EFFICIENT AND ROBUST LANGUAGE MODELING IN AN AUTOMATIC CHILDREN'S READING TUTOR SYSTEM

Xiaolong Li[1], Yun-Cheng Ju[2] , Li Deng[2] , Alex Acero[2]

1. Microsoft Education Incubation Group 2. Microsoft Research
One Microsoft Way, Redmond, WA 98052
{xiaolli, yuncj, deng, alexac}@microsoft.com

ABSTRACT

Recently, there has been a rapidly increasing interest in using ASR for children's language learning. An Automatic Reading Tutor system built with ASR technologies can track children's oral reading against story texts, detect reading miscues, and measure the level of reading fluency. They may even diagnose the nature of the miscues and provide feedback to improve reading skills. In such tasks, N-gram language models (LM) may be trained from the whole story text, or may be generated based on current story sentence with heuristic probabilities for both regular words in the sentence and explicitly predicted reading miscues. The disadvantages of those methods are either they require a relatively large text and are time-consuming, or a large-sized LM and complex processing are needed to accommodate all possible words in reading stories as well as in reading miscues. This paper proposes an efficient and robust LM which can be easily built on-the-fly with current reading sentences. With an additional parallel "garbage" model, the LM can also deal effectively with a wide range of reading miscues. Our experiments in a standard children's reading task show that the new LM reaches the state-of-the-art performance in detecting reading miscues with a fast speed while only a relatively simple children's acoustic model of speech was used.

Index Terms— Automatic Reading Tutor, Children's Speech Recognition, Language Model, ASR, Reading Miscues

1. INTRODUCTION

Recent years have witnessed a rapid increase in research activities by both academia and industry in using Automatic Speech Recognition (ASR) for children's language learning [1-5]. An Automatic Reading Tutor (ART) system built with state-of-the-art ASR technologies is capable of improving children's reading skills for several reasons [6]. First, it can track the children's oral reading against story text and provide suitable feedback to show the current word position. A commercial system [5] has shown that providing only simple tracking feedbacks to the child is already very helpful to keep his/her engagement in oral reading tasks. Second, an ART system can detect the reading miscues including mispronunciations, repetitions, deletions, filler pauses, and so on, which are very common for a starting-level reader. By correctly detecting those miscues, the system may provide appropriate help information to the child similar to a human tutor. Such information includes correcting the mispronunciation by providing pre-recorded or TTS-based pronunciations, and suggestions on specific practices with respect to difficult words or phonemes. Third, an ART system can measure the fluency level of a child and provide testing metrics similar to regular language testing. The advantages of the ART system in keeping children engaged in reading and learning for a longer time than human tutor are particularly beneficial. Not only does it save cost and time compared with human tutors, it also improves the reading proficiency with a faster speed than regular classroom study [3].

However, there are a number of technical challenges in building a robust and effective ART system that works well for children. First, the broad range of variants in children's speech makes the acoustic modeling particularly difficult [7]. Second, the system is required to detect as many as possible reading disfluencies or miscues, while not frustrating the child with too many false alarms at the same time. This is a dilemma by itself. Third, the system needs to consider very special user experience with respect to children, as children tend to play with the system with random actions. One typical example is that many children do not like to wear a headphone or cannot help playing with it while reading.

One unique feature of an ART system differentiating it from other speech applications is that the system knows the sentences the child will read. However, a brute-force force-alignment would work poorly in this scenario since children may skip, repeat, pause, jump around the words, or generate non-speech fillers. To deal with these reading miscues, many reported systems exploited an N-gram language model (LM) which may be trained from the whole story text (e.g. [4]), or may be generated based on current reading sentence with heuristic probabilities for both regular words in the sentence and explicitly predicted irregular words or mispronunciations ([3,8]). Regular language modeling methods such as smoothing can be used in this procedure. However, these approaches are either time-consuming in processing whole reading text, or vulnerable for Out-of-Vocabulary words or miscues. To overcome these difficulties, we in this paper propose an efficient language modeling approach using the interpolated N-gram Filler Model [9,10] which can be built on-the-fly based on the current reading sentences. With a built-in "garbage" path, it is possible to detect different kinds of miscues. Our initial experiments in a standard children's story-reading task (using the corpus from University of Colorado-Boulder [11]) show that the performance of detecting reading miscues is as good as the state-of-the-art ART system which built the LM with much more complex processing [4,11]. This performance is achieved using much simpler children's acoustic model than that reported in [11].

