Automated Essay Scoring Based on Finite State Transducer: towards ASR Transcription of Oral English Speech

Xingyuan Peng[*], Dengfeng Ke[*], Bo Xu[*†]
[*] Digital Content Technology and Services Research Center
[†] National Lab of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
No.95 Zhongguancun East Road, Haidian district, Beijing 100190, China
{xingyuan.peng,dengfeng.ke,xubo}@ia.ac.cn

Abstract

Conventional Automated Essay Scoring (AES) measures may cause severe problems when directly applied in scoring Automatic Speech Recognition (ASR) transcription as they are error sensitive and unsuitable for the characteristic of ASR transcription. Therefore, we introduce a framework of Finite State Transducer (FST) to avoid the shortcomings. Compared with the Latent Semantic Analysis with Support Vector Regression (LSA-SVR) method (stands for the conventional measures), our FST method shows better performance especially towards the ASR transcription. In addition, we apply the synonyms similarity to expand the FST model. The final scoring performance reaches an acceptable level of 0.80 which is only 0.07 lower than the correlation (0.87) between human raters.

1 Introduction

The assessment of learners' language abilities is a significant part in language learning. In conventional assessment, the problem of limited teacher availability has become increasingly serious with the population increase of language learners. Fortunately, with the development of computer techniques and machine learning techniques (natural language processing and automatic speech recognition), Computer-Assisted Language Learning (CALL) systems help people to learn language by themselves.

One form of CALL is evaluating the speech of the learner. Efforts in speech assessment usually focus on the integrality, fluency, pronunciation, and prosody (Cucchiarini et al., 2000; Neumeyer et al., 2000; Maier et al., 2009; Huang et al., 2010) of the speech, which are highly predictable like the exam form of the read-aloud text passage. Another form of CALL is textual assessment. This work is also named AES. Efforts in this area usually focus on the content, arrangement and language usage (Landauer et al., 2003; Ishioka and Kameda, 2004; Kakkonen et al., 2005; Attali and Burstein, 2006; Burstein et al., 2010; Persing et al., 2010; Peng et al., 2010; Attali, 2011; Yannakoudakis et al., 2011) of the text written by the learner under a certain form of examination.

In this paper, our evaluation objects are the oral English picture compositions in English as a Second Language (ESL) examination. This examination requires students to talk about four successive pictures with at least five sentences in one minute, and the beginning sentence is given. This examination form combines both of the two forms described above. Therefore, we need two steps in the scoring task. The first step is Automatic Speech Recognition (ASR), in which we get the speech scoring features as well as the textual transcriptions of the speeches. Then, the second step could grade the text-free transcription in an (conventional) AES system. The present work is mainly about the AES system under the certain situation as the examination grading criterion is more concerned about the integrated content of the speech (the reason will be given in subsection 3.1).

There are many features and techniques which are very powerful in conventional AES systems, but

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 50–59,
Jeju, Republic of Korea, 8-14 July 2012.

applying them in this task will cause two different problems as the scoring objects are the ASR output results. The first problem is that the inevitable recognition errors of the ASR will affect the performance of the feature extractions and scoring system. The second problem is caused by the special characteristic of the ASR result. As all these methods are designed under the normal AES situation that they are not suitable for the characteristic.

The impact of the first problem can be reduced by either perfecting the results of the ASR system or building the AES system which is not sensitive to the ASR errors. Improving the performance of the ASR is not what we concern about, so building an error insensitive AES system is what we care about in this paper. This makes many conventional features no longer useful in the AES system, such as spelling errors, punctuation errors and even grammar errors.

The second problem is caused by applying the bag-of-words (BOW) techniques to score the ASR transcription. The BOW are very useful in measuring the content features and are usually robust even if there are some errors in the scoring transcription. However, the robustness would not exist anymore because of the characteristic of the ASR result. It is known that better performance of ASR (reduce the word error rate in ASR) usually requires a strong constrain Language Model (LM). It means that more meaningless parts of the oral speeches would be recognized as the words quite related to the topic content. These words will usually be the key words in the BOW methods, which will lead to a great disturbrance for the methods. Therefore, the conventional BOW methods are no longer appropriate because of the characteristic of the ASR result.

