An Integrative and Discriminative Technique for Spoken Utterance Classification

Sibel Yaman, *Student Member, IEEE*, Li Deng, *Fellow, IEEE*, Dong Yu, *Senior Member, IEEE*, Ye-Yi Wang, *Senior Member, IEEE*, and Alex Acero, *Fellow, IEEE*

***Abstract*—Traditional methods of spoken utterance classification (SUC) adopt two independently trained phases. In the first phase, an automatic speech recognition (ASR) module returns the most likely sentence for the observed acoustic signal. In the second phase, a semantic classifier transforms the resulting sentence into the most likely semantic class. Since the two phases are isolated from each other, such traditional SUC systems are suboptimal. In this paper, we present a novel integrative and discriminative learning technique for SUC to alleviate this problem, and thereby, reduce the semantic classification error rate (CER). Our approach revolves around the effective use of the N-best lists generated by the ASR module to reduce semantic classification errors. The N-best list sentences are first rescored using all the available knowledge sources. Then, the sentence that is most likely to helps reduce the CER are extracted from the N-best lists as well as those sentences that are most likely to increase the CER. These sentences are used to discriminatively train the language and semantic-classifier models to minimize the overall semantic CER. Our experiments resulted in a reduction of CER from its initial value of 4.92% to 4.04% in the standard ATIS task.**

***Index Terms*—Automatic speech recognition (ASR), discriminative training, spoken language understanding (SLU), spoken utterance classification (SUC), statistical language modeling.**

I. Introduction

IN CONTRAST to automatic speech recognition (ASR), which converts a speaker's spoken utterance into a text string, spoken language understanding (SLU) aims at interpreting users' intentions from their speech utterances [1], [2]. As a special form of SLU, spoken utterance classification (SUC) has found many practical applications including call routing [3], dialog systems [4], [5], command and control [6], and speech-to-speech translation [7].

The ultimate objective of an SUC system is to reduce the classification error rate (CER). There are two kinds of observed errors: errors in ASR transcription and errors in utterance classification. Semantic classifiers typically require operation with significant freedom in utterance variations. In spite of such high degree of freedom of expression, the semantic classifiers should be able to interpret, for example, the two phrases "*Show all flights*" and "*Give me flights*" as variants of the same semantic class "Flight."

Traditional SUC techniques adopt a sequential scheme with two independent phases. In the first phase, an ASR module returns the *single*-most likely sentence for the observed acoustic signal. A semantic classifier then transforms the resulting sentence into a semantic class in the second phase. However, this approach is suboptimal because of the strong assumption that the ASR and semantic classification phases are fully independent. Relaxing this assumption by allowing the semantic classifiers to *directly* operate on the N-best sentences helps reduce CER by providing more information to the second phase. In addition, traditional SUC systems are such that CER is reduced in semantic classification phase, yet it is the word error rate (WER) that is reduced in the ASR phase. Ideally, both of the two phases should reduce the CER.

In [8], the authors investigated discriminative language modeling in a similar scenario. Their motivation is that a reduced WER provides with ASR transcriptions that are more likely to be classified correctly. Using the perceptron algorithm, the authors trained joint language and classifier models either independently or simultaneously, under various parameter update conditions. Although, authors also select a one-best sequence to train the system parameters, the discriminative language modeling in SUC achieved reductions in both the ASR WER and the semantic CER.

In [9], the authors describe how the parameters of an n-gram LM are trained to achieve minimum sentence error through improving the separation of the correct sentence from the competing sentences in the N-best lists. Our work described in this paper distinguishes itself in several ways. First of all, the goal of the training in [9] is to reduce WER, whereas the goal of the proposed method is to reduce CER. Second, in the proposed method, there is a second phase where the ASR transcriptions are used to train semantic classifiers. For this reason, the class-discriminant function we use includes information regarding the semantic class, whereas the discriminant function defined in [9] includes information regarding only the ASR phase. Third, in the proposed method, each N-best list sentence is matched with a semantic class. This helps improve CER by distinguishing the sentence that is most likely to yield a correct classification decision.

In this paper, we describe a discriminative training technique to tie these two phases so that each phase can use the output of

Manuscript received February 28, 2007; revised May 8, 2008. Published July 16, 2008 (projected). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Helen Meng.

S. Yaman is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: syaman@ece.gatech.edu).

L. Deng, D. Yu, Y.-Y. Wang, and A. Acero are with Microsoft Research, Microsoft Corporation, Redmond, WA 98052 USA (e-mail: deng@microsoft.com; dongyu@microsoft.com; yeyiwang@microsoft.com; alexac@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2001106

1558-7916/$25.00 © 2008 IEEE

the other phase to help reduce the classification error rate. This is achieved by training the language and the semantic classifier models in an SUC system to minimize the CER directly. We follow the minimum error classification (MCE) framework [10] to model the training objective as functions of system parameters. The N-best lists generated by the ASR phase are first rescored using all the available knowledge sources. The rescored N-best lists are then used to discriminatively learn the system parameters. Our motivation in this paper is to make *explicit* use of the N-best lists generated by the ASR systems in the semantic classification phase to reduce the CER. In similar studies, the N-best lists are used just to extract the sentence that gives the lowest WER. In our paper, the N-best lists are used not only to find the sentence that gives the lowest CER (instead of lowest WER) but also to find out those sentences which are most likely to yield incorrect decisions. The LM and classifier model parameters are trained so that the scores of those sentences that yield correct classification are increased, while the scores of those sentences that yield incorrect classification are reduced.

