CASA-NLU: Context-Aware Self-Attentive Natural Language Understanding for Task-Oriented Chatbots

Arshit Gupta[†] **Peng Zhang**[‡] **Garima Lalwani**[‡] **Mona Diab**[†]
[†]Amazon AI, Seattle [‡]Amazon AI, East Palo Alto
{arshig, pezha, glalwani, diabmona}@amazon.com

Abstract

Natural Language Understanding (NLU) is a core component of dialog systems. It typically involves two tasks - intent classification (IC) and slot labeling (SL), which are then followed by a dialogue management (DM) component. Such NLU systems cater to utterances in isolation, thus pushing the problem of context management to DM. However, contextual information is critical to the correct prediction of intents and slots in a conversation. Prior work on contextual NLU has been limited in terms of the types of contextual signals used and the understanding of their impact on the model. In this work, we propose a context-aware self-attentive NLU (CASA-NLU) model that uses multiple signals, such as previous intents, slots, dialog acts and utterances over a variable context window, in addition to the current user utterance. CASA-NLU outperforms a recurrent contextual NLU baseline on two conversational datasets, yielding a gain of up to 7% on the IC task for one of the datasets. Moreover, a non-contextual variant of CASA-NLU achieves state-of-the-art performance for IC task on standard public datasets - SNIPS and ATIS.

1 Introduction

With the advent of smart conversational agents such as Amazon Alexa, Google Assistant, etc., dialogue systems are becoming ubiquitous. In the context of enterprises, the majority of these systems target task oriented dialogues with the user trying to achieve a goal, e.g. booking flight tickets or ordering food. Natural Language Understanding (NLU) captures the semantic meaning of a user's utterance within each dialogue turn, by identifying **intents** and **slots**. An intent specifies the goal underlying the expressed utterance while slots are additional parameters for these intents. These tasks are typically articulated as intent classification (IC) coupled with sequence tagging task of slot labeling (SL).

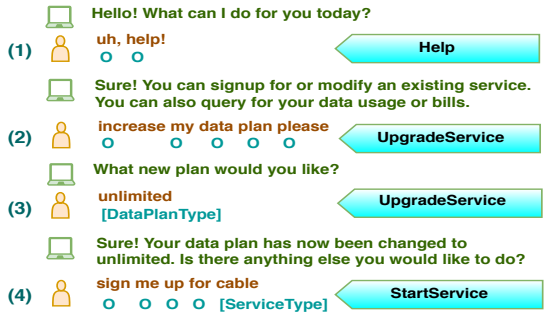

Figure 1: Snippet of a sample Cable data conversation with intents shown in boxes and slots aligned with user request

Over time, human-machine interactions have become more complex with greater reliance on contextual cues for utterance understanding (Figure 1). With traditional NLU frameworks, the resolution of contextual utterances is typically addressed in the DM component of the system using rule-based dialogue state trackers (DST). However, this pushes the problem of context resolution further down the dialogue pipeline, and despite the appeal of modularity in design, it opens the door for significant cascade of errors. To avoid this, end-to-end dialogue systems have been proposed (Wen et al., 2017; Bordes and Weston, 2016), but, to date, such systems are not scalable in industrial settings, and tend to be opaque where a level of transparency is needed, for instance, to understand various dialogue policies.

To address the propagation of error while maintaining a modular framework, Shi et al. (2015) proposed adding contextual signals to the joint IC-SL task. However, the contributions of their work were limited in terms of number of signals and how they were used, rendering the contextualization process still less interpretable. In this work, we present a multi-dimensional self-attention based contextual NLU model that overcomes prior work's shortcomings by supporting

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1285–1290, Hong Kong, China, November 3–7, 2019.

variable number of contextual signals (previous utterances, dialogue acts[1], intents and slot labels) that can be used concurrently over variable length of conversation context. In addition, our model allows for easy visualization and debugging of contextual signals which are essential, especially in production dialogue systems, where interpretability is a desirable feature. Our contributions are:

- Context-Aware Self-Attentive NLU (CASA-NLU) model that uses various contextual signals to perform joint IC-SL task, outperforming contextual and non-contextual NLU baselines by significant margins on two in-house conversational IC-SL datasets
- Propose a novel non-contextual variant of CASA-NLU that achieves SOTA performance on IC task for both SNIPS and ATIS datasets
- Analysis of the various contextual signals' contributions to model performance

2 Related Work

There have been numerous advancements in NLU systems for dialogues over the past two decades. While the traditional approaches used handcrafted features and word n-gram based features fed to SVM, logistic regression, etc. for IC task and conditional random fields (CRF) for SL task (Jeong and Lee, 2008; Wang1 et al., 2002; Raymond and Riccardi, 2007), more recent approaches rely on deep neural networks to jointly model IC and SL tasks (Yao et al., 2014a,b; Guo et al., 2014; Zhang and Wang, 2016; Liu and Lane, 2016c; Goo et al., 2018). Attention as introduced by Bahdanau et al. (2014) has played a major role in many of these systems (Liu and Lane, 2016a; Ma et al., 2017; Li et al., 2018a; Goo et al., 2018), for instance, for modeling interaction between intents and slots in (Goo et al., 2018).

