TOWARDS ACCURATE RECOGNITION FOR CHILDREN’S ORAL READING FLUENCY

Jian Cheng, Jianqiang Shen

Knowledge Technologies, Pearson
299 S. California Ave, Palo Alto, California 94306, USA
jian.cheng@pearson.com, jianqiang.shen@pearson.com

ABSTRACT

Systems for assessing and tutoring reading skills place unique requirements on underlying ASR technologies. This paper presents VersaReader, a system automatically measuring children’s oral reading fluency skills. Critical techniques that improve the recognition accuracy and make the system practical are discussed in detail. We show that using a set of linguistic rules learned from a collection of transcriptions, the proposed rule-based language model [1] outperformed traditional n-gram language models. Combined with a specific acoustic model with explicit long silence modeling, plus adaptation, a WER 7.25% was achieved in our test set. The impact of different kinds of rules on performance is also discussed. We demonstrate that VersaReader can provide highly accurate Words Correct Per Minute scores automatically, which are virtually indistinguishable from scores provided by careful human analysis.

Index Terms— oral reading fluency, language testing, literacy, language model

1. INTRODUCTION

Fluency is the ability to read text quickly, accurately, and with proper expression [2]. Fluency provides a bridge between word recognition and comprehension. Several decades of research has shown that the students’ level of reading fluency provides a highly reliable and valid measure of general reading achievement, including comprehension [3, 4, 5]. The national Research Council recommended assessing oral reading fluency regularly and systematically in the classroom to monitor student progress and providing effective reading instruction when disfluent reading is detected [6]. Lots of school districts in the United States choose to monitor the students’ level of reading fluency as an important indicator of overall student progress in reading.

The standard way for the oral reading fluency test is letting a child read a normed and unpracticed grade-level passage aloud for one minute and then calculating the number of words read correct per minute (WCPM). The general oral reading fluency (ORF) tests, such as the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) [7] and AIMSWeb [8], are paper/pencil based, administrated and graded by human. An automatic ORF grading system built with speech recognition techniques can conduct this assessment more efficiently and consistently than traditional methods. It can dramatically reduce the time needed by teachers to administer and score tests. Meanwhile it delivers more objective, consistent assessments without the variability typical of multiple teachers and administrations. It makes frequent tests possible so that teachers can tailor instructional approaches to individual students by quickly assessing what was working and what was not working.

There has been previous work [9, 10, 11] using ASR in reading passage, mainly in tutoring reading skills. Different modified n-gram language models have been proposed for such task. When applying to tutoring reading assessment, the main concern is to catch as many miscues as possible, at the same time to maintain a low false rejection rate. When applying ASR to children’s oral reading fluency, the main concern is to provide an accurate word correct per minute score, which relies on the low word error rate.

For the testing of children’s oral reading fluency, we know the exact sentences test takers are expected to read. One of the scores, the number of words read correctly, is measured based on the difference between the expected text and what the test taker says. The smaller the difference is, the better. A high accurate speech recognition engine is the key to develop a reliable machine ORF assessment system. In this paper, we discussed the methods Pearson Education used to achieve that target. We compared our methods with previously reported ones and showed that our rule-based language model plus the methods discussed in the paper provided a highly accurate result. The performance improvement from different kinds of rules was discussed.

The main contributions of this paper are:

• Introducing the VersaReader system developed by Pearson Education and demonstrating the feasibility of reliably measuring children’s oral reading fluency.
• Presenting the key technologies leading to the successful testing of children’s oral reading fluency. We discuss the effect of a specifically tuned acoustic model, a rule-based language model, long pause modeling and
adaptation in oral reading fluency tests.
- Showing that the proposed rule-based language model works well for oral reading fluency tests and outperforms traditional n-gram language models. We also analyze the effects of different kinds of language rules.

The remainder of this paper is organized as follows. We first give an introduction about the Pearson’s VersaReader system. We then introduce the rule-based language model and discuss the acoustic model used in the VersaReader system. After that, we present the recognition performance from different configurations, and show the correlation results between human rater and machine. The paper concludes with a discussion of the results.

2. THE VERSAREADER SYSTEM

The VersaReader testing system from Pearson Education combines applications of telephone, web, and advanced language processing technologies to measure the oral reading fluency of students in grade levels 1-8. It can test all types of reading passages used by publishers as well as oral reading fluency programs from the benchmark tests, which are used to assess reading performance at key times during the school year. It helps students to practice passages to build their fluency skills and teachers to monitor student progress and provide instruction support.

