How Time Matters: Learning Time-Decay Attention for Contextual Spoken Language Understanding in Dialogues

Shang-Yu Su[†] **Pei-Chieh Yuan**[†] **Yun-Nung Chen**[⋆]
[†]Department of Electrical Engineering
[⋆]Department of Computer Science and Information Engineering
National Taiwan University
{r05921117,b03901134}@ntu.edu.tw y.v.chen@ieee.org

Abstract

Spoken language understanding (SLU) is an essential component in conversational systems. Most SLU components treat each utterance independently, and then the following components aggregate the multi-turn information in the separate phases. In order to avoid error propagation and effectively utilize contexts, prior work leveraged history for contextual SLU. However, most previous models only paid attention to the related content in history utterances, ignoring their temporal information. In the dialogues, it is intuitive that the most recent utterances are more important than the least recent ones, in other words, time-aware attention should be in a decaying manner. Therefore, this paper designs and investigates various types of time-decay attention on the sentence-level and speaker-level, and further proposes a flexible universal time-decay attention mechanism. The experiments on the benchmark Dialogue State Tracking Challenge (DSTC4) dataset show that the proposed time-decay attention mechanisms significantly improve the state-of-the-art model for contextual understanding performance[1].

1 Introduction

Spoken dialogue systems that can help users to 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. Today, there are several virtual intelligent assistants, such as Apple's Siri, Google's Home, Microsoft's Cortana, and Amazon's Echo. Recent advance of deep learning has inspired many applications of neural models to dialogue systems (Wen et al., 2017; Bordes et al., 2017; Dhingra et al., 2017; Li et al., 2017).

A key component of a dialogue system is a spoken language understanding (SLU) module—it parses user utterances into semantic frames that capture the core meaning (Tur and De Mori, 2011). A typical pipeline of SLU is to first decide the domain given the input utterance, and based on the domain, to predict the intent and to fill associated slots corresponding to a domain-specific semantic template, where each utterance is treated independently (Hakkani-Tür et al., 2016; Chen et al., 2016b,a; Wang et al., 2016). To overcome the error propagation and further improve understanding performance, the contextual information has been shown useful (Bhargava et al., 2013; Xu and Sarikaya, 2014; Chen et al., 2015; Sun et al., 2016). Prior work incorporated the dialogue history into the recurrent neural networks (RNN) for improving domain classification, intent prediction, and slot filling (Xu and Sarikaya, 2014; Shi et al., 2015; Weston et al., 2015; Chen et al., 2016c). Recently, Chi et al. (2017) and Zhang et al. (2018) demonstrated that modeling speaker role information can learn the notable variance in speaking habits during conversations in order to benefit understanding.

In addition, neural models incorporating attention mechanisms have had great successes in machine translation (Bahdanau et al., 2014), image captioning (Xu et al., 2015), and various tasks. Attentional models have been successful because they separate two different concerns: 1) deciding which input contexts are most relevant to the output and 2) actually predicting an output given the most relevant inputs. For example, the highlighted current utterance from the tourist, "*uh on august*", in the conversation of Figure 1 is to respond the question about `WHEN`, and the content-

[1]The source code is at: https://github.com/MiuLab/Time-Decay-SLU.

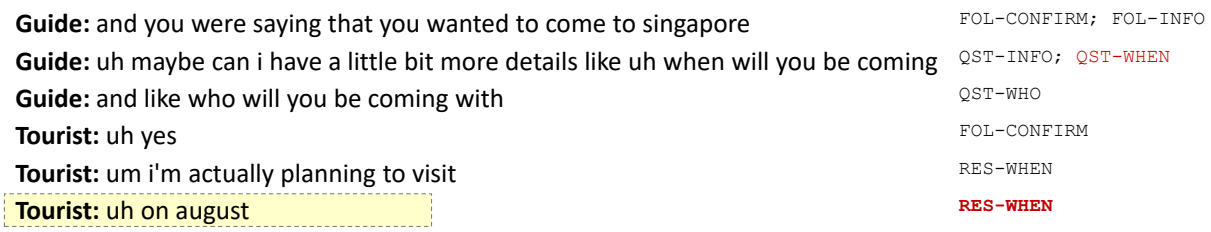

Figure 1: The human-human conversational utterances and their associated semantic labels from DSTC4.

