EXPLORING ASR-FREE END-TO-END MODELING TO IMPROVE SPOKEN LANGUAGE UNDERSTANDING IN A CLOUD-BASED DIALOG SYSTEM

Yao Qian, Rutuja Ubale, Vikram Ramanarayanan, Patrick Lange, David Suendermann-Oeft, Keelan Evanini, and Eugene Tsuprun

Educational Testing Service Research, USA

{yqian, rubale, vramanarayanan, plange, suendermann-oeft, kevanini, etsuprun}@ets.org

ABSTRACT

Spoken language understanding (SLU) in dialog systems is generally performed using a natural language understanding (NLU) model based on the hypotheses produced by an automatic speech recognition (ASR) system. However, when new spoken dialog applications are built from scratch in real user environments that often have sub-optimal audio characteristics, ASR performance can suffer due to factors such as the paucity of training data or a mismatch between the training and test data. To address this issue, this paper proposes an ASR-free, end-to-end (E2E) modeling approach to SLU for a cloud-based, modular spoken dialog system (SDS). We evaluate the effectiveness of our approach on crowdsourced data collected from non-native English speakers interacting with a conversational language learning application. Experimental results show that our approach is particularly promising in situations with low ASR accuracy. It can further improve the performance of a sophisticated CNN-based SLU system with more accurate ASR hypotheses by fusing the scores from E2E system, i.e., the overall accuracy of SLU is improved from 85.6% to 86.5%.

Index Terms— end-to-end, spoken language understanding

1. INTRODUCTION

Recent advances in deep learning with big data have significantly improved the performance of speech recognition, language understanding, and machine translation, which have in turn accelerated spoken dialog systems (SDSs) to move towards offering a more natural, intuitive, robust and effective interaction. However, when a cloud-based SDS is bootstrapped [1, 2], especially for realistic interactive speaking applications in the education domain, e.g., language learning, the system struggles to achieve high performance. This is because it is difficult to obtain large amounts of matched training data from real production environments when we are developing new applications from scratch, requiring them to be bootstrapped from relatively cleaner, but mismatched data. And as one might expect, spoken language understanding (SLU), which is the interpretation of the meaning conveyed by speech utterances, thereby playing a key role in deciding appropriate system actions in SDSs, can be adversely affected by data mismatch.

State-of-the-art SLU systems generally contain two components: the automatic speech recognizer (ASR), which decodes the input speech into text, and the natural language understanding (NLU) module, that transforms the ASR hypothesis into a concept or semantic label that can drive subsequent SDS behavior. Nowadays, the two components are typically based on statistical approaches trained on a large amount of data with various machine learning methods. Unlike NLU on written text, the efficiency of the SLU largely depends on the performance of ASR and its capability to handle errors and the vagaries of spontaneous speech, e.g., hesitations, corrections, repetitions, and other disfluencies.

Crowdsourcing techniques can allow us to rapidly and cheaply obtain data for bootstrapping a SDS. But dialog data crowdsourced from (i) *non-native English speakers* in (ii) *potentially adverse or uncontrolled audio environments* and (iii) collected over *poor internet connections* poses a significant challenge to SLU due to its dependence on ASR. We have observed a large variation in the quality of recorded speech due to the difficulty in controlling participants' recording equipment and environment. The poor audio quality could be either caused by wave distortions, e.g., clipping occurs when an amplifier is overdriven, or by packet loss resulting in dead silence when the internet transmission is unstable, or by large amounts of background noise resulting in low signal-to-noise ratio (SNR). Exacerbating this, as mentioned earlier, the non-native speech collected by SDS-based language learning applications may contain pronunciation errors, large numbers of disfluencies, ungrammatical phrases, loan words, etc., which make the ASR output even worse. Note that in such cases, human experts also find it difficult to transcribe such poor-quality non-native speech.

To address this issue, we propose an end-to-end modeling approach to SLU in a cloud-based SDS. The semantic labels are directly predicted from SLU models by using audio samples as inputs. Unlike two components used in the conventional SLU system, an ASR is not needed in our approach. The paper is organized as follows: We first review related work on SLU in SDS and the end-to-end approach in speech and language processing; After that, we introduce the open-source, cloud-based SDS we use to collect dialog data, along with the specific job interview conversational

task we analyze; Next, we present our approach to SLU for such an application together with an experiment design to validate the approach; Finally, we analyze the experimental results and discuss implications for future research and development.

