

# Towards The Automatically Semantic Scoring in Language Proficiency Evaluation

Jie Jiang, Bo Xu

*Institute of Automation, Chinese Academy of Sciences,  
Beijing, China*

*{jjiang, xubo}@hitic.ia.ac.cn*

## Abstract

*Many features have been proposed to evaluate examinees' language proficiency. However, few of them are semantic based. In this paper, a novel feature for semantic scoring is presented. It is designed for a typical question type in language tests, namely reading-answering-problem. The proposed feature extraction process involves several operations: transcribing the speech data, automatically tagging the transcribed text and scoring the tagged text. The pattern based tagging is performed on the pre-designed Finite State Machines (FSMs) and the scoring fusion is based on the semantic calculations in a knowledge database. Experiment on Mandarin data validates the effectiveness of the semantic feature in the language proficiency evaluation.*

## 1. Introduction

Computer aided language learning (CALL) systems which are designed to listen to examinees' speeches and to judge their language abilities are very valuable in foreign language learning and proficiency evaluation.

In the past decades, great achievements have been acquired in the field of CALL. Many features, such as the word posterior probability, timing and duration methods [1-4] are proved to be efficient to foretell the language proficiency. Methods which combine more or less of these features are also proposed to provide the machine scores automatically.

Although the performance of the above features is comparable to that of human raters in closed-question, namely passage reading and sentence repeating etc., they do not consider the examinees' semantic correctness. Therefore, it is not fit for the tests which aim to evaluate the examinees' understanding and expressing abilities.

In fact, in the real tests, reading-answering-problem is one of the most common question types. The reading-answering-problem is usually composed of a reading passage, a question, and an answer rubric. Its question is limited on the passage. In testing, examinees are ordered to answer the question after reading the passage. They are allowed to express their idea freely and their responses are graded according to the answer rubric. The answer rubric usually contains following specification: the standard answer text to the question, the pronunciation requests, and the tone requests etc.

Compared with the evaluation methods of the closed-question, there is one problem to be addressed with the consideration of the above characteristics of reading-answering-problem:

- Rejections on synonymous answers.

This problem arises when the examinees try to express the same idea in another way. If it is not thoroughly considered, the performance of CALL system will be decreased significantly.

To solve this problem, in this paper, a novel feature is presented to evaluate the semantic correctness of the answers to the reading-answering-problem. The question context and the knowledge database are incorporated into the evaluation process to solve the above problems. The question context is composed of the objective topic materials, namely the reading passage, the question and the answer rubric. It is utilized in the semantic scoring process. Furthermore, to evaluate the semantic correctness of the answer text, a pattern based method is devised to capture the answer scheme. It is implemented by Finite State Machines (FSMs), which can tag the answers by pre-designed patterns. Finally, the semantic feature is mapped from scores extracted based on the tagged answers.

The organization of this paper is as follows: Section 2 first gives an overview of our proposed feature extraction process, and then the main components of the model are explained in detail in Section 3-4,

including semantic evaluation and score mapping etc. In Section 5, the proposed feature is systematically evaluated with a standard corpus. Finally, a further discussion is given in Section 6.

## 2. Overview on semantic feature extraction

The extraction of the proposed semantic feature is depicted in Figure 2. First, the input WAV files are transcribed in the pre-processing module. Sequentially, two semantic scores are extracted from sentence tagging and text evaluation modules. At last, the two scores are mapped into the semantic feature. It is worth noting that the question context and a knowledge database are incorporated into the extraction process to address the problem mentioned in Section 1. Besides, Finite State Machines (FSMs) are generated from the question context to support the sentence tagging module.