aware contexts that can help current understanding are the first two utterances from the guide "*and you were saying that you wanted to come to singapore*" and "*un maybe can i have a little bit more details like uh when will you be coming*". Previous 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 (Chen et al., 2017). However, the time-aware attention function is an inflexible hand-crafted setting, which is a fixed function of time for assessing the attention.

This paper focuses on investigating various flexible time-aware attention mechanism in neural models with contextual information and speaker role modeling for language understanding. The contributions are three-fold:

- This paper investigates different time-aware attention mechanisms and provides guidance for the future research about designing the time-aware attention function.
- This paper proposes an end-to-end learnable universal time-decay mechanism with great flexibility of modeling temporal information for diverse dialogue contexts.
- The proposed model achieves the state-of-the-art understanding performance in the dialogue benchmark DSTC dataset.

2 The Proposed Framework

The model architecture is illustrated in Figure 2. 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 role levels. The whole model is trained in an end-to-end fashion, where the history summary vector and the attention weights 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, $p(\hat{\mathbf{y}} \mid \mathbf{x})$, by minimizing the cross-entropy loss.

2.1 Speaker Role Contextual Language Understanding

Given the current utterance $\mathbf{x} = \{w_t\}_1^T$, the goal is to predict the user intents of $\mathbf{x}$, which includes the speech acts and associated attributes. We apply a bidirectional long short-term memory (BLSTM) model (Schuster and Paliwal, 1997) to history encoding in order to learn the probability distribution of the user intents.

$$\mathbf{v}_{\text{cur}} = \text{BLSTM}(\mathbf{x}, W_{\text{his}} \cdot \mathbf{v}_{\text{his}}), \quad (1)$$

$$\mathbf{o} = \texttt{sigmoid}(W_{\text{SLU}} \cdot \mathbf{v}_{\text{cur}}), \quad (2)$$

where W_{his} is a weight matrix and $\mathbf{v}_{\text{his}}$ is the history summary vector, $\mathbf{v}_{\text{cur}}$ 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 dense 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 (Chi et al., 2017), we follow the prior work for utilizing the contexts from two roles to learn history summary representations, $\mathbf{v}_{\text{his}}$ in (1), in order to leverage the role-specific contextual information. Each role-dependent recurrent unit $\text{BLSTM}_{\text{role}_i}$ receives corresponding inputs, x_{t,role_i}, which includes multiple utterances u_i ($i = [1, ..., t-1]$) preceding the current utterance u_t from the specific role, role_i, and have been processed by an encoder model.