We have conducted experiments to evaluate the proposed technique on the standard DARPA Air Travel Information System (ATIS) task [11], [12]. Our experimental results have demonstrated significant improvement over the earlier best system reported in the literature on the identical task. We trained a baseline classification phase with the one-best output of the ASR phase that uses an LM trained using the manual transcriptions. The resulting CER was 4.92%. By training all system parameters to reduce CER, our method described in this paper yielded a WER of 4.04%. The reduction shows that the N-best list sentence that yields the lowest CER instead of lowest WER should be used in the classification phase.

This paper is organized as follows: In Section II, the basic building blocks in a traditional SUC system architecture are presented. In Section III, the proposed technique is described in detail. Experimental results are reported in Section IV. Finally, the concluding remarks and directions for future work are discussed in Section V.

II. General Spoken Utterance Classification System Architecture

An SUC system classifies the rth spoken speech utterance X_r, $r = 1, \ldots, R$, into one of M semantic classes, $\hat{C}_r \in \mathcal{C} = \{C_1, \ldots, C_M\}$. $\hat{C}_r$ is chosen so that the class-posterior probability given X_r, $P(C_r|X_r)$, is maximized. Formally

$$\hat{C}_r = \arg\max_{C_r} P(C_r|X_r). \tag{1}$$

As depicted in Fig. 1, standard SUC techniques typically involve a speech recognizer to transform the speech utterance into words, and a text classifier to transform the resulting sentence into a semantic class [13]–[17]. In most of the conventional SUC approaches, the ASR phase is designed so that the WER is reduced. Reducing the errors in the automatic transcription can improve the CER by virtue of providing better transcriptions to the semantic classifier. However, it has been reported that reductions in WER do not necessarily translate into reductions in CER [18]. This is because the sentence that gives the lowest WER might not be the sentence that gives the lowest CER. Hence, training the system parameters using the sentence that gives the lowest CER may result in an increased WER.

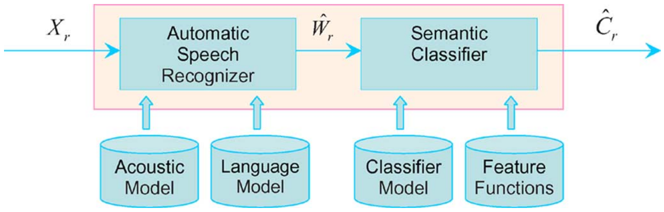

Fig. 1. Traditional spoken utterance classification system is composed of two isolated stages: an automatic speech recognition system is followed by a semantic classification system.

A. Automatic Speech Recognition (ASR)

Modern architectures for ASR [19], [20] aim to generate the mostly likely sentence hypotheses, $\hat{W}_r$ given X_r, i.e.,

$$\hat{W}_r = \arg\max_{W} P(W|X_r). \tag{2}$$

where $P(X_r|W)$ is the acoustic model (AM) score, and $P(W)$ is the language model (LM) score.

Current state-of-the-art ASR systems make use of word graphs and word lattices [21]–[23], and statistical n-gram LMs [24]. The AM and LM scores often have vastly different dynamic ranges. By introducing the so-called LM scaling factor L to account for this complication and using Bayes' rule, (2) is rewritten as

$$\hat{W} = \arg\max_{W} \left[P^{\frac{1}{L}}(X_r|W)P(W) \right]. \tag{3}$$

B. Semantic Classification

In this paper, we use binary n-gram features with $n = 1, 2, 3$ to capture the likelihood of the n-grams to be generated to express the user intent for the semantic class C. As an example, binary bigram feature functions are in the following form:

$$f^{BG}_{c,w_x w_y b}(C_r, W_r) = \begin{cases} 1, & \text{if } c = C_r \wedge w_x w_y b \in W_r \\ 0, & \text{otherwise.} \end{cases} \tag{3a}$$

That is, if the event $w_x w_y$ appears in the sentence W and if the class is the class of interest C_r, the binary feature function takes on a value of 1, and 0 otherwise. Once the features are extracted from the text, the task becomes a text classification problem, and traditional text categorization techniques are used to maximize the class-posterior probability $P(C_r|W_r)$, i.e., the probability of observing C_r given W_r [8], [25]–[29].

III. Integration of the ASR and Semantic Classification Phases

A block diagram of the proposed method is shown in Fig. 2. Instead of using the one-best sentence in the N-best list, the proposed method scores all of the N word sequences. First, the sentence that yields lowest CER instead of lowest WER is extracted. Then, the remaining sentences are scored so that those that are likely to yield incorrect classification are found.

One key contribution of this paper is the way in which the N-best hypotheses generated by the ASR module are used in

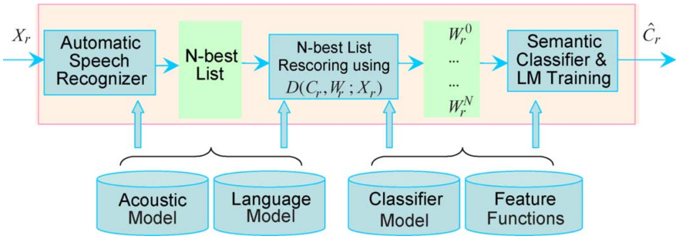

Fig. 2. ASR and semantic classification phases are integrated using the N-best lists.

the discriminative training of the system parameters. The proposed technique associates a score to each pair formed by the N-best list sentences and the semantic classes of interest. These $(M \cdot R)$ scores not only integrate information from all the available knowledge sources but also represent how likely it is that each sentence classifies X_r into the associated semantic class. Therefore, it also provides information about which N-best list sentences are most likely to yield the correct and incorrect classification decisions. The role of the discriminative training is to increase the scores of the sentence that yields lowest CER while reducing the scores of all other sentences.