2. RELATED WORK

Early attempts aimed at language understanding included computer programs such as STUDENT [3] which was developed at MIT to read and solve word problems found in high school algebra books and the chat-bot ELIZA [4] which used simple pattern matching to carry on a conversation on any topic. For most of the early language understanding systems, semantic parsers based on hand-crafted rules were widely used. In the 1990s, several research studies were carried out for the Airline Travel Information System (ATIS) project. The initial systems developed for this study used semantic rules to extract task specific information from slots in a frame. MIT's TINA [5], CMU's Phoenix [6, 7] and SRI's GEMINI [8] are examples of such knowledge-based systems. Although these systems were seen to perform very well, a major drawback of using hand-crafted rules is that it is time-consuming and laborious in terms of human effort to construct such rules. These rules are highly specific to the applications they were designed for and lack robustness to errors and irregularities. In real-world spoken dialog applications, new words and unseen speech utterances are encountered all the time thereby increasing the vocabulary and corpus size, and hand-crafted rules can result in misclassification for such utterances that are not covered by the fixed-grammar rules.

To reduce the amount of human effort in building SLU models, some statistical models were proposed such as AT&T's Chronus system [9] that applied a Markov model-based approach where a set of concepts corresponding to hidden states were used for semantic representation. Machine learning techniques were used in the BBN-HUM model [10] that was developed for the ATIS task for understanding sentences and extracting their meaning with respect to the preceding sentences. Some other statistical approaches to semantic parsing include semantic classification trees (SCTs) (decision trees with nodes representing regular expressions) in which semantic rules are learned automatically from the training corpus to build a natural language understanding system [11] or the application of a hidden vector state (HVS) model to hierarchical semantic parsing [12]. Most state-of-the-art techniques involve the use of deep learning for understanding based on transcriptions or ASR hypothesis [13, 14, 15, 16].

Recently, several research studies addressed the modeling of speech signals using end-to-end (E2E) optimization, which utilizes as little a prior knowledge as possible, e.g., using filter-bank features instead of MFCC [17] or directly using speech waveform [18]. Multiple studies have demonstrated that features automatically extracted by DNNs are far superior to those produced by feature-engineering techniques generally used in GMM-based acoustic modeling, e.g. [19]. E2E speech recognition systems have yielded competitive performance compared to conventional hybrid DNN-HMM systems [20, 21, 22]. E2E learning also has produced promising results on speaker verification [23], language identification [24], emotion recognition [25] and keyword search [26]. To the best of our knowledge, there is few research work exploring ASR-free E2E modeling for SLU, although several studies have tried to limit or suppress the need of ASR for performing classification [27, 28, 29, 30].

Table 1: An example of different responses (along with corresponding gold-standard semantic labels) to one particular dialog state ("mistake") in the job interview task that deals with how the interviewee would deal with a co-worker's mistake.

Question	Imagine you saw your coworker make a mistake. Which do you think would be better? To tell the co-worker about the mistake or to speak with your manager?
Response	I would talk to the team member and ask him to rectify their mistake and it is a better way of resolving the issue.
Semantic Label	coworker
Response	Speaking with the manager is the best thing I guess.
Semantic Label	manager
Response	Yeah if it is a normal issue, then I'll go and discuss with the uh uh coworker himself. If it is something big, then I'll go to manager and I will discuss with him and we will come to the solution.
Semantic Label	depends
Response	Uh uh currently I am staying in India. Eh.
Semantic Label	nomatch

3. SPOKEN DIALOG SYSTEM AND TASK

We use an SDS that leverages different open-source components to form a framework that is cloud-based, modular and standards-compliant. For more details on the architectural components, please refer to [31]. This framework is employed to develop conversational applications and collect data using a crowdsourcing setup. In this iterative data collection framework, the data logged to the database during initial iterations is transcribed, annotated, rated, and finally

Table 2: Dialog state and semantic labels

Dialog State	Semantic Labels
Mistake (MT)	coworker, depends, manager, nomatch
Part or Full (PF)	either, full time, nomatch, part time
Self or Group (SG)	both, group, nomatch,self
Work Experience (WE)	yes, no, nomatch

used to update and refine the conversational task design and models for speech recognition and spoken language understanding [1]. Since the targeted domain of the tasks in this study is conversational practice for English language learners, we restricted the crowdsourcing user pool to non-native speakers of English.