Dahlbäck and Jönsson (1989) and Bertomeu et al. (2006) studied contextual phenomena and thematic relations in natural language, thereby highlighting the importance of using context. Few previous works focused on modeling turn-level predictions as DST task (Williams et al., 2013). However, these systems predict the possible slot-value pairs at utterance level (Zhong et al., 2018), making it necessary to maintain ontology of all possible slot values, which is infeasible for certain slot types (e.g., restaurant names). In industry settings, where IC-SL task is predominant, there is also an additional effort involved to invest in rules for converting utterance level dialog state annotations to token level annotations required for SL. Hence, our work mainly focuses on the IC-SL task which eliminates the need for maintaining any ontology or such handcrafted rules.

Bhargava et al. (2013) used previous intents and slots for IC and SL models. They were followed by Shi et al. (2015) who exploited previous intents and domain predictions to train a joint IC-SL model. However, both these studies lacked comprehensive context modeling framework that allows multiple contextual signals to be used together over a variable context window. Also, an intuitive interpretation of the impact of contextual signals on IC-SL task was missing.

3 CASA-NLU: Context-Aware Self-Attentive NLU

Our model architecture is composed of three sub sections - signal encoding, context fusion and IC-SL predictions (Figure 2).

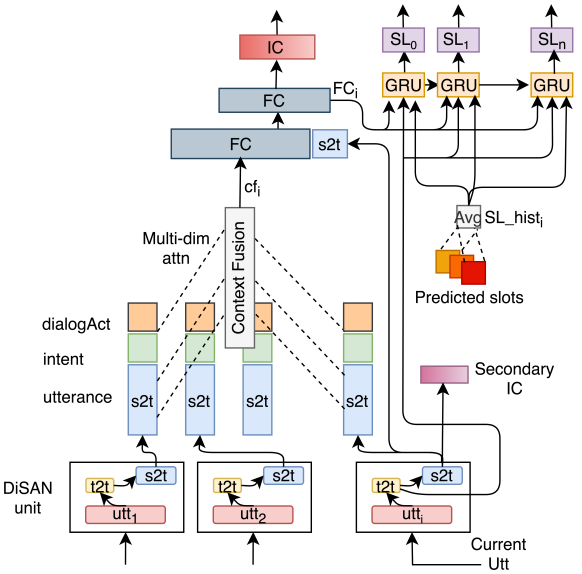

Figure 2: CASA-NLU model architecture for joint IC-SL

3.1 Signal Encoding

Utterance (Utt): For the utterance encoding, we adopt the directional self-attention (Shen et al., 2017) based encoder (DISAN), adding absolute position embedding (Gehring et al., 2017) to further improve the encoding performance.[2] DISAN unit consists of two types of multi-dimensional attention - word level token2token (t2t) attention fol-

[1]Dialog Act signifies the actions taken by the agent such as Close (when an intent is fulfilled), ElicitSlot, etc

[2]Details provided in Appendix A

lowed by sentence level source2token (s2t) attention. For turn i, the output of this unit is given by $\mathbf{h}(\text{Utt}_i) \in \mathbb{R}^{2d_h \times 1}$, where d_h is hidden layer size of the DISAN unit (Figure 2).

Intent / Dialog Act (DA) / Slot Label History: We pass the one-hot representation of intent ground truth labels through an embedding layer to get intent history representation $\mathbf{h}(\text{I}_i) \in \mathbb{R}^{d_I \times 1}$ for any previous turn i with d_I being the intent embedding dimension. Similarly, for DA history, $\mathbf{h}(\text{DA}_i) \in \mathbb{R}^{d_{DA} \times 1}$. We use a special dummy symbol for the intent and DA of the current turn. For slot label history, for turn i, we take the average of all slot embeddings that were observed in previous turns giving $\mathbf{h}(\text{SL\_hist}_i) \in \mathbb{R}^{d_{SL} \times 1}$.

