LEARNING CONTEXT-SENSITIVE TIME-DECAY ATTENTION FOR ROLE-BASED DIALOGUE MODELING

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

National Taiwan University, Taipei, Taiwan

shangyusu.tw@gmail.com 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 content in history utterances and ignored the temporal information. In dialogues, it is intuitive that the most recent utterances are more important than the least recent ones, hence time-aware attention should be in a decaying manner. Therefore, this paper allows the model to automatically learn a time-decay attention function 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 role-based context-sensitive time-decay attention mechanisms significantly improve the state-of-the-art model for contextual understanding performance[1].

Index Terms— Spoken language understanding, spoken dialogue systems, 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].

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 [5]. A typical pipeline of SLU is to first determine the domain given input utterances, and based on the domain to predict the intent and to fill the associated slots corresponding to a domain-specific semantic template. 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, the contextual information has been leveraged and shown useful [6, 7, 8, 9]. Prior work incorporating dialogue contexts into the recurrent neural networks (RNN) for improving domain classification, intent prediction, and slot filling [7, 10, 11, 12]. Recently, [13, 14, 15] demonstrated that modeling speaker role information can learn the notable variance in speaking habits during conversations for better performance.

Neural models incorporating attention mechanisms have had great successes in machine translation [16], image captioning [17], 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. In dialogues, although content-aware contexts may help understanding, 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 [14]. However, the time-aware attention function is an inflexible hand-crafted setting, which is a fixed function of time for assessing the attention.

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. THE PROPOSED 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 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, $p(\mathbf{y} \mid \mathbf{x})$, by minimizing the cross-entropy loss between the predicted distribution and the target distribution.

[1] The code will be released once accepted.

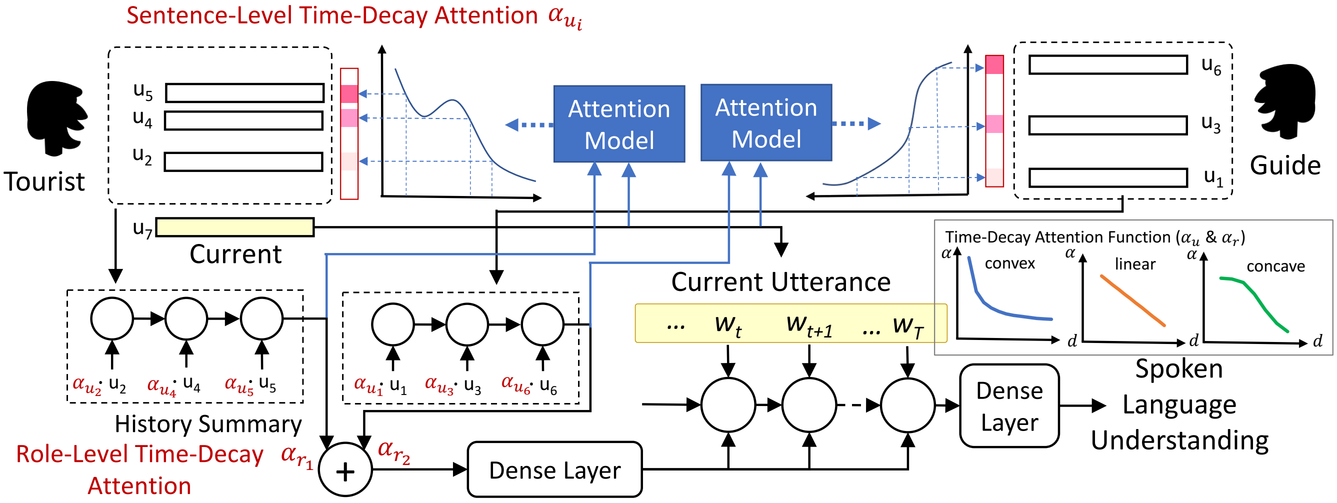

Fig. 1. Illustration of the proposed role-based context-sensitive time-decay attention contextual model.