This paper is organized as follows. In Section 2 the proposed language modeling approach is described. In Section 3 the "garbage" model for the detection of miscues is discussed. In

Section 4, we report the experiments on a standard public children reading task and draw conclusions in Section 5.

2. ON-LINE LANGUAGE MODELING FOR AUTOMATIC READING TUTOR

Many ART systems use N-gram based statistical language models (SLM) to improve the robustness of the system [3,4]. There are several approaches in the literature to using SLM for ART. First, a **general-domain** N-gram SLM (typically built on the desktop-dictation task) has no development cost for the ART system but the performance is generally not good enough. Second, **domain-specific** SLM (completely trained from reading texts) works well but it requires large data corpus and complex processing involving specially designed toolkits. Here we propose using an **on-line interpolated** LM [10] where a domain-specific core LM is trained completely from a limited set of training sentences (e.g., current story paragraph or sentences). The concept behind this LM is the fact that SLM can be implemented as a Context Free Grammar (CFG) [12], and hence the core LM can be constructed on-the-fly as a CFG given current story texts. At the same time, a general-domain "garbage" model N-Gram Filler [9] (typically a trimmed version of the dictation grammar) is attached through the unigram back-off state to improve robustness. The interpolation is achieved by reserving some unigram counts for the unseen garbage words. Fig. 1 illustrates the proposed LM for a given story paragraph or sentence. There are a target CFG and a garbage CFG, corresponding to the domain-specific LM and general-domain LM, respectively. These two paths are connected by a unigram back-off node from the target CFG. <S> and </S> are the entry and exit nodes for the grammar. The two weights shown in Fig. 1 control the possibilities of moving from the target CFG to the backoff node (w_1), and from the backoff node to the garbage CFG (w_2).

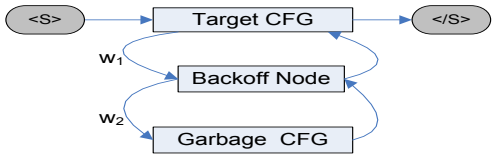

Fig. 1. An schematic illustration for proposed interpolated LM.

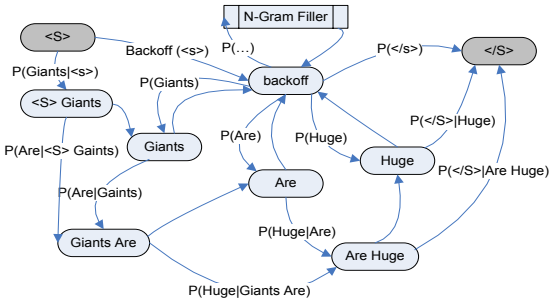

Fig. 2. Interpolated N-gram Model for a short sentence ("Giants are huge."). Transitions without labels are back-off transitions.

The size of the resulting LM above is small (typically a few Kilo-bytes) due to the small size of the current story text. The Garbage model (N-gram Filler) is relatively large --- 4 Mega bytes in our system, but there is no harm to the overall system since it can be shared by all different sentences or paragraphs. Fig. 2 depicts the binary CFG built from a single story sentence "Giants are huge". In addition to the three special states (<S>, </S>, and unigram backoff state), states with one word (e.g., "Giants") are bigram states and states with two words (e.g., "Giants are") are trigram states.

It is critical to use a small-sized LM since it gives fast response time for tracking, which is very important for keeping children's engagement. Some other systems can not provide such real-time tracking for each word (although they did provide sentence-level tracking) [3,5], partially because of the large size in their LMs. In contrast, our approach makes it relatively easy to achieve word-level real-time tracking since the LM can be constructed for each sentence with a very small overhead.

3. DETECTION OF READING MISCUES

Due to the "imperfect" nature of children's speech, most typically for those at the stage of acquiring reading ability, many mispronunciations, repetitions, and insertions, deletions, as well as filler pauses in the speech are expected. All these disfluent speech phenomena are called reading miscues [3,8]. One purpose of an ART system is to automatically detect all reading miscues and trigger necessary help information so that the struggling readers can benefit from the system's feedback with correct pronunciations or specific practices on difficult words [3]. One possible way to detect reading miscues is to predict possible miscues and add them into the lexicon and the LM. In [8], different types of miscues are obtained from a large miscues database and are combined with some truncated forms of regular words. And all those miscues have to be assigned with some small N-gram probabilities discounted from their corresponding regular words' N-grams. In [13], a special decoding engine was developed, where three cascaded lexical trees was built for subword units in addition to the regular lexical tree for normal words. This multi-layer method is targeting on the detection of subword in children's speech. In [14], a phonetic lattice was built based on FST, and the phoneme graph for each word is specially designed to allow partial-pronunciations. All these methods have the advantages in detecting sub-word level miscues, but they come with a cost of needing to re-design the decoding engine or re-train the LM.