Suppose that we are given the AM score $P(X_r|Wr)$, the class-posterior probability $P(C_r|Wr)$ and the LM probability $P(W_r)$. Then, the classification decision rule in (1) can be approximated as

$$\begin{aligned}\hat{C}_r &= \arg\max_{C_r} \sum_{W_r} P(C_r, W_r|X_r)\\ &= \arg\max_{C_r} \sum_{W_r} P(C_r|W_r, X_r) P^{\frac{1}{L}}(X_r|W_r)P(W_r)\\ &\cong \arg\max_{C_r} \sum_{W_r} P(C_r|W_r) P^{\frac{1}{L}}(X_r|W_r)P(W_r)\\ &\cong \arg\max_{C_r} \max_{W_r \in \aleph} P(C_r|W_r) P^{\frac{1}{L}}(X_r|W_r)P(W_r) \qquad (4)\end{aligned}$$

where the probability $P(C_r|X_r)$ is rewritten as the sum of $P(C_r, W_r|X_r)$ over all possible W_r. This summation is expanded in the second line, and $P(C_r|W_r, X_r)$ is substituted with $P(C_r|W_r)$ in the third line. In the last line, the summation term is replaced with the selection of the sentence among the N-best list, which we denote as $\aleph$.

Upon a closer look at (4), we see that the classification decision involves calculating the so-called *class-discriminant function* $D(C_r, W_r; X_r)$ for each pair W_r and C_r given by

$$D(C_r, W_r; X_r) = \log \left[P(C_r|W_r) P^{\frac{1}{L}}(X_r|W_r)P(W_r) \right]. \qquad (5)$$

$D(C_r, W_r; X_r)$ is an integrative score that shows how likely it is that W_r yields C_r for a given X_r. Since the computation of D integrates the acoustic, language and classifier scores, all available knowledge sources are being used to rescore and rerank the N-best sentences in a way that will be most useful for discrimination. Therefore, $D(C_r, W_r; X_r)$ is a natural choice to *discriminate* the sentence that is most likely to yield the correct classification decision.

A. Illustrative Example

In this section, we consider an example from the ATIS database in Table I to clarify our formulation.

In training stage, we aim to rescore the sentences and adjust the LM and classifier parameters. The goal is to ensure that the N-best list sentences are ranked with regard to their strength to yield the correct classification. For notational convenience, we reserve the superscript 0 for the correct class and its associated word sequence, whereas the superscript n, $n = 1, \ldots, N$, for the incorrect classes and their associated word sequences.

As a first step, the pair (C_r^0, W_r^0) is found. In this example, the correct class, C_r^0, is GROUND SERVICE, and W_r^0 is the sentence that maximizes the $D(.)$ score with C_r^0. Hence, it is the most likely sentence to yield correct classification decision. Following that, (C_r^1, W_r^1) is found, which is the most likely pair to yield an incorrect classification. For doing so, the sentence other than W_r^0 that has the highest $D(.)$ score with any class other than C_r^0 is found. This procedure is repeated until either all N-best list sentences or all the classes in $\mathcal{C}$ are used.

If this spoken utterance were used in the test stage, it would be assigned to the class that has the highest $D(C_r, W_r; X_r)$ score with any N-best list sentence. In this example, $D(C_r^0, W_r^0)$ is -20.04, whereas $D(C_r^1, W_r^1)$ is -17.56. The classification decision would yield FARE class, which implies a misclassification for X_r.

B. Rescoring in the Training Stage

For training, first of all, W_r^0 is found so that

$$W_r^0 = \arg\max_{W_r \in \aleph} \left[D\left(C_r^0, W_r; X_r\right)\right]. \qquad (6)$$

Because W_r^0 is extracted among all the N-best list sentences, it is the most likely sentence to yield the correct decision *independent of* whether or not it has the most correct ASR transcription. Hence, it may yield a higher WER than the sentence top-ranking in the N-best list.

After determining W_r^0, the remaining N-best list sentences are paired with the other semantic classes. Let C^n, $n = 1, \ldots, T$, denote the set of the semantic classes that are not yet paired with any word sequence, i.e., $C^n = C \setminus \{C_r^0, \ldots, C_r^{n-1}\}$. Also let $\aleph^n$ denote the set of the sentences in the N-best list that are not yet paired with any semantic class, i.e., $\aleph^n = \aleph \setminus \{W_r^0, \ldots, W_r^{n-1}\}$. Then, the classes in C^n and the sentences in $\aleph^n$ are paired according to the following rule:

$$C_r^n = \arg\max_{C_r \in \mathcal{C}^n} \left[\max_{W_r \in \aleph^n} D(C_r, W_r; X_r)\right].$$

We define the class-specific misclassification function $d_r(X_r)$ and a class-specific loss function $\ell_r(d_r(X_r))$ for each speech utterance X_r

$$\begin{aligned} d_r(X_r) =& -D\left(C_r^0, W_r^0; X_r\right)\\ &+ \log\left[\frac{1}{T-1}\sum_{n=1}^{T} \exp\left[\eta D\left(C_r^n, W_r^n; X_r\right)\right]\right]^{\frac{1}{\eta}}\\ \ell_r\left(d_r(X_r)\right) =& \frac{1}{1+\exp\left(-\alpha d_r(X_r)+\beta\right)}.\end{aligned}$$

TABLE I
ASSIGNMENT OF SEMANTIC CLASSES TO WORD SEQUENCES

SEMANTIC CLASS	WORD SEQUENCE	$D(.)$
C_r^0: GROUND SERVICE	W_r^0: WHAT IS THE TRANSPORTATION IN ATLANTA	-20.04
C_r^1: FARE	W_r^1: WHAT IS THE ROUND TRIP FARE FROM ATLANTA	-17.56
C_r^2: CITY	W_r^2: WHAT IS THE TRANSPORTATION ATLANTA	-25.46
C_r^3: FLIGHT	W_r^3: WHAT IS THE TRANSPORTATION AND ATLANTA	-28.49
C_r^4: FARE BASIS	W_r^4: WHAT IS THE ROUND TRIP FARE FROM THE ATLANTA	-27.98
C_r^5: AIRPORT SERVICE	W_r^5: WHAT IS THE TRANSPORTATION THE ATLANTA	-29.09
GROUND SERVICE	WHAT IS THE GROUND TRANSPORTATION IN ATLANTA	

where η, α, and β are the standard parameters in MCE training. The total classification loss approximating CER that the LM model $\Lambda_{\rm LM}$ and the semantic classifier model Λ_λ induce is then

$$L(\Lambda_W, \Lambda_\lambda) = \sum_r \ell_r\left(d_r(X_r)\right).$$

More details about training the parameters of Λ_W and Λ_λ are described in Sections III-D and III-E, respectively.