To tackle the two problems described above, we apply the FST (Mohri, 2004). As the evaluating objects are from an oral English picture composition examination, it has two important features that make the FST algorithm quite suitable.

- Picture composition examinations require students to speak according to the sequence of the pictures, so there is strong sequentiality in the speech.
- The sentences for describing the same picture are very identical in expression, so there is a hierarchy between the word sequences in the sentences (the expression) and the sense for the same picture.

FST is designed to describe a structure mapping two different types of information sequences. It is very useful in expressing the sequences and the hierarchy in picture composition. Therefore, we build a FST-based model to extract features related to the transcription assessment in this paper. As the FST-based model is similar to the BOW metrics, it is also an error insensitive model. In this way, the impact of the first problem could be reduced. The FST model is very powerful in delivering the sequence information that a meaningless sequence of words related to the topic content will get low score under the model. Therefore, it works well concerning the second problem. In a word, the FST model can not only be insensitive to the recognition error in the ASR system, but also remedy the weakness of BOW methods in ASR result scoring.

In the remainder of the paper, the related work of conventional AES methods is addressed in section 2. The details of the speech corpus and the examination grading criterion are introduced in section 3. The FST model and its improved method are proposed in section 4. The experiments and the results are presented in section 5. The final section presents the conclusion and future work.

2 Related Work

Conventional AES systems usually exploit textual features to assess the quality of writing mainly in three different facets: the content facet, the arrangement facet and the language usage facet. In the content facet, many existing BOW techniques have been applied, such as the content vector analysis (Attali and Burstein, 2006; Attali, 2011) and the LSA to reduce the dimension of content vector (Landauer et al., 2003; Ishioka and Kameda, 2004; Kakkonen et al., 2005; Peng et al., 2010). In arrangement facet, Burstein et al. (2010) modeled the coherence in student essays, while Persing et al. (2010) modeled the organization. In language usage facet, grammar, spelling and punctuation are common features in assessment of the writing competence (Landauer et al., 2003; Attali and Burstein, 2006), and so does the diversity of words and clauses (Lonsdale and Strong-Krause, 2003; Ishioka and Kameda, 2004). Besides

Grading levels		Content Integrity	Acoustic
(18-20)	passed	Describe the information in the four pictures with proper elaboration	Perfect
(15-17)		Describe all the information in all of the four pictures	Good
(12-14)		Describe most of the information in all of the four pictures	Allow errors
(9-11)	failed	Describe most of the information in the pictures, but lose about 1 or 2 pictures	--
(6-8)		Describe some of the information in the pictures, but lose about 2 or 3 pictures	
(3-5)		Describe little information in the four pictures	
(0-2)		Describe some words related to the four pictures	

Table 1: Criterion of Grading

the textual features, many methods are also proposed to evaluate the quality. The cosine similarity is one of the most common used similarity measures (Landauer et al., 2003; Ishioka and Kameda, 2004; Attali and Burstein, 2006; Attali, 2011). Also, the regression or the classification method is a good choice for scoring (Rudner and Liang, 2002; Peng et al., 2010). The rank preference techniques show excellent performance in grading essays (Yannakoudakis et al., 2011). Chen et al. (2010) proposed an unsupervised approach to AES.

As our work concerns more about the content integrity, we applied the LSA-SVR approach (Peng et al., 2010) as the contrast experiment, which is very effective and robust. In the LSA-SVR method, each essay transcription is represented by a latent semantic space vector, which is regarded as the features in the SVR model. The LSA (Deerwester et al., 1990) considers the relations between the dimensions in conventional vector space model (VSM) (Salton et al., 1975), and it can order the importance of each dimension in the Latent Semantic Space (LSS). Therefore, it is useful in reducing the dimensions of the vector by truncate the high dimensions. The support vector machine can be performed for the function estimation (Smola and Schölkopf, 2004). The LSA-SVR method takes the LSS vector as the feature vector, and applies the SVR for the training data to obtain the SVR model. Each test transcription represented by the LSS vector can be scored by the model.