3.2 Context Fusion

We combine the vectorized signals together in both spacial and temporal dimensions. For the former, we simply concatenate the contextual signals to get current turn feature vector, i.e. $\mathbf{T}_i = [\mathbf{h}(\text{Utt}_i); \mathbf{h}(\text{I}_i); \mathbf{h}(\text{DA}_i)]$. However, for the latter, concatenation becomes intractable if context window is large. To address this issue and automatically learn more relevant components of context for each turn, we add a source2token (Shen et al., 2017) multi-dimensional self-attention layer over the turn vectors. This is essentially a per dimension learned weighted average ($\mathbf{cf}_i$) over all the turn vectors within a context window (Equation 1). As shown later in Section 5, this enhances the model's robustness by learning different attention weights for different contextual signals.

$$\mathbf{cf}_i = \sum_{t=i-K}^{i} \mathbf{P}(\mathbf{T}_t) \odot \mathbf{T}_t \qquad (1)$$

where, K (= 3 in our experiments) is the context window, and $\mathbf{T}_t$ and $\mathbf{P}(\mathbf{T}_t)$ are the turn vector and attention weights for t^{th} time step respectively. We use padding tokens for $i - K < 0$.

One of the shortcomings with such attention mechanism is that it is position invariant. To address this problem, we add learned absolute position embedding $\mathbf{p}_c$ to the turn matrix $\mathbf{T}$ that learns temporal information across the turns.

3.3 IC-SL Predictions

Following (Liu and Lane, 2016b; Li et al., 2018b), we train a joint IC-SL model. To improve IC performance for our deep network, we also add a secondary IC loss function, $\text{L}_{\text{Sec\_IC}}$ at the utterance level (Figure 2). The new aggregated loss is:

$$\text{L} = \text{L}_{\text{IC}} + \alpha \times \text{L}_{\text{SL}} + \beta \times \text{L}_{\text{Sec\_IC}} \qquad (2)$$

IC: At turn i, we take the output of context fusion layer $\mathbf{cf}_i$, pass it through a fully connected layer and concatenate the output with the current utterance encoding $\mathbf{h}(\text{Utt}_i)$. This is then further projected down using a fully connected layer (FC_i) and finally fed into *softmax* layer to predict the intent.

SL: For turn i, the t2t attention output is first fused with the utterance embedding using a fusion gate (Hochreiter and Schmidhuber, 1997) to generate $\mathbf{h}_{ij}$ where j represents token index in the utterance. Then, for each token position in the utterance, we apply a sliding window, w (=3) over neighboring words that transforms each token's embedding space from $\mathbf{h}_{ij}$ to $w \times \mathbf{h}_{ij}$ (not shown in the Figure 2). To add contextual information to SL task, each token's dimension is augmented using slot history (SL_hist_i) as well as penultimate fully connected layer for IC task (FC_i), yielding a final dimension of $w \times \mathbf{h}_{ij} + \mathbf{h}(\text{SL\_hist}_i) + \mathbf{h}(\text{FC}_i)$. Finally, a Gated Recurrent Unit (GRU) renders the labels auto-regressive followed by *softmax* layer.

4 Experiments

Datasets: Since there are no existing public datasets for contextual IC and SL task, we use two in-house datasets for evaluation - **Booking** dataset,[3] which is a variation of DSTC-2 dataset (Williams et al., 2014) with intent, slot and dialog act annotations, and **Cable** dataset, a synthetically created conversational dataset. The Booking dataset contains 9,351 training utterances (2,200 conversations) and 6,727 test utterances, with 19 intents and 5 slot types. Cable dataset comprises 1856 training utterances, 1,814 validation utterances and 1,836 test utterances, with 21 intents and 26 slot types.[4] In addition, we also evaluate the model on non-contextual IC-SL public datasets - ATIS (Hemphill et al., 1990) and SNIPS (Coucke et al., 2018).

Experimental setup: To emphasize the importance of contextual signals in modeling, we first devise a non-contextual baseline of our CASA-NLU model, NC-NLU. It is similar to CASA-

[3] Dataset will be released to the public

[4] Detailed data stats provided in Appendix B

Model	ATIS	SNIPS
RNN-LSTM* (Hakkani-Tür et al., 2016)	92.6	96.9
Atten.-BiRNN* (Liu and Lane, 2016b)	91.1	96.7
LSTM+attn+gates (Goo et al., 2018)	94.1	97.0
Capsules Neural Network (Zhang et al., 2018)	95.0	97.7
NC-NLU (ours)	**95.4**	**98.4**

Table 1: Average IC accuracy scores (%) on non-contextual datasets. *: as reported in (Goo et al., 2018)

NLU in terms of model architecture except no contextual signals are used. Since neither datasets nor implementation details were shared in previous works on contextual NLU (Bhargava et al., 2013; Shi et al., 2015), we also implement another baseline, CGRU-NLU that uses GRU (Cho et al., 2014) instead of self-attention for temporal context fusion. For fair comparison to existing non-contextual baselines, no pre-trained embeddings are used in any of the experiments though the model design can easily benefit from pre-training.