C. Decision Rule and Rescoring in the Test Stage

In the test stage, the decision rule

$$\hat{C}_r = \arg\max_{C_r}\left[\max_{W_r \in \aleph} D(C_r, W_r; X_r)\right] \tag{7}$$

is implemented. Doing so requires rescoring of the N-best lists sentences using $D(C_r, Wr; Xr)$, as indicated by the max operation inside the square brackets of (7). In rescoring, $D(C_r, Wr; Xr)$ scores are computed for each W_r and for each of the M semantic classes. Then, each sentence W_r is paired with a semantic class C_j^r, where $j \in \{1, 2, \ldots, T = min(M; N)\}$, that gives the greatest $D(C_r^j, Wr; Xr)$. Eventually, the scores are ranked and the semantic class with the greatest $D(\hat{C}_r, \hat{W}_r; X_r)$ is selected as $\hat{C}_r$.

D. Discriminative Training of the LM Parameters

In this section, we analyze the training of only bigram probabilities to simplify the algorithm description. The described procedure is also valid for learning other n-grams, and our experimental design includes the use of unigrams and trigrams as well.

Let $p_{w_x w_y}$ denote the bigram log-probabilities defined as $p_{w_x w_y} = \log P(w_y | w_x)$. The LM model parameters are trained to minimize the total loss function $L(\Lambda_W, \Lambda_\lambda)$ using the steepest descent method. This training results in the following update rule for the bigram LM probabilities:

$$\begin{aligned} p_{w_x w_y}^{(t+1)} &= p_{w_x w_y}^{(t)} - \varepsilon_{\rm LM} \sum_r \frac{\partial \ell_r\left(d_r(X_r)\right)}{\partial p_{w_x w_y}} \\ &= p_{w_x w_y}^{(t)} - \varepsilon_{\rm LM}\alpha \sum_r \ell_r(d_r)\left[1 - \ell_r(d_r)\right] \frac{\partial d_r(X_r)}{\partial p_{w_x w_y}}. \end{aligned} \tag{8}$$

where $\varepsilon_{\rm LM}$ is an appropriately chosen step-size. Suppose that the bigram $w_{x_i} w_{y_i}$ appears $n(W_r^n, w_{x_i} w_{y_i})$ times in the sentenceW_r^n. Then

$$\begin{aligned} \log P\left(W_r^n\right) &= \log\left[\prod_i \left[P\left(w_{y_i} | w_{x_i}\right)\right]^{n\left(W_r^n, w_{x_i} w_{y_i}\right)}\right] \\ &= \sum_i n\left(W_r^n, w_{x_i} w_{y_i}\right) \log\left[P\left(w_{y_i} | w_{x_i}\right)\right]. \end{aligned}$$

This gives us

$$\frac{\partial d_r(X_r)}{\partial p_{w_x w_y}} = -n\left(W_r^0, w_x w_y\right) + \sum_{n=1}^{T} H_r^n n\left(W_r^n, w_x w_y\right)$$

where the weighting coefficients H_r^n are given by

$$H_r^n = \frac{\exp\left[\eta D\left(C_r^n, W_r^n; X_r\right)\right]}{\sum_{m=1}^{T} \exp\left[\eta D\left(C_r^m, W_r^m; X_r\right)\right]}. \tag{9}$$

When $\eta \to \infty$, only the correct and the most competitive hypothesis contribute in (8). The LM parameters corresponding to the bigrams that are present in W_r^0 but not in W_r^1 (in the example, THE TRANSPORTATION, TRANSPORTATION IN, IN ATLANTA) are increased. In contrast, the LM parameters corresponding to the bigrams in W_r^1 but not in W_r^0 (THE ROUND, ROUND TRIP, TRIP FARE, FARE FROM, FROM ATLANTA) are decreased. The updates for the bigrams common to both W_r^0 and W_r^1 (WHAT IS, IS THE) cancel out and the corresponding LM parameters are left unchanged.

E. Discriminative Training of the Classifier Parameters

Let λ_k denote the weight of the kth feature function $f_k(C_r, W_r)$. The weights are trained to minimize the total loss function $L(\Lambda_W, \Lambda_\lambda)$. This gives us the following update rule for the classifier parameters, λ_k, $k = 1, \ldots, K$:

$$\begin{aligned} \lambda_k^{(t+1)} &= \lambda_k^{(t)} - \varepsilon_\lambda \sum_r \frac{\partial \ell_r\left(d_r(X_r)\right)}{\lambda_k} \\ &= \lambda_k^{(t)} - \varepsilon_\lambda \alpha \sum_r \ell_r\left(d_r(X_r)\right)\left[1 - \ell_r\left(d_r(X_r)\right)\right] \\ &\quad \times \frac{\partial d_r(X_r)}{\partial \lambda_k}. \end{aligned}$$

where ε_λ is appropriately chosen step-size, and

$$\frac{\partial d_r(X_r)}{\partial \lambda_k} = \frac{\partial D\left(C_r^0, W_r^0; X_r\right)}{\partial \lambda_k} + \sum_{j=1}^{T} H_r^j \frac{\partial D\left(C_r^j, W_r^j; X_r\right)}{\partial \lambda_k}.$$

