Deep Learning for Dialogue Systems

Yun-Nung Chen
National Taiwan University
Taipei, Taiwan
yvchen@ieee.org

Asli Celikyilmaz
Microsoft Research
Redmond, WA
asli@ieee.org

Dilek Hakkani-Tür
Google Research
Mountain View, CA
dilek@ieee.org

Abstract

In the past decade, goal-oriented spoken dialogue systems have been the most prominent component in today's virtual personal assistants. The classic dialogue systems have rather complex and/or modular pipelines. The advance of deep learning technologies has recently risen the applications of neural models to dialogue modeling. However, how to successfully apply deep learning based approaches to a dialogue system is still challenging. Hence, this tutorial is designed to focus on an overview of the dialogue system development while describing most recent research for building dialogue systems and summarizing the challenges, in order to allow researchers to study the potential improvements of the state-of-the-art dialogue systems. The tutorial material is available at http://deepdialogue.miulab.tw.

1 Tutorial Overview

With the rising trend of artificial intelligence, more and more devices have incorporated goal-oriented spoken dialogue systems. Among popular virtual personal assistants, Microsoft's Cortana, Apple's Siri, Amazon Alexa, Google Assistant, and Facebook's M, have incorporated dialogue system modules in various devices, which allow users to speak naturally in order to finish tasks more efficiently.

The traditional conversational systems have rather complex and/or modular pipelines. The advance of deep learning technologies has recently risen the applications of neural models to dialogue modeling. Nevertheless, applying deep learning technologies for building robust and scalable dialogue systems is still a challenging task and an open research area as it requires deeper understanding of the classic pipelines as well as detailed knowledge on the benchmark of the models of the prior work and the recent state-of-the-art work.

The goal of this tutorial is to provide the audience with developing trend of the dialogue systems, and a roadmap to get them started with the related work. In the first section of the tutorial, we motivate the work on conversation-based intelligent agents, in which the core underlying system is task-oriented dialogue systems. The second and third sections describe different approaches using deep learning for each component in the dialogue system and how it is evaluated. The last two sections focus on discussing the recent trends and current challenges on dialogue system technology and summarize the challenges and conclusions. Then the detailed content is described as follows.

2 Outline

1. Introduction & Background [15 min.]
   - Brief history of dialogue systems
   - Summarized challenges of intelligent assistants
   - Task-oriented dialogue system framework
   - Neural network basics
   - Reinforcement learning (RL) basics
2. Deep Learning Based Dialogue System [75 min.]
   - Spoken/Natural language understanding (SLU/NLU)
     - Semantic frame representation
     - Domain classification
     - Slot tagging
     - Joint semantic frame parsing
     - Contextual language understanding

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-Tutorial Abstracts, pages 8–14
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-5004

    - Structural language understanding
  - Dialogue management (DM) – Dialogue state tracking (DST)
    - Neural belief tracker
    - Multichannel tracker
  - Dialogue management (DM) – Policy optimization
    - Dialogue RL signal
    - Deep Q-network for learning policy
    - Hierarchical RL for learning policy
  - Natural language generation (NLG)
    - Template-based NLG
    - Plan-based NLG
    - Class LM NLG
    - Phrase-based NLG
    - RNN-LM NLG
    - Semantic Conditioned LSTM
    - Structural NLG
    - Contextual NLG
3. Evaluation [10 min.]
  - Crowdsourcing
  - User simulation
4. Recent Trends on Learning Dialogues [45 min.]
  - End-to-end neural dialogue systems
    - Chit-chat seq2seq model
    - E2E joint NLU and DM
    - E2E supervised dialogue system
    - E2E memory network for dialogues
    - E2E RL-based *InfoBot*
    - E2E LSTM-based dialogue control
    - E2E RL-based task-completion bot
  - Dialogue breath
    - Domain adaptation
    - Intent expansion
    - Policy for domain adaptation
  - Dialogue depth
    - High-level intent for dialogue planning
    - Multimodality in dialogue systems
5. Challenges & Conclusions [5 mins]

3 Dialogue System Basics

This section will motivate the work on conversation-based intelligent agents, in which the core underlying system is task-oriented spoken dialogue systems.

