A FRAMEWORK OF PERFORMANCE EVALUATION AND ERROR ANALYSIS METHODOLOGY FOR SPEECH UNDERSTANDING SYSTEMS

Bor-shen Lin, Lin-shan Lee

Department of Electrical Engineering, National Taiwan University

Taipei, Taiwan, Republic of China

e-mail: bsl@speech.ee.ntu.edu.tw

ABSTRACT

With improved speech understanding technology, many successful working systems have been developed. However, the high degree of complexity and wide variety of design methodology make the performance evaluation and error analysis for such systems very difficult. The different metrics for individual modules such as the word accuracy, spotting rate, language model coverage and slot accuracy are very often helpful, but it is always difficult to select or tune each of the individual modules or determine which module contributed to how much percentage of understanding errors based on such metrics.

In this paper, a new framework for performance evaluation and error analysis for speech understanding systems is proposed based on the comparison with the 'best-matched' references obtained from the word graphs with the target words and tags given. In this framework, all test utterances can be classified based on the error types, and various understanding metrics can be obtained accordingly. Error analysis approaches based on an error plane are then proposed, with which the sources for understanding errors (e.g. poor acoustic recognition, poor language model, search error, etc.) can be identified for each utterance. Such a framework will be very helpful for design and analysis of speech understanding systems.

1. INTRODUCTION

With improved speech understanding technology in recent years, many spoken dialogue systems have been successfully developed [1]. In general, almost each system is associated with a performance evaluation and error analysis scheme. For example, the speech recognizers and language understanding modules were very often evaluated by different metrics [2], such as the spotting rate, the word accuracy, the language model coverage, the slot accuracy, etc., but it's always difficult to select among the many available acoustic/linguistic processing modules to achieve the best understanding results, and determine which module contributed to how much percentage of which type of understanding errors based on such metrics. This is apparently due to the high degree of complexity of such speech understanding systems and the wide variety of design methodology and system architecture. Some of the difficulties also come from the wide variety of application tasks of such systems. Take the example below. Three acoustic recognition modules are considered for the acoustic front end of a speech understanding system: module A with word accuracy 92%, module B with keyword spotting rate 90% and false alarm rate 7%, and module C with key phrase spotting rate 95% and false alarm rate 24%. All these metrics are helpful, but none of them can describe how the acoustic modules A, B or C can perform within a very complicated speech understanding mechanism and which one should be selected. For example, all the three modules can generate a word graph for each utterance for understanding purposes, but the characteristics of a word graph involved in the understanding processes include not only the word accuracies or spotting rates, but many other factors such as the discriminating functions of the acoustic scores, and the accuracies of the time spans for the word candidates in the graph. The interaction among such characteristics with the following language understanding mechanism is another key. As a result, the performance metrics for each individual module may not be very helpful in determining the final understanding performance. Also, the analysis of the understanding errors and the improvement of system performance based on error analysis become very difficult as well for the same reason.

In this paper, a new framework for performance evaluation and error analysis for speech understanding systems is proposed based on the comparison with the 'best-matched' references obtained from the word graphs with the target words and tags given. In this framework, all test utterances can be classified based on the error types, and various understanding metrics can be obtained accordingly. Error analysis approaches based on an error plane are then proposed, with which the sources for understanding errors (e.g. poor acoustic recognition, poor language model, search error, etc.) can be identified for each utterance. Such a framework will be very helpful for design and analysis of speech understanding systems.

2. THE PROPOSED FRAMEWORK AND THE BEST-MATCHED REFERENCES

In a speech understanding system as shown on the left-hand side of Figure 1, each input utterance u is usually first recognized by an acoustic front end to produce a set of promising word candidates located on different time spans, or a word graph. Some graph search algorithms such as the A* search is then performed on the graph based on some language models to find the desired hypothesis path[1] on the graph. Such a hypothesis path may contain different types of errors due to deleted, substituted or inserted words, poor acoustic or language modeling scores, and wrong time-alignments in the word graph. This hypothesis path is then transcribed into a hypothesis slot sequence or the understanding output (denoted as $P_h(u)$ in Figure 1) by some semantic processing approaches. The evaluation and error analysis method proposed in this paper is shown on the right –

[1] The word 'path' here may represent a word sequence as in the N-best interface or a tag sequence with associated parsing trees.