In this paper, the classifiers are initialized with maximum entropy training using the one-best ASR transcriptions as input. Given the classifier parameters λ_i and the lexical n-gram feature functions $f_i(C, W)$, the distributions yielding maximal entropy are in the following form [30]:

$$P(C_r|W_r) = \frac{1}{Z_\Lambda(W_r)} \exp\left[\sum_i \lambda_i f_i(C_r, W_r)\right]$$

where $Z_\Lambda(W_r)$ is a normalization factor

$$Z_\Lambda(W_r) = \sum_{C_r} \exp\left(\lambda_i f_i(C_r, W_r)\right).$$

Let $\varphi(C_r^j, W_r^j)$ denote the partial derivative of $P(C_r^j|W_r^j)$ with respect to λ_k. Noting that $P(X_r|W_r)$ and $P(W_r)$ do not depend on λ_k, we obtain

$$\begin{aligned}\varphi\left(C_r^j, W_r^j\right) &= \frac{\partial \log P\left(C_r^j|W_r^j\right)}{\partial \lambda_k}\\ &= f_k\left(C_r^j, W_r^j\right) - \sum_{\tilde{C}} \xi\left(\tilde{C}, W_r^j\right) f_k\left(\tilde{C}, W_r^j\right)\end{aligned}$$

where the weighting factors of the feature functions are

$$\xi\left(\tilde{C}, W_r^j\right) = \frac{\exp\left[\sum_i \lambda_i f_i\left(\tilde{C}, W_r^j\right)\right]}{\sum_{\hat{C}} \exp\left[\sum_i \lambda_i f_i\left(\hat{C}, W_r^j\right)\right]}.$$

As with LM parameters, when $\eta \to \infty$, the classifier parameters, λ_k, associated with the bigrams that are present in W_r^0 but not in W_r^1 are increased. In contrast, the classifier parameters λ_k associated with the bigrams that are present in W_r^1 but not in W_r^0 are decreased. In addition, the updates for the bigrams common to both W_r^0 and W_r^1 cancel out and the corresponding classifier parameters are left unchanged.

IV. Experiments

We used the ATIS database to evaluate the discriminative LM and semantic classifier model learning methods described in this paper. In this section, we first describe the details of the experimental setup and then report the results of the experiments.

The ATIS task involves typical air travel planning scenarios such as flight schedules, fares, and ground transportation that were obtained from a relational database using spoken natural language. Following [28], ATIS2 and ATIS3 Category A data are used for training (5798 utterances), ATIS3 1993 and 1994 Category A test set (914 utterances) for testing, and the ATIS3 development set (410 utterances). The task involves 14 semantic classes.

In the ASR stage, we use the recognizer that is provided as part of the Microsoft Speech API (SAPI) without adaptations to its acoustic model. Also, η was fixed at 1, α was fixed at a constant, and whereas β_m was heuristically adjusted for each class. The heuristics we used is such that the $\pi_m(x, \Lambda)$ that are small in magnitude are summed, averaged, and used as β_m. The idea behind the adopted heuristics is to associate more loss to those samples for which $\ell_m(d_m(x, \Lambda))$ is close to 0.5, which represent the more confusable examples to the classifier.

TABLE II
PERFORMANCE OF THE BASELINE SYSTEM ON TEXT INPUTS AND SPEECH INPUTS FOR ATIS DOMAIN

	Test WER (%)	Test CER (%)
Manual Transcription	0.00	4.81
ASR Transcription	4.82	4.92

TABLE III
PROPOSED DT METHOD IMPROVES CER IMPROVES. SIGNIFICANT IMPROVEMENT OVER THE BASELINE SYSTEM IS ACHIEVED

	Dev. Set		Test Set	
iteration	WER (%)	CER (%)	WER (%)	CER (%)
0	6.80	5.12	5.8	4.92
1	7.3	5.61	6.2	4.60
2	7.2	5.61	6.4	4.38
3	7.4	5.12	6.4	4.38
4	7.2	5.12	6.3	4.27
5	7.3	4.63	6.4	4.04
6	7.4	5.37	6.5	4.04

A. Baseline System Performance

The baseline training system is composed of two separate stages [28]. The first stage involves the extraction of the best matching word sequence $\hat{W}_r$ for each spoken speech utterance X_r. In the second stage, maximum entropy classifiers are trained with steepest gradient descent method. Both the classifier and the trigram LM for ASR are trained from the in-domain manual transcriptions. The best-scenario ASR WERs and CERs for the baseline system are tabulated in Table II. These results were the best on this standard SUC task in the literature prior to the work described in this paper.

B. Performance of the Proposed Method

In our experiments, a trigram LM was obtained using the CMU-Cambridge toolkit [31], and 14 semantic classifier models based on the maximum entropy principle [28] were trained using the manual transcriptions. These seed models are denoted as $\Lambda_{\mathrm{LM}}^{\mathrm{man}}$ and $\Lambda_\lambda^{\mathrm{man}}$, respectively. In the very first (0th) iteration, $\Lambda_{\mathrm{LM}}^{\mathrm{man}}$ is used in ASR stage to generate the N-best lists. $\Lambda_{\mathrm{LM}}^{\mathrm{man}}$ is also used as the initial LM for discriminative LM training, which then yields Λ_{LM}^0. Then, Λ_{LM}^0 and $\Lambda_\lambda^{\mathrm{man}}$ are used to rescore the N-best list for discriminative classifier training, which yields the classifier model Λ_λ^0.