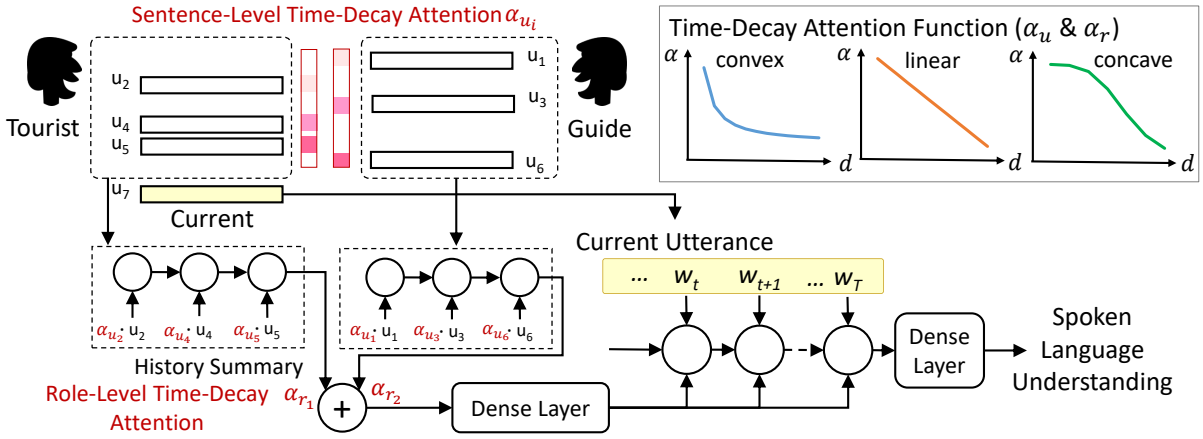

Figure 2: Illustration of the proposed time-aware attention contextual model with three types of time-decay attention functions.

$$\mathbf{v}_{\text{his}} = \sum_{\text{role}} \mathbf{v}_{\text{his,role}} \quad (3)$$
$$= \sum_{\text{role}} \text{BLSTM}_{\text{role}}(x_{t,\text{role}}),$$

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 $\mathbf{v}_{\text{cur}}$ from (1) would contain role-based contextual information.

2.2 Neural Attention Mechanism

One of the earliest work with a memory component applied to language processing is memory networks (Weston et al., 2015; Sukhbaatar et al., 2015), which encodes mentioned facts into vectors and stores them in the memory for question answering. The idea is to encode important knowledge and store it into memory for future usage with attention mechanisms. Attention mechanisms allow neural network models to selectively pay attention to specific parts. There are also various tasks showing the effectiveness of attention mechanisms (Xiong et al., 2016; Chen et al., 2016c). 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 based on the investigation of different time-decay functions, and further learning an universal time-decay function automatically. For time-aware attention mechanisms, we apply it using two levels, sentence-level and role-level structures, and Section 3 details the design and analysis of time-aware attention.

For the sentence-level attention, before feeding into the contextual module, each history vector is weighted by its time-aware attention α_{u_j} for replacing (3):

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

For the role-level attention, a dialogue is disassembled from a different perspective on which speaker's information is more important (Chi et al., 2017). The role-level attention is to decide how much to address on different speaker roles' contexts ($\mathbf{v}_{\text{his,role}}$) in order to better understand the current utterance. The importance of a speaker given the contexts can be approximated to the maximum attention value among the speaker's utterances, $\alpha_{\text{role}} = \max \alpha_{u_j}$, where u_j includes all contextual utterances from the speaker. With the role-level attention, the sentence-level history from (3) can be rewritten into

$$\mathbf{v}^R_{\text{his}} = \sum_{\text{role}} \alpha_{\text{role}} \cdot \mathbf{v}_{\text{his,role}} \quad (4)$$

for combining role-dependent history vectors with their attention weights.

2.3 End-to-End Training

The objective is to optimize SLU performance, predicting multiple speech acts and attributes described in Section 2.1. In the proposed model,

all encoders, prediction models, and attention weights can be automatically learned in an end-to-end manner.

3 Time-Decay Attention Learning

The decaying function curves can be easily separated into three types: *convex*, *linear*, and *concave*, illustrated in the top-right part of Figure 2, and each type of time-decay functions expresses a time-aware perspective given dialogue contexts. Note that all attention weights will be normalized such that their summation is equal to 1.

3.1 Convex Time-Decay Attention

A *convex* curve also known as "concave upward", in a simple 2D Cartesian coordinate system (x, y), a convex curve $f(x)$ means when x goes greater, the slope $f'(x)$ is increasing. Intuitively, recent utterances contain more salient information, and the salience decreases very quickly when the distance increases; therefore we introduce the time-aware attention mechanism that computes attention weights according to the time of utterance occurrence explicitly. We first define the time difference between the current utterance and the preceding sentence u_i as $d(u_i)$, and then simply use its reciprocal to formulate a convex time-decay function:

$$\alpha_{u_i}^{\text{conv}} = \frac{1}{a \cdot d(u_i)^b}, \tag{5}$$

where a and b are scalar parameters.

