INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Minimum Semantic Error Cost Training of Deep Long Short-Term Memory Networks for Topic Spotting on Conversational Speech

Zhong Meng, Biing-Hwang (Fred) Juang

School of Electrical and Computer Engnineering, Georgia Institute of Technology
75 5th Street NW, Atlanta, GA 30308, USA

zhongmeng@gatech.edu, juang@ece.gatech.edu

Abstract

The topic spotting performance on spontaneous conversational speech can be significantly improved by operating a support vector machine with a latent semantic rational kernel (LSRK) on the decoded word lattices (i.e., weighted finite-state transducers) of the speech [1]. In this work, we propose the minimum semantic error cost (MSEC) training of a deep bidirectional long short-term memory (BLSTM)-hidden Markov model acoustic model for generating lattices that are semantically accurate and are better suited for topic spotting with LSRK. With the MSEC training, the expected semantic error cost of all possible word sequences on the lattices is minimized given the reference. The word-word semantic error cost is first computed from either the latent semantic analysis or distributed vector-space word representations learned from the recurrent neural networks and is then accumulated to form the expected semantic error cost of the hypothesized word sequences. The proposed method achieves 3.5% - 4.5% absolute topic classification accuracy improvement over the baseline BLSTM trained with cross-entropy on Switchboard-1 Release 2 dataset.

Index Terms: topic spotting, rational kernels, long short-term memory, recurrent neural networks, discriminative training, automatic speech recognition

1. Introduction

Topic spotting on spontaneous conversational speech is an essential technique for spoken-dialog systems. The response of a spoken-dialog system is often guided by the topic category of the speakers utterance. The topic spotting is aimed at classifying an utterance into one of a pre-defined set of topics.

Conventionally, topic spotting is conducted by scoring the one-best transcription generated by a large-vocabulary continuous speech recognition (LVCSR) system based on a set of keywords selected according to their contributions for topic discrimination [2, 3, 4]. To take advantage of the multiple hypothesized word sequences on the decoded lattices, Cortes et al. proposed rational kernels [5], which are a series of kernels defined on the WFSTs. The topic classification is conducted via support vector machine (SVM) with the n-gram rational kernels which maps the WFSTs (lattices) to a high dimensional n-gram feature space and then employs an inner product for topic identification [6, 7]. However, the n-gram rational kernel assumes an exact match of the n-grams (words or phrases) and treats the contribution of each n-gram to the topic discrimination uniformly. To overcome this problem, Weng et al. [1, 8] proposed the latent semantic rational kernels (LSRK) for topic spotting on spontaneous speech. In the LSRK framework, the WFSTs (lattices) are mapped onto a reduced dimensional latent semantic space rather than the n-gram feature space. LSRK is generalized to incorporate external knowledge from several text analysis techniques such as WordNet [9], latent semantic analysis (LSA) [10].

However, the word lattices of the utterances in [1] are generated by a Gaussian mixture model (GMM)-hidden Markov model (HMM) based LVCSR system trained with maximum likelihood estimation (MLE). With MLE, the GMMs model the distribution of the speech frames given the senone (tri-phone state), which does not necessarily lead to minimized recognition error or maximized topic spotting accuracy. Many discriminatvie training methods such as MCE [11, 12, 13], maximum mutual information (MMI) [14, 15], minimum word error (MWE) [16], minimum phone error (MPE) [17], state-level minimum Bayes risk (sMBR) [18, 19] and boosted MMI [20] are proposed to further refine the acoustic model.

For topic spotting on conversational speech, the lattices are classified based on their semantic meaning rather than their spellings or pronunciations and the high phoneme or state accuracy of a sentence does not necessarily lead to the high semantic accuracy. For instance, the sentence *"This machine is productive."* is much more semantically correct than *"This machine is inefficient."* given the reference *"This machine is efficient."*, but its phoneme or state accuracy is much lower than the latter one. For the topic spotting task, the LVCSR is expected to generate word lattices that are accurate in terms of the semantic meanings instead of the pronunciations. Therefore, we propose the minimum semantic error cost (MSEC) training of the acoustic model, in which the expected semantic error cost of all possible word sequences on the lattices is minimized given the reference. The semantic error cost between a pair of words can be estimated via LSA or recurrent neural networks (RNN) learned vector space word representations. The expected semantic error cost of the hypothesized sentences can be obtained by accumulating the word-word semantic error costs on the lattices via forward-backward algorithm.