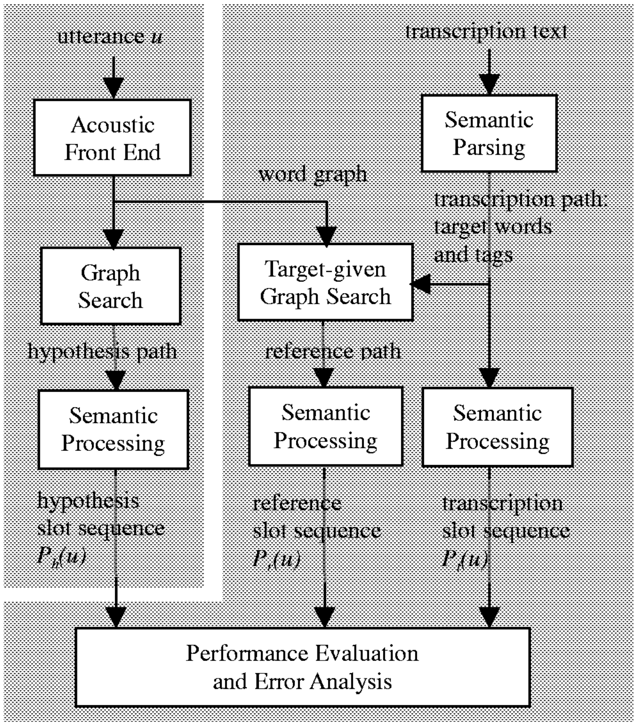
utterance u
transcription text
Acoustic Front End
Semantic Parsing
word graph
transcription path: target words and tags
Graph Search
Target-given Graph Search
hypothesis path
reference path
Semantic Processing
Semantic Processing
Semantic Processing
hypothesis slot sequence $P_h(u)$
reference slot sequence $P_r(u)$
transcription slot sequence $P_t(u)$
Performance Evaluation and Error Analysis

Figure 1. The proposed framework for performance evaluation and error analysis methodology

hand side of Figure 1. The transcribed text for the input utterance is first parsed by some semantic parsing algorithm to produce a transcription path. This path may include all the target words and tags. These target words and tags are then directly applied on the word graph obtained in the speech understanding system on the left-hand side of the figure to perform the target-given graph search, such that all the target words can be obtained as long as they are located in the word graph well. The output, the reference path as shown in the middle of Figure 1, is therefore the 'upper bound' of the hypothesis path because all the target words and tags are given. This reference path together with the transcription path obtained directly from the transcription text then go through the semantic processing to produce the reference slot sequence ($P_r(u)$ in Figure 1) and the transcription slot sequence ($P_t(u)$ in Figure 1). Again, the reference slot sequence $P_r(u)$ is the 'upper bound' of the hypothesis slot sequence $P_h(u)$. All the evaluation and error analysis proposed here in this paper is then based on the comparison among the three slot sequences: the hypothesis slot sequence $P_h(u)$, the reference slot sequence $P_r(u)$, and the transcription slot sequence $P_t(u)$.

In order to analyze how the defected word graphs actually constrained the understanding accuracy, the concept of 'best-matched' references is introduced here which includes the 'upper bound' reference path generated by the target-given graph search shown in the middle of Figure 1, and the 'upper bound' reference slot sequence $P_r(u)$ after the semantic processing. In principle, the 'best-matched' reference path is the 'best obtainable' path in the word graph prior to semantic processing, and the 'best-matched' reference slot sequence $P_r(u)$ gives the 'best obtainable' understanding of the utterance given the word graph, because in both cases the knowledge about the transcription path has been given. The above description can be summarized with the following inequality:

$$A_s(P_h(u), P_t(u)) \ \leq A_s(P_r(u), P_t(u)) \ \equiv \ B_r$$

where $A_s(P_i(u), P_j(u))$ denotes the slot accuracy by comparing the slots of $P_i(u)$ to those of $P_j(u)$, and B_r is the reference bound indicating the degree of understanding achievable from the given word graph or some kind of metric for 'the quality of the word graph' in terms of understanding. Such a metric is sometimes very useful because it provides more information than the acoustic metrics alone. For example, different acoustic front ends can be compared with this metric for understanding purposes. Also, with the proposed approach as shown in Figure 1, more detailed error analysis in terms of understanding performance also becomes possible as will be given below.