The increasing slopes of the decay-curve assert that importance of utterances should be attenuated rapidly, and the importance of a earlier history sentence would be considerably compressed. Note that Chen et al. used a fixed convex time-decay function ($a = 1, b = 1$) (Chen et al., 2017).

3.2 Linear Time-Decay Attention

A *linearly* decaying time-aware attention function should also be taken into consideration. In a simple 2D Cartesian coordinate system (x, y), the slopes of a linear function remain consistent when x changes. That is, the importance of preceding utterances linearly declines as the distance between the previous utterance and the target utterance becomes larger.

$$\alpha_{u_i}^{\text{lin}} = \max(e \cdot d(u_i) + f, 0), \tag{6}$$

where e and f are the slope and the α-intercept of the linear function. Note that when the distance $d(u_i)$ is larger than $-\frac{f}{e}$, we assign the attention value as 0.

3.3 Concave Time-Decay Attention

A *concave* curve also called "concave downward", in contrast to convex curves, in a simple 2D Cartesian coordinate system (x, y), a concave curve $f(x)$ means that the slope $f'(x)$ is decreasing when x goes greater. Intuitively, the attention weight decreases relatively slow when the distance increases. To implement this idea, we design a $Butterworth\ filter$-like low-distance pass filter (Butterworth, 1930) that is similar to the concave time-decay function in the beginning of the curve.

$$\alpha_{u_i}^{\text{conc}} = \frac{1}{1 + (\frac{d(u_i)}{D_0})^n}, \tag{7}$$

where D_0 is the cut-off distance and n is the order of filter. The decreasing slopes of the decay-curve assert that the importance of utterances should weaken gradually, and the importance of a earlier history sentence would still be considerably compressed. Moreover, it is more likely to preserve the information in the multiple recent utterances instead of focusing only on the most recent one.

3.4 Universal Time-Decay Attention

As mentioned previously, there are three types of decaying curves: convex, linear, concave, each type represents a different perspective on dialogue contexts and models different contextual patterns. However, because the contextual patterns may be diverse, a single type of function could not fit the complex behavior well. Hence, we propose a flexible and universal time-decay attention function by composing three types of attentional curves:

$$\begin{aligned} \alpha_{u_i}^{\text{univ}} &= w_1 \cdot \alpha_{u_i}^{\text{conv}} + w_2 \cdot \alpha_{u_i}^{\text{lin}} + w_3 \cdot \alpha_{u_i}^{\text{conc}} \\ &= \frac{w_1}{a \cdot d(u_i)^b} + w_2(e \cdot d(u_i) + f) \\ &\quad + \frac{w_3}{1 + (\frac{d(u_i)}{D_0})^n}, \end{aligned} \tag{8}$$

where w_i are the weights of time-decay attention functions. 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 proposed universal time-decay attention function expresses the flexibility of not being

LU Model			Sentence-Level Attention				Role-Level Attention			
			Conv.	Lin.	Conc.	Univ.	Conv.	Lin.	Conc.	Univ.
(a)	*DSTC4-Best*[2]		61.4							
(b)	Naïve LU		70.18							
(c)	No Attention Context		74.52							
(d)	Content-Aware Context		73.69				74.28			
(e)	Time-Aware	(Hand)	75.95†	74.12	74.26	76.41†	76.73†	76.11†	76.01†	76.68†
(f)		(E2E)	76.04†	74.25	74.32	**76.67**†	76.69†	76.26†	76.08†	**76.75**†
(g)	Content+Time	(Hand)	74.71†	73.40	73.28	75.48†	76.70†	76.24†	76.03†	76.61†
(h)		(E2E)	74.94†	73.79	73.47	**75.83**†	76.51†	75.76†	76.22†	**76.74**†

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. Hand: hand-crafted; E2E: end-to-end trainable.

strictly decaying; that is, the model can automatically learn a properly oscillating curve in order to model the diverse and complex contextual patterns using the attention mechanism.