3 Data

As characteristics of the data determine the effectiveness of our methods, the details of it will be introduced first. Our experimental data is acquired in an oral English examination for ESL students. Three classes of students participated in the exam and 417 valid speeches are obtained in the examination. As the paper mainly focuses on scoring the text transcriptions, we have two ways to obtain them. One is manually typing the text transcriptions which we regarded as the Correct Recognition Result (CRR) transcription, and another is the ASR result which we named ASR transcription. We use the HTK (Young et al., 2006), which stands for the state of art in speech recognition, to build the ASR system.

score	> 0	> 12	> 15	> 18
WER(%)	58.86	50.58	45.56	36.36
MR(%)	72.88	74.03	75.70	78.45

Table 2: WER and MR of ASR result

To better reveal the differences of the methods' performance, all the experiments will be done in both transcriptions. A better understanding of the difference in the CRR transcription and the ASR transcription from the low score to the high score is shown in Table 2, where WER is the word error rate and MR is the match rate which is the words' correct rate.

3.1 Criterion of Grading

According to the Grading Criterion of the examination, the score of the examination ranges from 0 to 20, and the grading score is divided into 7 levels with 3 points' interval for each level. The criterion mainly concerns about two facets of the speech: the acoustic level and the content integrity. The details of the criterion are shown in Table 1. The criterion indicates that the integrity is the most important part in rating the speech. The acoustic level only works well in excellent speeches (Huang et al., 2010). Therefore, this paper mainly focuses on the integrity

Correlation	R1	R2	R3	ES	OC
R1	-	0.8966	0.8557	0.9620	0.9116
R2	-	-	0.8461	0.9569	0.9048
R3	-	-	-	0.9441	0.8739
Average		0.8661		0.9543	0.8968

Table 3: Correlations of Human Scores

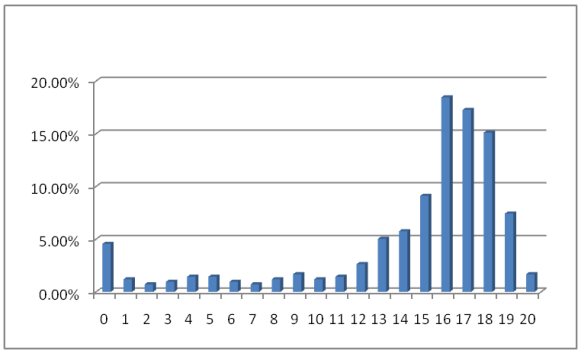

Figure 1: Distribution of Final Expert Scores

of content. The acoustic level as well as other levels such as grammar errors is ignored. Because the criterion is almost based on the content, our methods obtain good performance although we ignore some features.

3.2 Human Score Correlation and Distribution

Each speech in our experiments was scored by three raters. Therefore, we have three scores for each speech. The final expert score is the average of these three scores. The correlations between human scores are shown in Table 3.

R1, R2, and R3 stand for the three raters, and ES is the final expert score. The Open Correlation (OC) is the correlation between human rater scores and the final scores, which are not related to the human scores themselves (average of the other two scores).

As most students are supposed to pass the examination, the expert scores are mostly distributed above 12 points, as shown in Figure 1. In the range of the pass score, the distribution is close to normal distribution, while in the range of failed score except 0, the distribution is close to uniform distribution.

4 Approach

The approach used in this paper is to build a standard FST for the current examination topic. However, the annotation of the corpus is necessary before the building. After the annotation and the building, the features are extracted based on the FST. The automated machine score is computed from the features at last. Therefore, subsection 4.1 will show the corpus annotation, subsection 4.2 will introduce how to build the standard FST of the current topic, and subsections 4.3 and 4.4 will discuss how to extract the features, at last, an improved method is proposed in subsection 4.5.