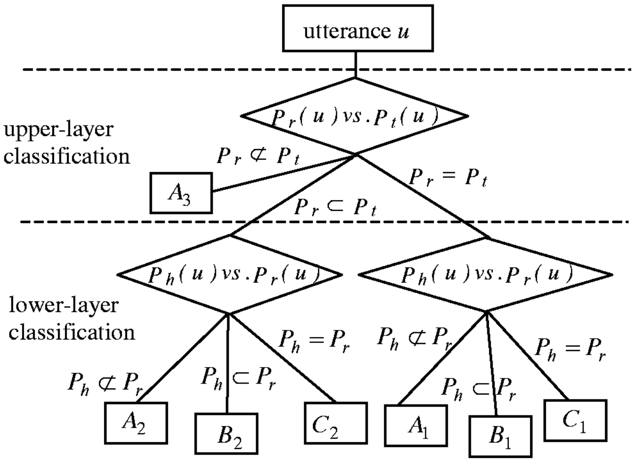

Figure 2. Utterance clustering according to the relationships of the slot sequences $P_h(u)$, $P_r(u)$, and $P_t(u)$

3. UTTERANCE CLASSIFICATION

Based on the relationships among the three slot sequences, $P_h(u)$, $P_r(u)$, and $P_t(u)$ for each utterance, all the utterances can be classified into seven different clusters, as shown in Figure 2. The relationship, '$P_i(u) = P_j(u)$' represents 'completely-matched', i.e., all the slots of $P_i(u)$ and $P_j(u)$ are identical, the relationship '$P_i(u) \subset P_j(u)$' represents 'partially-matched', i.e., all slots of $P_i(u)$ are also slots of $P_j(u)$ but the reverse is not true, and the relationship '$P_i(u) \not\subset P_j(u)$' represents 'mismatched' with insertion or substitution errors, i.e., some slots of $P_i(u)$ are not slots of $P_j(u)$. The relationships 'partially-matched' ($\subset$), and 'mismatched' ($\not\subset$) are separately considered, because the insertion or substitution errors (mismatched) usually lead to misunderstandings which either make the dialog fail or need to be corrected by more and complicated dialogs, while pure deletion errors (partially-matched) usually only lead to incomplete understanding. The classification processes in Figure 2 have two layers. For each utterance u the reference slot sequence $P_r(u)$ and transcription slot sequence $P_t(u)$ are first compared in the upper layer. If $P_r(u) \not\subset P_t(u)$, 'the quality of the word graph' is really poor. If $P_r(u) \subset P_t(u)$ or $P_r(u) = P_t(u)$, $P_h(u)$ and $P_r(u)$ are then compared in the lower layer. The seven utterance clusters obtained in Figure 2 is thus:

Region 1: between $S_a+S_l=0$ and $S_l=0$
poor acoustic score

Region 2: between $S_a=0$ and $S_l=0$
poor acoustic and language-modeling scores

Region 3: between $S_a+S_l=0$ and $S_a=0$
poor language-modeling score

Region 4: $S_a+S_l>0$
search errors

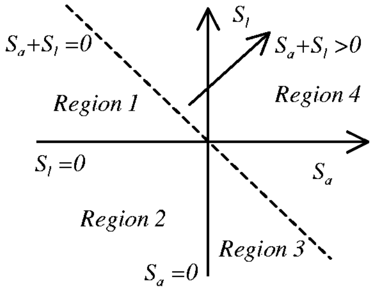

Figure 3. Error analysis based on the error plane

$$A_1=\{u: P_r(u) = P_t(u) \text{ and } P_h(u) \not\subset P_r(u)\}$$
$$B_1=\{u: P_r(u) = P_t(u) \text{ and } P_h(u) \subset P_r(u)\}$$
$$C_1=\{u: P_r(u) = P_t(u) \text{ and } P_h(u) = P_r(u)\}$$
$$A_2=\{u: P_r(u) \subset P_t(u) \text{ and } P_h(u) \not\subset P_r(u)\}$$
$$B_2=\{u: P_r(u) \subset P_t(u) \text{ and } P_h(u) \subset P_r(u)\}$$
$$C_2=\{u: P_r(u) \subset P_t(u) \text{ and } P_h(u) = P_r(u)\}$$
$$A_3=\{u: P_r(u) \not\subset P_t(u)\}$$

In this way, different types of errors can be well classified. For example, the cluster A_3 consists of all utterances that always lead to misunderstandings due to defected word graphs. The cluster B_1, on the other hand, consists of the utterances for which complete-understanding is in principle possible ($P_r(u) = P_t(u)$), but not achieved ($P_h(u) \subset P_r(u)$) because of poor acoustic and linguistic scores. Based on this clustering scheme, we can further define the following sets:

