DYNAMICALLY CONTEXT-SENSITIVE TIME-DECAY ATTENTION FOR DIALOGUE MODELING

Shang-Yu Su Pei-Chieh Yuan Yun-Nung Chen

National Taiwan University, Taipei, Taiwan

{f05921117, b03901134}@ntu.edu.tw y.v.chen@ieee.org

ABSTRACT

Spoken language understanding (SLU) is an essential component in conversational systems. Considering that contexts provide informative cues for better understanding, history can be leveraged for contextual SLU. However, most prior work only paid attention to the related history utterances and ignored the temporal information. In dialogues, prior work considers an inflexible decaying time-aware attention to allow the model to pay more attention to the most recent utterances than the least recent ones. To improve the flexibility of function design, this paper allows the model to automatically learn a time-decay attention function where the attentional weights can be dynamically decided based on the content of each role's contexts, which effectively integrates both content-aware and time-aware perspectives and demonstrates remarkable flexibility to complex dialogue contexts. The experiments on the benchmark Dialogue State Tracking Challenge (DSTC4) dataset show that the proposed dynamically context-sensitive time-decay attention mechanisms significantly improve the state-of-the-art model for contextual understanding performance.[1]

Index Terms— Spoken language understanding, dialogue modeling, contextual information, time-decay attention

1. INTRODUCTION

Spoken dialogue systems that can help users solve complex tasks such as booking a movie ticket have become an emerging research topic in artificial intelligence and natural language processing areas. With a well-designed dialogue system as an intelligent personal assistant, people can accomplish certain tasks more easily via natural language interactions. The recent advance of deep learning has inspired many applications of neural dialogue systems [1, 2, 3, 4, 5].

A key component of a dialogue system is a spoken language understanding (SLU) module—parsing user utterances into semantic frames that capture the core meaning [6]. A typical SLU first determines the domain given input utterances, predicts the intent, and then fill the associated slots [7, 8, 9, 10]. However, the above work focused on single-turn interactions, where each utterance is treated independently. To overcome the error propagation and further improve understanding performance, contextual information has been leveraged and shown useful [11, 12, 13, 14]. Prior work incorporated dialogue contexts into the recurrent neural networks (RNN) for improving understanding results [12, 15, 16, 17]. Recently, modeling speaker role information [18, 19, 20] has been demonstrated to learn the notable variance in speaking habits during conversations for better understanding performance.

[1]The source code is available at `https://github.com/MiuLab/CxtSen-SLU`.

Neural models incorporating attention mechanisms have advanced various tasks such as machine translation [21], image captioning [22], etc. Attentional models have been successful because they separate two different concerns: 1) deciding which input contexts are most relevant to the output and 2) predicting an output given the most relevant inputs. In dialogues, although content-aware contexts may help understanding [16, 17], the most recent contexts may be more important than others, so the temporal information can provide additional cues for the attention design. Prior work proposed an end-to-end time-aware attention network to leverage both contextual and temporal information for spoken language understanding and achieved the significant improvement, showing that the temporal attention can guide the attention effectively [19, 23]. However, the time-aware attention function is an inflexible, which is a fixed function of time for assessing the attention weights.

This paper focuses on learning a flexible time-aware attention mechanism in neural models, where the attention can be dynamically decided based on the contexts for better language understanding. This work is built on top of the role-based contextual model by modeling role-specific contexts differently to design the associated time-aware attention functions for improving system performance. The contributions are three-fold:

- The proposed end-to-end learnable attention has great flexibility of modeling temporal information for diverse dialogue contexts.
- This work investigates speaker role modeling in attention mechanisms and provides guidance for the future research about designing attention functions in dialogue modeling.
- The proposed model achieves the state-of-the-art understanding performance in the dialogue benchmark dataset.