This study examines a conversational task developed for English language learners that was designed to provide speaking practice for non-native speakers of English in the context of a simulated job interview. The conversation is set up as a system-initiated dialog in which a representative at a job placement agency interviews the language learner about the type of job they are looking for and their qualifications. An example of dialog states in the job interview task, including question, responses, and corresponding semantic labels, is shown in Table 1. Table 2 comprehensively lists the possible semantic labels associated with each dialog state. The ultimate aim of the task is to provide interactive feedback to language learners about whether they have demonstrated the linguistic skills necessary to provide appropriate, intelligible responses to the interviewer's questions and to complete the communicative task successfully.

4. END-TO-END MODELING OF SLU

The conventional ASR+NLU approach to SLU requires a decent ASR system. Generally it needs over a hundred hours of speech collected under real usage scenarios (along with associated transcriptions for acoustic and language modeling) to obtain a reasonable ASR system performance [32]. This is an important factor to take into consideration when one uses deep learning methods, the recognition performance increases monotonically with more training data [33]. Any new application can be continuously improved by using a cycle of data collection. In this study, we investigate the potentials to build an ASR-free end-to-end model for SLU.

The task of predicting semantic labels for spoken utterances from the job interview conversations can be treated as a semantic utterance classification task, which aims at classifying a given utterance into one of M semantic classes, $\hat{c^k} \in \{c_1^k, ..., c_M^k\}$, where k is the dialog state index. A straight-forward way to model semantic utterance classification is to use a sequence-to-tag function, where the input X is a sequence of speech feature vectors, $X = \{x_1, x_2, ..., x_T\}$; x_t is the speech feature vector, e.g.,

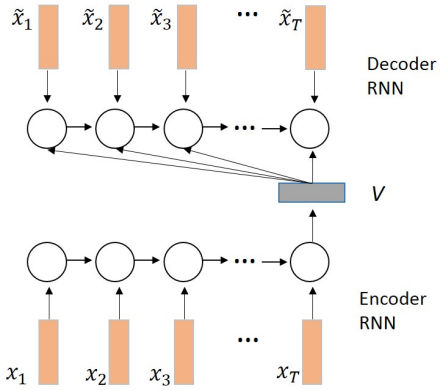

Fig. 1: RNN acoustic auto-encoder

MFCC, in t-th frame; T is the total number of frames in an utterance, and output C is the semantic label. Recurrent Neural Networks (RNNs) can use their internal memory to process an arbitrary length of inputs and are successfully applied to solve a wide range of machine learning problems that involve sequential data. We try to use RNNs to learn a sequence-to-tag function for predicting semantic labels from speech. Frame-level speech features are used as input layer. The output layer is a softmax layer which contains dialog-state-dependent semantic labels represented by a one-hot vector. However, the preliminary results are not promising. We conjecture that it is suffering from the limited training data. Speech acoustic features vary largely from the factors, e.g., age, gender, dialectal background and personal style. Even for the same speaker, the actual values change from time to time due to different phone sequences spoken. Therefore, a large number of spoken utterance with semantic categories is required for training to get a decent classifier.

We propose to use a compact representation for the utterance in variable length and then employ the resultant low-dimensional feature vector to do semantic label modeling. Our approach is inspired by two popular techniques: (i) pre-training [34, 35], which initializes DNN weights to a better starting point than random initialization prior to back-propagation (BP), which in turn helps facilitate a rapid convergence of the BP process, and (ii) the auto-encoder, which is used to learn a compact lower-dimensional feature representation of a higher-dimensional input feature vector sequence.

Two approaches of compact audio feature representation by using unsupervised learning are explored in this study. One is an RNN-based acoustic auto-encoder [26, 36] as shown in Figure 1. It depicts the structure of a sequence-to-sequence auto-encoder which contains two RNNs: Encoder RNN and Decoder RNN. The acoustic feature vector sequence $\{x_1, x_2, ..., x_T\}$ is mapped into a vector representation in a fixed dimensionality V by the Encoder RNN, and the Decoder RNN reconstructs another sequence $\{\hat{x}_1, \hat{x}_2, ..., \hat{x}_T\}$ from vector V to minimize the reconstruction error, generally measured as the mean squared error between X and $\hat{X}$.