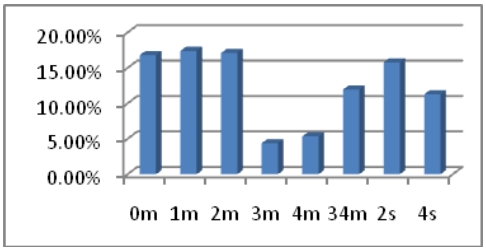

Figure 2: Distribution of Sentence Labels

4.1 Corpus Annotation

The definitions of the sequences and hierarchy in the corpus will be given before we apply the FST algorithm. According to the characteristics of the picture composition examination, each composition can be held as an orderly combination of the senses of pictures. The senses of pictures are called sense-groups here. We define a sense-group as one sentence either describing the same one or two pictures or elaborating on the same pictures. The description sentence is labeled with a tag 'm'(main sense of the picture) and the elaboration one is labeled with 's'(subordinate sense of the picture). The first given sentence in the examination is labeled with 0m and the other describing sentences for the 1 to 4 pictures are labeled with 1m to 4m, while the elaboration ones for the 4 pictures are labeled with 1s to 4s. Therefore, each sentence in the composition is labeled as a sense-group. For the entire 417 CRR transcriptions, we manually labeled 274 transcriptions whose scores are higher than 15 points. We gained 8 types of labels from the manually labeled results. They are 0m, 1m, 2m, 3m, 34m (one sentence describes both of the third and the fourth pictures), 4m, 2s and 4s. Other labels were discarded for the number of their appearance is very low. The distribution of sentences with each label is shown in Figure 2. There are 1679 sentences in the 274 CRR

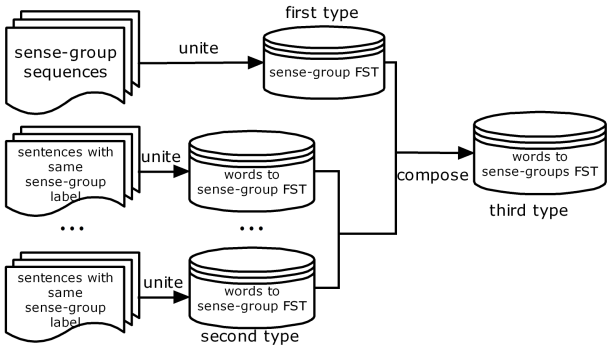

Figure 3: FST Building

transcriptions and 1667 are labeled in the eight symbols.

4.2 FST Building

In this paper, we build three types of FST to extract scoring features with the help of openFST tool (Allauzen et al., 2007). The first is the sense-group FST, the second is the words to each sense-group FST and the last is the words to all the sense-groups FST. They are shown in Figure 3.

The definition of the sense-group has been given in subsection 4.1. The sense-group FST can describe all the possible proper sense-group sequences of the current picture composition topic. It is also an acceptor trained from the labeled corpus. We use manually labeled corpus, which are the sequences of sense-groups of the CRR transcriptions with expert scores higher than 15 points, to build the sense-group FST. In the process, each CRR transcription sense-group sequence is a simple sense-group FST. Later, we unite these sense-group FSTs to get the final FST which considers every situation of sense-group sequences in the train corpus. Also, we use the operation of "determinize" and "minimize" in openFST to optimize the final sense-group FST that its states have no same input label and is a smallest FST.

The second type is the words to sense-group FST. It determines what word sequence input will result in what sense-group output. With the help of these FSTs, we can find out how students use language to describe a certain sense-group, or in other words, a certain sense-group is usually constructed with what kind of word sequence. All the different sentences with their sense-group labels are taken from the train corpus. We regard each sentence as a simple words to sense-group FST, and then unite these FSTs which have the same sense-group label. The final union FSTs can transform proper word sequence into the right sense-group. Like building the sense-group FST, the optimization operations of "determinize" and "minimize" are also done for the FSTs.

The last type of FST is a words to sense-groups FST. We can also treat it as a words FSA, because any word sequence accepted by the words to sense-groups FST is considered to be an integrated composition. Meanwhile, it can transform the word sequence into the sense-group label sequence which is very useful in extracting the scoring features (details will be presented in subsection 4.4). The FST is built from the other two types of FST that we made before. We compute the composition of all the words to each sense-group FSTs (the second type) and the sense-group FST (the first type) with the operations of "compose" in openFST. Then, the composition result is the words to sense-groups FST, the third type of FST in this paper.

4.3 Search for the Best Path in FST

Now we have successfully built the words to sense-groups FST, the third type described above. Just like the similarity methods mentioned in section 2 can score essays from a have-been-scored similar essay, we need to find the best path, which is closest to the to-be-scored transcription, in the FST. Here, we apply the edit distance to measure how best the path is. This means the best path is the word sequence path in the FST which has the smallest edit distance compared with the to-be-scored transcription's word sequences .

