Automatic Scoring in a Task of Retelling Stories for Language Learners

Meng Chen, Dean Luo, Lan Wang
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences/The Chinese University of HongKong
{chenmeng, da.luo, lan.wang}@siat.ac.cn

ABSTRACT

In this paper, we propose a novel method for automatic scoring of retelling stories. By taking consideration of possible paraphrases, new scoring features on contents are proposed with the use of ASR in addition to traditional evaluation measures. Linear models are introduced to combine different features for automatic scoring. We evaluated machine scores by correlating them with human scores which were manually rated by an expert. Experimental results show the correlation between machine scores and human scores can be improved.

Keywords: Automatic Scoring, Retelling, Computer-Assisted Language Learning

1. INTRODUCTION

Recently, systems for computer-assisted language learning (CALL) have shown great advantages over traditional methods. It would be a much cheaper alternative, which is accessible at any time and at any place, and certainly tireless. In these systems, the main task is how to provide the type of feedback that a human teacher would provide. From a pedagogical point of view, a score for the overall assessment of the language learners is important. In order to provide feedback of score without the presence of a human teacher, methods for automatic scoring are required.

Many researchers have studied automatic methods based on Automatic Speech Recognition (ASR) for evaluating the speaking ability of language learners. There are mainly two kinds of tasks in previous work. One is for the restricted speaking tasks such as reading aloud. For example, Neumeyer and Franco presented a system for automatic evaluation of the pronunciation quality in task of reading English texts aloud [1] [2]. Cucchiarini et al. developed a system for Dutch pronunciation scoring along similar lines [3] [4]. The other is for the unstructured, unrestricted, and spontaneous speech. For example, Educational Testing Service (ETS) investigates the automatic scoring of unrestricted, spontaneous speech of non-native speakers in the task of questions and answers [5].

In this study, we focus on the task of retelling stories, which has been proved an efficient way to improve oral proficiency of a language learner [6] [7]. In the test, students listen to a monologue of story (200~300 words) spoken by a native speaker, and then retell the story with their own words. The responses of students are spontaneous speech with lexical and syntactic errors. This means the vocabulary of ASR should include variations that are extended from the original story but semantically similar.

We approach automatic scoring by extracting scoring features from the output of ASR, and linear regression models to combine different scoring features. Considering the uniqueness of retelling, content related features are critical for the automatic scoring of the retelling task. Therefore, we compute the similarity to represent the content correctness of speech by comparing recognition hypothesis from ASR with all possible paraphrased expressions extended from original story. We also improve the feature of keyword coverage rate based on the vocabulary extended rules. Experiments showed a higher correlation between machine scores and human scores than the traditional methods.

In the remainder of this paper, the scoring features and linear regression model will be introduced in Section 2. The experiments and results will be presented in Section 3. Section 4 is the conclusions and discussions.

2. AUTOMATIC SCORING FEATURES AND METHODS

2.1 Scoring features

The traditional scoring features are mainly based on the intelligibility and fluency of non-native speech such as the global and local log-likelihood derived from the HMM log-likelihood [1], the rate of speech, and silence

length etc. However, content related scoring features are not taken enough consideration for automatic scoring. The possible reason may be it's not necessary for the task of reading text because the prompts are known to the learner. And for spontaneous speech in the task of questions and answers, the content of speech is difficult to predict. However, for the task of retelling, because the content of speech is related to the topic and the reference story, the extraction of content related scoring features is feasible and critical for automatic scoring.

For the task of retelling, the students are required to repeat the story as possible as what they hear, however, they would try to express with their own words when they can't remember clearly. That means the vocabularies and sentences in the speech of students are not the same as the original story. Therefore, the correct answers are not the only one. For the vocabulary, synonyms and near synonyms should be considered. And for the sentence, different sentence patterns but semantically similar should also be included.

According to this uniqueness, we made the extended rules to generate possible paraphrased expressions from the original story. Firstly, we extended the original story manually based on different sentence patterns. Then, we replaced the keywords with their synonyms and near synonyms automatically. As current stage, we didn't consider pruning during extending. After this process, the possible paraphrased expressions are generated, and used in the automatic scoring scheme.

Based on the extended rules above, we introduce new content-related scoring features, which are given as below,

1) Similarity, which shows the content correctness of speech,

$$Similarity = max(S_1, S_2, \dots S_i, \dots S_n), \qquad (1)$$

where S_i is the similarity between the speech of student with the i_{th} possible paraphrased expression.