4 Experiments

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

4.1 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, these 35 dialogs sum up to 31,034 utterances and 273,580 words (Kim et al., 2016). 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. Unlike previous DSTC series collected human-computer dialogues, human-human dialogues contain rich and complex human behaviors and bring much difficulty to all the tasks. Given the complex dialogue patterns and longer contexts, DSTC4 is a suitable benchmark dataset for evaluation. We randomly selected 28 dialogues as the training set, 5 dialogues as the testing set, and 2 dialogues as the validation set.

We choose the mini-batch *Adam* as the optimizer with the batch size of 256 examples. The size of each hidden recurrent layer is 128. We use pre-trained 200-dimensional word embeddings $GloVe$ (Pennington et al., 2014). We only apply 30 training epochs without any early stop approach. 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 five runs. We include the best understanding performance (row (a)) from the participants of DSTC4 in IWSDS 2016 for reference (Kim et al., 2016). The one-tailed t-test is performed to validate the significance of improvement, and the numbers with markers indicate the significant improvement with $p < 0.05$.

4.2 Effectiveness of Time-Decay Attention

To evaluate the proposed time-decay attention, we compare the performance with the naïve LU model without any contextual information (row (b)), the contextual model without any attention mechanism (row (c)), and the one using the content-aware attention mechanism (row (d)), where the attention can be learned at sentence and role levels. It is intuitive that the model without considering contexts (row (b)) performs much worse than the contextual ones for dialogue modeling. The proposed time-aware results are shown in the rows (e)-(h), where the rows (e)-(f) use only the time-aware attention while the rows (g)-(h) model both content-aware and time-aware attention mechanisms together. It is obvious that almost all time-aware results are better than three baselines.

In order to investigate the performance of various time-decay attention functions, for each curve we apply two settings: 1) **Hand**: hand-crafted hyper-parameters (rows (e) and (g)) and 2) **E2E**: end-to-end training for parameters (rows (f) and (h)). In the hand-crafted setting, the hyper-

parameters $a = 1, b = 1, e = -0.125, f = 1, D_0 = 5, n = 3$ are adopted[3]. Table 1 shows that among three types of the sentence-level time-decay attention, only the convex time-decay attention significantly outperforms the baselines, indicating that an unsuitable time-decay attention function is barely useful. For both settings, the convex functions perform best among the three types of time-decay functions. Also, the end-to-end trainable setting results in better performance for most cases.

For our proposed universal time-decay attention mechanism, the same settings are conducted: 1) composing fixed versions for three types of time-decay functions weighted by learned parameters w_i and 2) fully trainable parameters for all time-decay functions. These two settings provide different levels of flexibility in fitting dialogue contextual attention, and the experimental results show that two settings both outperform all other time-decay attention functions.

For sentence-level attention, the end-to-end trainable universal time-decay attention achieves best performance (rows (f) and (h)), where the flexible time-aware attention (rows (f) and (h)) obtains 2.9% relative improvement compared to the model without the attention mechanism (row (c)) and the model using content-aware attention only (row (d)). For role-level attention, all types of time-decay functions significantly improve the results. The probably reason may be that modeling temporal importance for each sentence is more difficult and less accurate, and speaker roles in the dialogues provide informative cues for the model to connect the temporal importance from the same speakers together; therefore, the conversational patterns can be considered to additionally improve the understanding results. The further analysis is discussed in Section 4.3. Similarly, the best results are also from the end-to-end trainable universal time-decay function.

The significant improvement achieved by the universal functions indicates that our model can effectively learn a suitable attention function through this flexible setting and derive a proper curve to fit the temporal tendency to help the model preserve the essence and drop unimportant parts in the dialogue contexts. To further investigate what the universal time-decay attention learns, we inspect the learned weights w_i and find that the convex attention function almost dominates the whole function. In other words, our model automatically learns that the convex time-decay attention is more suitable for modeling contexts from the dialogue data than the other two types. Therefore, we can conclude that in complex dialogues, the recent utterances contain majority of salient information for spoken language understanding, where the attention decay trend follows a convex curve.