Another approach is to represent a variable length

speech utterance into a low-dimensional subspace based on factor analysis. The t-th frame of an utterance, x_t, is sampled from the following distribution:

$$x_t \sim \sum_j \gamma_{jt} N(m_j + T_j V, \Sigma_j) \quad (1)$$

where m_j and Σ_j are the mean and covariance of the j-th Gaussian component if a Gaussian mixture model (GMM) is used to train a universal background model (UBM). A GMM is an efficient method for modeling an arbitary distribution of acoustic feature vectors in a unsupervised manner using the EM algorithm. γ_{jt} is the statistical alignment result of the frame x_t, i.e., the posterior probability calculated from a UBM; T_j is the total variability, a low-rank rectangular matrix which is estimated by using the EM algorithm; V is the utterance-specific standard normal distributed latent vector obtained by using maximum a posterior (MAP) estimation.

Transfer learning or multi-task learning [37] can exploit commonalities between the training data of different learning tasks so as to transfer learned knowledge from one to another. We use multi-task learning to the semantic utterance classification by assuming each dialog state as one task. The schematic diagram of our approach is shown in Figure 2 where the input layer is the fixed-dimensional vector V output from either RNN encoder or factor analysis as the representation of variable length acoustic feature vector sequence and output layer is the softmax layer with K one-hot vectors (each vector represents one dialog state).

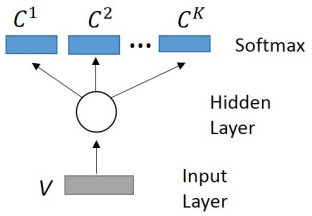

Fig. 2: Transfer learning with feedforward NN

5. EXPERIMENTS

Our ASR-free E2E modeling approach is evaluated in a spoken-dialog-based language learning application by comparing with the conventional approach of combining ASR and NLU.

5.1. Corpora

The application collected spoken dialog data via crowdsourcing by interacting with non-native interlocutors in a job interview task. The dialog state and the corresponding semantic labels are shown in Table 2. The collected dialog corpus consists of 4,778 utterances spoken by 1,179 speakers. 4,191 utterances are used as a training set and the rest of 586 utterances are used as a testing set. 200 utterances randomly selected from the corpus are used to manually check the audio quality by reading the waveform and spectrogram together with listening to the sound. We found the percentage of labels for bad quality (perceptibly clipping distortion, packet loss or large background noise), no voice and good quality are 62.5%, 8.5% and 29%, separately. The quality of transcriptions is also checked by Levenstein distance between the transcriptions from different transcribers for the same utterance, i.e., calculating word error rate (WER) by assuming one transcription is reference and another one is recognition hypotheses. It shows that the average inter-transcriber WER is 38.3% measured on 1,004 utterances/10,288 tokens. This corpus is hereafter referred to as the job interview task (JIT) corpus.

Two corpora are used to build our ASR system. One is drawn from a large-scale global assessment of English proficiency, which measures a non-native speaker's ability to use and understand English at the university level. The speaking tasks in this test elicit monologues of 45 or 60 seconds in duration; example tasks include expressing an opinion on a familiar topic or summarizing information presented in a lecture. It contains over 800 hours of non-native spontaneous speech covering over 100 L1s (native languages) across 8,700 speakers. This corpus is hereafter referred to as the non-native speech (NNS) corpus.

Another one is collected by our SDS via crowdsourcing for different spoken dialog based applications. Job interview task is one of the tasks based on language learning application. This corpus is collected under realistic usage scenarios. The acoustic environments and speaking styles were matched with the data of job interview task. It contains 41,185 utterances (roughly 50 hours). This corpus is hereafter referred to as the SDS corpus.