No such a cost incurs in our proposed method, where the N-gram garbage model shown in Fig. 1 is used for the detection of reading miscues. This garbage model is obtained from a general-domain N-gram model but is trimmed down so that it includes only the most common 1600 words. To save the CPU/memory, only bigram and unigram are used in the garbage CFG (actually we also found that using trigram-based garbage model did not provide much improvement). We gain three advantages with the above miscue detection scenario. First, it can be easily built on-the-fly. Second, there is no cost to change decoding engine, and it can work with any commercial ASR system. Finally, by changing weights w_1, and w_2 (shown in Fig. 1) it is easy to obtain a ROC curve instead of a fixed point of detection rate and false alarm [3, 4]. This final advantage can be illustrated in Fig. 3, where an equivalent two-path grammar is drawn for each word. Weight w_0 is the equivalent garbage model weight which can be calculated based on w_1, w_2 and the Target CFG in Fig. 1. Given a speech segment *X*, the target word *T*, and the garbage word(s)[1] *G*, we obtain a hypothesis testing scenario as follows:

H_0: Target Word *T* exists;

H_1: Target Word *T* does not exist;

[1] There may be more than one garbage words corresponding to a target word.

Then the decision rule is given by

H_0: when $\frac{P(H_0 \mid X)}{P(H_1 \mid X)} = \frac{(1-w_0)P(X \mid T)P(T)}{w_0 P(X \mid G)P(G)} > 1$

H_1: otherwise

where $P(X \mid T)$ and $P(X \mid G)$ are the acoustic score for the target and garbage words, respectively, and $P(T)$ and $P(G)$ are LM scores for the target and garbage words, respectively. The above decision rule is equivalent to the following decision rule:

H_0: when $\frac{P(X \mid T)P(T)}{P(X \mid G)P(G)} > \lambda = \frac{w_0}{1-w_0}$

H_1: otherwise,

Here λ is a threshold as an explicit function of w_0. As can be seen, this detection scenario is equivalent to regular hypothesis testing in the utterance verification problem.

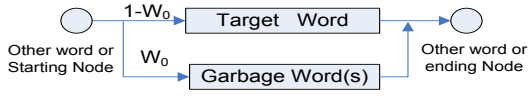

Fig. 3. The equivalent grammar for each single word.

Our detection technique presented above may appear to be able to detect word-level instead of subword-level miscues. However, in our experiments, we have found many detected subword-level miscues. Detailed analysis indicates that this rather remarkable ability is achieved by virtue of outputing short words contained in the garbage model that are acoustically similar to the subwords.

4. EXPERIMENTS

In this section, experiments with proposed language model are reported on a published story-reading task of children. The training and testing data are from three US-English kids speech corpora: the Kid's Prompted & Read Speech corpus (A) and Kid's Read & Summarized Story corpus (B), both from University of Colorado-Boulder [4,11], and the read speech from CMU children speech corpus[2] (C). Table 1 lists the speaker number, grades, utterance number, and duration for each part of data used in training and testing.

	Source	#Spkr	Grades	#utterance	Time (h)
Train-1	A	665	K~5	39006	26.7
Train-2	B	221	1,2	28829	24.5
Train-3	C	76	K~5	5180	9.1
Subtotal	ABC	962	K~5	73015	60.3
Test	B	105	3~5	105	12.4

Table 1. The training and testing data used in the experiments. A is the Kids' Prompt & Read Speech corpus, B is the Kid's Read & Summarized Story corpus, C represents CMU Kid's speech corpus.

The testing data consists of 105 stories read by 105 children in grades 3, 4, and 5 (17 speakers in grade 3, 28 in grade 4 and 60 in grade 5). The data is collected in a quiet room with desktop microphone. The sampling rate is 16 KHz. There are totally 10 different stories and each story contains an average 1054 words (ranging from 532 to 1926 words). Each story also comes with a spontaneous summary generated by the child, but we did not use those data for this study.