In addition, the GMM-HMM acoustic model with diagonal covariance matrices in [1] are not good at handling highly correlated frames and the concatenation of neighboring frames will inevitably bring about the curse of dimensionality issue during model training. Therefore, we introduce the deep bi-directional long short-term memory (BLSTM)-HMM for acoustic modeling. The cycles in a BLSTM allows it to store and update the context information about the past and future inputs in its internal state for an amount of time that is not fixed a priori, but rather depends on its weights and on the input data [21]. The deep BLSTMs are able to exploit the long-term temporal context information within a dynamically changing window over the input speech sequence. We define the MSEC objective function and derive the backpropagation error for the BLSTM. The BLSTM is then optimized using backpropagation through time and stochastic gradient decent. With the MSEC training of BLSTM acoustic model, the LVCSR decoder is able to gener-

http://dx.doi.org/10.21437/Interspeech.2017-590

ate word lattices with significantly reduced semantic error cost for the subsequent topic spotting within the LSRK framework. The BLSTM acoustic is trained with Switchboard-1 Release 2 dataset, which is a large-scale spontaneous conversational telephone speech (CTS) dataset. The proposed method achieves 3.5% - 4.5% absolute improvement over BLSTM baseline system trained with cross-entropy criterion for the topic classification task on a subset of Switchboard-1 Release 2.

2. Deep BLSTM acoustic model

The LSTM network, a special kind of RNN with purpose-built memory cells to store information, have been successfully applied to many sequence modeling tasks. Recently, LSTMs based acoustic modeling has achieved improved performance over DNNs [22, 23] and conventional RNNs [24, 25] for LVCSR as they are able to model temporal sequences and long-range dependencies more accurately than the others especially when the amount of training data is large. LSTM has been successfully applied in both the LSTM-HMM hybrid systems [26, 27, 28] and the end-to-end system [29, 30, 31].

For acoustic modeling, the LSTM takes in a sequence of input speech frames $X = \{x_1, \ldots, x_T\}$ and computes the hidden vector sequence $H = \{h_1, \ldots, h_T\}$ by iterating the equation below

$$h_t = \text{LSTM}(x_t, h_{t-1}) \tag{1}$$

where $\text{LSTM}(\cdot)$ denote the hidden layer function of the LSTM. In this paper, we implement Eq. (1) with the LSTM introduced in [32]

A deep LSTM stacks multiple LSTM hidden layers on top of each other, The deep BLSTM processes the sequence of speech frames from both directions. It computes the forward hidden vector sequence $\overrightarrow{H} = \{\overrightarrow{h}_1, \ldots, \overrightarrow{h}_T\}$, the backward hidden vector sequence $\overleftarrow{H} = \{\overleftarrow{h}_1, \ldots, \overleftarrow{h}_T\}$ and the output sequence of senone posterior $Y = \{y_1, \ldots, y_T\}$ by iterating the backward layer from $t = T$ to 1, the forward layer from $t = 1$ to T and then updating the output layer. For each time, the output from both the forward and backward hidden layers are concatenated and then fed as the input of the next forward and backward hidden layers or the output layer.

To reduce the number of trainable parameters and alleviate the computational complexity, we introduce a separate linear projection layer after each BLSTM layer as in [33]. The deep BLSTM acoustic model in this work is formulated as follow.

$$\overrightarrow{h}_t^1 = \text{LSTM}_1^{\text{forward}}(x_t, \overrightarrow{h}_{t-1}^1) \tag{2}$$

$$\overleftarrow{h}_t^1 = \text{LSTM}_1^{\text{backward}}(x_t, \overleftarrow{h}_{t+1}^1) \tag{3}$$

$$\overrightarrow{h}_t^n = \text{LSTM}_n^{\text{forward}}(p_t^{n-1}, \overrightarrow{h}_{t-1}^n), \quad n = 2, \ldots, N \tag{4}$$

$$\overrightarrow{p}_t^n = W_{\overrightarrow{h}^n, \overrightarrow{p}^n} \overrightarrow{h}_t^n, \quad n = 1, \ldots, N \tag{5}$$

$$\overleftarrow{h}_t^n = \text{LSTM}_n^{\text{backward}}(p_t^{n-1}, \overleftarrow{h}_{t+1}^n), \quad n = 2, \ldots, N \tag{6}$$