5.2. ASR+NLU

ASR systems are constructed by using the tools from Kaldi [38]. A GMM-HMM is first trained to obtain senones (tied tri-phone states) and the corresponding aligned frames for DNN training. The input feature vectors used to train the GMM-HMM contain 13-dimensional MFCCs and their first and second derivatives. Contextual dependent phones, tri-phones, are modeled by 3-state HMMs and the pdf of each state is represented by a mixture of 8 Gaussian components. The splices of 9 frames (4 on each side of the current frame) are projected down to 40-dimensional vectors by linear discriminant analysis (LDA), together with maximum likelihood linear transform (MLLT), and then used to train the GMM-HMM by using maximum likelihood estimation. Concatenated MFCC features and i-vector features, which is a promising approach to speaker adaptation for speech recognition, are used for DNN training. The input features stacked over a 15 frame window (7 frames to either side of the center frame for which the prediction is made) are used as the input layer of DNN. The output layer of the DNN consists of the senones of the HMM obtained by decision-tree based clustering. The input and output feature pairs are obtained by frame alignment for senones with the GMM-HMM. The DNN has 5 hidden layers, and each layer contains 1,024 nodes. The Sigmoid activation function is used for all hidden layers. All the parameters of the DNN are firstly initialized by pre-training, then

trained by optimizing the cross-entropy function through BP, and finally refined by sequence-discriminative training, state-level minimum Bayes risk (sMBR).

The Bag of Words model is used as a feature for training NLU. In this model, a text string (recognized hypothesis sequence) is represented as a vector based on the occurrence of each word. Dialog state dependent models are trained to perform multi-class classification of Bag of Words features using decision tree classifier. Apart from the conventional NLU method, the approach of using convolutional neural networks (CNNs) is also investigated in this study. The input tokenized text string is firstly converted to a 2D tensor with the shape (maximum length of word * the dimension of word-embedding), and then fed into a 1D convolution network with multipe filters, finally the maximum values from all filters via max pooling are formalized as a vector to predict the sematic labels by softmax output layer. The CNN is constructed using the Keras Python package[1] and the structure of CNN is configured as follows: 300-dim word embedding vectors trained from Google news; the rectified linear unit (ReLU) acitivation function and dropout with (p=0.5); categorical cross-entropy loss function and Adadelta optimizer used in the training.

5.3. ASR-free E2E

To overcome vanishing gradient problem occurred in RNN-based machine learning, long short-term memory (LSTM) [39] RNN is used for RNN encoder-decoder. The input features to the LSTM-RNN is 13-dim static MFCCs without delta features and stacked frame window since RNN architecture already captures the long-term temporal dependencies between the sequential events. The silences at the beginning and ending of utterances are deleted through an energy-based voice activity detection (VAD) method. A two layer stacked LSTM is employed. The number of LSTM cell is 640. We unfolded both encoder and decoder RNN for 10 seconds or 1,000 time steps. 10 seconds is the median length of utterances in our corpus. All feature sequences are either padded or downsampled to make their length equal to 1,000 time steps. A linear layer with 400 nodes, i.e., the dimension of V in the Figure 1 is 400, is used to compute the embedding from the final hidden layer of the encoder RNN. A back-propagation through time (BPTT) learning algorithm is used to train LSTM-RNN parameters.

The acoustic features used for factor analysis contain 13 dimensional MFCCs along with their first and second derivatives. Non-speech segments within utterances were deleted through the same VAD method used in auto-encoder. Utterance-based cepstral mean normalization was performed on the acoustic feature vectors. A GMM with 1,024 components and a full covariance matrix was trained as the UBM. To make a fair comparison to LSTM-RNN encoder-decoder based feature representation, the same dimensional latent vector, i.e., 400-dim V used in Equation 1, is extracted from T-matrix trained by EM algorithm with the training set.

Two hidden layers, each layer with 128 nodes, are used for multitask learning with feedforward NN. Input layer of NN is 400-dim V and output layer of NN has 15 nodes separated by four tasks. All parameters of NN are trained by optimizing cross-entropy function through BP. The parameters in hidden layers are updated by using all data in the training set of JIT corpus while the corresponding dialog state dependent data is used to update the parameters in the top layer of NN.

5.4. Fusion of ASR-free E2E system and ASR+NLU system

ASR-free E2E system predicts the semantic labels from low-level raw acoustic features while ASR+NLU system predict the semantic labels from high-level word hypotheses. These two systems can compensate for each other. We adopt score-level fusion by using the semantic label posterior outputs generated from two neural networks as the input features to a support vector classifier to predict the semantic labels again.