1) Parameter Refinement After an ASR Phase: After the 0th iteration, there are two ways to proceed. First, we can use Λ_{LM}^0 for another speech recognition step, and follow the same steps as in 0th iteration to get Λ_{LM}^1 and Λ_λ^1, and so on. The WERs and CERs on both the development set and the test set at each iteration are listed in Table III, where we set $\eta = 1.0$, $\varepsilon_{\mathrm{LM}} = 0.001$, $\varepsilon_\lambda = 0.03$, $\alpha = 0.5$, and $L = 1$.

In the 0th iteration, the WER of 5.8% was obtained with $\Lambda_{\mathrm{LM}}^{\mathrm{man}}$ and the CER of 4.92% was obtained with using Λ_{LM}^0 and Λ_λ^0. In the first iteration, the WER increased to 6.2% by using the discriminatively trained Λ_{LM}^0. However, the CER reduced to 4.60% by using Λ_{LM}^1 and Λ_λ^1.

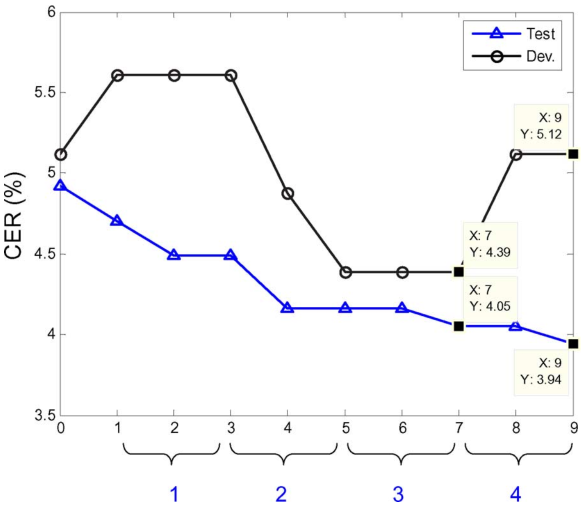

Fig. 3. With four outer and two inner iterations.

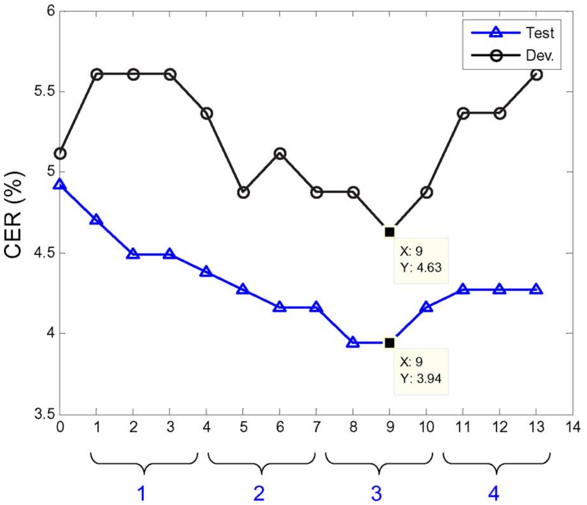

Fig. 4. With four outer and three inner iterations.

The CER has been reduced from 4.92% to 4.04% at the fifth iteration when the CER for the development set is at the minimum.

2) Parameter Refinement Without an ASR Phase: We can also schedule inner iterations, which do not require repeating the ASR phase. Since the classifier model $\Lambda_{\lambda}^{\text{man}}$ is replaced with Λ_{LM}^{0} and Λ_{λ}^{0} at the end of 0th iteration, we can use Λ_{λ}^{0} to rescore the N-best list for training Λ_{LM}^{1} *without* performing speech recognition step. This is followed by training a new classifier model Λ_{λ}^{1}, and so on.

In Figs. 3 and 4, we plotted our experimental results where we run four outer iterations with two and three inner iterations, respectively. First of all, the inner iterations provide the advantage of refining the models without repeating the (expensive) ASR step. (In the ATIS dataset, the ASR step takes two hours of CPU time for the training data using Microsoft SAPI, whereas discriminative training takes only 15 to 20 min.) For larger data sets, the saving in computation would be even greater. In addition, we observe that with three inner iterations, as shown in Fig. 4, the lowest CER of 3.92% in the test data is achieved with the lowest CER of 4.63% in the development data. We conclude that inner iterations can yield as good and even better performance while offering the advantage of reduction in computational load.

V. Conclusion and Future Work

Our motivation in this paper was to make an *explicit* use of the N-best lists generated by the ASR systems to alleviate the problems due to the isolation of the ASR and semantic classification phases in traditional SUC methods. For this purpose, the N-best lists generated in the ASR phase are rescored using all available knowledge sources—in this paper, the language, acoustic, and classifier models—and are utilized to discriminatively learn system parameters. More specifically, the sentence that is most likely to yield correct classification decision as well as those sentences that are more likely to yield incorrect decisions are found. These sentences are then used to discriminatively train the language and semantic-classifier models to minimize the overall semantic CER.

Our experiments demonstrate that when ASR transcriptions are sufficiently accurate, the proposed scheme can reduce CER obtained with ASR word transcriptions compared to the CER obtained with manual word transcriptions. Our experimental results on the standard ATIS SUC task demonstrated significant performance improvement, measured by the reduced amount of CER, from the earlier best system on the identical task. Specifically, the CER was reduced from 4.92% to 3.94%.

There are many directions of research that can be undertaken. One such direction is extending the current implementation to the estimation of other SUC system parameters. In this paper, we assumed the lexical, pronunciation, and acoustic model were fixed. The method described in this paper can be extended to updating the lexical and acoustic model parameters as well. Using the proposed method for lexical modeling introduces an additional term into the class discriminant function. This would mean the n-best sentences would be listed considering the lexical effects as well. This would also mean that the lexical model parameters would be updated so as to reduce CER.