Here, we modify the Wagner-Fischer algorithm (Wagner and Fischer, 1974), which is a Dynamic Programming (DP) algorithm, to quest the best path in the FST. A simple example is illustrated in Figure 4. The best path can be described as

$$path = \underset{\substack{path \in \\ allpath}}{\arg\min} EDcost(path, transcription) \quad (1)$$

$$EDcost = ins + del + sub \quad (2)$$

EDcost is the edit distance from the transcription to the paths which start at state 0 and end at the end

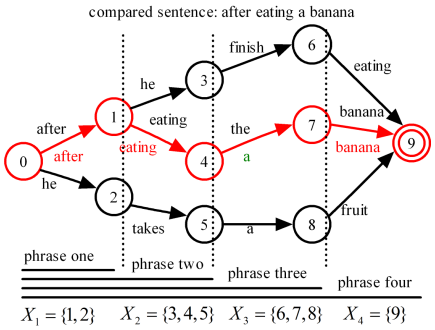

Figure 4: Search the Best Path in the FST by DP

state. The DP process can be described by equation (3):

$$\min EDcost(i) = \underset{\substack{j \in \\ X_1,\ldots,X_{p-1}}}{\arg\min} \left(\min EDcost(j) + cost(j,i)\right) \tag{3}$$

The minEDcost(j) is the accumulated minimum edit distance from state 0 to state j, and the cost(i,j) is the cost of insertion, deletion or substitution from state j to state i. The equation means the minED of state i can be computed by the accumulated minEDcost of state j in the phase p. The state j belongs to the have-been-calculated state set $\{X_0,\ldots,X_{p-1}\}$ in phase p. In phrase p, we compute the best path and its edit distance from the transcription for all the to-be-calculated states which is the X_p shown in Figure 4. After computing all the phrases, the best path and its edit distances of the end states are obtained. Then the final best path is the one with the smallest edit distance.

4.4 Feature Extraction

After building the FST and finding the best path for the to-be-scored transcription, we can extract some effective features from the path information and the transcription. Inspired by the similarity scoring measures, our proposed features represent the similarity between the best path's word sequence and the to-be-scored transcription.

The features used for the scoring model are as follows:

- **The Edit Distance (ED):**

  The edit distance is the linear combination of the weights of insertion, deletion and substitution. The relation is shown in equation (2), where ins, del and sub are the appearance times of insertions, deletions and substitutions, respectively. Normally, we set the cost of each to be 1.

- **The Normalized Edit Distance(NED):**

  The NED is the ED normalized with the transcription's length.

$$NEDcost = EDcost/length \tag{4}$$

- **The Match Number(MN):**

  The match number is the number of words matched between the best path and the transcription.

- **The Match Rate(MR):**

  The match rate is the match number normalized with the transcription's length.

$$MR = MN/length \tag{5}$$

- **The Continuous Match Value(CMV):**

  Continuous match should be better than the fragmentary match, so a higher value is given for the continuous situation.

$$CMV = \sum OM + 2\sum SM + 3\sum LM \tag{6}$$

  where OM (One Match) is the fragmentary match number, SM (Short Match) is the continuous match number which is no more than 4, and LM (Long Match) is the continuous match number which is more than 4.

- **The Length(L):**

  The length of transcription. Length is always a very effective feature in essay scoring (Attali and Burstein, 2006).

- **The Sense-group Scoring Feature(SSF):**

  For each best path, we can transform the transcription's word sequence into the sense-group label sequence with the FST. Then, the words match rate of each sense-group can be computed. The match rate of each sense-group can be regarded as one feature so that all the sense-group match rate in the transcription will be combined to a feature vector (called the Sense-group Match Rate vector (SMRv)), which is an 8-dimensional vector in the present experiments. After that, we applied the SVR algorithm to train a sense-group scoring model with the vectors and scores, and the transcription gets its SSF from the model.

4.5 Extend the FST model with the similarity of synonym

Because the FST is trained from the limited corpus, it does not contain all the possible situations proper for the current composition topic. To complete the current FST model, we add the similarity of synonym to extend the FST model so that it can handle more situations.