The children's acoustic model was trained with HTK tools based on EM algorithm. The language model was built with proposed method. Since each story was recorded in one wav file, an interpolated N-gram was built based on every story instead of every sentence, with a garbage model built from 1600 word based N-gram dictation model in Wall Street Journal domain. The N-gram was built on-line so there is no overhead for language model training as in some other systems. It is also possible to build sentence-level interpolated N-gram, but much more efforts are needed to cut out the large waveform files into sentences.

The similar methodologies as [11] were used in performance measurement. The word error rate was computed by aligning the recognition hypothesis with human transcription. The detection rate and false alarms were computed by firstly aligning story text with human-transcription (*story-trans*) and then aligning story text with recognition hypothesis (*story-hyps*), based on well-known NIST alignment algorithm. A reading miscue is defined as any of cases of insertion, deletion and substitution appeared in *story-trans* alignment. If the same reading miscue also appears in *story-hyps* alignment, it is regarded as a detected miscue; on the other hand, if there is one error in *story-hyps* alignment but no miscue in the same position of *story-trans* alignment, it is regarded as a false alarm. The miscue detection rate is defined as the number of detected miscues divided by total number of miscues; the false alarm is defined as the number of false alarms divided by total number of correctly read words. For the details of those procedures, please refer to [11].

Miscue Categories	Definition
REP	Word Repetition
BR	Breath
PW	Partial Word
PS	Pause
HS	Hesitation or Elongation
WW	Wrong Word
MP	Mispronunciation
BN	Background Noise
IJ	Interjection/Insertion
NS	Non-Speech Sound
OA	Over-articulation

Table 2. The categories and definitions of reading miscues.

Table 2 gives the categories and definitions for all kinds of reading miscues we used in the experiment. It should be noted that different from [11], for substitution errors (e.g., WW and MP), a simple rule was used to judge if two words are the same by directly comparing their spellings, with only one exception for the affix "-s" (e.g., "Cat"→ "Cats" was treated as no error). Compared with "soft decision" used in [11], this hard decision may be more desirable in real application but is believed to somehow reduce the detection rate or increase the false alarm rate.

Table 3 gives the results using above error analyses methodology. A ROC curve is also illustrated in Figure 4. As can be seen from Table 3, when miscue detection rate increases (by increasing garbage model weight), the false alarm rate also increases, and at the same time the word error rate (WER) also goes up, which is mainly because more garbage words were generated. Please note that miscues in function words are also considered here.

Table 3 also gives the system's overall Real-Time Factor (RTF) for different operating points. As can be seen, with larger weight of garbage model, a little bit more running time is needed, but the overall system runs very fast (less than 0.2 xRT). This also proves our previous analyses that this language model has small overhead. Some other experiments also indicated that when the language

[2] http://www.ldc.upenn.edu/Catalog/CatalogList/LDC97S63/

model is built on sentence level instead of story level, it will be even more efficient.

Compared with the recently published results on the same testing task [11], although here the word error rate is worse ([11] reported 8% WER with complex processing in both acoustic and language modeling), the detection rate and false alarm are the same or even better, which indicates our proposed language model may detect more miscues at the same level of false alarm. It should be noted that in this initial experiment, no special technologies were used in improving children's acoustic model, such as VTLN, online adaptation, and Speaker-Adaptation Training, etc. By improving the acoustic model with those technologies in the near future, we can improve recognition accuracy, which may further improve the detection rate and false alarm.

On the other hand, the above result suggests that WER may be not the best metrics for Reading Tutor tasks since different WER may still result in the same detection rate and false alarm. In fact, paper [6] also believes WER is not good enough in this type of task because even when the recognizer makes a mistake, it is still able to detect the miscues.

Detection Rate (%)	False Alarm (%)	WER (%)	RTF
73.40	5.16	12.04	0.10
74.76	6.36	13.02	0.11
75.81	7.56	14.12	0.13
76.31	9.11	15.52	0.17
76.64	11.69	17.98	0.18
76.93	15.15	21.31	0.19

Table 3. Experimental results with proposed N-gram model on children's story-reading task. The weight of garbage model increased from the top row to the bottom row.

