A Chinese-English Interpretation Testing Autoscoring System Based on Semantic Analysis

Xinguang Li[1,2, a] , Jiyou Xu[2,b *] , Shengbin Zhang[2,c] and Zhiming He[2,d]

[1]Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China

[2]Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China

[a]lxggu@163.com, [b]jiyou.xu@foxmail.com, [c]578318529@qq.com, [d] 277573564@qq.com

Keywords: Speech Recognition, Semantic Analysis, Speech Autoscoring, Wordnet

Abstract. This paper aims to find out an approach to autoscoring the Chinese-English interpretation testing speech, which is one of the important research areas, meets the demand of computer-assistant English testing and adds value to machine intelligence. We proposed a method to evaluate the semantic similarity based on Wordnet combining words and sentences, which was applied to recognizing the keywords and its synonym in testing speech via HMM model and calculating the score of semantic. In addition, speech fluency was also considered in our developed system and has a contribution to the final score. This paper collected 300 testing speech data from a course final examination in Guangdong University of Foreign Studies to examine the autoscoring system. Result has shown well performance of the average rate of speech recognition and agreement of system score and human score.

Introduction

With the quantity of English learners growing continues in China, all kinds of English speech testing exist in China like the final exam in campus, college entrance examination, College English Testing (CET) and so on. Faced to the large-scale examination, it is a great challenge to score all the testing speech by human, who cannot well control the consistency, accuracy, objectivity and reliability. The situations above promote the research of autoscoring by computer.

Most of the existed English speech autoscoring systems focus on the pronunciation mainly, which separated from semantic analysis was. Pronouncing accuracy is the top priority in evaluating the speech pronunciation. Stanford Research Institute (SRI) of Unite State first initiated the frame wrapping log posterior probability based on Hidden Markov Model (HMM) as core algorithm to measure the pronouncing accuracy [1], followed by pursuer and employed in different language. Zhang[2] studied the feature of pronunciation in Chinese English learner and developed phone-based automatic score for L2 speech quality. Yan[3,4] introduced the linguistic feature to the algorithm and designed an autoscoring system to evaluate the speech, applied to the type of following reading in Chinese English speech testing. Furthermore, some studies have established the evaluating method of rhythm[5], stress[6] and intonation[7], which are three aspects of prosody, along with the fluency.

Chinese-English interpretation is one of the most common types set in English testing, refers to translating the Chinese text into English in oral within the limited time. The testing speech should be evaluated not only the speech fluency, but also the content given by examinee. The Chinese source text used in the testing was either single sentence or paragraph in short text. Thus, words played an important role in scoring. Human raters would received the suggest sentence or words and were asked to well understand the testing speech to check that the spoken words had matched the suggested word or its synonym.

The main task in this paper is as following. First is to study the method of searching for synonym of given keywords and estimate their likeness. Secondly, recognize the keywords in the speech. Thirdly, design an algorithm to calculating the semantic similarity of sentences compared. At last, score the speech weighted by semantic and speech fluency.

Method

Wordnet WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets[8] are interlinked by means of conceptual-semantic and lexical relations. At present, the WordNet has become an international standard, which rationality of its framework has been accepted and used widely in the research field of English lexical semantics and computational lexicography[9]. The method to evaluate lexical semantic similarity proposed in this paper makes use of the lexical taxonomy and data in WordNet.

Semantic Similarity between Words Semantic relevance refers to the relationship between concept (term or word) and can be classified into three kinds of pattern like hyponymy, antonymy and functional relationship. Semantic similarity is one of the practical cases which belong to hyponymy. Figure 1 shows an example from lexical taxonomy in Wordnet. “nickel” and “dime” are two different concepts in the tree, whose lowest common subsumer is “coin”. “Medium of exchange” is the lowest common subsumer of “nickel” and “credit card”. As is well known that “dime” is more similar to “nickel” than “credit card” and “Medium of exchange” is situated in the upper in contrast to “coin”. Thus it reveals a law in Wordnet that the lowest common subsumer of two concepts in the lexical taxonomy determines the similarity of two concepts. On the basis of Wordnet, the studies to calculate the concepts semantic similarity is to find the feature to indicate its lowest common subsumer and turn it into quantified. Most of the existed algorithm is belong to two types, “Network Distance Models” and “Information Theoretic Models”.