2. END-TO-END SLU FRAMEWORK

The model architecture is illustrated in Figure 1. First, the previous utterances are fed into the contextual model to encode into the history summary, and then the summary vector and the current utterance are integrated for helping understanding. The contextual model leverages the attention mechanisms highlighted in red, which implements different attention functions for sentence and speaker levels. The whole model is trained in an end-to-end fashion, where the history summary vector and the attention functions are automatically learned based on the downstream SLU task. The objective of the proposed model is to optimize the conditional probability of the intents given the current utterance, $p(\mathbf{y} \mid \mathbf{x})$, by minimizing the cross-entropy loss between prediction and target $q(\mathbf{y} \mid \mathbf{x})$:

$$\mathcal{L} = -\sum_{k}\sum_{z} q(y_k = z \mid \mathbf{x}) \log p(y_k = z \mid \mathbf{x}), \qquad (1)$$

where the labels $\mathbf{y}$ are the labeled intent tags for understanding.

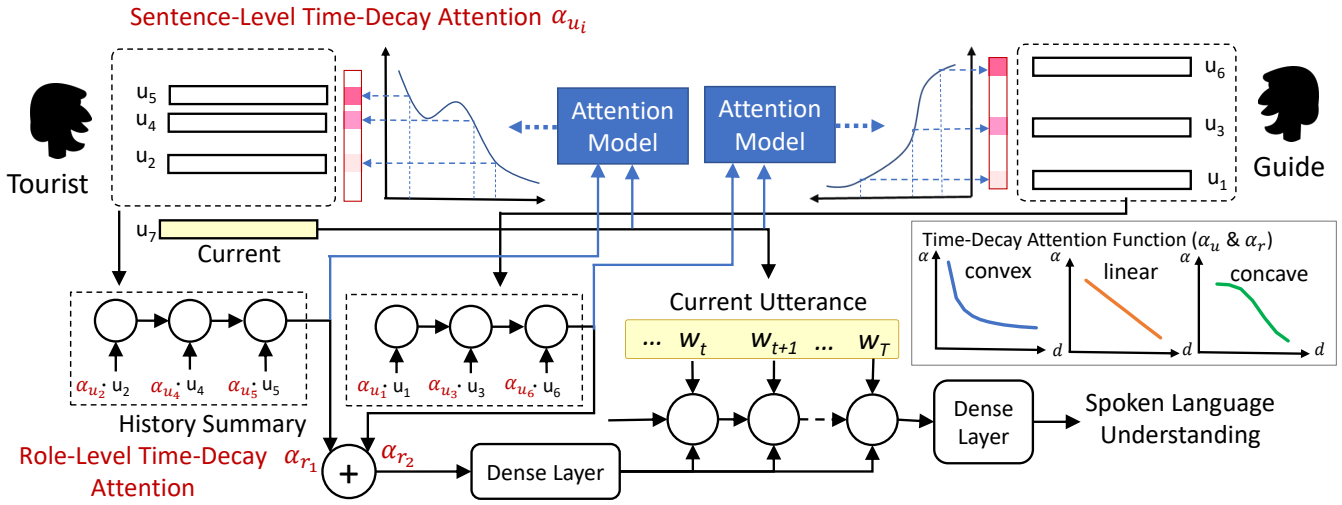

Fig. 1. Illustration of the proposed role-based context-sensitive time-decay attention contextual model.

2.1. Attentional Speaker-Aware Contextual SLU

Given the current utterance $\mathbf{x} = \{w_t\}_1^T$, the goal is to predict the user intents of $\mathbf{x}$, including speech acts and associated attributes. We apply the bidirectional long short-term memory (BLSTM) model [24] to context encoding to learn the probability distribution of user intents.

$$\mathbf{v}_{\text{o}} = \text{BLSTM}(\mathbf{x}, W_{\text{his}} \cdot \mathbf{v}_{\text{his}}), \quad (2)$$

$$\mathbf{o} = \texttt{sigmoid}(W_{\text{SLU}} \cdot \mathbf{v}_{\text{o}}), \quad (3)$$

where W_{his} and W_{SLU} are weight matrices and $\mathbf{v}_{\text{his}}$ is the history summary vector. $\mathbf{v}_{\text{o}}$ is the context-aware vector of the current utterance encoded by the BLSTM, and $\mathbf{o}$ is the intent distribution. Note that this is a multi-label and multi-class classification, so the `sigmoid` function is employed for modeling the distribution after a linear layer. The user intent labels are decided based on whether the value is higher than a threshold tuned by the development set.