$S_{comp}= C_1$ *(set of complete-understanding)*
$S_{part}= B_1+ B_2+ C_2$ *(set of partial-understanding)*
$S_{mis}= A_1 + A_2+ A_3$ *(set of misunderstanding)*
$S_{corr}= S_{comp}+ S_{part}$ *(set of correct understanding)*
$S_{err}= S_{mis} +S_{part}= A_1+ A_2+ A_3+B_1+ B_2+ C_2$ *(error set)*

Furthermore, some meaningful understanding metrics can be defined as follows.

$$R_{comp} = \frac{N(S_{comp})}{N(S_{All})} \quad \textit{complete understanding rate}$$

$$R_{part} = \frac{N(S_{part})}{N(S_{All})} \quad \textit{partial understanding rate}$$

$$R_{mis} = \frac{N(S_{mis})}{N(S_{All})} \quad \textit{misunderstanding rate}$$

$$R_{corr} = \frac{N(S_{corr})}{N(S_{All})} \quad \textit{correct understanding rate}$$

where $N(S_j)$ *is the number of utterances in* S_j
S_{All} *is the union of all seven clusters*

The errors occurring in the different clusters in the error set S_{err} can be further analyzed, which will be discussed in the next section.

4. ERROR ANALYSIS

Unlike the error analysis usually performed in large vocabulary speech recognizers [3], the analysis for understanding errors here emphasizes on the reference path for comparison rather than the transcription path. In other words, instead of paying great attention to analyzing the differences between the hypothesis path and the transcription path, here in this approach the differences between the hypothesis path and the 'upper bound' reference path are analyzed with more attention. This is because the function of the graph search here is not to find the transcription path, but instead to achieve best possible understanding out of the given word graph. This is also the way to separate the effect of a defected word graph from other understanding mechanism. First of all, two clusters A_3 and C_2 should be used to analyze the acoustic front end module and/or reestimate the acoustic models, because A_3 is the set with seriously defected word graphs, and the graph search performed on utterances in C_2 are in fact completely correct. Excluding the set C_1 with completely correct processing, further error analysis can be performed on the four clusters of utterances: A_1, A_2, and B_1, B_2. Two very useful parameters S_a and S_l are first defined for each utterance as follows:

$$S_a = S_{a,r} - S_{a,h}, \quad S_l = S_{l,r} - S_{l,h}$$

where $S_{a,h}$ and $S_{l,h}$ are the acoustic recognition and language-modeling scores for the hypothesis path respectively, and $S_{a,r}$ and $S_{l,r}$ are the acoustic recognition and language-modeling scores of the reference path respectively. By normalizing with the scores for the hypothesis path, the scores S_a and S_l represents the differences between the hypothesis path and the 'upper bound' path in terms of the graph search. They have also been normalized in such a way that the values of S_a and S_l are comparable and additive. Using (S_a ,S_l) as the coordinates, each utterance in the set A_1, A_2, B_1, B_2 can be located on an error plane as shown in Figure 3. Different types of errors can be easily identified by the error regions defined on the error plane in Figure 3. Region 3 in Figure 3, for example, is the region enclosed by the lines $S_a+ S_l=0$ and $S_a=0$, is the region in which the reference path has a better acoustic score ($S_a>0$), but loses in the total score ($S_a + S_l<0$) due to poor language-modeling score. Therefore, the utterances located into region 3 can be used to update the language models, etc. In this way, all utterances with understanding errors can be properly analyzed, and the real source causing each of the errors can be easily identified. It should be pointed out that the language models should be updated by not only the transcription texts but the reference paths, so as to improve the grammar coverage.

	hypothesis path		reference path
	without n-gram	with n-gram	
A_s	68.34%	71.28%	74.67%
R_{mis}	37.60%	24.30%	19.44%
R_{part}	7.42%	12.53%	13.55%
R_{comp}	54.99%	63.17%	67.01%
R_{corr}	62.40%	75.70%	80.56%

Table 1. Understanding rates and slot accuracy for the example experiment

		A_1	A_2	A_3	B_1	B_2	C_1	C_2	Total
Understanding type		mis	mis	mis	partial	partial	complete	partial	
hypothesis paths without n-gram	# of sentences	41	30	76	7	0	215	22	391
	# of slots	64	84	134	14	0	318	65	679
	# of slot errors	45	37	92	7	0	0	34	215
	error rate contributions	6.63%	5.45%	13.55%	1.03%	0.00%	0.00%	5.01%	31.66%
hypothesis paths with n-gram	# of sentences	15	4	76	1	0	247	48	391
	# of slots	21	10	134	2	0	373	139	679
	# of slot errors	18	10	99	1	0	0	67	195
	error rate contributions	2.65%	1.47%	14.58%	0.15%	0.00%	0.00%	9.87%	28.72%

