items for that student.

Feature type	Subtype
Syntax	ditransitive, copula, SIM, COM
Test scores	final, speaking
Syllable count	manually determined
Syntax	DES, SQS
Syllable count	automatically determined
Test score	listening
Syntax	NFC, SUB
Syntax	overall metric
Test scores	grammar
Syntax	VIN
Test score	reading
Morphology	overall metric
Semantics	overall metric
Syntax	TRN
Student info	gender, L1, birth country
	birth year, nationality
Vocabulary	overall metric
Test scores	listening, grammar, reading
Syntax	PPT

Table 3. Contribution of features for guessing EI item score.

The simulations were run on a 512-node, 6144-core Intel Westmere supercomputer with 12,288 GB of memory. Not all 800-plus students for whom we have data were simulated; we selected 124 students at random for this computation. Again, the predictions were made using TiMBL.

Figure 2 shows the results for the EI testing administration simulations. The vertical axis indicates the accuracy of each guess, and the horizontal axis shows the number of items tested. Not surprisingly, the first guess fares poorly, but soon the values jitter around the mid-80% range. By about 20 items the system is able to predict with almost 90% accuracy on average the score for any item for any student. Interestingly, after about 60 items the curve starts falling slightly; this is due to data sparseness since only a few students have been tested on more than 60 items. Note also that this simulation computed accuracy on an exact match with the actual 4-score; the system would exhibit even higher accuracy scores if using within-1 scoring.

In scoring the responses this test used all available information—several hundred thousand previously scored items and dozens of features for each. This would not be practical in an adaptive testing environment, where response latency must be minimized and where local computing power is more modest.

2.3. Automatically sequencing test items

Another requirement for adaptive testing is to choose the sequence of items more intelligently. The next set of experiments is designed to explore how effectively this can be per-

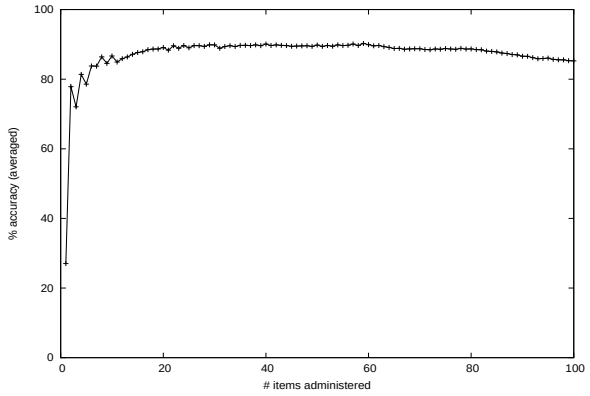

Fig. 2. Predictions for sentence items: accuracy scores averaged across 124 randomly chosen students tested.

formed. We performed an item response theory analysis [14] on all of the English EI items to determine their relative difficulty. Then the items were categorized into 5 groups: the bottom 20% are the easiest, the top 20% are the hardest, and so on. Based on the algorithm discussed above in subsection 1.3, several experiments were run with various scoring algorithms, with up to 60 items adaptively selected for presentation per test.

Algorithm	Human-scored	ASR
avg. scores	95.7%	77.0%
avg. IRT level	99.4%	79.6%
avg. (IRT+scores)/2	99.7%	89.1%

Table 4. Accuracy of simulated adaptive test using various algorithms (809 subjects, up to 60 questions, within-1 scores).

Table 4 shows results for various simulated adaptive test administrations using different algorithms. Line 1 simply averages the result scores across each subject's answers to the questions selected at run time. Line 2 averages the IRT level of items selected and presented to the subject. Line 3 is the average of the IRT level and scores across all presented items. All three of these procedures are grossly oversimplified—even crude, but are useful for illustrative purposes. Presumably it would be possible to compute a regression or some other finer-grained scoring regime to get even better scores.

When only *a posteriori* human scores were taken into consideration, scoring accuracy is very high for all three rudimentary algorithms. Note, though, that we are benefitting here from hindsight; in a truly adaptive system the test subject's results will not be available for comparison.

The last column of the table shows how the adaptive algorithm fares when only ASR scoring is done on the items. This is a more natural assumption since it does not depend on human-generated scores. When the adaptive test items' IRT scores and ASR scores are averaged together, performance achieves a respectable 89.1%. Again, this is only taking into account one feature, a scoring metric; presumably using all or some of the other features would help improve the score, but at the risk of increasing response time.