5 Results and Analysis

Table 1 shows the performance of our non-contextual model on two datasets, ATIS and SNIPS. As shown, we obtain a **new SOTA** for IC on both the datasets.[5] We hypothesize that the high performance is due to the utterance-level position-aware multi-dimensional self-attention.

As shown in Table 2, CASA-NLU model outperforms non-contextual NLU on both the Booking and Cable datasets. Further, CASA-NLU model significantly **outperforms** CGRU-NLU on the Cable dataset by 7.26% on IC accuracy and 5.31% on SL F1 absolute, respectively. We believe the reason for strong performance yielded by the CASA-NLU model is due to its multi-dimensional nature, where we learn different weights for different dimensions within the context feature vector $\mathbf{T}_i$. This enables the model to learn different attention distributions for different contextual signals leading to more robust modeling compared to CGRU-NLU model. Table 3 gives further breakdown of the results by showing performance on first vs. follow-up turns in a dialogue. For the more challenging follow-up turns, CASA-NLU yields significant gains over the baseline IC performance.

Table 4 shows impact of some of the contextual signals on model performance for the Booking validation dataset. As expected, contextual signals improve IC and SL performance (Configs - II-V). We observe that adding intent history (I_{hist}) leads to highest gains in IC accuracy (Config - IV). At the same time, we see that slot history (SL_{hist}) has minimal impact on SL performance for this dataset. Exhaustive experiments showed that the choice of contextual signals is dependent upon the dataset. Our model facilitates switching these contextual signals on or off easily.

Model	Booking		Cable	
	IC	SL	IC	SL
NC-NLU	91.11	88.21	44.98	27.34
CGRU-NLU	94.86	88.47	66.68	51.58
CASA-NLU	**95.16**	**88.80**	**73.94**	**56.97**

Table 2: IC accuracy and SL F1 scores (%) for the three models NC-NLU, CGRU-NLU, CASA-NLU on the 2 contextual datasets - Booking and Cable.

Model	Booking		Cable	
	Ft	FU	Ft	FU
NC-NLU	95.67	89.51	40.35	48.33
CGRU-NLU	98.93	94.24	46.6	72.22
CASA-NLU	**99.70**	**94.45**	**47.44**	**81.25**

Table 3: IC accuracy scores (%) on first (Ft) and follow-up (FU) turns in contextual datasets - Booking and Cable.

Config	I_{hist}	SL_{hist}	Utt_{hist}	DA_{hist}	IC	SL
I	x	x	x	x	89.44	97.62
II	x	x	✓	x	92.52	98.09
III	x	✓	✓	✓	92.35	98.33
IV	✓	✓	✓	x	95.27	98.34
V	✓	✓	✓	✓	96.53	98.25

Table 4: Impact of contextual signals on IC accuracy and SL F1 scores (%) on Booking validation set for CASA-NLU

Qualitative Analysis: Using example conversation in Figure 1, we highlight the relevance of contextual information in making intent predictions by visualizing attention weights $\mathbf{P}(\mathbf{T})$ (Equation 1) for different contextual signals as shown in Figure 3.[6] At user turn 2, the *intent* is switched to *UpgradeService* which the model successfully interprets by paying less attention to previous intents. At turn 3, however, contextual information is critical as user responds to elicitation by agent and hence model emphasizes on last utterance and intent rather than the current or other previous turns.

[5] Since we compute token-level F1, SL performance is not compared to results reported in previous work

[6] For each context signal, attention weights are averaged across all its feature dimensions in $\mathbf{P}(\mathbf{T})$

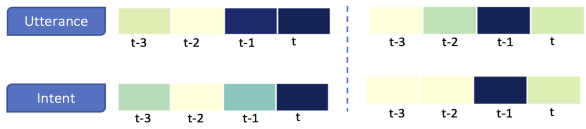

Figure 3: **Visualization of attention weights** given to two different context signals - previous utterances (top) and previous intents (bottom) for turns t=2 (left) and t=3 (right) from Figure 1; darker colors reflect higher attention weights.