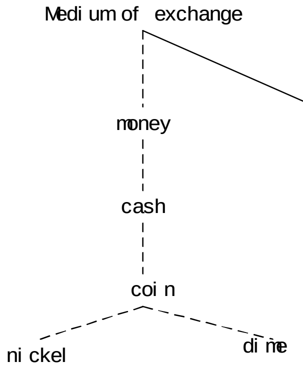

Figure 1. Part of the lexical taxonomy in Wordnet

Network Distance Models It is a method which combines the path distance linking two concepts and depth of its common subsumer both in the lexical taxonomy of Wordnet. The output of the method is associated with semantic similarity. The smaller the path distance value, the more similar two concepts. Two representative algorithms[10,11] were promoted as shown in Eq.1 and Eq.2 respectively.

$$\mathrm{Sim}_{\mathrm{WP}}(C_1,C_2)=\frac{2\times \mathrm{depth}(\mathrm{lso}(C_1,C_2))}{\mathrm{len}(C_1,\mathrm{lso}(C_1,C_2))+\mathrm{len}(C_2,\mathrm{lso}(C_1,C_2)+2\times \mathrm{depth}(\mathrm{lso}(C_1,C_2)))} \qquad (1)$$

$$\mathrm{Sim}_{\mathrm{LC}}(C_1,C_2)=\text{-}\log\frac{\mathrm{len}(C_1,C_2)}{2\times \max\limits_{c\in Wordnet}\mathrm{depth}(c)} \qquad (2)$$

lso(C_1, C_2) represents the lowest common subsumer of C_1 and C_1. len(C_1, C_2) is the shortest path distance from C_1 and C_2. Function depth(c) is used to measure the depth of concept c in the lexical taxonomy of Wordnet.

Information Theoretic Models Resnik(1995) was the first to put forward information of concepts as feature for basis of evaluating, which related the semantic similarity with the shared information of the two concepts. It is well known that the lowest common subsumer in the lexical taxonomy determines the similarity of two concepts. If p(c) represents the probability of concept c in the Wordnet, the Information Content (IC) was defined as $-\lg \mathrm{p}(c)$.The defining equation of Resnik Model is given below.

$$\mathrm{Sim}_{\mathrm{RE}}(C_1,C_2) = -\lg \mathrm{p}(lso(C_1,C_2)) \tag{3}$$

The semantic similarity of concept C_1 and concept C_2 has been transformed to the IC of common subsumer. That means the calculated value $-\lg \mathrm{p}(\text{"coin"})$ can measure the similarity of “dime” and “nickel”. Jiang-Conrath[13] developed the algorithm like the Eq. 4 shows.

$$\mathrm{Sim}_{\mathrm{JC}}(C_1,C_2) = 2\lg p(lso(C_1,C_2)) - (\lg p(C_1) + \lg p(C_2)) \tag{4}$$

Algorithm of Lin In this paper we use the algorithm proposed by Lin[14], which is also belong to the type of Information Theoretic Models. The algorithm of Lin considers the information content of lowest common subsumer and the two compared concepts as shown below.

$$\mathrm{Sim}_{\mathrm{LC}}(C_1,C_2) = \frac{2 \times \lg \mathrm{p}(\mathrm{lso}(C_1,C_2))}{\lg \mathrm{p}(C_1) + \lg \mathrm{p}(C_2)} \tag{5}$$

Semantic Similarity between Sentences The method to evaluate the similarity between two compared sentences was proposed in the paper, which is based on the similarity between the words appeared in the sentences. In addition, the part of speech of each word was also taken into the consideration and added into the final algorithm as calculable weighted value. Adjusted cosine similarity algorithm was applied to evaluate the semantic similarity after both the two sentences have been transfer into vector space model. The processing of the method was presented as following.