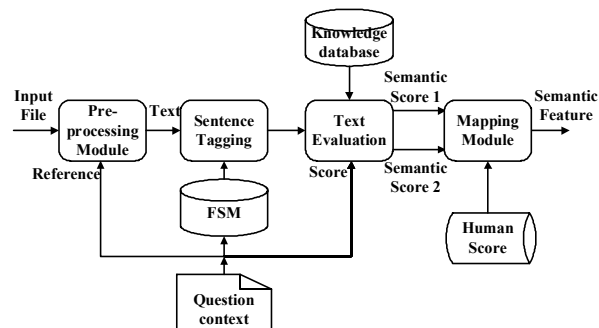


Figure 1. System structure for reading-answering-problem

In the pre-processing module, the examinees' speeches are transcribed on the reference of the question context. The transcriptions of the speeches are used in the latter processing units. Thus, the performance of the proposed method can be considered as text dependent only.

The follow sections will concentrate on the details of the feature extraction process.

## 3. Semantic Evaluation

The aim of semantic evaluation is to score the transcriptions according to the answer rubric. Since the examinees are allowed to present their ideas freely, the transcriptions are diverse in the sentence structures. Therefore, the key to this problem is how to validate the transcribed sentences given the objective topic materials.

The proposed method is to divide this process into two operations: sentence tagging and text evaluation. The first operation is trying to identify the structure of the transcribed sentences by a pattern based method,

and the second operation is trying to scoring the tagged sentences by knowledge databases.

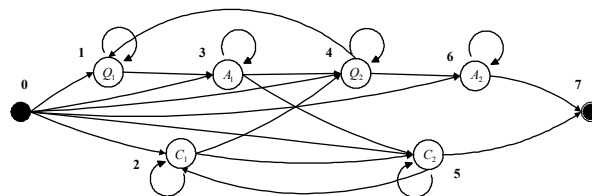
## 3.1. Sentence Tagging

In the sentence tagging module, the structure of the examinees' answers is identified by some predefined tags. Suppose the examinees' answers are structural related to the question context, the tags are involved to describe the answer patterns.

In reading-answering-problem, given  $M$  questions  $Q_i$  and  $M$  related standard answers  $A_i$ ,  $1 \leq i \leq M$ , where  $M$  is the number of questions. Thus, the goal is to tag the sentences by the tags in Table 1:

Table 1: Tags for sentences

Tag	Description
$Q_i$	sentence is related to the question $Q_i$
$A_i$	sentence is related to the answer $A_i$
$C_i$	sentence is related to the composition of question $Q_i$ and answer $A_i$



Transition Matrix

Node	0	1	2	3	4	5	6	7
0	0	1	1	1	1	1	1	0
1	0	1	1	1	0	0	0	1
2	0	0	1	1	1	1	0	1
3	0	1	0	1	1	1	0	1
4	0	0	0	0	1	1	1	1
5	0	0	0	0	0	1	0	1
6	0	0	0	0	1	0	1	1
7	0	0	0	0	0	0	0	0

Figure 2. FSM for  $M=2$

To capture the topological structure of the answers, the patterns are formulated into a Finite State Machine (FSM). The topology of the FSM is typically drawn from the transcriptions of sample database, and organized by experts manually. One FSM represents the typical answer structure to a certain question and is only dependent on the number of questions. Evaluation model will draw the appropriate FSM from the database according to the question numbers. One sample of an FSM is depicted in Figure 2, for the case of  $M=2$ . Note that, for concision, some of the edges in Figure 2 are omitted.

Then, our task is to find the best tag-path to describe the structure of the transcribed text. Denote the tag sequence set generate by an FSM as  $T$ , given the sentence number of the transcribed text  $L$ , the goal

is to find the best match in tag sequences in  $T_L$ , where  $T_L$  is the subset of  $T$  with the sequence length  $L$ . To solve it, each of the sequence in  $T_L$  is expanded into a sentence set  $S_L$  by their tags. Each element of  $S_L$  is composed of a sequence of questions, answers or compositions correspondent to the element in  $T_L$ . Denote the transcribed text as  $r$ , the objective function can be formulated as:

$$\arg \max_{s_i \in S_L} \{Sim(r, s_i)\} \quad (1)$$

where  $Sim$  is the similarity between two sequences of sentence. Suppose  $r$  is composed of  $L$  sentences named  $r_j$ , and  $s_i$  is composed of  $L$  sentences named  $s_{ij}$ , where  $1 \leq j \leq L$ . Function (1) can be approximated as:

$$\begin{aligned} & \arg \max_{s_i \in S_L} \{Sim(r, s_i)\} \\ & \approx \arg \max_{s_i \in S_L} \left\{ \sum_{j=1}^L \lambda_j Sim(r_j, s_{ij}) \right\} \end{aligned} \quad (2)$$

where  $\lambda_j$  is the alignment weight determined by the tag correspondent to  $s_{ij}$ , and  $Sim(r_j, s_{ij})$  is the similarity calculation between two sentences in section 4.3. Thus the problem becomes computationally tractable, and the best tag sequence for the transcribed text could be found.

### 3.2. Text Evaluation

In this module, the tagged transcribed text is scored. Since the tags indicate the answer patterns of the transcribed text, scoring can be implemented by calculating the semantic similarities. Here both the goodness of structure and correctness of the answer are considered respectively. The first is treated as the structure score, and second is treated as the correctness score.

Given each tag a structure weight  $w_j$ , and the sequence of sentences  $s_l$  which are related to the best tag sequence  $I$ , the structural goodness of  $r$  is measured by:

$$Struct(r) \approx \sum_{j=1}^L w_j Sim(r_j, s_{ij}) \quad (3)$$

where  $w_j$  is the scoring weight determined by the tag correspondent to  $s_{ij}$ , and  $Sim(r_j, s_{ij})$  is the similarity calculation between two sentences in section 4.3. Generally, we suppose the structure weight  $w_j$  equals to one.

To get the correctness score, tag set  $G$  is divided into  $M+1$  subsets  $g_k$  ( $1 \leq k \leq M+1$ ) by tag descriptions. For example, all questions  $Q_i$  ( $1 \leq i \leq M$ ) are in subset  $g_1$ , and  $A_i$  and  $C_i$  are in subset  $g_{i+1}$ . Given each subset a score weight  $h_k$ , the correctness of  $r$  is given by:

$$\begin{aligned} & Correct(r) \\ & \approx \sum_{g_k \in G} h_k \frac{\sum_{j=1}^L Sim(r_j, s_{ij}) \delta(s_{ij}, g_k)}{1 + \sum_{j=1}^L \delta(s_{ij}, g_k)} \end{aligned} \quad (4)$$

where  $\delta(s_{ij}, g_k)$  is a indicator function defined by:

$$\delta(s_{ij}, g_k) = \begin{cases} 1, & \text{the tag of } s_{ij} \text{ is in set } g_k \\ 0, & \text{the tag of } s_{ij} \text{ is NOT in set } g_k \end{cases} \quad (5)$$

Thus, formula (3) and (4) are used to calculate the most two important parameters for the semantic feature.

The first parameter, goodness of the structure, is given by formula (3). It describes how well the examinee's answer matches the given FSM. For example, if the structure of the examinee's answer is generally acceptable, the objective function (1) will reach a high score, which will result in a high structural score (3). Otherwise, if the expression is poor, the outcome of formula (3) will be very low.

The second parameter, correctness of the answer, is represented by formula (4). It reflects how closely the examinee's answer resembles the given rubrics. As shown by formula (4), it is a similarity score of predefined answers. Only if the examinees' answer is both structurally and semantically correct, the result of the formula (4) will reach a high score.

### 3.3. Similarity Calculation

The method to calculate the similarity between two sentences is presented. It is considered to be an alignment problem similar to that in Machine Translation [5]. Given a pair of sentences, the task is to choose the word alignment that maximizes the semantic similarity over all possible alignment. It is formulated as:

$$\arg \max_A \{Sim(A | S_1, S_2)\} \quad (6)$$

where  $S_1$  and  $S_2$  are two sentences to be aligned and  $A$  is an alignment. An alignment  $A$  is a set consisting of  $W_1 \Leftrightarrow W_2$  pairs where each  $W_1$  or  $W_2$  is a word. The approximation is taken by accumulate the similarity of aligned words:

$$\begin{aligned} & Sim(A | S_1, S_2) \\ & \approx \sum_{(W_1 \Leftrightarrow W_2) \in A} Sim(W_1 \Leftrightarrow W_2 | S_1, S_2) \end{aligned} \quad (7)$$