5.5. Experimental results and analysis

We employ WER and prediction accuracy to evaluate the performance of ASR and SLU on the testing set of JIT corpus. Our proposed ASR-free E2E approach does not require any transcriptions from the three corpora mentioned in Section 5.1. The acoustic features extracted from NNS and SDS corpora are used to train LSTM-RNN auto-encoder and the GMM for factor analysis in the sense of unsupervised learning. The preliminary results show that the performance of LSTM-RNN and factor analysis has no significant difference regarding the extraction of compact representation V from variable length of utterance in our tasks. V extracted from factor analysis slightly outperforms that extracted from LSTM-RNN. So we show its results as the results of ASR-free E2E approach hereafter.

The performance of different ASR systems, in terms of WER, on the testing set of JIT corpus is shown in Table 3. The WERs are broken down by dialog state as well as those of overall (All) and the reference (Ref), which are tested on the matched data sets. The state-of-the-art DNN-based ASR trained on the Fisher corpus [40] using Kaldi can achieve 22.2% WER on its own testing set [41]. Although the Fisher corpus is a collection of conversational telephone speech, it still has a significant mismatch with the speech collected by SDS and results in a very high WER. The DNN-based ASR system with i-Vector based speaker adaptation technology trained on the NNS corpus (which is also a collection of non-native speakers' speech), can obtain the WERs of 18.5% and 23.3% on monologue and dialogue data sets respectively (using LM interpolation technology to compensate for the speaking style difference across tasks) [42]. However, when it is applied to recognize the data collected by SDS, the WER is degraded to 55.5% even if we use the transcriptions from the training set of the JIT corpus for language model adaptation. Using data collected by SDS or combining NNS corpus with SDS corpus can significantly improve the performance of ASR, i.e., the WERs on JIT testing set are reduced to 49.4% and 43.5%, respectively. While this is still a very large WER value, we should

[1] https://keras.io

Table 3: WER(%) of ASR systems built with different corpora

Dialog State	PF	WE	SG	MT	All	Ref
Fisher	86.4	88.8	84.2	90.9	88.1	22.2
NNS	54.3	62.6	52.0	54.8	55.5	18.5
SDS	35.4	55.8	45.0	55.1	49.4	N/A
NNS+SDS	35.8	50.1	39.5	46.1	43.5	N/A

Table 4: Accuracy(%) of different SLU systems

Dialog State	PF	WE	SG	MT	All
E2E (JIT)	56.3	79.4	53.7	75.4	64.1
E2E (NNS)	62.3	81.2	54.1	76.4	66.7
E2E (NNS+SDS)	63.0	82.4	54.9	76.8	67.4
ASR+NLU (NNS), DT+BW	64.6	75.5	63.4	72.5	68.0
ASR+NLU (SDS), DT+BW	79.0	87.3	59.2	75.4	74.0
ASR+NLU (NNS+SDS), DT+BW	81.8	87.3	67.1	77.5	77.6
ASR+NLU (Transcription), DT+BW	82.9	92.2	57.9	53.6	70.6
ASR+NLU (NNS+SDS), CNN	89.0	91.2	79.3	84.8	85.6
Fusion (CNN, E2E))	89.0	94.1	79.9	85.5	86.5
Majority Vote	53.6	79.4	45.7	70.3	59.8

contexualize this result in light of the fact that the average inter-transcriber WER is also quite large at 38.3 %. Reducing both the ASR and inter-transcriber WERs for such data are crucial to improving system performance in real-world environmental conditions and usage scenarios, and pose an interesting challenge to the speech processing research community going forward.

Table 4 shows the performance of SLU in terms of semantic prediction accuracy from different systems. The corpora in the bracket for E2E systems indicate the corpora used to train total variability matrix, which is employed to project the variable length utterance to fixed length feature vector V. Our E2E approach performs much better than the majority vote baseline, i.e., the accuracy is improved from 59.8% to 64.1% and there is no degradation for dialog state-dependent performance. SDS and NNS corpora can cover large amount of acoustic variations and V extractor trained on them can yield superior discrimination for semantic classification. The overall accuracy of E2E (NNS+SDS) is improved by 3.3%, comparing with that of E2E (JIT), where the V extractor is trained on JIT corpus, and the dialog state of PF (Part or Full) achieves the largest gains among the four dialog states, i.e., the accuracy is improved by 6.7%.