Considering that speaker role information is shown to be useful for better understanding in complex dialogues [18, 20], we utilize the contexts from two roles to learn role-specific history summary representations, $\mathbf{v}_{\text{his}}$ in (2). Each role-dependent recurrent unit $\text{BLSTM}_{\text{role}}$ receives corresponding inputs, $x_{t,\text{role}}$, which includes multiple utterances u_i ($i = [1, ..., t-1]$) preceding the current utterance u_t from the specific role, and have been processed by an encoder model.

There are various tasks showing the effectiveness of attention mechanisms [25, 17]. Recent work showed that two attention types (content-aware and time-aware) and two attention levels (sentence-level and role-level) significantly improve the understanding performance for complex dialogues. This paper focuses on expanding the time-aware attention by learning dynamically context-sensitive time-decay functions in an end-to-end fashion. For time-aware attention mechanisms, we apply it using two levels, sentence-level and role-level structures.

For the sentence-level attention, before feeding into the contextual module, each history vector is weighted by its time-aware attention α_{role_i}:

$$\mathbf{v}_{\text{his}}^U = \sum_{\text{role}} \text{BLSTM}_{\text{role}}(x_{t,\text{role}}, \{\alpha_{u_j} \mid u_j \in \text{role}\}), \quad (4)$$

where $x_{t,\text{role}}$ are vectors after one-hot encoding that represent the annotated intent and the attribute features. Note that this model requires the ground truth annotations for history utterances for training and testing. Therefore, each role-based contextual module focuses on modeling role-dependent goals and speaking style, and $\mathbf{v}_{\text{o}}$ from (2) would contain role-based contextual information.

2.2. Universal Time-Decay Attention

Because we assume that the most recent contexts are more important in dialogues, a time-aware attention should be a decaying function. Considering that the contextual patterns may be diverse, a flexible and universal time-decay attention function that composes three types of attentional curves is formulated [23]:

$$\begin{aligned}\alpha_{u_i}^{\text{univ}} &= w_1 \cdot \alpha_{u_i}^{\text{conv}} + w_2 \cdot \alpha_{u_i}^{\text{lin}} + w_3 \cdot \alpha_{u_i}^{\text{conc}} \\ &= \frac{w_1}{a \cdot d(u_i)^b} + w_2(e \cdot d(u_i) + f) + \frac{w_3}{1 + (\frac{d(u_i)}{D_0})^n},\end{aligned} \quad (5)$$

where w_i are the weights of time-decay attention functions, including three types [23]: *convex*, *linear*, and *concave*, illustrated in the top-right part of Figure 1. Note that all attention weights will be normalized such that their summation is equal to 1.

- **Convex** α_u^{conv}: Intuitively, recent utterances contain more salient information, and the salience decreases very quickly when the distance increases.
- **Linear** α_u^{lin}: The importance of preceding utterances linearly declines as the distance between the previous utterance and the target utterance becomes larger.
- **Concave** α_u^{conc}: Intuitively, the attention weight decreases relatively slow when the distance increases.

Each of three types of decaying curves represents a different perspective on dialogue contexts and models different contextual patterns following the design in the prior work [23].

Because the framework can be trained in an end-to-end manner, all parameters (w_i, a, b, e, f, D_0, n) can be automatically learned to construct a flexible time-decay function. With the combination of different curves and the adjustable weights, the model can automatically learn a properly oscillating curve in order to model the diverse and complex contextual patterns using the attention mechanism.