Note that, for simplicity, the inner sentence context is not considered, so the similarity between two words is formulated as:

$$Sim(W_1 \Leftrightarrow W_2 | S_1, S_2) = Sim(W_1, W_2) \quad (8)$$

Given an alignment  $A$ , once the similarity between two words can be calculated, the similarity of the two sentences can be measured. Since the lengths of the sentences are relatively small in this problem, the best

alignment  $A^*$  can be easily found out by enumeration and the correspondent sentence similarity can be computed by formula (6) with  $A^*$ .

It is worth noting that the calculation of words similarities  $Sim(W_1, W_2)$  has already been investigated by Q. Liu [6] and P. Rosso [7], for Mandarin and English respectively. Both of their methods are based on knowledge databases. Q. Liu’s method for Mandarin is adopted here. The method utilizes a Chinese knowledge database named *Hownet* [8] to calculate the semantic similarities between two Chinese words. In the experiment, the implementation is based on Q. Liu’s software available at the CNLP website [9].

#### 4. Score mapping

In this module, the structural goodness in formula (3) and the correctness in formula (4) are mapped into the semantic feature. The linear regression algorithm [2] is employed to perform the mapping. It is given by:

$$Y = \beta_0 x_0 + \beta_1 x_1 + \varepsilon \quad (9)$$

Where  $Y$  is the semantic feature,  $x_0$  and  $x_1$  are the goodness and correctness, respectively.  $\beta_0$  and  $\beta_1$  are the linear regression coefficients and  $\varepsilon$  is the residue.

Note that,  $\beta_0$  and  $\beta_1$  are pre-trained from the human scores under the MSE criteria. Thus, the output of the text evaluation module is mapped into the final semantic feature.

### 5. Experiments results

In this section, the proposed feature will be evaluated on a standard corpus.

#### 5.1. Experiment Corpus

The experiment corpus consists of 397 examinees’ speeches in PCM format. It is collected from Mandarin tests. The corpus is separated into two parts: a training set (198 speeches) and a testing set (199 speeches). Parameters in feature extraction process are trained on the training set, and the effectiveness of the feature is validated in the testing set.

#### 5.2. Performance Measure

To quantitatively assess the effectiveness of the proposed semantic feature, the correlation between the human scores and the mapping scores is adopted as the performance measure [1-4]. The correlation could be formulated as:

$$Correlation(a, b) = \frac{\sum_{i=1}^N [(a_i - \bar{a}) \times (b_i - \bar{b})]}{\sqrt{\sum_{i=1}^N (a_i - \bar{a})^2 \times \sum_{i=1}^N (b_i - \bar{b})^2}} \quad (10)$$

where  $N$  is the dimension of the two score vectors,  $\bar{a}$  is the mean of vector  $a$ , and  $\bar{b}$  is the mean of vector  $b$ .

#### 5.3. Human scores

In order to measure the scoring consistency, the testing set has been annotated by five human judges. Table 2 shows the correlation among all experts and Table 3 displays the open correlation between an expert and the mean score of all others.

Table 2: Correlations of human scores

Expert ID	1	2	3	4	5
1	\	0.74	0.68	0.75	0.70
2	\	\	0.80	0.77	0.76
3	\	\	\	0.77	0.78
4	\	\	\	\	0.82
5	\	\	\	\	\

Table 3: Open Correlations of human scores

Exp. ID	1	2	3	4	5	Avg.
Corr.	0.79	0.85	0.84	0.86	0.84	0.84

From the data shown in Table 2 and Table 3, it appears that the experts manage to evaluate the semantic correctness and pronunciation quality with a certain degree of reliability. However, the agreement is not very high, probably due to the examinees’ diversities in expression and the experts’ subjectivities in judgment. Since the degree of agreement for human scores does reach a high level in common language tests, in the present task the evaluation is more difficult for the machines.