$$\overleftarrow{p}_t^n = W_{\overleftarrow{h}^n, \overleftarrow{p}^n} \overleftarrow{h}_t^n, \quad n = 1, \ldots, N \tag{7}$$

$$p_t^n = (\overleftarrow{p}_t^n, \overrightarrow{p}_t^n), \quad n = 1, \ldots, N \tag{8}$$

$$y_t = \text{softmax}(W_{p^N y} \tanh(p_t^N) + b_y) \tag{9}$$

where $\text{LSTM}_n^{\text{forward}}(\cdot)$ and $\text{LSTM}_n^{\text{backward}}(\cdot)$ denote the forward and backward n^{th} hidden layer functions of the LSTM respectively. $\overrightarrow{p}_t^n$ and $\overleftarrow{p}_t^n$ are the projection vectors of forward and backward hidden vectors $\overrightarrow{h}_t^n$ and $\overleftarrow{h}_t^n$ respectively at the n^{th} layer. $W_{\overrightarrow{h}^n, \overrightarrow{p}^n}$ and $W_{\overleftarrow{h}^n, \overleftarrow{p}^n}$ are the projection matrices. p_t^n is the concatenation of forward and backward projection vectors $\overrightarrow{p}_t^n$ and $\overleftarrow{p}_t^n$. y_t is the senone posterior output vector given input speech frame x_t.

3. Minimum Semantic Error Cost Training of BLSTM Acoustic Model for Topic Spotting

With cross-entropy criterion, the BLSTM acoustic models are trained to model the senone distribution given an input speech frame, which do not necessarily lead to minimized recognition error rate in LVCSR tasks. Improved performance can be achieved by discriminatively training the DNN [34] and LSTM [35] acoustic model with MPE, MWE, sMBR and etc.

For topic spotting on conversational speech, the speech signal is first decoded to a word lattice by an LVCSR system and the SVM with LSRK then operates on the decoded lattices to predict the topic category. The lattices are classified based on their semantic meaning rather than their spellings or pronunciations and the high phone or state accuracy of a sentence does not necessarily lead to the high semantic accuracy.

To improve the topic spotting accuracy, the LVCSR system is expected to generate word lattices that are accurate in terms of the semantic meanings rather than the pronunciations. This motivate us to device a new objective function for discriminatively training the BLSTM acoustic model so that the LVCSR can generate word lattices with reduced *semantic error cost*. Therefore, we propose the *minimum semantic error cost (MSEC) training of BLSTM* acoustic model for topic spotting with LSRK.

We first define the word-word semantic error cost $C(i, j)$ of mistakenly recognizing one word with index i to another with index j as the *negative of the semantic similarity* between the words i and j denoted by $S_{i,j}$, i.e., $C(i, j) = -S_{i,j}$. The expected semantic error cost of all hypothesized word sequences on the lattice with respect to the reference can be accumulated from the semantic error cost of the words.

Assume that the training data is given by training utterances $r = \{1, \ldots, R\}$. $X_r = \{x_{r1}, \ldots, x_{rT_r}\}$ is the sequence of observations for utterance r, W_r is the word sequence in the reference (label transcription) for utterance r. W is one of all the word sequences in the decoded speech lattice for utterance r. $S_W = \{s_{W1}, \ldots, s_{WT}\}$ is the senone sequence corresponding to W, where s_{Wt} is the senone which frame x_{rt} is aligned with.

The MSEC is aimed at minimizing the expected semantic errors cost of all possible word sequences given the reference. The objective function is formulated as

$$\begin{aligned} \mathcal{L}_{\text{MSEC}} &= \sum_{r=1}^{R} \sum_{W} P(W|X_r) C(W, W_r) \\ &= \sum_{r=1}^{R} \frac{\sum_W P(W|X) P(W) C(W, W_r)}{\sum_{W'} P(X_r|W') P(W')} \end{aligned} \tag{10}$$

where $C(W, W_r)$ is the semantic error cost of mis-recognizing the reference W_r as the hypothesis sentence W.