The section starts with an overview of the standard pipeline framework for dialogue system illustrated in Figure 1 (Tur and De Mori, 2011). Basic components of a dialog system are automatic speech recognition (ASR), language understanding (LU), dialogue management (DM), and natural language generation (NLG) (Rudnicky et al., 1999; Zue et al., 2000; Zue and Glass, 2000).. This tutorial will mainly focus on LU, DM, and NLG parts.

Language Understanding Traditionally, domain identification and intent prediction are framed as utterance classification problems, where several classifiers such as support vector machines and maximum entropy have been employed (Haffner et al., 2003; Chelba et al., 2003; Chen et al., 2014). Then slot filling is framed as a word sequence tagging task, where the IOB (in-out-begin) format is applied for representing slot tags, and hidden Markov models (HMM) or conditional random fields (CRF) have been employed for slot tagging (Pieraccini et al., 1992; Wang et al., 2005; Raymond and Riccardi, 2007).

Dialogue Management A partially observable Markov decision process (POMDP) has been shown to be beneficial by allowing the dialogue manager to be optimized to plan and act under the uncertainty created by noisy speech recognition and semantic decoding (Williams and Young, 2007; Young et al., 2013). The POMDP policy controlling the actions taken by the system is trained in an episodic reinforcement learning (RL) framework whereby the agent receives a reinforcement signal after each dialogue (episode) reflecting how well it performed (Sutton and Barto, 1998). In addition, the dialogue states should be tracked in order to measure the belief of the current situation during the whole interaction (Young et al., 2010; Sun et al., 2014).

Natural Language Generation There are two NLG approaches, one focuses on generating text using templates or rules (linguistic) methods, the another uses corpus-based statistical techniques (Oh and Rudnicky, 2002). Oh and Rudnicky showed that stochastic generation benefits from two factors: 1) it takes advantage of the practical language of a domain expert instead of the developer and 2) it restates the problem in terms of classification and labeling, where expertise is not required for developing a rule-based generation system.

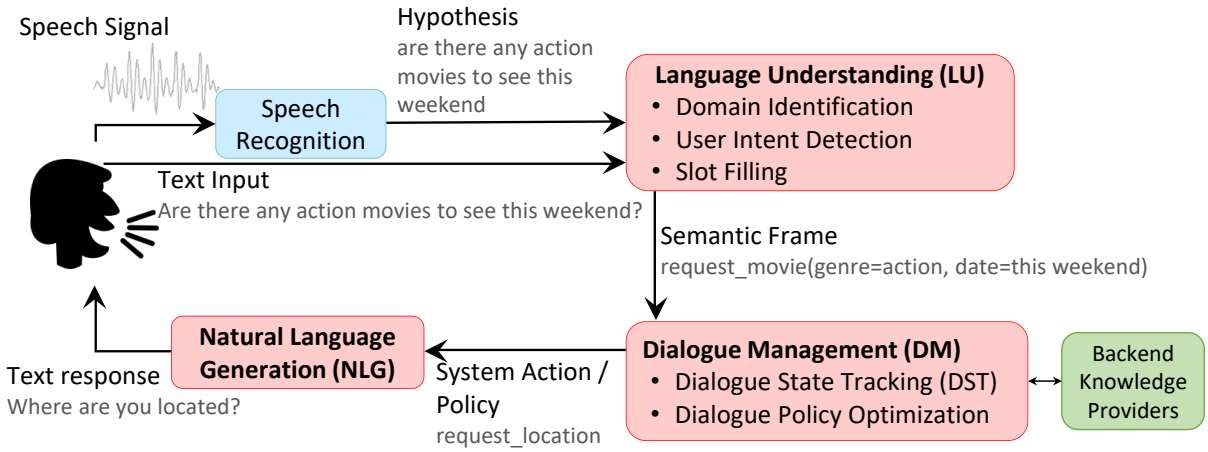

Figure 1: Pipeline framework of spoken dialog system.

W	find	action	movies	this	weekend
	↓	↓	↓	↓	↓
S	O	B-genre	O	B-date	I-date
I	find_movie				

Figure 2: An example utterance with annotations of semantic slots in IOB format (S) and intent (I), B-date and I-date denote the date slot.

4 Deep Learning Based Dialogue System

With the power of deep learning, there is increasing research work focusing on applying deep learning for each component.