Figure 1 shows the procedure of VersaReader. Teachers initiate a telephone call to a student from the online class grade book. Once the student finishes reading a passage, the recorded speech is sent to the grading system for analysis.

The scoring system scores the test and results are available online within minutes. Teachers are able to listen to voice recordings, generate reports, competitive graphs and charts, and compare the results with national norms. A single test can be administered within 2-5 minutes, depending on the length of the passage. Each testing session produces a score based on the number of words read correctly in a minute of reading (Words Correct Per Minute, or WCPM), an informative measure of oral reading fluency. The system can also provide an expressiveness score.

VersaReader uses a web-based, permission-controlled interface to administer tests, manage data, and produce reports. The system is designed to permit hierarchical access to functions. The levels of access in descending order are: District, School, and Classroom. Individuals at “higher” levels are able to view and manipulate information and settings at “lower” levels. For example, a District Administrator (at the highest level) has permission to make changes to account information associated with a Teacher, but the reverse is not permitted. Users can also view reports of performance from students, classrooms, schools, and district, with customizable comparison groups as granted by user level permissions.

3. THE RULE-BASED LANGUAGE MODEL

In a bigram language model [12], only one previous word $w_{i-1}$ will be used to estimate the likelihood of the current word. A trigram language model considers only two previous words $w_{i-1}, w_{i-2}$. The rule-based language model [1] considers much longer sequential dependencies. The basic idea
for the rule-based language model is that for each specified passage, a simple direct graph is built that has a path from the first word in the reading passage to the last word. Different direct arcs are added to represent different classes of errors made by the readers, such as skipping, repeating, inserting, and substituting words. For each arc, a probability is assigned to represent the chance that the arc will be chosen. We use a knowledge-based approach, which includes a list of linguistic rules, such as she may be substituted by he, a single noun may be substituted by a plural noun, to represent some potential arcs. The arc itself can remember which rule it stands for. If such a graph represents many of the renditions encountered, then it should be a useful language model.

The language rules are extracted from many transcriptions of spoken responses to various passages iteratively by starting from four simple rules: any word can be substituted by any word with a very low probability; any word can be inserted after any word with a very low probability; any word can be skipped with a small probability; any word can be repeated with a small probability. They are the only rules that allow out-of-vocabulary words to appear, and their probabilities are fixed to the lowest level no matter how we update other probabilities. They will never be fired unless there is no other choice. By collecting those garbage model firing patterns, we can cluster similar cases and propose the further potential rules iteratively. The probabilities for different rules are estimated using the maximum likelihood method. After we built this knowledge base, it can be directly applied to any other new reading passages.

The RBLM essentially is a Markov chain. Every state in the language model corresponds to the equivalence class of preceding words as classified by the language model. The probability that a speaker will choose the $i^{th}$ word depends on a certain state. Between that state and the start node a path that covers the previous words should exist.

The rule-based language model can be used to infer underlying knowledge about the speech, such as which rule has been fired. The errors which are linguistically acceptable or linguistically unacceptable may be treated differently. The further analysis about the rule-firing detail may provide extra diagnostic linguistic information about the children’s reading habits that can be reported and analyzed.

4. ACOUSTIC MODEL

It is well-known that children’s speech is substantially harder to be recognized than adults’ [13, 14]. In our previous study [1], we used a non-native, telephone band, intra-word triphone, children’s acoustic model to do the speech recognition. This acoustic model was trained from a widely representative sample of non-native children’s spoken materials collected by Pearson. Most of the training materials were from children’s responses in repeating single sentences. After more analysis we discovered that these training materials are not a good fit for our oral reading fluency testing purposes. Therefore, we specifically collected large quantities of recordings from children reading longer ORF passages to train a new similar acoustic model. We noticed a significant performance improvement when using training materials that are similar to the target response materials.

Three children’s oral reading data sets were used in this paper (Table 1). All these data were collected by using landline telephone or cell phones’ speaker mode at 8 kHz. Set 1 consists of 3592 valid oral reading transcribed responses from 902 elementary school students (third, fourth and fifth graders) who read 4 out of 24 total passages. Set 2 consists of a total of 1242 valid oral reading transcribed responses from 414 elementary school students (first, third and fifth graders) who read 3 out of 18 possible passages. Set 3 consists of data from 164 elementary school students (46 first graders, 62 third graders, and 56 fifth graders) recruited from different parts of the United States, from a range of ethnic and linguistic backgrounds. Roughly half of the students were male and half were female. Each student took the grade-appropriate Benchmark test (3 passages), yielding 492 valid oral reading transcribed responses. The average length of a response was around 80 seconds. The passages used in Set 1 are totally different from those in Set 2 and Set 3. Set 2 and Set 3 used the same passages. No student appearing in Set 3 appears in Set 1 or Set 2. The acoustic model was trained on Set 1 and Set 2. The total duration for the training speech is around 108 hours. Set 3 was used as the test set. Machine performance reported in this paper is based on the test set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Grades</th>
<th>#Children</th>
<th>#Passages</th>
<th>#Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>3, 4, 5</td>
<td>902</td>
<td>4/24</td>
<td>3592</td>
</tr>
<tr>
<td>Set 2</td>
<td>1, 3, 5</td>
<td>414</td>
<td>3/18</td>
<td>1242</td>
</tr>
<tr>
<td>Set 3</td>
<td>1, 3, 5</td>
<td>164</td>
<td>3/18</td>
<td>492</td>
</tr>
</tbody>
</table>