Take the derivative of derivative of Eq. (10) with respect to

the activation $a_{rt}(s)$ for senone s at the output layer is

$$\frac{\partial \mathcal{L}_{MSEC}(\Lambda)}{\partial a_{rt}(s)} = \sum_u \frac{\partial \mathcal{L}_{MSEC}(\Lambda)}{\partial \log p(x_{rt}|u)} \frac{\partial \log p(x_{rt}|u)}{\partial a_{rt}(s)}$$
$$= \gamma_{rt}^{W}(s) \left\{ E_{P(W|s_{Wt}=s,X_r)}[C(W,W_r)] - E_{P(W|X_r)}[C(W,W_r)] \right\} \quad (11)$$

where $\gamma_{rt}^{W \neq W_r}(s)$ is the posterior of being in senone s at time t, computed over the denominator lattice of the utterance r excluding the path corresponding to the word sequence W_r, $\log p(x_{rt}|u)$ is the log-likelihood of x_{rt} given senone u obtained by subtracting the log senone prior $\log p(u)$ from the log senone posterior $\log(y_t)$ in Eq. (9).

Eq. (11) is the error to be backpropagated through time to derive the gradients for all the parameters of BLSTM, in which $E_{P(W|s_{Wt}=s,X_r)}[C(W,W_r)]$ and $E_{P(W|X_r)}[C(W,W_r)]$ are obtained by accumulating the word-word semantic error costs $C(i,j)$ (i.e., $-S_{i,j}$) through performing *forward-backward algorithm* on the decoded word lattice. The decoded lattices of the utterance X_r can be viewed as a directed graph WFST(X_r). For a state q of the WFST(X_r), $\mathcal{DP}(q)$ denote the set of direct predecessors of q and $\mathcal{DS}(q)$ denote the set of direct successor of q. The arrow (arc) directed from state p to state q is denoted by $l_{p,q}$, the weight on $l_{p,q}$ is denoted by $g_{p,q}$, the word (input label) of the $l_{p,q}$ is denoted by $m_{p,q}$ and the time for $l_{p,q}$ in X_r is $t_{p,q}$. $\mathcal{F}(X_r)$ is the set of final states in the WFST(X_r) and g_f is the final weight of the final state f. W_{rt} is the word at time t of the reference word sequence W_r. The first round of forward-backward is performed with forward and backward likelihood $\alpha_q^{(1)}$ and $\beta_q^{(1)}$.

$$\alpha_q^{(1)} = \sum_{p \in \mathcal{DP}(q)} \alpha_p^{(1)} g_{p,q}, \qquad \beta_q^{(1)} = \sum_{p \in \mathcal{DS}(q)} \beta_p^{(1)} g_{q,p} \quad (12)$$

$$P(W|X_r) = \sum_{f \in \mathcal{F}(X_r)} \alpha_f^{(1)} g_f \quad (13)$$

The second round of forward-backward is formulated as

$$\alpha_q^{(2)} = \frac{1}{\alpha_q^{(1)}} \sum_{p \in \mathcal{DP}(q)} \alpha_p^{(1)} g_{p,q} [\alpha_p^{(2)} + C(m_{p,q}, W_{rt_{p,q}})] \quad (14)$$

$$\beta_q^{(2)} = \frac{1}{\beta_q^{(1)}} \sum_{p \in \mathcal{DS}(q)} \beta_p^{(1)} g_{q,p} [\beta_p^{(2)} + C(m_{q,p}, W_{rt_{q,p}})] \quad (15)$$

$$E_{P(W|X_r)}[C(W,W_r)] = \frac{\sum_{f \in \mathcal{F}(X_r)} \alpha_f^{(1)} g_f \alpha_f^{(2)}}{P(W|X_r)} \quad (16)$$

where $\alpha_q^{(2)}$ and $\beta_q^{(2)}$ are the average cost of the partial state sequences preceding and following q respectively. Assume that senone s is on the arrow directed from state p_s to state q_s of the WFST, we have

$$E_{P(W|s_{Wt}=s,X_r)}[C(W,W_r)] = \alpha_{p_s}^{(2)} + C(m_{p_s,q_s}, W_{rt_{p_s,q_s}}) + \alpha_{q_s}^{(2)} \quad (17)$$

$$\gamma_{rt}^{W \neq W_r}(s) = \frac{\alpha_{p_s}^{(1)} g_{p_s,q_s} \beta_{q_s}^{(1)}}{P(W|X_r)} \quad (18)$$