We analyze the content-aware attention impact by comparing the results between time-aware only (rows (e)-(f)) and content and time-aware jointly (rows (g)-(h)). The content-aware attention (row (d)) fails to focus on the important contexts for improving understanding performance in the complex dialogues and even performs slightly worse than the contextual model without attention (row (c)). Without a delicately-designed attention mechanism, it is not guaranteed that incorporating an additional content-aware attention would bring better performance and the experimental results show that a simple and coarse content-aware attention barely provides any usable information given the complex dialogues. Therefore, we focus on whether our time-aware attention mechanisms can compensate the poor attention learned from the content-aware model. In other words, we are not going to verify whether our time-aware attention mechanisms could collaborate with the content-aware attention mechanism, instead, we focus on examining how much our proposed time-aware attention could mitigate the detriment of the content-aware attention. By comparing the results between time-aware only (rows (e)-(f)) and content and time-aware jointly (rows (g)-(h)), we find that our universal time-decay attention keeps the improvement without too much performance drop by involving the learned temporal attention. Namely, our proposed attention mechanism can capture temporal information precisely, and it therefore can counteract the harmful impact of inaccurate content-aware attention.

4.3 Effectiveness of Role-Level Attention

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 (d)). However, linear and concave time-decay functions do not provide addi-

[3]The chosen parameters are based on the domain knowledge about dialogue properties.

LU Model	Context Length		
	3	5	7
No Attention Contextual	74.75	74.69 (-)	74.52 (-)
Content-Aware Contextual	74.04	73.90 (-)	73.69 (-)
Time-Aware (Hand)	76.05	76.34 (+)	76.41 (+)
Time-Aware (E2E)	76.26	76.43 (+)	76.67 (+)
Content+Time (Hand)	75.16	75.27 (+)	75.48 (+)
Content+Time (E2E)	75.82	75.92 (+)	75.83 (-)

Table 2: The sentence-level performance reported on F1 of the proposed universal time-decay attention under different context length settings (%). The symbols '+' and '-' indicate the performance trends.

Parameter	Time-Aware (E2E Trainable)	
	Sentence	Role
w_1	0.758	1.078
w_2	0.544	-0.378
w_3	-0.302	0.300
a	0.888	0.841
b	0.969	1.084
e	-0.320	-0.129
f	0.640	0.993
D_0	4.873	4.980
n	2.977	2.755

Table 3: The converged values of end-to-end trainable parameters from the proposed universal time-decay attention models. The values are averaged over five runs.

tional improvement when we model the attention at the sentence level. The probable reason may be that it is difficult to model attention for individual sentences given the unsuitable time-decay functions. That is, if designs of attention functions are unsuitable for dialogue contexts, the encoded sentence embeddings would be weighted by improper attention values. On the other hand, for role-level attention, each speaker role is assigned an attention value to represent their importance in the conversational interactions. Previous work (Chi et al., 2017; Chen et al., 2017) also demonstrated the effectiveness of considering speaker interactions for better understanding performance. By introducing role-level attention, the sentence-level attentional weights can be smoothed to avoid inappropriate values. Surprisingly, even though learning sentence-level temporal attention is difficult, our proposed universal time-decay attention can achieve similar performance for sentence-level and role-level attention (76.67% and 76.75% from the row (f)), further demonstrating the strong adaptability of fitting diverse dialogue contexts and the capability of capturing salient information.