Language Understanding With the advances on deep learning, deep belief networks (DBNs) with deep neural networks (DNNs) have been applied to domain and intent classification tasks (Sarikaya et al., 2011; Tur et al., 2012; Sarikaya et al., 2014). Recently, Ravuri and Stolcke (2015) proposed an RNN architecture for intent determination. For slot filling, deep learning has been viewed as a feature generator and the neural architecture can be merged with CRFs (Xu and Sarikaya, 2013). Yao et al. (2013) and Mesnil et al. (2015) later employed RNNs for sequence labeling in order to perform slot filling. Such architectures have later been extended to jointly model intent detection and slot filling in multiple domains (Hakkani-Tür et al., 2016; Jaech et al., 2016). End-to-end memory networks have been shown to provide a good mechanism for integrating longer term knowledge context and shorter term dialogue context into these models (Chen et al., 2016b,c). In addition, the importance of the LU module is investigated in Li et al. (2017a), where different types of errors from LU may degrade the whole system performance in an reinforcement learning setting.

Dialogue Management The state-of-the-art dialog managers focus on monitoring the dialog progress by neural dialog state tracking models. Among the initial models are the RNN based dialog state tracking approaches (Henderson et al., 2013) that has shown to outperform Bayesian networks (Thomson and Young, 2010). More recent work on Neural Dialog Managers that provide conjoint representations between the utterances, slot-value pairs as well as knowledge graph representations (Wen et al., 2016; Mrkšić et al., 2016) demonstrate that using neural dialog models can overcome current obstacles of deploying dialogue systems in larger dialog domains.

Natural Language Generation The RNN-based models have been applied to language generation for both chit-chat and task-orientated dialogue systems (Vinyals and Le, 2015; Wen et al., 2015b). The RNN-based NLG can learn from unaligned data by jointly optimizing sentence planning and surface realization, and language variation can be easily achieved by sampling from output candidates (Wen et al., 2015a). Moreover, Wen et al. (2015b) improved the prior work by adding a gating mechanism for controlling the dialogue act during generation in order to avoid semantics repetition, showing promising results.

5 Recent Trends and Challenges on Learning Dialogues

This part will focus on discussing the recent trends and current challenges on dialogue system technology.

End-to-End Learning for Dialogue System With the power of neural networks, there are more and more attempts for learning dialogue systems in an end-to-end fashion. Different learning frameworks are applied, including supervised learning and reinforcement learning. This part will discuss the work about end-to-end learning for dialogues (Dhingra et al., 2016; Wen et al., 2016; Williams and Zweig, 2016; Zhao and Eskenazi, 2016; Li et al., 2017b).

Recent advance of deep learning has inspired many applications of neural models to dialogue systems. Wen et al. (2016) and Bordes and Weston (2016) introduced a network-based end-to-end trainable task-oriented dialogue system, which treated dialogue system learning as the problem of learning a mapping from dialogue histories to system responses, and applied an encoder-decoder model to train the whole system. However, the system is trained in a supervised fashion, thus requires a lot of training data, and may not be able to explore the unknown space that does not exist in the training data for an optimal and robust policy.

Zhao and Eskenazi (2016) first presented an end-to-end reinforcement learning (RL) approach to dialogue state tracking and policy learning in the DM. This approach is shown to be promising when applied to a task-oriented system, which is to guess the famous person a user thinks of. In the conversation, the agent asks the user a series of Yes/No questions to find the correct answer. Dhingra et al. (2016) proposed an end-to-end differentiable KB-Infobot to improve the flexibility of question types and robustness. Li et al. (2017b) further presented an end-to-end neural dialogue system for completing tasks, which supported flexible question types, allowed user-initiated requests during conversation, and finally achieved better robustness.

Dialogue Breath In order to extend the coverage of the systems, transfer learning has been applied to different extended systems in order to proceed to a multi-domain scenario. Chen et al. (2016a) transfered the dialogue acts across different domains so that the performance of the newly-developed domain can be boosted. Kim et al. proposed to learn a domain-specific and domain-independent information in order to transfer the shared knowledge more efficiently and effectively. In addition, Gašić et al. (2015) presented the policy committee in order to boost the performance for policy training in a new domain. All above work extended the dialogue coverage using different directions.

Dialogue Depth Most current systems focus on knowledge-based understanding, but there are hierarchical understanding according to the dialogue complexity. For example, an intent about party scheduling may include restaurant reserving and invitation sending. Sun et al. (2016) learned the high-level intentions that span on multiple domains in order to achieve common sense understanding. Moreover, a more complex dialogue such as "*I feel sad...*" requires empathy in order to generate the suitable response. Fung et al. (2016) first attempted to leverage different modalities for emotion detection and built an emotion-aware dialogue system.

