Asian Journal of Information Technology 6 (5): 647-651, 2007
© Medwell Journals, 2007

Multiple Hypotheses in Confidence Scoring Measures for Spoken Dialog System

[1]V. Meenakshi and [2]V. Shanthi
[1]SRM Arts and Science College, Chennai, India
[2]St. Joseph's Engineering College, Chennai, India

Abstract: This study provides improved confidence assessment for detection of word-level speech recognition errors, out of domain utterances and incorrect concepts. New features from the speech understanding component are proposed for confidence annotation at utterance and concept levels. In addition, the confidence measures are used to improve the word recognition accuracy. Several hypotheses from different speech recognizers are compiled into a word-graph. The word-graph is searched for the hypothesis with the best confidence.

Key words: Confidence scoring, classification error rate, LM feature, decoder feature

INTRODUCTION

Today's speech recognition research conducts research leading to the development of conversational systems for human-machine interaction. These systems must not only recognize the words which are spoken by a user but also understand the user's query and respond accordingly. To achieve this goal, accurate automatic speech recognition is a necessity. The presence of incorrectly recognized words may cause the system to misunderstand a user's request, possibly resulting in the execution of an undesirable action. Unfortunately today's speech recognition technology is far from perfect and errors in recognition must be expected. Speech recognition technology h as advanced to the stage where real-world applications are feasible. However, due to the current imperfect nature of speech recognition, confidence scoring has emerged as an important component of current systems. Confidence scoring attempts to assign trust to the hypotheses produced by speech recognition systems. In a spoken dialogue system we can define 3 different levels for confidence measures:

Word level: In this level, confidence measures provide an idea about the accuracy of each recognized word. For this level we use decoder and Language Model (LM) features.

Utterance level: Here, the target is the detection of out of domain utterances. In this level we use acoustic, LM and parsing features.

Concept level: The end-to-end system performance does not change in cases where the phrases in a sentence that belong to concepts are correctly recognized while the filler words or phrases are not correctly recognized. In this level, we focus on parts of phrases that are meaningful to the task. Decoder, LM and parsing features will be used to tag the concepts with the confidence measures. The system combines continuous speech recognition, natural language understanding and flexible dialogue control to enable natural conversational interaction by telephone callers to access information about airline flights, hotels and rental cars.

MATERIALS AND METHODS

The data used for the experiments has been obtained during the telephone data collection. Over 500 calls were collected during this period totaling approximately 9,500 utterances. We have randomly split the data into three sets, 60% of the data for training, 20% for evaluation and 20% for testing. We have repeated it 6 times providing 6-Round Robin data sets to verify the results. The results presented are the average of these experiments.

Word level: For word level confidence we investigated a subset of the most promising features which were considered in (Chase, 1997; Kamppari and Hazen, 2000). We consider decoder and LM features.

Decoder features

Normalized score: Acoustic score of the word divided by the number of frames that it spans.

Count in the Nbest: Percentage of times the word appearsin the 100-best hypotheses in similar position.

Lattice density: Number of alternative paths to the word considered in the word-graph generated in the recognizer.

Corresponding Author: V. Meenakshi, SRM Arts and Science College, Chennai, India

Phone perplexity: Average number of phones searched along the frames where the recognized word has been active in the decoding process.

Language model features

Language model back-off behavior: Back-off behavior of an N-gram language model along a 5 word context.

Language model score: The log-probability for each word in a sequence as computed from a back-off language model.

Feature combination: In this study, the features were quantized with 115 binary inputs. 10 bits per feature were used except for the 5-context LM back-offs where it was necessary to use only 5 bits to code all possible situations. We have coded the features considering more resolution in ranks with more training data. The hidden layer consisted of 30 units and one output node was used to model the word-level confidence. During weight estimation, a target value of 1 is assigned when the decoder correctly recognizes the word and a value of 0, is assigned during incorrect recognition (e.g., substitutions and insertions).