4.4 Robustness to Context Lengths

It is intuitive that longer context brings richer information; however, it may obstruct the attention learning and result in poor performance because more information should be modeled and accurate estimation is not trivial. Because when modeling dialogues, we have no idea about how many contexts are enough for better understanding, the robustness to varying context lengths is important for the contextual model design. Here, we compare the results using different context lengths (3, 5, 7) for detailed analysis in Table 2, where the number is for each speaker. The models without attention and the content-aware models become slightly worse with increasing context lengths. However, our proposed universal time-decay attention model mostly achieves better performance when including longer contexts, demonstrating not only the flexibility of adapting diverse contextual patterns but also the robustness to varying context lengths.

4.5 Universal Time-Decay Attention Analysis

This paper proposes a flexible time-decay attention mechanism by composing three types of time-aware attention functions in different decaying tendencies, where each decaying curves reflect a specific perspectives on distribution over salient information in dialogue contexts. The proposed universal time-decay attention shows great capability of modeling diverse dialogue patterns in the experiments and therefore proves that our proposed method is a general design of time-decay attention. In our design, we endow the attention function with flexibility by employing many trainable parameters and hence it can automatically learn a properly decaying curve for fitting the dialogue contexts better.

To further analyze the combination of different time-decay attention functions, we inspect the converged values of the trainable parameters from the proposed universal time-decay attention models in Table 3. Under the end-to-end trainable setting, the initialization of the trainable parameters are the same as the hand-crafted ones ($w_i = 1, a = 1, b = 1, e = -0.125, f = 1, D_0 = 5, n = 3$). In the experiments, the models automatically figure out that convex time-decay attention function should have a higher weight than others for both sentence-level or role-level models ($w_1 > w_2$ and $w_1 > w_3$). Namely, in dialogue contexts, the recent utterances contain most information related

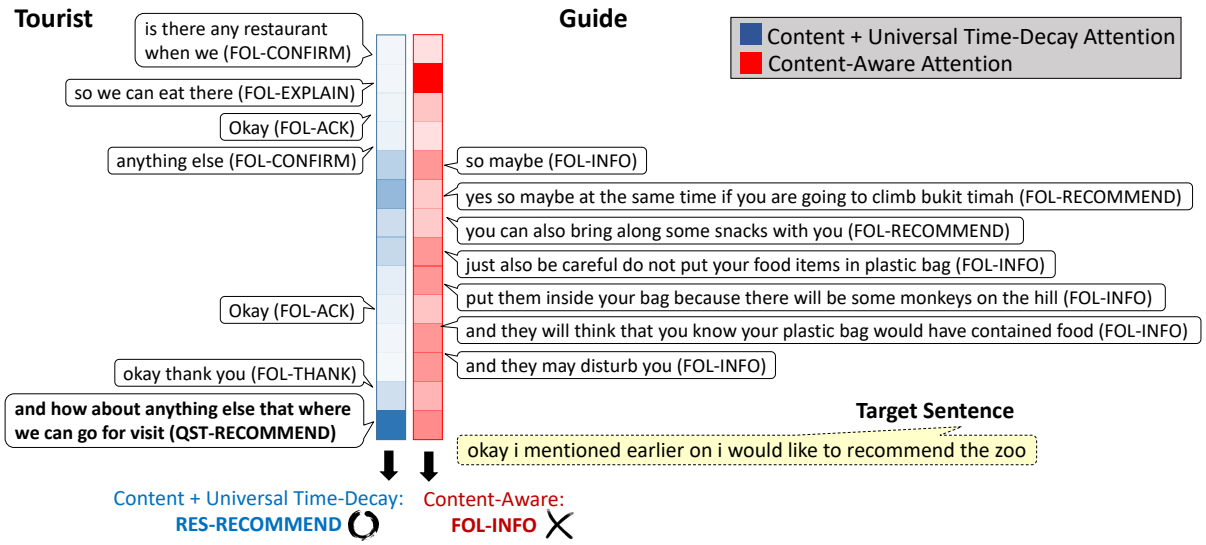

Figure 3: The visualization of the attention weights enhanced by the proposed time-decay function compared with the weights learned by the content-aware attention model.

to the current utterance, which is aligned with our intuition.