It is also possible to develop a more general method using a noise robust front-end, and analyze the sensitivity of the system performance to the set of system parameters that are difficult to learn automatically. Furthermore, we plan to extend our approach in the context of SUC as discussed in this paper to more general spoken language understanding tasks, including slot filling, which is a sequence labeling task, using conditional random fields (CRFs). Finally, the application of our integrative technique to large-scale, realistic tasks will potentially create a more significant impact of the research presented in this paper. One such realistic research scenario is the investigation of what would happen when there are thousands of semantic classes instead of only 14 in the ATIS task or when there are thousands of hours of spoken speech data available.

References

[1] Y.-Y. Wang, L. Deng, and A. Acero, “Spoken language understanding,” *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 16–31, Sep. 2005.

[2] G. Tur, D. Hakkani-Tür, and G. Riccardi, “An active approach to spoken language processing,” *IEEE Signal Process. Mag.*, vol. 3, no. 3, pp. 1–31, Oct. 2006.

[3] H.-K. J. Kuo, I. Zitouni, E. Fosler-Lussier, E. Ammicht, and C.-H. Lee, "Discriminative training for call classification and routing," in *Proc. Int. Conf. Speech Lang. Process.*, Denver, CO, 2002, pp. 1145–1148.

[4] M. A. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. O. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passoneau, S. Roukos, G. Sanders, S. Seneff, and D. Stallard, "DARPA communicator evaluation: Progress from 2000 to 2001," in *Proc. Int. Conf. Speech Lang. Process.*, Denver, CO, 2002, pp. 273–276.

[5] S. Seneff, E. Hurley, R. Lau, C. Pao, P. SChmid, and V. Zue, "GALAXY-II: A reference architecture for conversational system development," in *Proc. Int. Conf. Speech Lang. Process.*, Sydney, Australia, 1998, pp. 931–934.

[6] J. R. Bellegarda, "Semantic inference: A data driven solution for NL interaction," in *Proc. Int. Conf. Speech Lang. Process.*, Denver, CO, 2002.

[7] A. Weibel, "Interactive translation of conversational speech," *Computer*, vol. 29, no. 7, pp. 41–48, 1996.

[8] M. Saraclar and B. Roark, "Joint discriminative language modeling and utterance classification," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, Mar. 2005, pp. 561–564.

[9] H. Jiang, H.-K. Kuo, E. Fosler-Lussier, and C.-H. Lee, "Discriminative training of language models for speech recognition," in *Proc. Int. Conf. Speech Lang. Process.*, Orlando, FL, 2002, pp. I-325–I-328.

[10] W. Chou, "Discriminant-function-based minimum recognition error rate pattern-recognition approacht to speech recognition," *Proc. IEEE*, vol. 88, no. 8, pp. 1201–1224, Aug. 2000.

[11] L. Hirschman, M. Bates, D. Dahl, W. Fisher, J. Garofolo, K. Hunicke-Smith, D. Pallett, C. Pao, P. Price, and A. Rudnicky, "Multi-site data collection for a spoken language corpus," in *Proc. DARPA Speech Natural Lang. Workshop*, Harriman, New York, 1992, pp. 7–14.

[12] L. Hirschman, M. Bates, D. Dahl, and W. Fisher, "Multi-site data collection and evaluation in spoken language understanding," in *Proc. ARPA Human Lang. Technol. Workshop*, Princeton, NJ, 1994, pp. 19–24.

[13] S. Young, "The statistical approach to the design of spoken dialogue systems," Cambridge Univ. Eng. Dept., Cambridge, U.K., Tech. Rep. CUED/F-INFENG/TR.433, Sep. 2002.

[14] S. Boyce, "Natural spoken dialogue systems for telephony applications," *Commun. ACM*, vol. 43, no. 9, pp. 29–34, Sep. 2000.

[15] Y. Wang, L. Deng, and A. Acero, "Spoken language understanding—An introduction to the statistical framework," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 16–31, Sep. 2005.

[16] Y.-Y. Wang and A. Acero, "Discriminative models for spoken language understanding," in *Proc. Int. Conf. Speech Lang. Process.*, Sep. 2006, pp. 2426–2429.

[17] Y. Wang, A. Acero, M. Mahajan, and J. Lee, "Combining statistical and knowledge-based spoken language understanding in conditional models," in *Proc. COLING/ACL*, Jul. 2006, pp. 882–889.

[18] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding?," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, Virgin Islands, Dec. 2003, pp. 577–582.

[19] L. Rabiner and B.-H. Juang*, Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[20] F. Jelinek*, Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1998.

[21] C.-H. Lee and L. R. Rabiner, "A network-based frame-synchronous level building algorithm for connected word recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, New York, 1988, pp. 410–413.

[22] V. Steinbiss, "Sentence-hypotheses generation in continous-speech recognition systems," in *Proc. Eur. Conf. Speech Commun. Technol.*, Paris, 1989, pp. 51–54.

[23] R. Schwartz and Y.-L. Chow, "A comparison of several approximate algorithms for finding multiple (n-best) sentence hypotheses," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Toronto, ON, Canada, 1991, pp. 701–704.

[24] C. D. Manning and H. Schütze*, Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

[25] J. R. Bellegarda and K. E. A. Silverman, "Natural language spoken interface control using data-driven semantic inference," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 267–277, May 2003.

[26] C. Chelba, M. Mahajan, and A. Acero, "Speech utterance classification," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, Apr. 2003.

[27] G. Tur, D. Hakkani-Tur, and G. Riccardi, "Extending boosting for call classification using word confusion networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2004, pp. 437–440.

[28] Y.-Y. Wang, J. Lee, and A. Acero, "Speech utterance classification model training without manual transcriptions," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, pp. I-553–I-556.

[29] R. C. Rose, H. Yao, G. Riccardi, and J. Wright, "Integration of utterance verification with statistical language modeling and spoken language understanding," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1998, pp. 237–240.