The SLU performance of conventional ASR+NLU systems are also shown in Table 4. The corpora in the bracket for ASR+NLU systems indicate the corpora used for building ASR system. Clearly, the observed trend is that the lower the ASR WER, the higher the accuracy of the SLU. It is interesting that the SLU trained on transcription does not be able to outperform the SLU trained on the hypothesis produced by ASR system. We suspect that it might be caused by the inconsistency and the ambiguity present in the human transcriptions we commissioned. The E2E SLU system of E2E (NNS), which doesn't require any transcription for modeling and ASR system for transcribing the spoken utterance into text, can be on par with the system of ASR+NLU (NNS).

The decision tree classifier and the Bag of Words features (DT+BW) are used as the conventional NLU in this study due to the latency issue, which requires a fast response in a cloud-based dialog system. Following our research curiosity, we tried CNN-based semantic utterance classification approach and fused it with ASR-free E2E approach. The experimental results show that the semantic label prediction accuracy can be significantly improved, i.e., the overall accuracy is improved from 77.6% to 85.6%, by CNN approach, and the score-level fusion by using posteriors output from these two approaches can further improve the accuracy from 85.6% to 86.5%.

6. CONCLUSIONS

In this paper, we developed an automatic speech recognition (ASR)-free end-to-end spoken language understanding (SLU) module for a job interview-based language learning dialog application. Given an utterance, we first projected a variable-length sequence of acoustic feature vectors onto a low-dimensional fixed-length vector and then fed the resulted vector into a feedforward neural network trained in the sense of transfer learning to predict its semantic category. The evaluation results show that our SLU approach of directly predicting semantic labels from speech is a promising alternative to traditional methods when a decent ASR system for more realistic, noisy usage scenarios is unavailable. In addition, since no ASR is required, we found the training time and semantic decoding time of our proposed approach to be much faster than conventional approaches.

7. REFERENCES

[1] V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, A. V. Ivanov, K. Evanini, Z. Yu, E. Tsuprun, and

Y. Qian, "Bootstrapping development of a cloud-based spoken dialog system in the educational domain from scratch using crowdsourced data," *ETS Research Report Series, Wiley. doi: 10.1002/ets2.12105*, 2016.

[2] V. Ramanarayanan, D. Suendermann-Oeft, A. V. Ivanov, and K. Evanini, "A distributed cloud-based dialog system for conversational application development," *In 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, page 432.*, 2015.

[3] D. Bobrow, "Natural language input for a computer problem solving system.," 1964.

[4] J. Weizenbaum, "ELIZA - A computer program for the study of natural language communication between man and machine.," *Communications of the ACM, 9(1):36-45.*, 1966.

[5] S. Seneff, "TINA: A natural language system for spoken language applications.," *Computational linguistics, 18(1):61-86.*, 1992.

[6] S. Issar and W. Ward, "Cmu's robust spoken language understanding system.," *In Proceedings of Eurospeech, volume 93.*, 1993.

[7] W. Ward and S. Issar, "Recent improvements in the cmu spoken language understanding system.," *In Proceedings of the workshop on Human Language Technology, pages 213-216. Association for Computational Linguistics.*, 1994.

[8] J. Dowding, J. Gawron, J. Bear D. Appelt, L. Cherny, R. Moore, and D. Moran, "Gemini: A natural language system for spoken-language understanding.," *In Proceedings of the 31st annual meeting on Association for Computational Linguistics, pages 54-61. Association for Computational Linguistics.*, 1993.

[9] E. Levin and R. Pieraccini, "Concept-based spontaneous speech understanding system.," *In Fourth European Conference on Speech Communication and Technology.*, 1995.

[10] R. Schwartz, S. Miller, D. Stallard, and J. Makhoul, "Language understanding using hidden understanding models.," *In Spoken Language. ICSLP 96. Proceedings, Fourth International Conference on, volume 2, pages 997-1000. IEEE.*, 1996.

[11] R. Kuhn and R. De Mori, "The application of semantic classification trees to natural language understanding.," *IEEE transactions on pattern analysis and machine intelligence, 17(5):449-60.*, 1995.

[12] Y. He and S. Young, "A data-driven spoken language understanding system.," *IEEE transactions on pattern analysis and machine intelligence, 17(5), pages 449-460.*, 2003.

[13] G. Mesnil, X. He, L. Deng, and Y.Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding.," *In Interspeech, pages 3771-3775.*, 2013.