2.3. Dynamically Context-Sensitive Attention

As described in the previous sections, the proposed time-decay attention mechanisms have parameters (a, b, e, f, D_0, n) to determine the shapes of curves. In addition to the time-decaying property, we further improve our design to dynamically encode context-sensitive characteristics into the associated attention weights. The feature vector $\mathbf{v}_{\text{cur}}$ of the current utterances $\mathbf{x}$ can be extracted by BLSTM or use the mean vector among pre-trained word embeddings of the current utterance.

Considering that different speakers may have totally different speaking behaviors [18, 19, 20], a role-based context-sensitive attention is proposed. To better model the attention curve, the contextual information is also encoded by the BLSTM model, where the preceding utterances from different speakers are encoded by different modules.

$$\mathbf{v}_{\text{his,role}} = \text{BLSTM}_{\text{role}}(x_{t,\text{role}}), \tag{6}$$

$$\mathbf{p}_{\text{role}} = W_{\text{p,role}} \cdot (\mathbf{v}_{\text{his,role}}, \mathbf{v}_{\text{cur}}) + bias, \tag{7}$$

where the speaker-specific contextual encoding $\mathbf{v}_{\text{his,role}}$ is fed along with the feature of the current utterance ($\mathbf{v}_{\text{cur}}$) into fully-connected layers to predict the parameters $\mathbf{p}_{\text{role}} \in \{a, b, e, f, D_0, n \mid \text{role}\}$ to determine the tendency of the attention curve. Because the parameters $\mathbf{p}_{\text{role}}$ are determined by the output of neural attention models without any clipping or projection and some of these uncontrolled real number are exponents, therefore the following two regularization terms are introduced as soft constraints,

$$-\alpha \cdot \min(\mathbf{p}_{\text{role}}, 0) + \beta \cdot \sum \mathbf{p}^2_{\text{role}}. \tag{8}$$

The first loss term is to encourage the model to output a positive number, and the second term is to facilitate the model to predict numbers with small absolute values, where α and β are the weights to adjust the intensity of regularization. Note that not all attention models use both regularization terms, while we endow the models with maximum flexibility and add constraints only if necessary. For example, if the cut-off distance D_0 of the concave time-decay attention is negative, the denominator $1 + (d(u_i)/D_0)^n$ would easily become complex number, which is not applicable. To make $D_0 \geq 0$, we use the model output as the exponent of the exponential function with e as the base. In order to further facilitate the concave decaying manner, the first term is applied; on the other hand, to prevent explosion, the second regularization term is utilized.

3. EXPERIMENTS

To evaluate the proposed model, we conduct the language understanding experiments on human-human conversational data.

3.1. Setup

The experiments are conducted using the DSTC4 dataset, which consist of 35 dialogue sessions on touristic information for Singapore collected from Skype calls between 3 tour guides and 35 tourists, including 31,034 utterances and 273,580 words [26]. All recorded dialogues with the total length of 21 hours have been manually transcribed and annotated with speech acts and semantic labels at each turn level. The speaker information (guide and tourist) is also provided. The human-human dialogues contain rich and complex human behaviors and bring much difficulty to all tasks. We randomly selected 28 dialogues as the training set, 5 dialogues as the testing set, and 2 dialogues as the validation set.

We focus on predicting multiple labels including intents and attributes, so the evaluation metric is an average F1 score for balancing recall and precision in each utterance. The experiments are shown in Table 1, where we report the average results over more than three runs for both tourists and guides. In all experiments, we use mini-batch *Adam* as the optimizer with the batch size of 32 examples. The size of each hidden recurrent layer is 128 or 64; since the proposed approach uses additional attention models to predict parameters of decaying curves, to fairly verify the effectiveness of the proposed method, smaller hidden recurrent layers (size = 64) are utilized in the proposed model (row (h)) and bigger ones are conducted in others (rows (b)-(g)). We use pre-trained 200-dimensional word embeddings $GloVe$ [27]. We only apply 40 training epochs without any early stop approach.

In the training process, we can assign the attention models random targets to incorporate the supervised loss during the first few epochs to accelerate training. This paper simply sets a integer target for the attention model at the very beginning. Note that experiments show that our attention model can be train from scratch in an end-to-end manner without any supervised signal and achieve the same performance.