2) Keyword Coverage Rate (KCR),

$$KCR = \frac{\sum_{i=0}^{n} cover(k_i)}{n}, \qquad (2)$$

where k_i is the i_{th} keyword and n is the number of all keywords, and

$$cover(k_i) = \begin{cases} 1, & if\ k_i\ occured\ in\ the\ speech \\ 0, & if\ k_i\ didn't\ occur\ in\ the\ speech \end{cases}. \qquad (3)$$

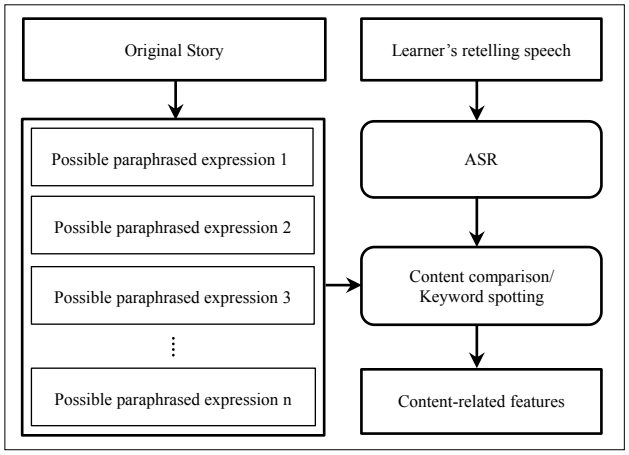

Figure 1: Block diagram of content-related features extraction.

Figure 1 shows the block diagram of content-related feature extraction. We proposed to compute the similarity between recognition hypothesis and all possible paraphrased expressions. Based on the vocabulary extended rules and the content of the presented story, we extended the possible paraphrased expressions first. Then a Dynamic Programing-based string matching was processed between the recognition hypothesis and each possible paraphrased expression (including the original story) for similarity computation. The maximum value was chosen to represent the similarity between the recognition hypothesis and paraphrased expressions, which showed the content correctness of the speech.

We also proposed a novel way of keyword spotting to calculate keyword coverage rate. We extended the keyword set based on the vocabulary extended rules to include more possible keywords which might occur in the speech of retelling task. We also used content-related features of word number and unique word number.

Our scoring feature set, extracted from the ASR results, can be categorized as follows: (1) Content-related, (2) Intelligibility, and (3) Fluency. Table 1 shows a complete list of the scoring features we computed, along with a brief explanation.

2.2 Linear regression model

In order to estimate learners' proficiency, linear regression models are used to combine different features. We established various independent variables $\{x_i\}$ as parameters and the value Y as the human's score, and

Table 1 List of features with definitions

Feature name	Description	Category
Similarity	Similarity between recognition hypothesis and possible paraphrased expression	Content-related
KCR	Keyword coverage rate	Content-related
WN	Number of words in recognition hypothesis	Content-related
UWN	Number of unique words in recognition hypothesis	Content-related
GL	Global log-likelihood	Intelligibility
LL	Local log-likelihood	Intelligibility
ROS	Rate of speech in phoneme level	Fluency
SN	Number of silences in recognition hypothesis	Fluency

the linear regression model was defined as:

$$Y = \sum_i \alpha_i \times x_i + \varepsilon, \quad (4)$$

where ε is the intercept. The coefficients $\{\alpha_i\}$ were estimated by the optimal least-squares of ε. Once the coefficients and intercept were determined, the machine score could be predicted with extracted features.

3. EXPERIMENT

3.1 Speech recognition system

Our speech recognition system is based on speaker independent continuous Hidden Markov Models. We used the TIMIT/WSJ database to train the acoustic models [8] [9]. The speech is downsampled to 16kHz and preemphasized, and then a Hamming window with a width of 25ms is applied every 10ms. The acoustic features include 12 MFCCs (Mel Frequency Cepstrum Coefficient), energy, delta and acceleration features. The HMMs are composed of three emitting states, each of which has sixteen mixed Gaussian distributions with full covariance matrices. The inter-word triphone HMMs are trained.

In order to enhance the robustness of our ASR system, we conducted the acoustic model adaptation and language model adaptation based on the specific task of retelling.

Based on the extended rules, we could predict the vocabulary in the speech of students. One class is synonymy of the words in the original story, which are correctly used but not occurred in the original story. The other class is wrong variants of the words in the original story such as the wrong tense of verbs, the wrong singular or plural forms, and the wrong adjectives or adverbs etc. Although these wrong variants are not used correctly, they are much likely to occur in the speech of non-native speaker. With the predicted vocabulary from the original story, we could adapt the general language model to our specific task of retelling.

3.2 Database

We used RETELL data set, which contains 280 responses from 280 speakers and each response is two minute long, for our evaluation experiment. For each response, there are 100-200 words in total. All the speakers are native Chinese high school students. The data was collected in classrooms when the students were using our application for retelling task. An overall proficiency score is given for each learner by an expert in English education. A discrete score scale (from 1 to 6) indicates the overall proficiency (from least proficient to most proficient). These scores are used as reference scores.