The extension of the FST model is mainly reflected in calculation of the edit distance of the best path. The previous edit distance, in equation (2), refers to the Levenshtein distance in which the insertions, deletions and substitutions have equal cost, but in the edit distance in this section, the cost of substitutions is less than that of insertions and deletions. Here, we assume that the cost of substitutions is based on the similarity of the two words. Then with the help of different cost of substitutions, each word edge is extended to some of its synonym word edges under the cost of similarity. The new edit distance is calculated by equation (7) as follows:

$$EDcost = ins + del + sub \times (1 - sim) \qquad (7)$$

where, sim is the similarity of two words.

We used the Wordnet::Similarity software package (Pedersen et al., 2004) to calculate the similarity between every two words at first. However, the performance's reduction of the AES system indicates that the similarity is not good enough to extend the FST model. Therefore, we seek for human help to accurate the similarity calculation. We manually checked the similarity, and deleted some improper similarity. Thus the final similarity applied in our experiment is the Wordnet::Similarity software computing result after the manual check.

5 Experiments

In this section, the proposed features and our FST methods will be evaluated on the corpus we mentioned above. The contrasting approach, the LSA-SVR approach, will also be presented.

5.1 Data Setup

The experiment corpus consists of 417 speeches. With the help of manual typing and the ASR system, 417 CRR transcriptions and 417 ASR transcriptions are obtained from the speeches after preprocessing which includes the capitalization processing and the stemming processing. We divide them into 3 sets by the same distribution of their scores. Therefore, there are totally 6 sets, and each of them has 139 of the transcriptions. The FST building only uses the CRR transcriptions whose expert scores are higher than 15 points. While treating one set (one CRR set) as the FST building train set, we get the ED, NED, MN, MR, CMV features and the SMR vectors for the other two sets(could be either CRR sets or ASR sets). Then, the SSF is obtained by another set as the SVR train set and the last set as the test set. The parameters of the SVR are trained through the grid search from the whole data sets (ASR or CRR sets) by cross-validation. Therefore, except the length feature, the other six features of each set can be extracted from the FST model.

FST build	SVR train	SVR test	CRR transcription	ASR transcription
Set2	Set3	Set1	0.7999	0.7505
Set3	Set2		0.8185	0.7401
Set1	Set3	Set2	0.8557	0.7372
Set3	Set1		0.8111	0.7257
Set1	Set2	Set3	0.9085	0.8086
Set2	Set1		0.8860	0.8086

Table 4: Correlation Between the SSF and the Expert Scores

Also, we presented the result of using LSA-SVR approach as a contrast experiment to show the improvement of our FST model in scoring oral English picture composition.

To quantitatively assess the effectiveness of the methods, the Pearson correlation between the expert scores and the automated results is adopted as the performance measure.

5.2 Correlation of Features

The correlations between the seven features and the final expert scores are shown in Tables 4 and 5 on the three sets.

The MN and CMV are very good features, while the NED is not. This is mainly due to the nature of the examination. When scoring the speech, human raters concern more about how much valid information it contains and irrelevant contents are not taken for penalty. Therefore, the match features are more reasonable than the edit distance features. This im-

Script	Train	Test	L	ED	NED	MN	MR	CMV
CRR	Set2	Set1	0.7404	0.2410	-0.6690	0.8136	0.1544	0.7417
	Set3			0.3900	-0.4379	0.8316	0.1386	0.7792
	Set1	Set2	0.7819	0.4029	-0.7667	0.8205	0.4904	0.7333
	Set3			0.4299	-0.5672	0.8370	0.5090	0.7872
	Set1	Set3	0.8645	0.4983	-0.7634	0.8867	0.2718	0.8162
	Set2			0.3639	-0.6616	0.8857	0.3305	0.8035
	Average		0.7956	0.3877	-0.6443	0.8459	0.3158	0.7769
ASR	Set2	Set1	0.1341	-0.2281	-0.6375	0.7306	0.6497	0.7012
	Set3			-0.1633	-0.5110	0.7240	0.6071	0.6856
	Set1	Set2	0.2624	-0.0075	-0.4640	0.6717	0.5929	0.6255
	Set3			0.0294	-0.4389	0.6860	0.6259	0.6255
	Set1	Set3	0.1643	-0.1871	-0.5391	0.7419	0.6213	0.7001
	Set2			-0.1742	-0.4721	0.7714	0.6199	0.7329
	Average		0.1869	-0.1218	-0.5104	0.7209	0.6195	0.6785