2.1. Speaker Role Contextual SLU

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, such as QST_WHERE. We apply the bidirectional long short-term memory (BLSTM) model [18] to context encoding to learn the probability distribution of user intents.

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

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

where W_{his} and W_{SLU} are dense 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 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 [13, 15], we utilize 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 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.

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

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 (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 [11, 19], which encode 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, which allow the networks to selectively pay attention to specific parts. There are also various tasks showing the effectiveness of attention mechanisms [20, 12]. 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 flexible 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, 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_{role_i} for replacing (3):

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

For the role-level attention, a dialogue is disassembled from a different perspective on which speaker's information is more important [13]. 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 role. With the role-level attention, the sentence-level history from (3) can be rewritten into

$$\mathbf{v}_{\text{his}}^R = \sum_{\text{role}} \alpha_{\text{role}} \cdot \mathbf{v}_{\text{his,role}} \quad (5)$$

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

3. 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. The decaying function curves can be easily separated into three types: *convex*, *linear*, and *concave*, illustrated in the top-right part of Figure 1, 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:

$$\alpha_{u_i}^{\text{conv}} = \frac{1}{a \cdot d(u_i)^b} \tag{6}$$

where $d(u_i)$ denotes the time difference between the current utterance and the preceding one, u_i, and a and b are scalar parameters. Note that [14] used a fixed convex time-decay function ($a = 1, b = 1$).

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.

$$\alpha_{u_i}^{\text{lin}} = \max(e \cdot d(u_i) + f, 0) \tag{7}$$

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 [21] that is similar to the concave time-decay function in the beginning of the curve.

$$\alpha_{u_i}^{\text{conc}} = \frac{1}{1 + (\frac{d(u_i)}{D_0})^n}, \tag{8}$$

where D_0 is the cut-off distance and n is the order of filter. 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

There are three types of decaying curves: convex, linear, concave, where 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, a flexible and universal time-decay attention function that composes three types of attentional curves is formulated:

$$\begin{aligned} \alpha_{u_i}^{\text{univ}} &= w_1 \cdot \alpha_{u_i}^{\text{conv}} + w_2 \cdot \alpha_{u_i}^{\text{lin}} + w_3 \cdot \alpha_{u_i}^{\text{conc}} \\ &= \frac{w_1}{a \cdot d(u_i)^b} + w_2(e \cdot d(u_i) + f) + \frac{w_3}{1 + (\frac{d(u_i)}{D_0})^n}, \end{aligned} \tag{9}$$

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 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.

3.5. Role-Based 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 encode context-sensitive characteristics into the attention mechanisms. 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 speaker may have totally different speaking behavior [13, 14, 15], 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.

$$\mathbf{v}_{\text{his,role}} = \text{BLSTM}_{\text{role}}(x_{t,\text{role}}), \tag{10}$$

$$\mathbf{p}_{\text{role}} = W_{\text{p,role}} \cdot (\mathbf{v}_{\text{his,role}}, \mathbf{v}_{\text{cur}}) + bias, \tag{11}$$

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,

$$-\alpha \cdot \min(\mathbf{p}_{\text{role}}, 0) + \beta \cdot \sum \mathbf{p}_{\text{role}}^2. \tag{12}$$

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.

3.6. End-to-End Training

The objective is to optimize SLU performance, predicting multiple speech acts and attributes. In the proposed model, all encoders, prediction models, and attention models can be automatically learned in an end-to-end manner.

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, including 31,034 utterances and 273,580 words [23]. 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 $GloVe$ [24]. 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.

4.2. 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, where (d)-(e) use only the time-decay attention, while (f)-(g) model both content-aware and time-decay attention mechanisms together. 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 [22]. Our proposed context-sensitive time-decay attention model is shown in the row (h).