6 Conclusion

We proposed CASA-NLU model that uses a variable context and various contextual signals in addition to the current utterance to predict the intent and slot labels for the current turn. CASA-NLU achieves gains of over 7% on IC accuracy for Cable dataset over CGRU-NLU baseline, and almost 29% over non-contextual version. This highlights the importance of using contextual information, meanwhile showing that, learning correct attention is also vital for NLU systems.

7 Implementation Details

We use hidden layer size of 56 with dropout probability of 0.3. Context history window K was varied from 1 to 5 and the optimal value of 3 was selected. Word embeddings are trained from scratch using an embedding layer size of 56. Adam (Kingma and Ba, 2014) algorithm with initial learning rate of 0.01 gave the optimal performance. Concatenation window size w of 3 is used. α and β in loss objective are set to 0.9. Early stopping is used with patience of 10 and threshold of 0.5. Each model is trained for 3 seeds and scores averaged across the seeds are reported.

8 Acknowledgements

The authors would like to thank the entire AWS Lex Science team for having insightful discussions and providing feedback with experiments. The authors would also like to express their gratitude to Yi Zhang for his generous help and suggestions for this work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Núria Bertomeu, Hans Uszkoreit, Anette Frank, Hans-Ulrich Krieger, and Brigitte Jörg. 2006. Contextual phenomena and thematic relations in database qa dialogues: results from a wizard-of-oz experiment. In *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006*, pages 1–8.

A. Bhargava, Asli elikyilmaz, Dilek Z. Hakkani-Tür, and Ruhi Sarikaya. 2013. Easy contextual intent prediction and slot detection. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8337–8341.

Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR*, abs/1605.07683.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Nils Dahlbäck and Arne Jönsson. 1989. Empirical studies of discourse representations for natural language interfaces. In *Proceedings of the fourth conference on European chapter of the Association for Computational Linguistics*, pages 291–298. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *NAACL-HLT*.

Daniel Guo, Gökhan Tür, Wen tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 554–559.

Dilek Hakkani-Tür, Gökhan Tür, 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 *Interspeech*, pages 715–719.

C.T. Hemphill, J.J. Godfrey, and G.R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Proceedings of the DARPA Speech and Natural Language Workshop*, page 96101.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Minwoo Jeong and Gary Geunbae Lee. 2008. Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:1287–1302.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Changliang Li, Liang Li, and Ji Qi. 2018a. A self-attentive model with gate mechanism for spoken language understanding. In *EMNLP*.

Changliang Li, Liang Li, and Ji Qi. 2018b. A self-attentive model with gate mechanism for spoken language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3833.

Bing Liu and Ian Lane. 2016a. Attention-based recurrent neural network models for joint intent detection and slot filling. *ArXiv*, abs/1609.01454.

Bing Liu and Ian Lane. 2016b. Attention-based recurrent neural network models for joint intent detection and slot filling. In *INTERSPEECH*.

Bing Liu and Ian Lane. 2016c. Joint online spoken language understanding and language modeling with recurrent neural networks. *ArXiv*, abs/1609.01462.

Mingbo Ma, Kai Zhao, Liang Huang, Bing Xiang, and Bowen Zhou. 2017. Jointly trained sequential labeling and classification by sparse attention neural networks. *ArXiv*, abs/1709.10191.

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

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2017. Disan: Directional self-attention network for rnn/cnn-free language understanding. *arXiv preprint arXiv:1709.04696*.

Yangyang Shi, Kaisheng Yao, Hu Chen, Yi-Cheng Pan, Mei-Yuh Hwang, and Baolin Peng. 2015. Contextual spoken language understanding using recurrent neural networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5271–5275.

Ye-Yi Wang1, Alex Acero, Ciprian Chelba, Brendan Frey, and Leon Wong. 2002. Combination of statistical and rule-based approaches for spoken language understanding. In *Seventh International Conference on Spoken Language Processing*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei hao Su, Stefan Ultes, David Vandyke, and Steve J. Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413.

Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. 2014a. Spoken language understanding using long short-term memory neural networks. *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 189–194.

Kaisheng Yao, Baolin Peng, Geoffrey Zweig, Dong Yu, Xiaolong Li, and Feng Gao. 2014b. Recurrent conditional random field for language understanding. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4077–4081.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S Yu. 2018. Joint slot filling and intent detection via capsule neural networks. *arXiv preprint arXiv:1812.09471*.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive dialogue state tracker. *CoRR*, abs/1805.09655.