Input: Sentence $T_1(t_{11},t_{12},...,t_{1n})$ and sentence $T_2(t_{21},t_{22},...,t_{2m})$.

Output: The quantified ***value*** to indicate the similarity between T_1 and T_2.

{

Step 1: construct the vector space model $W(w_1,w_2,...,w_k)=T_1 \cup T_2$, the each element w_i is the word appeared in the T_1 or T_2

Step 2: mark the part of speech of every word in the W, and set the weight value vector $(l_1,l_2,...,l_k)$[15].

Step 3: construct the semantic vector space model of T_1 and T_2: $X(x_1,x_2,...,x_k)$ and $Y(y_1,y_2,...,y_k)$. The pseudocode of algorithm to construct X is shown as following. Similarly the Y was. The threshold V used in the algorithm was adjusted by data experiment and set in the system.

```
for (i; i<=k; i++)
{
    for (j; j<=n; j++){ $a_j = \mathrm{Sim}(w_i,t_{1j})$ ;}
    if ( $\mathrm{Max}(A(a_1,a_2,...,a_n)) \geq V$ )
        then $x_i = \mathrm{Max}(A(a_1,a_2,...,a_n))$;
        else $x_i = 0$;
}
output the $X(x_1,x_2,...,x_k)$
```

Step 4: calculate the similarity between X and Y which was equated with the similarity between T_1 and T_2 in Eq. 6 where $l(l_1,l_2,...,l_k)$ was needed.

$$\mathrm{Sim}(T_1,T_2) = \mathrm{Sim}(X,Y) = \frac{\sum_1^{n+m} l_i x_i y_i}{\sqrt{\sum_1^{n+m} l_i x_i^2} \times \sqrt{\sum_1^{n+m} l_i y_i^2}} \tag{6}$$

}

The Design of a Chinese-English Interpretation Testing Autoscoring System

The system designed in this paper consists of three major modules as shown in Figure 2. The first is to search the synonym corresponding to the suggested keyword and construct keywords group. The second is to recognize the keywords in the testing speech. The last is to calculate the total score weighted by sentiment and fluency two parameter.

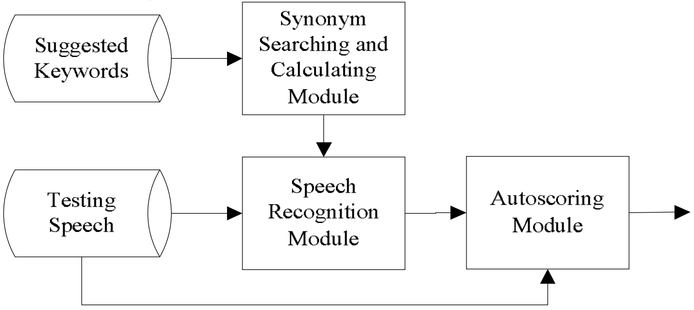
Suggested Keywords
Synonym Searching and Calculating Module
Testing Speech
Speech Recognition Module
Autoscoring Module

Figure 2. Three Module of the Autoscoring System.

Synonym Searching and Calculating Module All the suggested keywords which have been collected before the testing task would be typed into the system one after another to find out its synonym group in Wordnet. Using the Lin algorithm to calculating the similarity (s_{in}) between every single synonym $(C_{i1}, C_{i2}, ..., C_{in})$ and its corresponding suggested keyword (C_{i1}), system got the similarity vector result $(<C_{i1}, s_{i1}>, <C_{i2}, s_{i2}>, ..., <C_{in}, s_{in}>)$. The whole processing has been shown like the Figure 3.