Table 5: Correlations Between the Six Features and the Expert Scores

Script	Method	Set1	Set2	Set3	Average
CRR	Length	0.7404	0.7819	0.8645	0.7956
	LSA-SVR	0.7476	0.8024	0.8663	0.8054
	FST	0.8702	0.8852	0.9386	**0.8980**
ASR	Length	0.1341	0.2624	0.1643	0.1869
	LSA-SVR	0.5975	0.5643	0.5907	0.5842
	FST	0.7992	0.7678	0.8452	**0.8041**

Table 6: Performance of the FST Method, the LSA-SVR Approach and the Length Feature

pact is similar to the result displayed by the ASR output performance in Table 2 in section 3, where the WER has significant difference from the low score speeches to the high score ones while the MR does not, and the MR is much better than the WER.

As the length feature is a strong correlation feature in CRR transcription, the MR feature, which is normalized by the length, is strongly affected. However, with the impact declining in the ASR transcription, the MR feature performs very well. This also explains the reason of different correlations of ED and NED in CRR transcription.

The SSF is entirely based on the FST model, so the impact of the length feature is very low. The decline of it in different transcriptions is mainly because of the ASR error.

5.3 Performance of the FST Model

For each test transcription, it has 12 dimensions of FST features. The ED, NED, MN, MR and CMV features have two dimensions of each as trained from two different FST building sets. The SSF needs two train sets as there are two train models: one is for the FST building model and another is for the SVR model. As different sets for different models, it also has two dimension features. We use the linear regression to combine these 12 features to the final automated score. The linear regression parameters were trained from all the data by cross-validation. After the weight of each feature and the linear bias are gained, we calculate the automated score of each transcription by the FST features. The performance of our FST model is shown in Table 6. Compared with it, the performance of the LSA-SVR algorithm, the baseline in our paper, is also shown. As a usual best feature for AES, the length shows its outstanding performance in CRR transcription. However, it fails in the ASR transcription.

As we have predicted above, the BOW algorithm (the LSA-SVR) performance declines drastically in the ASR transcription, which also happens to the length feature. By contrast, the decline of the performance of our FST method is acceptable considering the impact of recognition errors in the ASR system. This means the FST model is an error insensitive model that is very appropriate for the task.

5.4 Improvement of FST by Adding the Similarity

The improved FST extends the original FST model by considering the word similarity in substitutions. In the extension, the similarities of the synonyms

Script	Method	Set1	Set2	Set3	Average
CRR	FST	0.8702	0.8852	0.9386	0.8980
	IFST	0.8788	0.8983	0.9418	0.9063
ASR	FST	0.7992	0.7678	0.8452	0.8041
	IFST	0.8351	0.7617	0.8168	0.8045

Table 7: Performance of the FST Method and the Improved FST Method

describe the invisible (extended) part of the FST, so it should be very accurate for the substitutions cost. Therefore, we added manual intervention to the similarity result calculated by the wordnet::similarity software packet.

After we added the similarity of synonym to extend the FST model, the performance of the new model increased stably in the CRR transcription. However, the increase is not significant in the ASR transcription (shown in Table 7). We believe it is because the superiority of the improved model is disguised by the ASR error. In other words, the impact of ASR error under the FST model is more significant than the improvement of the FST model. The performance correlation of our FST model in the CRR transcription is about 0.9 which is very close to the human raters' (shown in Table 3). Even though the performance correlation in the ASR transcription declines compared with that in the CRR transcription, the FST methods still perform very well under the current recognition errors of the ARS system.

6 Conclusion and Future work

The aforementioned experiments indicate three points. First, the BOW algorithm has its own weakness. In regular text essay scoring, the BOW algorithm can have excellent performance. However, in certain situations, such as towards ASR transcription of oral English speech, its weakness of sequence neglect will be magnified, leading to drastic decline of performance. Second, the introduced FST model is suitable in our task. It is an error insensitive model under the task of automated oral English picture composition scoring. Also, it considers the sequence and the hierarchy information. As we expected, the performance of the FST model is more outstanding than that of the BOW metrics in CRR transcription, and the decline of performance is acceptable in ASR transcription scoring. Third, adding the similarity of synonyms to extend the FST model improves the system performance. The extension can complete the FST model, and achieve better performance in the CRR transcription.