RESULTS AND DISCUSSION

Table 1 summarizes the correct detection rates for word-level recognition errors at false rejection rates of 2.5 and 5.0%. In this Table 1 we also present the minimum classification error and the baseline error that corresponds with the recognition rate. It can be seen that LM features provide better indicators for word-level confidence than decoder features. For example, using LM features alone, 42.0% of mis-recognized words were detected at a false rejection rate of 5%, similar to San-Segundo *et al.* (2000). Using decoder features we only reject 28.5% of the mis-recognized words.

The best results are obtained by combining all features. In this case we reject more than half of the incorrect words at 5% false rejection. These features reduce the classification error by 6.2%.

Combining multiple hypotheses: In the next sections we present several methods to use confidence measures for combining hypotheses from different decoders to improve speech recognition accuracy.

Table 1: Correct detection of mis-recognized words ata 2.5 and 5.0% False Rejection (FR) rate. Minimum classification error and baseline error (Substitutions and Insertions) are also shown

	Correct detection rates		
	---------------------------------		Classification error
Word level	2.5% FR	5.0% FR	Baseline 19.0%
Decoder features	16.9	28.5	17.5
LM features	28.3	42.0	15.0
All features	39.0	53.2	12.8

Flat List Confidence Rescoring (FLCR): For each hypothesis output from the bank of decoders, we calculate the average confidence along the whole sentence. The hypothesis with the highest average confidence value is selected as the best hypothesis.

Word Graph Confidence Rescoring (WGCR): In this study the key is to build a Word-Graph with all hypotheses and find the path along the graph with the highest average confidence value. This path can produce a new hypothesis, different from those used to build the graph. The idea is to pick up the best path from different sentences.

Word-graph generation: For each decoder hypothesis we tag each word with its confidence value. For example, consider Fig.1 that shows three hypotheses with confidence values shown underneath. In this figure we represent each word as an edge and we put a node between 2 consecutive words.

Next we join nodes from different hypotheses to build the graph. We join the beginning and end nodes of all the hypotheses and initial nodes of all words from different hypotheses situated in similar positions in the phrase. Finally we prune parallel transitions by picking the transition with the maximum confidence. The final word-graph is shown in Fig. 2.

Best path calculation: Using dynamic programming, the best path along the graph can be calculated. In this case our heuristic has been the Accumulated Average Confidence, i.e., the average confidence from the initial node to the current node. In Fig. 2 we can see the sentence finally obtained. Following the bolded edges we can see how each word was obtained from each hypothesis. In this study the resulting phrase matches perfectly with the reference and it is different from each individual hypothesis.

I WANT TO GO FROM CHENNAI TO KOLKATTA							
I	WANT	TO	GO	FROM	CHENNAI	TO	KOLKATTA
0.92	0.93	0.91	0.93	0.96	0.95	0.98	0.95
I	WANNA	GO	FROM		CHENNAI	TO	KOLKATTA
0.92	0.78	0.93	0.96		0.95	0.98	0.95
I	GANNA	GO	FROM		CHENNAI	TO	KOLKATTA
0.92	0.73	0.93	0.96		0.95	0.98	0.95

Fig. 1: Reference utterance and 3 alternatives

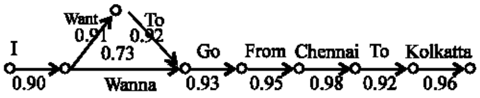

Fig. 2: Word-graph generated for the example in Fig. 2. The bold arrows indicate the best hypothesis

Experiments: We conducted experiments with 4 system configurations. In the baseline system that was evaluated by NIST (Pellom *et al*., 2000) our system initially runs two decoders in parallel. One decoder uses speaker independent models while the other uses female adapted models.

After 500 frames of input (5 sec) the system selects one best decoder to use for the remainder of the telephone dialog. In this research, we have considered running decoders continuously in parallel throughout the dialog.