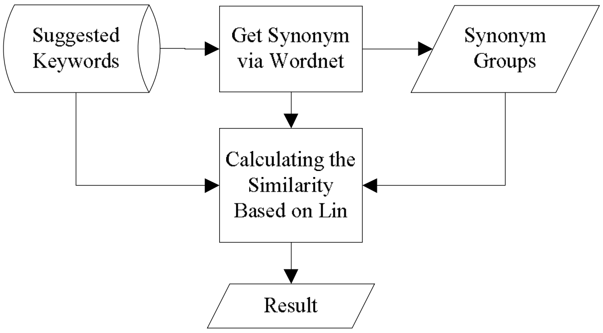

Figure 3. Process of Synonym Searching and Calculating.

Speech Recognition Base on HMM All the suggested keywords and its synonym named keywords group obtained from the previous module were recorded to train the HMM. The System took the testing speech signal into the pre-processing of pre-emphasis, framing and Hamming window orderly, and extracted the MFCC features from every single word, which was segmented by a double-threshold algorithm. Each word in the testing speech will be mapped to the single keyword with a similarity output probability by the recognition algorithm via HMM. The words will only be outputted as the final recognition result T_1 if its similarity output probability value was greater than a set threshold.

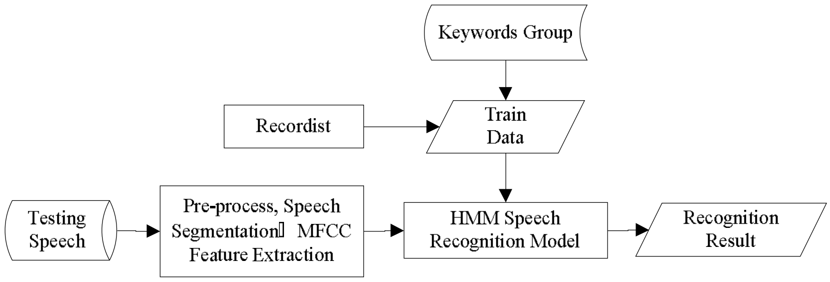
Keywords Group
Recordist
Train Data
Testing Speech
Pre-process, Speech Segmentation MFCC Feature Extraction
HMM Speech Recognition Model
Recognition Result

Figure 4. Process of Speech Recognition

Autoscoring Firstly, calculating the similarity between the result of speech recognition (T_1) and suggested keywords (T_2), $value_1 = \text{sim}(T_1, T_2)$ can be obtained step by step as shown in Chapter 2.3. Secondly, the testing speech was segmented into silence part and non-silence part by a double-threshold algorithm, which is based on short-time average energy and short-time average zero-crossing rate. The better the speech fluency the non-silence part is longer. Therefore *value₂* was calculated as shown in Eq.1, where *len₁* and *len₂* represents the duration of silence part and non-silence part separately.

$$value_2 = \frac{len_2}{len_1 + len_2} \qquad (7)$$

The weighted score of testing speech was calculated as shown in Eq.8. e is an adjusted parameter that trained by data experiment.

$$score = value_1 \times 1.5 + value_2 \times 0.5 + e \qquad (8)$$

Result and Discussion

This paper selected 300 English sentences speech, collected from one of a course final examination in Guangdong University of Foreign Studies, which were marked like 2, 1.5, 1.0, 0.5 and 0 for its full mark was 2. The suggested sentence was: “A provisional Method for the trial Establishment of sino-Foreign Joint-Venture Travel Agencies was introduced with the approval of the State Council.” The suggested keywords have been given following: {provisional, establishment, Sino-Foreign, Joint-Venture, travel Agencies, introduce, approval, the State Council}. We search for the synonym in Wordnet via our developed system and gained the group like that: {short-time, transitory, temporary, institution, association, affiliation, agency, present, agreement, probationary, provisionary, tentative, constitution, formation, organization, beginning, start, commencement, present, acquaint, inform, blessing, approving, support}. We invited 30 people to record all the words in 2 groups which trained the speech recognition model based on HMM. All the testing speech had been translated from speech to text before the experiment in order to calculate the recognition rate. The average recognition rates of these 32 keywords are 95%, where the lowest was 82.6% and highest was 100%. The scoring experiment data in Table 1 indicates the difference between S_h (score of human) and S_s (score of system). $|S_h - S_s| = 0$ meant that the human score and system score are the same.