The word-word semantic similarity $S_{i,j}$ (i.e, $-C(i,j)$) can be computed from LSA. In LSA, a document is first represented by a column vector d indexed by the word in the vocabulary and the corpus of documents is represented by a word-document matrix $D = [d_1, \ldots, d_m]$. The columns of D are indexed by the documents. $D_{i,j}$ describes the number of occurrence of word i in document j. With singular value decomposition and low-rank matrix approximation, we have $D \approx U_K \Sigma_K V_K^\top$, where Σ_K contains only the largest K singular values in Σ and the U_K, V_K contain the K left and right singular vectors corresponding to Σ_K. Under LSA framework, the *semantic similarity matrix* S is formulated as

$$S = U_K \Sigma_K^{-1} \Sigma_K^{-1} U_K^\top \quad (19)$$

where S is a square matrix with a dimension equal to the number of words in the vocabulary and the element $S_{i,j}$ of S is the *semantic similarity* between word i and word j. We set the diagonal elements of S to be the term frequency-inverse document frequency (tf-idf) [36] weights of the corresponding words and scale the non-diagonal elements proportionally.

The word-word semantic similarity $S_{i,j}$ can also be obtained from the distributed vector representations of words learned by RNN (e.g., continuous bag-of-words model and continuous skip-gram model [37]) from a large amount of text data. By training a RNN language model (LM), we obtains not only the model itself but also the vector-space word representations that are implicitly learned by the input layer weights. These word representations encode precise syntactic and semantic word relationships as well as linguistic regularities and patterns [38]. Suppose w_i is a column vector representation for the word i, the word vocabulary can be represented by a matrix $W = [w_1, \ldots, w_v]^\top$. With SVD and low-rank approximation, the similarity matrix S can be formulated as

$$S = WW^\top = W_K I_K W_K^\top \quad (20)$$

In this work, topic classification is performed by a multi-class SVM with LSRK that takes the decoded lattices of the utterances as the input. In the WFST framework, the LSRK is formulated as [1]

$$k_n(L_1, L_2) = w[(L_1 \circ T) \circ \text{WFST}(S) \circ (T^{-1} \circ L_2)] \quad (21)$$

where $w[B]$ denotes the shortest distance from the start state to the set of final states of the transducer B. The transducer T is used to extract all words. Transducer WFST(S) encodes the *semantic similarity matrix* S. The LSRK can be generalized with respect to the semantic similarity matrix S to incorporate external knowledge from LSA and RNN LM as in Eq. (19) and Eq. (20) to achieve better performance for topic spotting.

4. Experiments

4.1. Dataset Description

We evaluate the performance of the proposed framework on a large-scale CTS task, i.e., the 300 hours Switchboard-1 Release 2 (LDC97S62). It consists of 2348 two-sided telephone conversations from 543 speakers (302 males and 241 females) in the United States. One topic is assigned to each of the conversation between two callers and about 70 topics in total are provided in the corpus.

However, a large number of utterances within the 300 hours Switchboard data do not fit into a clear topic and are not suitable for the topic spotting task (e.g., "Oh yeah", "Um-hum", "You are right."). The selection of utterances are based on the

length of the transcriptions after filtering out the filler words, functional words and stop words. We keep the utterances whose transcriptions have more than 20 words after the filtering. The threshold is set based on the trade-off between the utterance duration and the number of remaining utterances. After the first round of filtering, we have 9192 utterances left. We further sift out the topics that have less than 200 utterances and finally have 4405 utterances on 19 different topics for topic spotting task. From each topic, we randomly select 90% utterances for training and 10% for testing. (The 3955 training and 450 test utterances used for topic spotting are exactly the same as the ones used in [8].) The rest of the Switchboard-1 corpus is used for training the acoustic model.

4.2. Discriminative training of BLSTM acoustic model for lattice generation (WFST)

We first train a BLSTM acoustic model for the LVCSR system to generate word lattices that are suited for topic spotting. 36 dimensional log Mel filterbank features are extracted and then concatenated with 3 dimensional pitch features (consisting of probability of voicing, log pitch and delta log pitch) [39] to form a 39 dimensional "log Mel filterbank + pitch" feature. To build a BLSTM, we stack 4 hidden layers and add a softmax output layer on the top to represent the 8784 senones posteriors. Each forward or backward hidden layer has 1024 hidden units and is connected to a 512 dimensional recurrent projection layer. The forward and backword projection layers are concatenated together (to form a 1024 dimensional vector) and fed as the input of the next BLSTM hidden layer. After appending delta and delta-delta coefficients to the 39 dimensional "log Mel filterbank + pitch" features, we use the 117 dimensional features with globally normalized zero mean and unit variance as the input to the BLSTM. The BLSTM is randomly initialized and then trained (initial learning rate 0.00002) to minimize the cross-entropy (CE) criterion using senone-level forced alignment generated by a GMM-HMM system as the target.