Table 1. The speech corpus used in this paper. The acoustic model was trained on Set 1 and Set 2. Set 3 was the test set.

5. RECOGNITION PERFORMANCE

To compare the performance, a naive system that does not involve any speech recognition was designed. It assumes that the number of words ($n$) read by a test taker is known. The first $n$ words from the reading passage were taken as the result. We treat this as a baseline system. We got this $n$ by counting the number of words in the transcription. Only real words were counted. Hesitation and mouth noise were ignored. The WER for such a naive system is 14.46% in Set 3. This WER basically gives a difficult measurement about the testing materials. Using the same method as in [8], we built a bigram and a trigram language model for each passage only based on the corresponding reading passage. No
transcription was used. We noted them as \textit{BLMs (prompts)} and \textit{TLMs (prompts)} in Figure 2. We didn’t treat the sentence boundary specially as [8]. For comparison, we built a bigram and a trigram language model for each passage based on the corresponding transcriptions from Set 2 and noted them as \textit{BLMs (transcriptions)} and \textit{TLMs (transcriptions)} in Figure 2. Unlike the reading tutor for children in [8] that detects voice activity and uses dynamic programming search to align each partial hypothesis, we processed the recording for a passage as a whole. The long pauses between words were handled by language models. The line skipping or repeating was handled by the specific rules in RBLMs.

5.1. Long pauses

We noticed that a long pause from children could happen in any position disregarding the punctuation. This problem was mentioned also in [8]. The recognition error analysis indicated that the general short pause model (SP) has difficulty in handling this problem correctly and introduces errors. We solved this problem by adding the long silence (SIL) as an alternative to the short pause after every word. Using the trigram language models based on reading passages, with the alternative, the WER is 8.94%, and without the alternative, the WER is 11.18%. Using the RBLMs based on training set transcriptions, with the alternative, the WER is 7.54%, and without the alternative, the WER is 8.65%. As we can see, adding the long silence as an alternative to the short pause after every word can decrease the WER significantly. Trigram language models benefit more than RBLMs, 20.0% relatively improvement versus 12.8%. It could be that a small error introducing by SP may cause trigram language models to have difficulty in returning back to the right path. All the performance discussed later used this method.

5.2. Performance comparison

From Figure 2 we can see, the RBLMs are significantly better than trigram language models. For the case without any knowledge of transcriptions for the specified passages, trigram language models achieved WER at 8.94% and RBLMs achieved WER at 7.83%, a 12.4% relatively improvement. In this case, the RBLMs used the task-independent rules and probabilities learned from the NAAL project directly [1]. For the case using the rules for the specified passages (task-dependent), trigram language models achieved WER at 8.47% and RBLMs achieved WER at 7.54%, a 11.0% relatively improvement. The performance improvement by using transcriptions for trigram language models is 5.3% relatively and for RBLMs is 3.5%. Overall, the performance improvement by using passage specified transcriptions is not so big. For oral reading recognition, we may build the language models from the passage prompts directly.

Fig. 2. The recognition performance for different configurations. BLM stands for bigram language model and TLM stands for trigram language model. 40% of OOV words in BLMs (prompts) and TLMs (prompts), 52% of OOV words in BLMs (transcriptions) and TLMs (transcriptions), 30% of OOV words in RBLMs (NAAL rules), and 42% of OOV words in RBLMs (transcriptions) are partial words.

5.3. The role of rules in RBLMs

In this subsection we explored the performance contributions from different rules. All the rules used in RBLMs were classified as five groups: 1. the different kinds of skip/repeat rules; 2. the rules using part-of-speech tagging information, such as plural noun becomes singular noun; 3. the rules adding partial words into language models, such as a partial word could be inserted in the beginning of the original word or replace the original word; 4. the general word level rules, such as a specific word could be substituted by another word (there → where, that → the, this → his, etc.); 5. the hesitation and mouth noise rules.