$\|S_h - S_s\|$			
=0	=0.5	=1	=1.5
70.42%	26.25%	3.33%	0%

Table 1 Experiment Result of the Agreement between Human Scoring and Autoscoring

Among the 300 experiment data, 70.42% of the score of system were as same as score of human. There was no has greater than 1.5. The result has proved its reliability as a computer-assistant English testing scoring system.

Conclusion

Autoscoring the speech by computer is a new research direction, the system will be more widely applied and make more practical value. This paper designed a method to recognize the synonym and measure the semantic similarity between compared sentences based on Wordnet. In addition, the system which can be employed to autoscoring the Chinese-English interpretation testing speech also was developed and introduced in this paper. It has been proved from the well performed average rate of keywords recognition and agreement that system can help human rater in the task of scoring. On the other size, due to limited corpus, the score results are influenced by subjective opinions of the experts to a certain extent. In order to reduce the personal and subjective opinion, the accumulation of corpus will be increased. In the future, more experiment data will be made use of to calculate a more objective and reasonable target weight.

Acknowledgments

This research was financially supported by a grant from Guangdong Natural Science Foundation (2014A030313576), a Guangzhou Science Technology Project (2014J4100025), an MOE (Ministry of Education in China) Project of Humanities and Social Sciences (13YJAZH046) and a Guangdong University of Foreign Studies postgraduate innovation research project (15GWCXXM-46).

Reference

[1] L. Neumeyer, H. Franco. Automatic scoring of pronunciation quality [J] .Speech Communication 30, 2000: 83- 93.

[2] Zhang J, Pan P, Yan Y. Automatic Scoring on English Passage Reading Quality[J]. Procedia Engineering, 2012, 29: 2744-2748.

[3] Yan K, Wei S, Dai L, et al. Pronunciation evaluation based on a phoneme-dependent posterior probability transformation[J]. Journal of Tsinghua University Science and Technology, 2011, 51(9).

[4] YAN K, WEI S, DAI L. Acoustic Model Refining Algorithm for Pronunciation Quality Evaluation[J]. Journal of Chinese Information Processing, 2013, 1: 013.

[5] Zhang J, Zhang M. Speech Material Recognition Technology on an Objective Evaluation System for the Rhythm of English Sentences[J]. Advances in Computer Science, Intelligent Systems and Environment: Vol. 1, 2011, 104: 501.

[6] Xinguang L, Guizhen W, Sizhe Y, et al. Research on objective evaluation system of English sentences based on stressed syllables and prosody[J]. Computer Engineering and Application, 2013, 49(8):105-104.

[7] Li X G, Li S M, Jiang L R, et al. Study of English Pronunciation Quality Evaluation System with Tone and Emotion Analysis Capabilities[J]. Applied Mechanics and Materials, 2014, 475: 318-323.

[8] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[9] Lingling M. Research on Semantic Similarity Metric Based on WordNet and Its Application in Query Suggestion[D]. East China Normal Univesity, 2014.

[10] Wu Zhibiao,Palmer M.Verb.semantics and lexical selection[C]. Proccedings of 32nd Annual Meeting of the Association for Computational Linguistics,Las Cruces,NM,June 1994:133-138.

[11] Claudia L,Chodorow M.Combining localcontext and WordNet similarity for word sense identification[M].Fellbaum C.Word-Net:An Electronic Lexical Database.Cambridge,MA:The MIT Press,1998:265-283.

[12] Resnik P. Using information content to evaluate semantic similarity in a taxonomy[J]. arXiv preprint cmp-lg/9511007, 1995.

[13] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy[J]. arXiv preprint cmp-lg/9709008, 1997.

[14] Lin D. An Information2Theoretic Definition of Similarity Semantic distance in WordNet[C]. Proceedings of the Fifteenth International Conference on Machine Learning .1998.

[15] Qiang L V, Deng W. Semantic similarity computation between sentences[J]. Computer Engineering & Applications, 2010.