In the second configuration, we select the hypothesis with the best path score as output from the decoder search. Finally, we consider using the proposed FLCR and WGCR methods. From these results we can affirm that confidence measures provide an important role in reducing the WER by combining hypotheses from different decoders. The WGCR method proposed in this paper produces a reduction of 3.8% in WER (14% relative) and performs better than the FLCR method. In these experiments the difference between FLCR and WGCR is low because the number of decoders run in parallel is small, only two and the average utterance length is also small i.e., 2.4 words per utterance (Table 2).

Utterance level: In this level we use all sources of information: decoder, LM and parsing features: Decoder and LM features:

Average word confidence: This is the average word confidence along the sentence.

Parsing features

Number of words parsed in the sentence: Number of words from the sentence belonging to a concept or a rule used to parse a concept.

Number of words that can be parsed: Number of words from the sentence belonging to a concept or any rule in the task grammar.

Number of concepts: Number of concepts obtained in the sentence.

Average count in the 100-best: Average percentage of times that a concept appears in the 100-best hypotheses.

Percentage of hypotheses in the 100-best with any concept: With this feature we want to represent how many hypotheses parse with at least one extracted concept. For combining all features we have considered an MLP (Multi-Layer Perceptron). In this study, the features were not quantized and they were used as input to the MLP. Because of this, a preprocessing is required to limit the dynamic range of each feature to the (0,1) interval. Here, the normalization consists of scaling utilizing the minimum and maximum value obtained for each feature in the training set.

Table 2: Word error rate results

Word error rates for four system configurations

Method	WER
Baseline	27.2%
Best path score	26.2%
FLCR	24.2%
WGCR	23.4%

Table 3: Correct detection of out of domain utterances at a 2.5% and 5.0% False Rejection (FR) rate. Minimum classification error and baseline error are also shown

	Correct detection rates		
Utterance level	2.5% FR	5.0% FR	Classification error baseline 4.8%
Decoder+LM features	41.1	49.8	4.2
Parsing features	43.8	52.8	4.1
All features	45.7	53.2	4.0

Experiments: Table 3 summarizes the correct detection rates for out of domain utterances at false rejection rates of 2.5 and 5.0%. It can be seen that parsing features provide better indicators for utterance-level confidence than decoder and LM features (San-Segundo *et al*., 2000).

Concept level: We have calculated correct and incorrect concepts considering hypotheses and references, passing both through the parser and comparing them using a dynamic programming algorithm. Similar to the WER it is possible to calculate insertions, deletions and substitutions for concepts. In our system, we have a Concept Error Rate (CER) of 27.9%. The amount of incorrect concepts obtained is 16.5% (substitutions and insertions). In Hazen *et al*. (2000) we can see similar work in this level. In this level we use all sources of information: Decoder, LM and parsing features.

Average word confidence in the words belonging to the rule used to get the concept: In this study we calculate the average word confidence obtained following the definition in along the rule applied to get the concept.

Average word confidence for the value of the concept: Average word confidence for the words that can be considered as the value of the concept. Considering the sentence: I want to go to Chennai to Kalcutta: The phrase go to Chennai contain the words belonging to the rule applied to get the concept: (Arrival City) and Chennai is the value for this concept.

Number of words in the rule
Number of words in the concept value
Concept count in the 100-best: Each hypothesis in the 100-best are parsed and the percentage of times that a concept appears in the hypotheses are counted.

Concepts and value count in the 100-best: in this case we consider the number of times that a concept appears with the same value along the 100-best hypotheses.

These two features are useful when a confusable pair appears in the hypotheses. For example, consider two cities with high confusion between them: Chennai and Mumbai. When we have the sentence: I want to go from Chennai to Kalkutta in the 100- best we observe many times how the word Chennai is substituted by Mumbai. In this study the concept (Departing City) has high confidence because it appears in almost all the hypotheses but its value changes considerably, so this is a good measure of the concept value confidence. Obtaining large differences between these values means that the concept will probably be right but its value is confused. The next 2 features are obtained considering a Concept Language Model. In our study we have trained a 3-gram LM considering the concepts obtained from the references in the training set. Similar to the case for word level, we have considered:

Language model back-off behavior: Back-off behavior of an N-gram language model

Language model score: The log-probability for each concept in a sequence as computed from a back-off language model. For combining all features we have considered a MLP with the same characteristics described in the previous section.