Table 2. Distributions of hypothesis paths for their respective clusters

	A_3	C_2	$A_1+A_2+B_1+B_2$				Total
			Region 1	Region 2	Region 3	Region 4	
error sources.	defected word graph	defected word graph	poor acoustic score	poor acoustic and LM scores	poor LM score	search error	
# of sentence	76	48	7	13	0	0	144
% of sentence	52.78%	33.33%	4.86%	9.03%	0.00%	0.00%	100.00%
# of errors	99	67	10	19	0	0	195
% of errors	50.77%	34.36%	5.13%	9.74%	0.00%	0.00%	100.00%

Table 3. Error analysis for the data with n-gram language models

5. AN EXAMPLE

The above evaluation and error analysis scheme was applied to a train ticket reservation system, which provides the user with a spoken dialogue interface in Mandarin Chinese such that the train schedule information retrieval and ticket reservation can be easily performed with voice. The acoustic front end is a key phrase spotter [4] which generates key phrase graphs, while the language understanding module includes a hierarchical tag-graph search [5], in which the n-gram language models are used and the tag sequence with associated parsing trees is generated for semantic processing. 391 utterances in 45 real dialogs generated by four male and four female speakers were used for the test in the development phase of the system. The various understanding rates as defined in section 3 together with the slot accuracy for the hypothesis paths (with and without language models) and the reference paths are shown in Table 1. As shown in the table, the 74.67% reference path slot accuracy constrains the overall sentence understanding rate to 80.56%. The improvement of slot accuracy for the hypothesis paths by n-gram language models, from 68.34% to 71.28%, is not significant because the defected word graphs have seriously limited the functions of the language models, although the slot accuracy of 71.28% is not too far from the ‘upper bound’ of 74.67%. The detailed distributions of all utterances in different clusters and corresponding test data are shown in Table 2. The n-gram language models, as shown in the table, have reduced the numbers of errors in the clusters A_1, A_2 and B_1, while increased that in A_3. This is because the utterances in cluster A3 with seriously destroyed word graphs get no benefits from the n-gram constraints. Furthermore, the reduction of error sentences in the clusters A_1, A_2 and B_1 leads to the increase of correct and partial understanding sentences in the clusters C_1 and C_2. Finally, the sources causing the understanding errors with n-gram language models applied are analyzed in Table 3. It can be found in Table 3 that in this case 50.77% of the understanding errors come from defected word graphs (A_3), which lead to fatal misunderstanding errors, while 34.36% of the understanding errors come from imperfect word graphs (C_2), which are not very serious because the ‘upper bound’ reference paths with partial understanding are achieved. The other 14.87% of the understanding errors are due to poor acoustic and/or language modeling scores. Though no search errors are observed in this case, the approach proposed here is able to handle them if any search errors are identified.

6. CONCLUDING REMARKS

It’s really difficult to evaluate a speech understanding system and analyze the different types of errors precisely. The traditional metrics developed for speech recognition are helpful but not necessarily able to provide a direct insight into the understanding mechanism, while the slot accuracy can’t indicate how and why the understanding errors occurred. In this paper, a best-matched path is derived from a target-given graph search, and a framework for performance evaluation and error analysis is proposed accordingly. This framework has the potential to become powerful tool for the design, analysis and evaluation of speech understanding systems.

7. REFERENCES

[1] Victor Zue, etc., “From Interface to Content: Translingual Access and Delivery of On-line Information”, *Proc. Eurospeech*, 2227-2230, 1997.

[2] Harald Aust, Hermann Ney, “Evaluating Dialogue Systems Used in Real World”, *Proc. ICASSP*, Vol. 2, 1053-1056, 1998.

[3] Lin Chase, “Blame Assignment For Errors Made By Large Vocabulary Speech Recognizers”, *Proc. Eurospeech*, 815-818, 1997.

[4] Berlin Chen, etc., “A*-Admissible Key-Phrase Spotting with Sub-Syllable Level Utterance Verification”, *to appear in Proc. ICSLP*, 1998.

[5] Bor-shen Lin, etc., “Hierarchical Tag-Graph Search For Spontaneous Speech Understanding in Spoken Dialog Systems”, *to appear in Proc. ICSLP*, 1998.