Given two branches of development, the ultimate goal is to build an open-domain dialogue system (coverage) with all levels of understanding (depth).

6 Instructors

Yun-Nung (Vivian) Chen is currently an assistant professor at the Department of Computer Science, National Taiwan University. She earned her Ph.D. degree from Carnegie Mellon University, where her research interests focus on spoken dialogue system, language understanding, natural language processing, and multi-modal speech applications. She received the Google Faculty Research Awards 2016, two Student Best Paper Awards from IEEE SLT 2010 and IEEE ASRU 2013, a Student Best Paper Nominee from Interspeech 2012, and the Distinguished Master Thesis Award from ACLCLP. Before joining National Taiwan University, she worked in the Deep Learning Technology Center at Microsoft Research Redmond. More information about her can be found at `http://vivianchen.idv.tw`.

Asli Celikyilmaz is currently a researcher at the Deep Learning Technology Center at Microsfot Research. Previously, she was a Research Scientist at Microsoft Bing from 2010 until 2016 focusing on deep learning models for scaling natural user interfaces to multiple domains. She has worked as a Postdoc Researcher at the EECS Department of the UC Berkeley from 2008 until 2010. She has worked with researchers at ICSI @ Berkeley during her postdoc research study. She

has earned her Ph.D. from University of Toronto, Canada in 2008. Asli's research interests are mainly machine learning and its applications to conversational dialogue systems, mainly natural language understanding and dialogue modeling. In the past she has also focused on research areas including machine intelligence, semantic tagging of natural user utterances of human to machine conversations, text analysis, document summarization, question answering, co-reference resolution, to name a few. Currently she is focusing on reasoning, attention, memory networks as well as multi-task and transfer learning for conversational dialogue systems. She has been serving as area chair, co-organizer of numerous NLP and speech conferences, such as ACL, NAACL, Interspeech, and IEEE Spoken Language Technologies (SLT). She co-organized a NIPS workshop on Machine Learning for Spoken Language Understanding and Interactions in 2015.

Dilek Hakkani-Tür is a research scientist at Google Research. Prior to joining Google, she was a researcher at Microsoft Research (2010-2016), International Computer Science Institute (ICSI, 2006-2010) and AT&T Labs-Research (2001-2005). She received her BSc degree from Middle East Technical Univ, in 1994, and MSc and PhD degrees from Bilkent Univ., Department of Computer Engineering, in 1996 and 2000, respectively. Her research interests include natural language and speech processing, spoken dialogue systems, and machine learning for language processing. She has over 50 patents that were granted and co-authored more than 200 papers in natural language and speech processing. She is the recipient of three best paper awards for her work on active learning for dialogue systems, from IEEE Signal Processing Society, ISCA and EURASIP. She was an associate editor of IEEE Transactions on Audio, Speech and Language Processing (2005-2008), member of the IEEE Speech and Language Technical Committee (2009-2014), area editor for speech and language processing for Elsevier's Digital Signal Processing Journal and IEEE Signal Processing Letters (2011-2013), and currently serves on ISCA Advisory Council (2015-2018). She is a fellow of IEEE and ISCA.

References

Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683* .

Ciprian Chelba, Monika Mahajan, and Alex Acero. 2003. Speech utterance classification. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP)*. IEEE, volume 1, pages I–280.

Yun-Nung Chen, Dilek Hakkani-T, Xiaodong He, et al. 2016a. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 6045–6049.

Yun-Nung Chen, Dilek Hakkani-Tur, and Gokan Tur. 2014. Deriving local relational surface forms from dependency-based entity embeddings for unsupervised spoken language understanding. In *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pages 242–247.

Yun-Nung Chen, Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Jianfeng Gao, and Li Deng. 2016b. Knowledge as a teacher: Knowledge-guided structural attention networks. *arXiv preprint arXiv:1609.03286* .

Yun-Nung Chen, Dilek Hakkani-Tür, Gokhan Tur, Jianfeng Gao, and Li Deng. 2016c. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Proceedings of the Interspeech*.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2016. End-to-end reinforcement learning of dialogue agents for information access. *arXiv preprint arXiv:1609.00777* .

Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. 2016. Towards empathetic human-robot interactions. *arXiv preprint arXiv:1605.04072* .

M Gašić, N Mrkšić, Pei-hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2015. Policy committee for adaptation in multi-domain spoken dialogue systems. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, pages 806–812.

Patrick Haffner, Gokhan Tur, and Jerry H Wright. 2003. Optimizing svms for complex call classification. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP)*. IEEE, volume 1, pages I–632.

Dilek Hakkani-Tür, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM. In *Proceedings of the Interspeech*. San Francisco, CA.

Matthew Henderson, Blaise Thomson, and Steve Young. 2013. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*. pages 467–471.

A. Jaech, L. Heck, and M. Ostendorf. 2016. Domain adaptation of recurrent neural networks for natural language understanding. In *Proceedings of the Interspeech*. San Francisco, CA.

Young-Bum Kim, WA Redmond, Karl Stratos, and Ruhi Sarikaya. ???? Frustratingly easy neural domain adaptation.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017a. Investigation of language understanding impact for reinforcement learning based dialogue systems. *arXiv preprint arXiv:1703.07055* .

Xuijun Li, Yun-Nung Chen, Lihong Li, and Jianfeng Gao. 2017b. End-to-end task-completion neural dialogue systems. *arXiv preprint arXiv:1703.01008* .

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(3):530–539.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2016. Neural belief tracker: Data-driven dialogue state tracking. *arXiv preprint arXiv:1606.03777* .

Alice H Oh and Alexander I Rudnicky. 2002. Stochastic natural language generation for spoken dialog systems. *Computer Speech & Language* 16(3):387–407.

Roberto Pieraccini, Evelyne Tzoukermann, Zakhar Gorelov, Jean-Luc Gauvain, Esther Levin, Chin-Hui Lee, and Jay G Wilpon. 1992. A speech understanding system based on statistical representation of semantics. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, volume 1, pages 193–196.

Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Christian Raymond and Giuseppe Riccardi. 2007. Generative and discriminative algorithms for spoken language understanding. In *INTERSPEECH*. pages 1605–1608.

Alexander I Rudnicky, Eric H Thayer, Paul C Constantinides, Chris Tchou, R Shern, Kevin A Lenzo, Wei Xu, and Alice Oh. 1999. Creating natural dialogs in the carnegie mellon communicator system. In *Eurospeech*.

Ruhi Sarikaya, Geoffrey E Hinton, and Anoop Deoras. 2014. Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22(4):778–784.

Ruhi Sarikaya, Geoffrey E Hinton, and Bhuvana Ramabhadran. 2011. Deep belief nets for natural language call-routing. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 5680–5683.

Kai Sun, Lu Chen, Su Zhu, and Kai Yu. 2014. A generalized rule based tracker for dialogue state tracking. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, pages 330–335.

Ming Sun, Yun-Nung Chen, and Alexander I Rudnicky. 2016. An intelligent assistant for high-level task understanding. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, pages 169–174.

Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Blaise Thomson and Steve Young. 2010. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech and Language* 24(4):562–588.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Gokhan Tur, Li Deng, Dilek Hakkani-Tür, and Xiaodong He. 2012. Towards deeper understanding: Deep convex networks for semantic utterance classification. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 5045–5048.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* .

Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken language understanding. *IEEE Signal Processing Magazine* 22(5):16–31.

Tsung-Hsien Wen, Milica Gašic, Dongho Kim, Nikola Mrkšic, Pei-Hao Su, David Vandyke, and Steve Young. 2015a. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. page 275.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562* .

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015b. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745* .

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language* 21(2):393–422.

Jason D Williams and Geoffrey Zweig. 2016. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269* .

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pages 78–83.

Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. 2013. Recurrent neural networks for language understanding. In *INTERSPEECH*. pages 2524–2528.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language* 24(2):150–174.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE* 101(5):1160–1179.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *arXiv preprint arXiv:1606.02560* .

Victor Zue, Stephanie Seneff, James R Glass, Joseph Polifroni, Christine Pao, Timothy J Hazen, and Lee Hetherington. 2000. JUPITER: a telephone-based conversational interface for weather information. *IEEE Transactions on speech and audio processing* 8(1):85–96.

Victor W Zue and James R Glass. 2000. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE* 88(8):1166–1180.