Table 4 summarizes the correct detection rates of incorrect concepts at false rejection rates of 2.5 and 5.0%. We have run three experiments: considering the Average Word Confidence features (AWCs) separated, the rest of the features and all together. It can be seen that using only the AWC features we get slightly improved results than considering the remaining proposed features. For example, 47.1% of the incorrect concepts were detected at a false rejection rate of 5%. Using the rest features we reject 40.1% for the same false rejection. The best results are obtained combining all features. In this study we reject more than 50% of incorrect concepts at 5% false rejection. Considering that we have 16.5% of the incorrect concepts, these features reduce the classification error in 4.5%.

Table 4: Wrong concepts correct detection at a 2.5 and 5.0% False rejection Rate (FR) rate. Minimum classification error and baseline error are also shown

	Correct detection rates		Classification error
Concept level	2.5% FR	5.0% FR	Baseline 16.5%
AWCs	31.1	47.1	12.8
Parsing features	29.3	40.1	13.5
All features	35.9	50.1	12.0

CONCLUSION

In this study we present an analysis for confidence annotation for spoken dialogue systems at different levels: word, utterance and concept levels. At the word level we detect 53.2% of mis-recognized words at a 5% false rejection rate reducing the classification error by 6.2%. From the experimental observations, features coming from the LM perform better than decoder features. The best results are obtained by combining all features together. We propose the use of confidence measures as heuristic to combine several hypotheses from different recognizers. We analyze two options for combining the hypotheses: the Flat List Confidence Rescoring (FLCR) method and the Word-Graph Confidence Rescoring (WGCR) method. The word-graph generation algorithm is described and results with this method are reported. Using the WGCR method we reach a 14% relative word error rate reduction. For the utterance level combining all features, 53.2% of out of domain utterances are detected at 5% false rejection rate. At the concept level a new set of features are proposed detecting more than 50% of incorrect concepts at a 5% false rejection rate. New features from the speech understanding component are proposed for confidence annotation at utterance and concept levels.

Future work: In a future work, we will analyze separately all features proposed from the speech understanding component, in order to obtain each contribution. For the utterance level we will consider the detection of, not only out of domain, but also misunderstanding utterances because of poor recognition. We also will consider the compilation of N-best lists from different coders into a single word graph to get a richer collection of hypotheses. Finally we will use the concept confidence to reduce the semantic errors that have more impact on the system's end-to-end performance.

REFERENCES

Chase, L., 1997. Error-Responsive Feedback Mechanisms for Speech Recognizers, Ph.D. thesis, Carnegie Mellon University, Technology Report CMU-RI-TR.

Chase, L., 1997. Word and acoustic confidence annotation for large vocabulary speech recognition. In: Proc. Eurospeech, Rhodes, Greece, pp: 815-818.

Hazen, T., T. Burianek., J. Polifroni and S. Seneff, 2000. Recognition Confidence Scoring for Speech Understanding Systems, Proc. of the ISCA ITRW ASR2000 Workshop on Automatic Speech Recognition: Challenges for the new Millenium, Paris France, pp: 213-220.

Kamppari, S.O. and T.J. Hazen, 2000. Word and phone level acoustic confidence scoring. Proc. ICASSP, Istanbul. Turkey, pp: III-1799, III-1802.

Pellom, B., W. Ward and S. Pradhan, 2000. The CU Communicator: an architecture for Dialogue Systems. In: Proceeding ICSLP, Beijing, China.

San-Segundo, R., B. Pellom, W. Ward and J.M. Pardo, 2000. Confidence measures for dialogue management in the CU communicator system. Proc. ICASSP, Istanbul. Turkey, pp: III-1237, III-1240.