3.2. Effectiveness of Time-Decay Attention

To evaluate the proposed time-decay attention, we compare the performance with the naïve SLU model without any contextual information (row (a)), the contextual model without any attention mechanism (row (b)), and the one using the content-aware attention mechanism (row (c)), where the attention can be learned at sentence and role levels. It is intuitive that the model without considering contexts (row (a)) performs much worse than the contextual ones for dialogue modeling. The rows (d)-(h) utilized the time-decay attention; rows (d)-(e) use only the time-decay attention; rows (f)-(g) model both content-aware and time-decay attention mechanisms together, where content-aware attention is directly estimated by concatenation of each context and the current utterance by a NN module. There are two settings for time-decay attention learning: 1) **Hand**: hand-crafted hyper-parameters (rows (d) and (f)) and 2) **E2E**: end-to-end training for parameters (rows (e) and (g)). In the hand-crafted setting, the hyper-parameters $a = 1, b = 1, e = -0.125, f = 1, D_0 = 5, n = 3$ are adopted, the parameters are chosen to examine the effectiveness of each type of decaying curve, where we choose the parameters such that the effectiveness of each type of decaying manner could be properly investigated (the linear one will be located between the two curves). In the end-to-end setting, all parameters are learnable parameter initialized as the hyper-parameters described above and fine-tuned by end-to-end learning. The row (e) previously achieves the state-of-the-art performance [23]. Our proposed context-sensitive time-decay attention model is shown in the row (h).

Table 1 shows that all models with the time-decay attention (row (d)-(g)) outperform the model without temporal modeling. However, row (c) performs worse than the one without any attention mechanism (row (b)), and rows (f)-(g) are slightly worse than the ones with only time-decay attention (rows (d)-(e)), revealing that without a delicately-designed attention mechanism, it is not guaranteed that incorporating an additional content-aware attention would bring improvement.

3.3. Analysis of Context-Sensitive Attention

Prior work (rows (f) and (g)) integrated both content-aware and time-decay attention to demonstrate the capability of mitigating the neg-

	SLU Model		Sentence-Level	Role-Level	Context Length 3	5	7
(a)	Naïve SLU		70.18		–		
(b)	No Attention Contextual Model		74.52		74.75	74.69 (-)	74.52 (-)
(c)	Content-Aware Contextual Model [18]		73.69	74.28	74.04	73.90 (-)	73.69 (-)
(d)	Time-Decay Attentional Model [23]	Hand	76.41†	76.68†	76.05	76.34 (+)	76.41 (+)
(e)		E2E	76.67†	76.75†	76.26	76.43 (+)	76.67 (+)
(f)	Content-Aware + Time-Decay Attention [23]	Hand	75.48†	76.61†	75.16	75.27 (+)	75.48 (+)
(g)		E2E	75.83†	76.74†	75.82	75.92 (+)	75.83 (-)
(h)	**Context-Sensitive Time-Decay Attention**		**77.05**†	**76.87**†	76.62	76.96 (+)	77.05 (+)

Table 1. The understanding performance reported on F-measure in DSTC4, where the context length is 7 for each speaker (%). † indicates the significant improvement compared to all baseline methods ($p < 0.05$ on the one-tailed t-test). Hand: hand-crafted; E2E: end-to-end trainable.

ative effect by the coarse design of content-aware attention model, but leveraging both attention types ironically results in worse performance than using single time-decay attention (row (d)-(e)) [23]. The reasons may be that: 1) the harmful impact of low-quality content-aware attention is overwhelming, 2) the interaction between two types of attention during learning is not cooperative enough. Even though the row (g) in Table 1 learns both content- and time-aware attention functions, the time-decay attention curve is fixed after training; in other words, it is not content-responsive. If a history sentence contains salient information, it would be weighted by a small attention value from the time-decay attention curve regarding the large time difference.