#### 5.4. Performance of the semantic feature

To extract the semantic feature, five experts are involved in the transcribing of speeches. Examinees’ speeches are transcribed into sentences, including ill pronunciations and murmurs. The transcriptions are checked among experts to validate their correctness. On the other hand, the human scores of the training set are used to calculate the mapping coefficients.

The performance of the semantic feature is shown in Table 4 and Table 5. Table 4 shows the correlations between the semantic feature and each expert. Table 5 shows the correlation between the semantic feature and the mean score of all experts, and the performance lost compared with the human raters.

Table 4: *Correlations between semantic feature and human raters*

	<i>Exp. ID</i>				
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>
<i>S. F.</i>	0.69	0.70	0.70	0.72	0.75

Table 5: *Performance of Semantic feature*

	<i>Corr. with Hum. Avg.</i>	<i>Perf. Lost</i>
<i>S. F.</i>	0.79	5.95%

As is depicted in Table 4, most of the correlations are a bit lower than those among humans, probably due to the ignorance of pronunciation quality in feature extraction. However, as is shown in Table 5, only about 6% performance lost is observed without the consideration of pronunciation quality. Since the pronunciation variation is diverse in the testing corpus, the lost in correlation is acceptable. Thus, as a single feature, it is shown to be effective enough to provide a judgment on the semantic correctness in the evaluation.

## 6. Conclusion

Automatically evaluate the semantic correctness in language test is a new challenge to researchers. In this paper, a novel method is presented to extract a semantic feature in the reading-answering-problem. Experiment shows the effectiveness of the feature as an indication of semantic correctness. Although the experiment is performed on Mandarin data, the method is language independent. Once the background resources, such as knowledge database, were changed for another target language, the feature could be applied directly.

Future work will concentrate on the optimization of the semantic extraction process. Furthermore, the widely used features, such as goodness of pronunciation [1], and other mapping methods, such as neural network [2], are also considered to be integrated into the reading-answering-problem evaluation.

## 7. References

- [1] Silke Witt, "Use of Speech Recognition in Computer-Assisted Language Learning", *PhD thesis, Cambridge University Engineering Department*, Cambridge, UK, 1999.
- [2] Horacio Franco, Leonardo Neumeyer, Vassilios Digalakis, and Orith Ronen, "Combination of Machine Scores for Automatic Grading of Pronunciation Quality", *Speech Communication*, 2000, pp. 121-130.
- [3] Yasushi Tsubota, Tatsuya Kawahara, Masatake Dantsuji, "Practical Use of English Pronunciation System for Japanese Students in the CALL Classroom", *Proc. ICSLP*, (2004), pp. 849-852.
- [4] Chin-Hui Lee, Haizhou Li, Li-Shan Lee, Ren-Hua Wang & Qiang Huo, *Advance in Chinese Spoken Language Processing*, Publisher: World Scientific Publishing Co. 2006, pp. 407-429.
- [5] Dekai Wu, "Alignment Parallel English-Chinese text Statistically with Lexical Criteria", *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, 1994, pp. 80-87.
- [6] Qun Liu, Sujian Li, "Word Similarity Computing Based on How-net", *Computational Linguistics and Chinese Language Processing, China, Taiwan*, 2002, pp. 59-76.
- [7] Paolo Rosso, Francesco Masulli, Davide Buscaldi, Ferran Pla, Antonio Molina, "Automatic Noun Sense Disambiguation", *Computational Linguistics and Intelligent Text Processing*, Publisher: Springer Berlin/Heidelberg, vol. 2588/2003, 2003, pp. 323-343
- [8] Zhendong Dong, Qiang Dong, "HowNet - a hybrid language and knowledge resource", *Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering*, 26-29 Oct. 2003, pp. 820-824.
- [9] CNLP (Chinese Natural Language Processing). <http://www.nlp.org.cn>