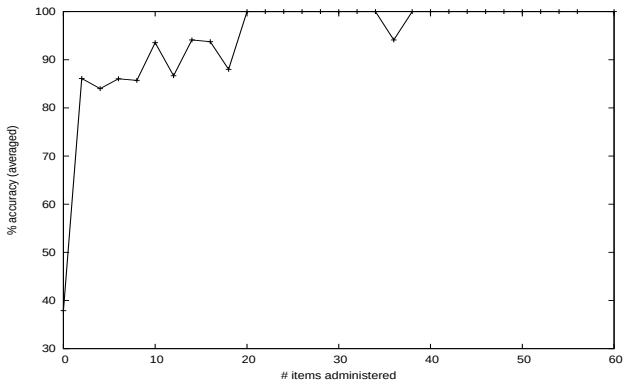

Fig. 3. Incremental accuracy of adaptive system with ASR grading.

2.4. Incremental ASR accuracy

The last results we present assesses how well the adaptive algorithm performs over time. Figure 3 plots the accuracy (i.e. percent of correct guesses overall) as the number of items presented grows. Scoring is entirely with ASR values. The accuracy of the guesses increases steeply and jitters until about 20 questions, at which point a high rate of accuracy is achieved. Again, adding more features should improve these results.

3. CONCLUSIONS

We have described in this paper groundwork for gauging the feasibility of an adaptive test for ASR-scored EI response items. Using machine learning we have assessed the relative importance of various features that characterize EI items. We were also able to perform large-scale simulations of test sessions based on data collected from prior administrations. This showed that—given essentially unlimited resources—we could effectively leverage the store of all previously scored items, along with the features of the test item, to predict how well a given student would perform on that item. We then implemented an algorithm to enact adaptive testing based on pairs of EI items chosen randomly from within a target range based on the student's performance on recent items. We demonstrated how the algorithm performs against human-scored and ASR-scored items, and illustrated incremental behavior of the prototype system.

Acknowledgments

We express appreciation to the members of the BYU PSST research group[2] for programming and data analysis support, to the BYU English Language Center where our tests were administered, and to the BYU supercomputer facility for generous use of resources for running simulations.

4. REFERENCES

[1] T. Vinther, "Elicited imitation: a brief overview," *International Journal of Applied Linguistics*, vol. 12, no. 1, pp. 54–73, 2002.

[2] L. Jessop, W. Suzuki, and Y. Tomita, "Elicited imitation in second language acquisition research," *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, vol. 64, pp. 215–238, 2007.

[3] R. Bley-Vroman and C. Chaudron, "Elicited imitation as a measure of second-language competence," in *Research methodology in second language acquisition*, E.E. Tarone, S. Gass, and A.D. Cohen, Eds., pp. 245–261. Lawrence Erlbaum, Hilldale, 1994.

[4] F. Wet and C. van der Walt, "Automatic assessment of oral language proficiency and listening comprehension," *Speech Communication*, vol. 51, no. 10, pp. 864–974, 2009.

[5] Y. Liu and C. Yang, "Automatic pronunciation scoring for Mandarin proficiency test based on speech recognition," in *2009 International Symposium on Intelligent Ubiquitous Computing and Education*, 2009, pp. 590–601.

[6] J.D. Brown, "Computers in language testing: present research and some future directions," *Language Learning & Technology*, vol. 1, no. 1, pp. 44–59, 1997.

[7] J.M. Linacre, "Development of computerised middle school achievement test (translation from Korean)," in *Computer-adaptive testing: a methodology whose time has come*, S. Chae, U. Kang, E. Jeon, and J.M. Linacre, Eds. Komesa Press, Seoul, 2000, MESA Memorandum No. 69.

[8] C. Roever, "Web-based language testing," *Language Learning & Technology*, vol. 5, no. 2, pp. 84–94, 2001.

[9] Carl Christensen, Ross Hendrickson, and Deryle Lonsdale, "Principled construction of elicited imitation tests," in *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC '10)*, N. Calzolari, K. Choukri, B. Maegaard, J Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Taipas, Eds., Valetta, Malta, 2010, pp. 233–238, European Language Resources Association (ELRA).

[10] C.R. Graham, D. Lonsdale, C. Kennington, A. Johnson, and J. McGhee, "Elicited imitation as an oral proficiency measure with ASR scoring," in *Proceedings of the 6th International Language Resources and Evaluation Conference (LREC '08)*, N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, Eds., Marrakech, Morocco, 2008, European Language Resources Association (ELRA).

[11] Kevin Cook, Jeremiah McGhee, and Deryle Lonsdale, "Elicited Imitation for Prediction of OPI Scores," in *Proceedings of the ACL/HLT 2011 Workshop on Building Educational Applications*, 2011, (in print).

[12] B. Freed, D. Dewey, and N. Segalowitz, "The language contact profile," *Studies in Second Language Acquisition*, vol. 26, pp. 349–356, 2004.

[13] W. Daelemans, J. van der Sloot, and A. van den Bosch, "TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide," Tech. Rep., 2010, ILK Research Group Technical Report Series no. 10-01.

[14] F.M Lord, *Applications of item response theory to practical testing problems*, Erlbaum, Mahwah, NJ, 1980.

[2] See http://psst.byu.edu.