The BLSTM in baseline system is then trained with the MSEC criterion as described in Section 3. The semantic similarity matrix S_{MSEC} is constructed with LSA using Eq. (19) with a rank of 750. The word-document matrix D is formed with the transcriptions of a subset of training utterances (2438 in total) in the topic spotting task. The diagonal elements of S_{MSEC} is set to the tf-idf weights of the corresponding word and the non-diagonal elements are scaled proportionally. Since the size of S_{MSEC} is very large (over $30,000 \times 30,000$), we prune the non-diagonal elements by keeping only the largest $80,000$ elements and setting the rest to zeros. For comparison, we also discriminatively train the baseline BLSTM with MWE and sMBR.

4.3. LSRK for topic spotting

We generate the decoded lattices for the 3955 training utterances using BLSTM based LVCSR systems trained in Section 4.2 and train a multi-class SVM with LSRK using these lattices for topic spotting. The LSRK is constructed with a semantic similarity matrix S_{LSRK} derived from the RNN LM learned vector-space word representations in Eq. (20). Each word in vocabulary is represented by a 4000 dimensional vector learned from a skip-gram model [37]. The skip-gram model is trained on billion bytes of the English Wikipedia[1] using the word2vec toolkit [37]. The out-of-vocabulary words are represented by zero vectors. We approximiate the S_{LSRK} with a matrix of rank K and pruned it to have M non-zero non-diagonal elements. The S_{LSRK} is scaled and pruned in the same way as we did for S_{MSEC}.

The topic classification accuracies for lattices generated by LVCSR trained with different objective function is shown in Table 1. The lattices generated by MSEC trained BLSTM achieve the best topic classification accuracy 60.00% when $M = 160,000; K = 400$ or $M = 240,000; K = 800$ or $M = 320,000; K = 800$, which is 3.5% - 4.5% absolutely higher than the baseline BLSTM trained with cross-entropy criterion. The BLSTM trained with MSEC also achieves 2%- 3% absolute accuracy gains over other discriminative training objectives. The classification accuracies vary slightly when M and K go beyond $160,000$ and 400 respectively.

Table 1: *The topic classification accuracies (%) of BLSTM-HMM systems trained with different objectives on the subset of Switchboard-1 Release 2. M is the number of non-zero non-diagonal elements left in S_{LSRK} after pruning and K is the rank of S_{LSRK} after low rank approximation.*

M	K	Objective Functions			
		CE	sMBR	MWE	MSEC
160,000	400	56.89	58.00	57.56	**60.00**
	800	56.67	58.00	57.56	59.33
	1200	56.67	57.78	57.33	58.56
240,000	400	55.78	58.00	57.78	59.56
	800	56.22	57.56	57.33	**60.00**
	1200	56.00	57.33	57.56	59.33
320,000	400	56.22	58.44	57.78	59.56
	800	55.56	58.00	57.78	**60.00**
	1200	56.00	58.00	57.33	59.56

4.4. Large-vocabulary continuous speech recognition

We evaluate LVCSR performance of the MSEC-trained BLSTM on the Switchboard portion of the 2000 HUB 5 English (LDC2002S09) and compare it with the other objectives. The ASR is conducted with the same acoustic and language models as the ones used in Sections 4.2 and 4.3 for topic spotting. The BLSTM trained with MSEC achieves 13.9% word error rate (WER), which is 0.7% and 0.8% absolutely lower than the BLSTMs trained with sMBR and MWE respectively as in Table 2. The degradation of LVCSR performanece is expected since MSEC is designed to minimize the expected semantic error cost instead of expected state or word errors as in sMBR or MWE.

Table 2: *The LVCSR performance of BLSTM-HMM systems trained with different objectives on the Switchboard portion of the 2000 HUB 5 English dataset.*

Objectives	CE	sMBR	MWE	MSEC
WER (%)	14.6	13.1	13.2	13.9

5. Conclusion

In this work, we successfully conduct the MSEC training of BLSTM to generate lattices for the topic spotting with LSRK on conversational spontaneous speech. LSA and RNN LM learned word representations are used to model the semantic similarities between words and sentences. We show that MSEC training of BLSTM can help LVCSR to generate lattices that are more semantically accurate and thus leads to higher topic classification accuracy than other training objectives.