Our proposed context-sensitive attention model effectively integrates time-aware and content-aware perspectives, where instead of training the content-aware and time-aware attention separately, we utilize contextual information to dynamically construct the time-decay attention curves. The results show that proposed role-based context-sensitive attention model (row (h)) outperform all compared baselines, yielding 9.7% improvement over the Naïve baseline (row (a)). As mentioned above, one can control the level of flexibility in the time-decay attention at will, it is possible that the combination may interfere attention model learning. Surprisingly, experiments show that the universal models outperform the models with a single time-decay attention type, demonstrating the positive interaction between attention functions and efficacy of our design.

3.4. Speaker Role in Attention Modeling

For role-level attention, Table 1 shows that all results with various time-decay attention mechanisms are better than the one with only content-aware attention (row (c)). Considering the benefit of considering speaker interactions [18, 19], therefore instead of weighting each utterance by its sentence-level attention, our model computes a representative attention value for each speaker by using the most important, representative utterances among what the speaker said. Namely, for role-level attention, each speaker role is assigned an attention value to represent the importance from the conversational interactions. By introducing role-level attention, the sentence-level attention weights can be smoothed to avoid inappropriate values and benefit language understanding. Surprisingly, even though learning sentence-level temporal attention is difficult, the proposed context-sensitive time-decay attention (row (h)) is the only one whose sentence-level results are better, further demonstrating the strong adaptability of fitting diverse dialogue contexts and the capability of capturing salient information.

The proposed methods are built on top of the role-base contextual framework, which utilizes separate modules to learn speaker-specific features to improve understanding. However, the prior time-decay attention models (rows (d)-(g)) are speaker-independent, where different speakers share the same decaying attention curve. To further investigate the effectiveness of the speaker role in attention modeling, we make the proposed context-sensitive attention speaker-dependent, so-called "role-based context-sensitive attention". The result (row (h)) shows that role-based attention modeling is promising, of which the universal design performs best. In sum, our attention model design not only elegantly combines content-aware and time-aware perspectives but effectively integrates the concept of speaker role modeling into attention mechanisms.

3.5. Robustness to Context Lengths

It is intuitive that longer context abounds richer information; however, it may obstruct attention learning and result in poor performance due to too much information for digesting and more noises for inaccurate estimation. Because when modeling dialogues, we have no idea about how many contexts are enough for better understanding, the robustness to varying context lengths becomes an important issue for contextual SLU. Here, we compare the results using different context lengths (3, 5, 7) for detailed analysis in Table 1, where the number is for each speaker. The results show that: 1) the models without attention and content-aware attention become slightly worse with increasing context lengths; 2) the time-decay attention models from the rows (d)-(g) in the Table 1 mostly achieve better performance when conducting longer contexts, where the model leveraging content-aware and time-aware attention by end-to-end learning outperforms the one under handcrafted setting whereas it weakens as context lengths become longer, showing less robustness to context lengths; 3) the proposed context-sensitive method performs the best for all context length settings, demonstrating not only the *flexibility* of adapting diverse contextual patterns but also the *robustness* to varying context lengths.

4. CONCLUSION

This paper designs a role-based context-sensitive time-decay attention functions based on an end-to-end contextual language understanding model, where different perspectives on dialogue contexts are analyzed. The experiments on a benchmark human-human dialogue dataset show that the understanding performance can be boosted by introducing the proposed attention mechanisms which elegantly integrate content-aware, time-ware, speaker-role perspectives. Furthermore, the proposed method is easily extensible to multi-party conversations and showing the potential of integrating temporal and contextual information in NLP tasks of dialogues.

5. REFERENCES

[1] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young, "A network-based end-to-end trainable task-oriented dialogue system," in *Proceedings of EACL*, 2017, pp. 438–449.

[2] Antoine Bordes, Y-Lan Boureau, and Jason Weston, "Learning end-to-end goal-oriented dialog," in *Proceedings of ICLR*, 2017.

[3] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng, "Towards end-to-end reinforcement learning of dialogue agents for information access," in *Proceedings of ACL*, 2017, pp. 484–495.