3.3 Results

We used leave-one-out cross validation method for experiments on our dataset. Each time we used the data of one learner as test data, and the others for training the scoring models. The Pearson coefficient of correlation between machine scores and human scores is used here as the measure of the agreement between raters (human or machines).

Table 2 shows the correlation of each feature with human scores. From this table, we can see the feature of Similarity, KCR (keyword coverage rate), and UWN (unique word number) have the highest correlation with human scores compared with the traditional features. Table 3 shows the correlations between machine scores and human scores with different combinations of features. In Table 3, we can see the performance was improved when the features of Similarity, KCR, UWN and WN (word number) were chosen. And we obtained higher correlation of 0.621 with human scores by combining all the features. Figure 2 illustrates the percentage of students for different absolute score difference between machine scores and human scores. According to Figure 2, there are 92.9% students whose absolute score difference of machine scores and human scores is less than 2. The average absolute score difference is 0.92. This confirms the effectiveness of our method.

Table 2 List of features and correlation with human scores

Feature name	Correlation
Similarity	0.462
KCR	0.496
WN	0.285
UWN	0.481
GL	0.231
LL	0.306
ROS	0.328
SN	-0.148

Table 3 Correlation between machine scores and human scores for different feature set

Feature set	Correlation
GL, LL, ROS, SN	0.484
GL, LL, ROS, SN, Similarity, KCR, WN, UWN	0.621

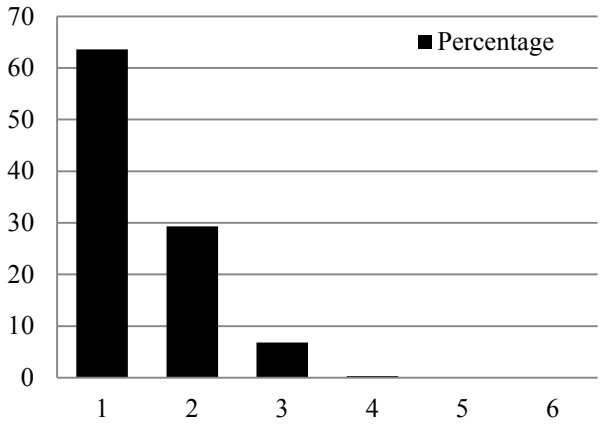

Figure 2 Percentage of students for different absolute score difference between machine scores and human scores

4. CONCLUSIONS

We presented a novel approach to automatic scoring of non-native English speech by taking consideration of the uniqueness of retelling. With improvements on content-related scoring features, we obtained a high correlation of 0.621, which is rather higher than the traditional methods.

An important step for future work will focus on improving speech recognition. We plan to adapt the acoustic model with more data of non-native speakers. And more spoken styles will also be considered in the adaptation of language model. Furthermore, scoring features that evaluate grammar correctness of speech in retelling task will be explored in order to obtain a broader coverage of communicative competence.

5. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (No.90920002), and The Knowledge Innovative Project of Chinese Academy of Sciences (No.KJCX2-YW-617).

6. REFERENCES

[1] L.Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech", in Proc. of ICSLP96: 1457-1460, 1996.

[2] L.Neumeyer, H. Franco, Vassilios Digalakis and M. Weintraub, "Automatic scoring of pronunciation quality". Speech Communication 30(2-3): 83-93. 2000.

[3] Catia Cucchiarini, Helmer Strik, and Boves L., "Automatic evaluation of Dutch pronunciation by using speech recognition technology". In Proc. of Automatic Speech Recognition and Understanding, IEEE Workshop. 1997.

[4] Catia Cucchiarini, Helmer Strik, and Boves L., "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms". Speech Communication 30(2-3): 109-119, 2000.

[5] Derrick Higgins, Xi Xiaoming, Klaus Zechner and David Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses". Computer Speech & Language 25(2): 282-306. 2011.

[6] Morrow L, "Retelling stories: A strategy for improving young children's comprehension, concept of story structure, and oral language complexity". Elementary School Journal, 647-661. 1985.

[7] Stein, Nancy L., Glenn, Christine G, "An analysis of story comprehension in elementary school children". In R. Freedle (ED), New Directions in Discourse Processes, Vol. 2, pp.53-120. Norwood, NJ:Ablex. 1979.

[8] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM". NTIS order number PB91-100354, 1993.

[9] D. Paul, J. Baker, "The Design for the Wall Street Journal-based CSR Corpus". DARPA Speech & Nat. Lang. Workshop, Arden House, NY, Feb. 1992.