[1] English Wikipedia is available on http://mattmahoney.net/dc/textdata.html

6. References

[1] C. Weng, D. L. Thomson, P. Haffner, and B. H. F. Juang, "Latent semantic rational kernels for topic spotting on conversational speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1738–1749, Dec 2014.

[2] J. H. Wright, M. J. Carey, and E. Parris, "Improved topic spotting through statistical modelling of keyword dependencies," in *Proceedings of ICASSP*, vol. 1, May 1995, pp. 313–316 vol.1.

[3] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may i help you?" *Speech Commun.*, vol. 23, no. 1-2, pp. 113–127, Oct. 1997.

[4] K. Myers, M. J. Kearns, S. P. Singh, and M. A. Walker, "A boosting approach to topic spotting on subdialogues," in *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, p. 655–662.

[5] C. Cortes, P. Haffner, and M. Mohri, "Rational kernels: Theory and algorithms," *J. Mach. Learn. Res.*, vol. 5, pp. 1035–1062, Dec. 2004.

[6] ——, "Lattice kernels for spoken-dialog classification," in *Proceedings of ICASSP*, vol. 1, April 2003, pp. I–628–31 vol.1.

[7] T. J. Hazen, F. Richardson, and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *2007 ASRU*, Dec 2007, pp. 659–664.

[8] C. Weng and B. H. Juang, "Latent semantic rational kernels for topic spotting on spontaneous conversational speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8302–8306.

[9] C. Fellbaum, *WordNet*. John Wiley & Sons, Inc., 2012.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, p. 391, Sep 01 1990, last updated - 2013-02-24.

[11] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, May 1997.

[12] Z. Meng and B.-H. Juang, "Non-uniform boosted mce training of deep neural networks for keyword spotting," in *Proceedings of INTERSPEECH*, 2016, pp. 770–774.

[13] ——, "Non-uniform mce training of deep long short-term memory recurrent neural networks for keyword spotting," in *Proceedings of INTERSPEECH*, 2017.

[14] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *ICASSP*, vol. 11, Apr 1986, pp. 49–52.

[15] V. Valtchev, J. Odell, P. Woodland, and S. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303 – 314, 1997.

[16] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 2005.

[17] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proceedings of ICASSP*, vol. 1, May 2002, pp. I–105–I–108.

[18] M. Gibson and T. Hain, "Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition." in *INTERSPEECH*. Citeseer, 2006.

[19] J. Kaiser, B. Horvat, and Z. Kacic, "A novel loss function for the overall risk criterion based discriminative training of hmm models," in *Sixth International Conference on Spoken Language Processing*, 2000.

[20] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted mmi for model and feature-space discriminative training," in *Proceedings of ICASSP*, March 2008, pp. 4057–4060.

[21] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, Mar 1994.

[22] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[23] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan 2012.

[24] C. Weng, D. Yu, S. Watanabe, and B. H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proceedings of ICASSP*, May 2014, pp. 5532–5536.

[25] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust asr," in *Proceedings of ICASSP*, March 2012, pp. 4085–4088.

[26] A. Graves, N. Jaitly, and A. r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 273–278.

[27] Z. Meng, S. Watanabe, J. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, March 2017.

[28] H. Erdogan, T. Hayashi, J. R. Hershey, T. Hori, C. Hori, W.-N. Hsu, S. Kim, J. Le Roux, Z. Meng, and S. Watanabe, "Multi-channel speech recognition: Lstms all the way through," in *The 4th International Workshop on Speech Processing in Everyday Environments*, 2016.

[29] A. Graves, A. r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6645–6649.

[30] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[31] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

[32] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[33] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *INTERSPEECH*, 2014, pp. 338–342.

[34] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *INTERSPEECH*, 2013, pp. 2345–2349.

[35] Andrew, H. Sak, F. de Chaumont Quitry, T. Sainath, and K. Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *2015 ASRU*, Dec 2015, pp. 604–609.

[36] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," in *INFORMATION PROCESSING AND MANAGEMENT*, 1988, pp. 513–523.

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[38] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *Proceedings of NAACL HLT*, vol. 13, 2013, pp. 746–751.

[39] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of ICASSP*, May 2014, pp. 2494–2498.