[4] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz, "End-to-end task-completion neural dialogue systems," in *Proceedings of The 8th International Joint Conference on Natural Language Processing*, 2017.

[5] Shang-Yu Su, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Yun-Nung Chen, "Discriminative deep dyna-q: Robust planning for dialogue policy learning," 2018.

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

[7] Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang, "Multi-domain joint semantic frame parsing using bi-directional rnn-lstm.," in *Proceedings of INTERSPEECH*, 2016, pp. 715–719.

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

[9] Yun-Nung Chen, Dilek Hakanni-Tür, Gokhan Tur, Asli Celikyilmaz, Jianfeng Guo, and Li Deng, "Syntax or semantics? knowledge-guided joint semantic frame parsing," in *Proceedings of 2016 IEEE Spoken Language Technology Workshop*, 2016, pp. 348–355.

[10] Zhangyang Wang, Yingzhen Yang, Shiyu Chang, Qing Ling, and Thomas S Huang, "Learning a deep l encoder for hashing," in *Proceedings of IJCAI*, 2016, pp. 2174–2180.

[11] Anshuman Bhargava, Asli Celikyilmaz, Dilek Hakkani-Tur, and Ruhi Sarikaya, "Easy contextual intent prediction and slot detection," in *Proceedings of ICASSP*, 2013, pp. 8337–8341.

[12] Puyang Xu and Ruhi Sarikaya, "Contextual domain classification in spoken language understanding systems using recurrent neural network," in *Proceedings of ICASSP*, 2014, pp. 136–140.

[13] Yun-Nung Chen, Ming Sun, Alexander I. Rudnicky, and Anatole Gershman, "Leveraging behavioral patterns of mobile applications for personalized spoken language understanding," in *Proceedings of ICMI*, 2015, pp. 83–86.

[14] Ming Sun, Yun-Nung Chen, and Alexander I. Rudnicky, "An intelligent assistant for high-level task understanding," in *Proceedings of IUI*, 2016, pp. 169–174.

[15] Yangyang Shi, Kaisheng Yao, Hu Chen, Yi-Cheng Pan, Mei-Yuh Hwang, and Baolin Peng, "Contextual spoken language understanding using recurrent neural networks," in *Proceedings of ICASSP*, 2015, pp. 5271–5275.

[16] Jason Weston, Sumit Chopra, and Antoine Bordesa, "Memory networks," in *Proceedings of ICLR*, 2015.

[17] Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng, "End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding.," in *Proceedings of INTERSPEECH*, 2016, pp. 3245–3249.

[18] Ta-Chung Chi, Po-Chun Chen, Shang-Yu Su, and Yun-Nung Chen, "Speaker role contextual modeling for language understanding and dialogue policy learning," in *Proceedings of IJCNLP*, 2017.

[19] Po-Chun Chen, Ta-Chung Chi, Shang-Yu Su, and Yun-Nung Chen, "Dynamic time-aware attention to speaker roles and contexts for spoken language understanding," in *Proceedings of ASRU*, 2017.

[20] Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev, "Addressee and response selection in multi-party conversations with speaker interaction rnns," in *Proceedings of AAAI*, 2018.

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

[22] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of ICML*, 2015, pp. 2048–2057.

[23] Shang-Yu Su, Pei-Chieh Yuan, and Yun-Nung Chen, "How time matters: Learning time-decay attention for contextual spoken language understanding in dialogues," in *Proceedings of NAACL-HLT*, 2018.

[24] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[25] Caiming Xiong, Stephen Merity, and Richard Socher, "Dynamic memory networks for visual and textual question answering," *arXiv preprint arXiv:1603.01417*, 2016.

[26] Seokhwan Kim, Luis Fernando DHaro, Rafael E Banchs, Jason D Williams, and Matthew Henderson, "The fourth dialog state tracking challenge," in *Proceedings of IWSDS*, 2016.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation.," in *Proceedings of EMNLP*, 2014, vol. 14, pp. 1532–1543.