The baseline system here is RBLMs using all the usable rules learned from transcriptions. Its WER is 7.54%. If we only used the different kinds of skip/repeat rules and ignored others, it becomes 7.82%. Using the different kinds of skip/repeat rules is needed to build a feasible recognition path for the responses. So, when we tested the performance contributions from different rules, we always included the rule set 1. We listed the different configuration results in Figure 3. From the figure we can see, using the different kinds of skip/repeat rules only, the performance is already significantly better than using trigram language models based on the transcriptions, and is close to the performance using all the rules. Letting that RBLMs can handle partial words, hesitation and mouth noise actually decrease the recognition performance a little bit in this case although not so significantly. Adding them together may increase the performance. Hesitation and mouth noise may be handled correctly by the implemented long pause model. The general word level rules
5.4. Adaptation

Adaptation using maximum likelihood linear regression (MLLR) can generally produce substantial gains in recognition accuracy [15]. We explored iterative unsupervised MLLR adaptation. In each iteration, MLLR was used to adapt the general models for each test taker so that they better matched this test taker and then the data was re-recognized with the new models. We used the same decision tree that was used in training the general acoustic models to combine Gaussian mixtures into one base class. With three iterations of adaptations, the WER is reduced from 7.54% to 7.25%. More iterations do not help.

5.5. Decoding speed

RBLMs consider much longer sequential dependencies than trigram language models. Thus, RBLMs could detect bad hypotheses more easily and allow the pruning to happen in an early stage. We noticed that with the same setting, the decoding speed by using RBLMs is significantly faster than using trigram language models. With our current setting, the VersaReader system averagely decodes a 90-second-long speech in around 30 seconds.

6. VALIDATION OF THE VERSAREADER SYSTEM

Human expert evaluators with expertise in reading education were employed to analyze and score the speech files in Set 3. Evaluators provided Accuracy Scores by analyzing and calculating the word-accuracy of each reading. Evaluators logged into a web application that presented audio recordings of the students readings one passage at a time. Evaluators were provided with printed score sheets for each passage on which to annotate student performance. That is, evaluators scored the recorded readings just as they would have had they been in the room with the student performing the reading, with the benefit of being able to replay portions of the recording. This feature was valuable because it allowed the evaluators to listen extremely carefully to the passages, especially when the child’s reading was difficult to hear or understand. Each word on the printed score sheets was numbered to allow evaluators to easily locate the first and last word attempted. Evaluators were trained to use a specialized set of proof marks to indicate reading errors such as word-substitutions, omissions, insertions, reversals, hesitations, and self-corrected instances of all of these; these proof marks were consistent with those used in AIMSweb [8] and other standardized oral reading fluency tests. Evaluators entered the numbers and types of errors into the database as well as the index of the first and last words attempted by the reader. Finally the base measure of Words Correct was used to calculate Words Correct Per Minute (WCPM) scores. The correlation between human and human is 1.00.

The machine WCPM scores were generated by the VersaReader system. It used the acoustic model and RBLMs discussed in this paper. The RBLMs were built using the rules learned from the NAAL project directly. The adaptation was not implemented in the VersaReader system for the consideration of efficiency. A scatter plot of human and machine-generated WCPM scores appears in Figure 4 below. Pearson’s product moment coefficient between human-generated scores and machine-generated scores yielded a very high correlation ($r = 0.997$). Basically there is no obvious outlier in the validation set.

We checked Pearson’s correlation coefficients at each grade level. They are all 0.996. The high correlations for individual grades and the entire (combined) validation set
demonstrate that the WCPM scores generated by the VersaReader system are virtually indistinguishable from scores provided by careful human analysis.

7. CONCLUSIONS

We presented VersaReader, a system automatically scoring children’s oral reading fluency skills. We explored different methods to improve the recognition performance. We showed that the proposed rule-based language model outperformed traditional n-gram language models. Training an acoustic model using similar speaking materials is important. Adding the long silence as an alternative to the short pause after every word helps improve the recognition performance of oral reading significantly. The rules used to deal with partial words, mouth noise or hesitation specifically may not be important for recognizing ORF responses. Using RBLMs plus carefully handling long pauses provides a highly accurate result. A WER 7.25% was achieved in our test set. The impact of different kinds of rules on performance was discussed. We demonstrated that the VersaReader system can provide accurate WCPM scores automatically, which are virtually indistinguishable from scores provided by careful human analysis.

8. REFERENCES