4.6 Qualitative Analysis

From the above experiments, the proposed time-decay attention mechanisms significantly improve the performance on both sentence and role levels. To further understand how the time-decay attention changes the content-aware attention, we dig deeper into the learned attentional values for sentences and illustrate the visualization in Figure 3. The figure shows a partial dialogue between the tourist (left) and the guide (right), where the color shades indicate the learned attention intensities of sentences. It can be found that the learned content-aware attention (red; row (c)) focuses on the incorrect sentence ("*so we can eat there*" (FOL-EXPLAIN)) and hence predicts the wrong label, FOL-INFO. The reason may be that with a coarse and simple design of content-aware attention mechanism, the attention function may not provide additional benefit for improvement. By additionally leveraging our proposed universal time-decay attention methods, the result (blue; row (g)) shows that the adjusted attention pays the highest attention on the most recent utterance and thereby predicts the correct intent, RES-RECOMMEND. It can be found that our proposed time-decay attention can effectively turn the attention to the correct contexts in order to correctly predict the dialogue act and attribute. Therefore, the proposed attention mechanisms are demonstrated to be effective for improving understanding performance in such complex human-human conversations.

5 Conclusion

This paper designs and investigates various time-decay attention functions based on an end-to-end contextual language understanding model, where different perspectives on dialogue contexts are analyzed and a flexible and universal time-decay attention mechanism is proposed. The experiments on a benchmark human-human dialogue dataset show that the understanding performance can be boosted by simply introducing the proposed time-decay attention mechanisms for guiding the model to focus on the salient contexts following a convex curve. Moreover, the proposed universal time-decay mechanisms are easily extensible to multi-party conversations and showing the potential of leveraging temporal information in NLP tasks of dialogues.

Acknowledgements

We would like to thank reviewers for their insightful comments on the paper. The authors are supported by the Ministry of Science and Technology of Taiwan, Google Research, Microsoft Research, and MediaTek Inc..

References

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

Anshuman Bhargava, Asli Celikyilmaz, Dilek Hakkani-Tur, and Ruhi Sarikaya. 2013. Easy contextual intent prediction and slot detection. In *Proceedings of ICASSP*. pages 8337–8341.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of ICLR*.

Stephen Butterworth. 1930. On the theory of filter amplifiers. *Wireless Engineer* 7(6):536–541.

Po-Chun Chen, Ta-Chung Chi, Shang-Yu Su, and Yun-Nung Chen. 2017. Dynamic time-aware attention to speaker roles and contexts for spoken language understanding. In *Proceedings of ASRU*. pages 554–560.

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

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, Gökhan Tür, Jianfeng Gao, and Li Deng. 2016c. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Proceedings of INTERSPEECH*. pages 3245–3249.

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

Ta-Chung Chi, Po-Chun Chen, Shang-Yu Su, and Yun-Nung Chen. 2017. Speaker role contextual modeling for language understanding and dialogue policy learning. In *Proceedings of IJCNLP*. pages 163–168.

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

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 *Proceedings of INTERSPEECH*. pages 715–719.

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

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of The 8th International Joint Conference on Natural Language Processing*. pages 733–743.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*. volume 14, pages 1532–1543.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

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

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Proceedings of NIPS*. pages 2431–2439.

Ming Sun, Yun-Nung Chen, and Alexander I. Rudnicky. 2016. An intelligent assistant for high-level task understanding. In *Proceedings of IUI*. pages 169–174.

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

Zhangyang Wang, Yingzhen Yang, Shiyu Chang, Qing Ling, and Thomas S Huang. 2016. Learning a deep l encoder for hashing. In *Proceedings of IJCAI*. pages 2174–2180.

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

Jason Weston, Sumit Chopra, and Antoine Bordesa. 2015. Memory networks. In *Proceedings of ICLR*.

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

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

Puyang Xu and Ruhi Sarikaya. 2014. Contextual domain classification in spoken language understanding systems using recurrent neural network. In *Proceedings of ICASSP*. pages 136–140.

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