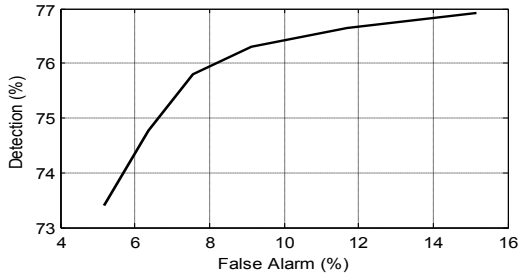

Fig. 4. The ROC curve for the detection rate and false alarm by changing the weight of garbage model.

5. CONCLUSION

This paper proposes an efficient and robust language modeling approach for building an Automatic Reading Tutor system. This language model is based on recently proposed interpolated LM approach by Microsoft Research. The advantages of this method exist in four folds. Firstly, it dramatically reduces the effort for building a new ART system which has new and special challenges considering children as target users. Secondly, it provides robustness to detect different types of reading miscues, without predicting which kind of miscues would be generated by children. Thirdly, by changing the weights of garbage model, it is easy to obtain a ROC curve instead of a fixed point. Finally, it is very efficient, almost no overhead to the system.

In the future, besides improving the children's acoustic model with more training data and better training approaches, we will continue improving the ability of this language model in detecting more reading miscues (at the same or lower false alarm). One possible direction is to use phoneme-level garbage model instead of word-level garbage model as used here, which may be more robust to sub-word level miscues.

6. ACKNOWLEGEMENTS

The authors would thank Margaret Johnson and John Paquin at Microsoft Education Incubation Group for encouragements and supports to this work, and Dr. Bryan Pellom and Dr. Ron Cole (University of Colorado) for advices in using the Kids' Read & Summarized Story corpus. We also thank Dr. Peng Liu from Microsoft Research Asia for insightful discussions.

7. REFERENCES

[1] V. Zue, S. Seneff, J. Polifroni, H. Meng, J. Glass, "Multilingual Human-Computer Interface Interactions: From Information Access to Language Learning," *ICSLP*, Philadelphia, PA, USA, 1996.

[2] M. Russell, C. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, P. Barker, "Applications of Automatic Speech Recognition to Speech and Language Development in Young Children", *ICSLP*, Philadelphia, PA, USA, 1996.

[3] J. Mostow, S. Roth, A. Hauptmann, M. Kane, "A Prototype Reading Coach that Listens", *AAAI*, Seattle, WA, USA, 1994, pp. 785-792.

[4] A. Hagen, B. Pellom, R. Cole, "Children's Speech Recognition with Application to Interactive Books and Tutors," *ASRU*, St. Thomas, USA, 2003.

[5] www.soliloquylearning.com

[6] J. Mostow, "Is ASR accurate enough for automated reading tutors, and how can we tell?" *InterSpeech*, Pittsburgh, PA, 2006, pp. 837-pp.840.

[7] S. Lee, A. Potamianos, S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *JASA*, vol. 105, pp. 1455-1468, Mar. 1999.

[8] S. Banerjee, J. Beck, J. Mostow, "Evaluating the Effect of Predicting Oral Reading Miscues," *InterSpeech*, Geneva, Switzerland, 2003, pp. 3165-3168.

[9] D. Yu, Y. C. Ju, Y. Y. Wang, A. Acero, "N-Gram Based Filler Model for Robust Grammar Authoring," *ICASSP*, Toulouse, France, 2006.

[10] Y.C., Ju., Y. Y. Wang, A. Acero, "Call Classification by Categorizing Utterances and Using Non-Speech Features," *ICASSP*, Toulouse, France, 2006.

[11] K. Lee, A. Hagen, N. Romanyshyn, S. Martin, B. Pellom, "Analysis and Detection of Reading Miscues for Interactive Literacy and Detection of Reading Miscues for Interactive Literacy Tutors," *20th International Conference on Computational Linguistics* (COLING), Geneva, Switzerland, 2004.

[12] G. Riccardi, R. Pieraccini, Bocchieri, "Stochastic Automata for Language Modeling," *Computer Speech and Language*, Vol. 10, pp. 265-293, 1996.

[13] A. Hagen, B. Pellom, "A Multi-Layered Lexical-Tree Based Token Passing Architecture for Efficient Recognition of Subword Speech Units," 2nd Language Technology Conference, Poznan, Porland, 2005.

[14] J. Duchateau, Wigham, K. Demuynck, H. Van Hamme, "A Flexible Recognizer Architecture in A Reading Tutor for Children," the ITRW on Speech Recognition and Intrinsic Variation, Toulouse, France, May 2006, pp. 59-64.