The future work may focus on three facets. First, as the extension of the FST model is a preliminary study, there is much work that can be done, such as calculating the similarity more accurately without manual intervention, or finding a balance between the original FST model and the extended one to improve the performance in ASR transcription. Second, as the task is speech evaluation, considering the acoustic features may give more information to the automated scoring system. Therefore, the features at the acoustic level could be introduced to complete the scoring model. Third, the decline of the performance in ASR transcription is derived from the recognition error of ASR system. Therefore, improving the performance of the ASR system or making full use of the N-best lists may give more accurate transcription for the AES system.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 90820303 and No. 61103152). We thank the anonymous reviewers for their insightful comments.

References

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut and Mehryar Mohri. 2007. OpenFst: a general and efficient weighted finite-state transducer library. In Proceedings of *International Conference on Implementation and Application of Automata*, 4783: 11-23.

Yigal Attali. 2011. A differential word use measure for content analysis in automated essay scoring. ETS research report, ETS RR-11-36.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater®V.2. *The Journal of Technology, Learning, and Assessment*, 4(3), 1-34.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human Language Technologies: The Annual Conference of the North American Chapter of the ACL*, 681-684.

Chih-Chung Chang, Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, Vol. 2.

Yen-Yu Chen, Chien-Liang Liu, Chia-Hoang Lee, and Tao-Hsing Chang. 2010. An unsupervised automated essay scoring system. *IEEE Intelligent Systems*, 61-67.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Acoustical Society of America*, 107(2): 989-999.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*. 41(6): 391-407.

Shen Huang, Hongyan Li, Shijin Wang, Jiaen Liang and Bo Xu. 2010. Automatic reference independent evaluation of prosody quality using multiple knowledge fusions. In *INTERSPEECH*, 610-613.

Tsunenori Ishioka and Masayuki Kameda. 2004. Automated Japanese essay scoring system: jess. In Proceedings of the International Workshop on database and Expert Systems applications.

Tuomo Kakkonen, Niko Myller, Jari Timonen, and Erkki Sutinen. 2005. Automatic essay grading with probabilistic latent semantic analysis. In Proceedings of *Workshop on Building Educational Applications Using NLP*, 29-36.

Thomas K. Landauer, Darrell Laham and Peter Foltz. 2003. Automatic essay assessment. *Assessment in Education: Principles, Policy and Practice* (10:3), 295-309.

Deryle Lonsdale and Diane Strong-Krause. 2003. Automated rating of ESL essays. In Proceedings of *the HLT-NAACL Workshop: Building Educational Applications Using NLP.*

Andreas Maier, F. Hönig, V. Zeissler, Anton Batliner, E. Körner, N. Yamanaka, P. Ackermann, Elmar Nöth 2009. A language-independent feature set for the automatic evaluation of prosody. In *INTERSPEECH*, 600-603.

Mehryar Mohri. 2004. Weighted finite-state transducer algorithms: an overview. *Formal Languages and Applications*, 148 (620): 551-564.

Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, Mitchel Weintraub. 2000. Automatic scoring of pronunciation quality. *Speech Communication*, 30(2-3): 83-94.

Ted Pedersen, Siddharth Patwardhan and Jason Michelizzi. 2004. WordNet::Similarity - measuring the relatedness of concepts. In Proceedings of *the National Conference on Artificial Intelligence*, 144-152.

Xingyuan Peng, Dengfeng Ke, Zhenbiao Chen and Bo Xu. 2010. Automated Chinese essay scoring using vector space models. In Proceedings of *IUCS*, 149-153.

Isaac Persing, Alan Davis and Vincent Ng. 2010. Modeling organization in student essays. In Proceedings of *EMNLP*, 229-239.

Lawrence M. Rudner and Tahung Liang. 2002. Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2):3-21.

G. Salton, C. Yang, A. Wong. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11): 613-620.

Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14(3): 199-222.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. Journal of the *ACM*, 21(1):168-173.

Helen Yannakoudakis, Ted Briscoe and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In Proceedings of *ACL*, 180-189.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland. 2006. The HTK book (for HTK version 3.4). Cambridge University Engineering Department.