[14] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, D. Yu G. Tur, and G. Zweig, "Using recurrent neural networks for slot filling in spoken language understanding.," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 23(3), pages 530-539.*, 2015.

[15] P. Xu and R. Sarikaya, "Convolutional neural network based triangular crf for joint intent detection and slot filling.," *In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pages 78–83. IEEE.*, 2013.

[16] G. Tur, D. Hakkani-Tur, L. Heck, and S. Parthasarathy, "Sentence simplification for spoken language understanding.," *In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 5628-5631. IEEE.*, 2011.

[17] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," *in Proc. ICML, Beijing, China, volume 14, pages 1764-1772.*, 2015.

[18] N. Jaitly and G. Hinton, "Learning a better representation of speech sound waves using restricted boltzmann machines," *in Proc. ICASSP, pages 5884-5887*, 2011.

[19] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," *in Proc. of IEEE ASRU, pages 24–29.*, 2011.

[20] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," *in Proc. ASRU. IEEE, pages 167-174*, 2015.

[21] D. Bahdanau, J. Chorowski, D. Serdyuk, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *in Proc. ICASSP. IEEE, pages 4945-4949*, 2016.

[22] M. Bhargava and R. Rose, "Architectures for deep neural network based acoustic models defined over windowed speech waveforms," *in Proc. INTERSPEECH, pages 6-10*, 2015.

[23] G. Heigold, I. Mereno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," *In Proc. ICASSP, pages 5115-5119.*, 2016.

[24] W. Geng, W. Wang, Y. Zhao, X. Cai, and B. Xu, "End-to-end language identification using attention-based recurrent neural networks," *In Proc. Interpseech, pages 2944-2948.*, 2016.

[25] G. Trigeorigis, F. Ringeval, R. Brucekner adn E. Marchi, M.A. Nicolaou, B. Schuller, and S. Zafeirou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," *In Proc. ICASSP, pages 5200-5204.*, 2016.

[26] K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, and B. Kingsbury, "End-to-end asr-free keyword search from speech," *in Proc. of ICASSP.*, 2017.

[27] Q. Huang and S. Cox, "Task-independent call-routing," *Speech Communication, volume 48 (3), pages 374-389.*, 2006.

[28] A. L. Gorin, D. Petrovska-Delacretaz, G. Riccardi, and J. H. Wright, "Learning spoken language without transcriptions," *In Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, Volume 99*, 1999.

[29] H. Alshawi, "Effective utterance classification with unsupervised phonotactic models," *In Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, page 1-7,Association for Computational Linguistics.*, 2003.

[30] Y. Y. Wang, J. Lee, and A. Acero, "Speech utterance classification model training without manual transcriptions," *In Proc. ICASSP, Volume 1, pp. 553-556*, 2006.

[31] V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, R. Mundkowsky, A. Ivanov, Z. Yu, Y. Qian, and K. Evanini, "Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System," in *Multimodal Interaction with W3C Standards*, pp. 295–310. Springer, 2017.

[32] G. Hinton, L. Deng, D. Yu, E. George Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. 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, 2012.

[33] J. Cheng, X. Chen, and A. Metallinou, "Deep neural network acoustic models for spoken assessment applications," *Speech Communication, volume 73, pages 1427.*, 2015.

[34] D. Yu, L. Deng, and G. Dahl, "Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition," *in Proc. of NIPS Workshop on Deep Learning and Unsupervised Feature Learning.*, 2010.

[35] D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, , and S. Bengio, "Why does unsupervised pre-training help deep learning?," *JMLR.*, 2010.

[36] Y. Chung, C. Wu, C. Shen, H. Lee, and L. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *in Proc. of Interspeech.*, 2016.

[37] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence, 35(8):1798-1828.*, 2013.

[38] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," *In IEEE 2011 workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584). IEEE Signal Processing Society.*, 2011.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation, 9(8):1735-1780.*, 1997.

[40] J. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text," *in LREC, volume 4, pages 69-71*, 2004.

[41] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," *in Proc. of INTERSPEECH, pages 3214-3218*, 2015.

[42] Y. Qian, X. Wang, K. Evanini, and D. Suendermann-Oeft, "Self-adaptive dnn for improving spoken language proficiency assessment," *in Proc. of Interspeech, 2016*, 2016.