[30] A. Berger, S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, pp. 39–71, 1996.

[31] A. Ward, "Recent improvements in the CMU spoken language understanding system," in *Proc. Human Lang. Technol. Workshop*, Plainsboro, NJ, 1994, pp. 213–216.

Sibel Yaman (S04) received the B.S. degree in electrical and electronic engineering from Bilkent University, Ankara, Turkey, in 2002. She is currently pursuing the Ph.D. degree in electrical engineering at the Georgia Institute of Technology, Atlanta.

She is a recipient of the Microsoft Research (Redmond) Graduate Fellowship for 2006–2007. Her research interests include text categorization, automatic language identification, multi-objective optimization techniques for classification, and language modeling.

Li Deng (M'86–SM'91–F'05) received the Ph.D. degree in electrical engineering from the University of Wisconsin-Madison in 1986.

In 1989, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, as an Assistant Professor, where he became Full Professor in 1996. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, and from 1997 to 1998, at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA, as Senior Researcher, where he is currently Principal Researcher. He is also an Affiliate Professor in Electrical Engineering at the University of Washington, Seattle. His research interests include acoustic–phonetic modeling of speech, speech and speaker recognition, speech synthesis and enhancement, speech production and perception, auditory speech processing, noise-robust speech processing, statistical methods and machine learning, nonlinear signal processing, spoken language systems, multimedia signal processing, and multimodal human–computer interaction. In these areas, he has published over 250 refereed papers in leading international conferences and journals, 12 book chapters, and has given keynotes, tutorials, and lectures worldwide. He has been granted 16 U.S. or international patents in acoustics, speech, and language technology, and signal processing. He coauthored the book *Speech Processing A Dynamic and Optimization-Oriented Approach* (Marcel Dekker, 2003) and authored the book *Dynamic Speech Models Theory, Algorithms, and Applications* (Morgan and Claypool, 2006).

Dr. Deng served on the Education Committee and the Speech Processing Technical Committee of the IEEE Signal Processing Society from 1996–2000 and was an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 2002–2005. He currently serves on the Society's Multimedia Signal Processing Technical Committee, on the Editorial Board of IEEE SIGNAL PROCESSING LETTERS, and as Area Editor for the IEEE SIGNAL PROCESSING MAGAZINE. He was a Technical Chair of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), and is the General Chair of the IEEE Workshop on Multimedia Signal Processing in 2006. He is a Fellow of the Acoustical Society of America.

Dong Yu (M'97–SM'06) received the B.S. degree (with honors) in electrical engineering from Zhejiang University, Hanzhou, China, the M.S. degree in electrical engineering from the Chinese Academy of Sciences, Beijing, the M.S. degree in computer science from Indiana University, Bloomington, and the Ph.D. degree in computer science from the University of Idaho, Moscow.

He joined Microsoft Research, Redmond, WA, in 1998, and Microsoft Speech Research Group in 2002. His research interests include speech processing, pattern recognition, and computer and network security. He has published more than 30 papers in these areas.

Ye-Yi Wang (M'98–SM'05) received the B.S. and M.S. degrees in computer science from Shanghai Jiao Tong University, Shanghai, China, in 1985 and 19988, respectively, the M.S. degree in computational linguistics from Carnegie Mellon University, Pittsburgh, PA, in 1992, and the Ph.D. degree in human language technology from Carnegie Mellon University in 1998.

He joined Microsoft Research, Redmond, WA, in 1998. His research interests include spoken dialog systems, natural language processing, language modeling, statistical machine translation and machine learning. He was on the editorial board of the Chinese Contemporary Linguistic Theory Series. He is the coauthor of *Introduction to Computational Linguistics* (Chinese Social Science Publishing House) and has published over 40 technical papers.

Alex Acero (S'85–M'90–SM'00–F04) received the M.S. degree from the Polytechnic University of Madrid, Madrid, Spain, in 1985, the M.S. degree from Rice University, Houston, TX, in 1987, and the Ph.D. degree from Carnegie Mellon University, Pittsburgh, PA, in 1990, all in electrical engineering.

He worked in Apple Computer's Advanced Technology Group from 1990 to 1991. In 1992, he joined Telefonica $I + D$, Madrid, as a Manager of the Speech Technology Group. In 1994, he joined Microsoft Research, Redmond, WA, where he became a Senior Researcher in 1996 and Manager of the Speech Research Group in 2000. Since 2005, he has been a Research Area Manager directing an organization with over 60 engineers conducting research in speech technology, natural language, computer vision, communication, and multimedia collaboration. He is currently an Affiliate Professor of Electrical Engineering at the University of Washington, Seattle. He is author of the books *Acoustical and Environmental Robustness in Automatic Speech Recognition* (Kluwer, 1993) and *Spoken Language Processing* (Prentice-Hall, 2001), has written invited chapters in four edited books and over 150 technical papers. He holds 35 U.S. patents. His research interests include speech and audio processing, natural language processing, image understanding, multimedia signal processing, and multimodal human–computer interaction.

Dr. Acero has served the IEEE Signal Processing Society as Vice President Technical Directions (2007–2009), 2006 Distinguished Lecturer, member of the Board of Governors (2003–2005), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (2003–2005), and the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2005–2007), and member of the editorial board of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING (2006–2008) and the IEEE SIGNAL PROCESSING MAGAZINE (2008–2010). He also served as member (1996–2002) and Chair (2000–2002) of the Speech Technical Committee of the IEEE Signal Processing Society. He was Publications Chair of ICASSP'98, Sponsorship Chair of the 1999 IEEE Workshop on Automatic Speech Recognition and Understanding, and General Co-Chair of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding. He is member of the editorial board of Computer Speech and Language.