Towards End-to-End Modeling of Spoken Language Understanding in a Cloud-based Spoken Dialog System

Yao Qian, Rutuja Ubale, Vikram Ramanaryanan, Patrick L. 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

We present an ASR-free end-to-end modeling approach to spoken language understanding for a cloud-based modular spoken dialog system. 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 performs almost as well as the traditional baseline of ASR-based semantic classification and is particularly promising in situations with low ASR accuracy.

1 Introduction

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).

Recently, several research studies have investigated models of the speech signal using an end-to-end (E2E) approach that utilizes as little a priori knowledge as possible, e.g., by using filter-bank features instead of MFCCs (Graves and Jaitly, 2015) or by directly using speech waveforms (Jaitly and Hinton, 2011). E2E speech recognition systems have yielded competitive performance compared to conventional hybrid DNN-HMM systems (Miao et al., 2015) and E2E models have also produced promising results on speaker verification (Heigold et al., 2016) and language identification (Geng et al., 2016). However, to the best of our knowledge, no studies have yet explored ASR-free E2E modeling for the task of SLU.

2 Methodology

Our experiments use an SDS that leverages a variety of open source components in a framework that is cloud-based, modular and standards compliant; (Ramanarayanan et al., 2017) provides further details about the SDS architecture. This study examines an interactive conversational task for English language learners designed to provide speaking practice in the context of a simulated job interview. The conversation is structured as a system-initiated dialog in which a representative at a job placement agency interviews the language learner about his or her job interests and qualifications.

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 classisfying 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. This study explores two approaches to compact audio feature representation using unsupervised learning. In the first approach, an RNN-based acoustic autoencoder maps the acoustic feature vector sequence onto a fixed-dimensional vector. In the second approach, factor analysis is used to transform the

Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue, August 15-17, 2017, Saarbrücken, Germany.

variable length spoken utterance into a low-dimensional subspace. As shown in Figure 1, the fixed-dimensional vector, V, generated by either the RNN encoder or factor analysis, is the input layer to the SLU model; the output layer is the softmax layer with K one-hot vectors (each vector represents one dialog state); Multi-task learning is used here by assuming each dialog state as one task; K=4 and M=3 or 4 are used in this study.

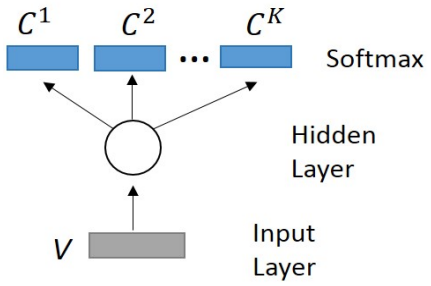

Figure 1: Transfer learning with feedforward NN

3 Experimental Results

A corpus of 4,778 utterances for the job interview task provided by 1,179 speakers was collected via crowdsourcing. 4,191 utterances are used as the training set and the remaining 586 utterances are used as the test set. Based on 1,004 utterances (10,288 tokens), the inter-transcriber word error rate (WER) is 38.3%. It is largely suffering from the poor audio quality, which could be either caused by waveform distortions, e.g., clipping occurs when an amplifier is overdriven, or dead silence caused by packet loss when the internet transmission is unstable, or low signal-to-noise ratio (SNR) in general due to large amounts of background noise. Two corpora are used to build our ASR system. One corpus (NNS) is drawn from a large-scale global assessment of English proficiency and contains over 800 hours of non-native spontaneous speech covering over 100 native languages across 8,700 speakers. Another corpus (SDS) was collected using our SDS via crowdsourcing with several different spoken dialog applications, including the job interview conversation task, and contains approximately 50 hours of speech. The experimental results show that there was no significant difference between the performance of the autoencoder and factor analysis approaches to extracting compact representations of the audio signal. Table 1 presents the performance of the ASR and SLU systems (E2E and ASR+NLU) on the test set and shows that the E2E system's accuracy is closest to the ASR+NLU system's accuracy when the ASR WER is the highest. NLU system performs multi-class classification of Bag of Words features extracted from the recognized hypotheses using decision tree classifier. As a reference, the SLU accuracy of a majority vote baseline is 59.8%.

Table 1: WER and SLU accuracy using three ASR systems and two SLU systems (E2E and ASR+NLU)

Corpus	**ASR**	**E2E**	**ASR+NLU**
NNS	55.5	64.1	68.0
SDS	49.4	66.7	74.0
NNS + SDS	43.5	67.4	77.6

References

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

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

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

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

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

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