Table 1 shows that almost all models with the time-decay attention (row (d)-(e)) 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. By introducing time-decay attention, the experimental results show that all models with role-level attention and the some with sentence-level attention obtain the improvement, where the universal design of the proposed context-sensitive attention model (row (h)) achieves the state-of-the-art performance.

4.3. 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 negative 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)) [22]. 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 for training a single attention model for constructing the time-decay attention curves, so-called "context-sensitive time-decay attention".

Considering that the parameters used to determine the tendency of decaying curves are predicted by neural models, we set two fully-connected layers of which size is 64 or 16 in the attention models, where the universal time-decay attention uses smaller size due to the number of attention models, so that the results can be fairly compared. The results show that most time-decay attention in the proposed role-based context-sensitive attention model (row (h)) outperform all compared baselines. Furthermore, the context-sensitive universal time-decay attention achieves the best performance, yielding 9.7% improvement over the Naïve baseline (row (a)). Note that although the additional attention models and recurrent units are conducted, we cut the hidden layers of the LSTM units into half-size to keep the same model size.

4.4. Speaker Role in Attention Modeling

For role-level attention, Table 1 shows that almost all results with various time-decay attention mechanisms are better than the one with only content-aware attention (row (c)). However, linear and concave time-decay functions in the rows (d)-(g) do not provide additional improvement when modeling the sentence-level attention over the baselines (row (c)) and even perform worse than the one without any attention mechanism (row (b)). The probable cause may be that it is

	SLU Model		Sentence-Level Attention				Role-Level Attention			
			Conv.	Lin.	Conc.	Univ.	Conv.	Lin.	Conc.	Univ.
(a)	Naïve SLU		70.18							
(b)	No Attention Contextual Model		74.52							
(c)	Content-Aware Contextual Model [13]		73.69				74.28			
(d)	Time-Decay Attentional Model [22]	Hand	75.95†	74.12	74.26	76.41†	76.73†	76.11†	76.01†	76.68†
(e)		E2E	76.04†	74.25	74.32	**76.67**†	76.69†	76.26†	76.08†	**76.75**†
(f)	Content-Aware + Time-Decay Attention [22]	Hand	74.71†	73.40	73.28	75.48†	76.70†	76.24†	76.03†	76.61†
(g)		E2E	74.94†	73.79	73.47	**75.83**†	76.51†	75.76†	76.22†	**76.74**†
(h)	**Context-Sensitive Time-Decay Attention**		74.89†	74.33	74.33	**77.05**†	74.24	74.76	74.69	**76.87**†

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.

difficult to model attention for individual sentences by means of the unsuitable time-decay functions, so that all results using sentence-level attention are worse than the ones with role-level attention in the baselines (rows (c)-(g)). In other words, if attention function designs are unsuitable for dialogue contexts, input sentences would be weighted by improper attention values.

Considering the benefit of considering speaker interactions [13, 14], 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 attentional 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 universal 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.

4.5. 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 2, 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 universal 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 (Content+Time (E2E)) outperforms the one under handcrafted setting (Content+Time (Hand)) 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.

SLU Model	Context Length		
	3	5	7
No Attention Context	74.75	74.69 (-)	74.52 (-)
Content-Aware Context	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 (-)
Context-Sensitive	76.62	76.96 (+)	77.05 (+)

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

4.6. Analysis of Universal Function

The proposed universal time-decay attention mechanism flexibly composes three types of time-aware attention functions in different decaying tendencies, each of which reflects a specific perspectives on distribution over salient information in dialogue contexts. It shows great capability of modeling diverse dialogue patterns in the experiments and therefore empirically shows that the proposed method is a general design of time-decay attention. The design of universal time-decay attention is simple, general and easily-extensible, the weights and the parameters in each type of attention can be assigned via hyperparameters (row (d) and (f)), initialized by hyperparameters and fine-tuned by end-to-end training (row (e) and (g)), or learned by fully end-to-end learning (row (h)).

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 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 $w_1 > w_3$). In other words, it reflects that the majority of salient in-

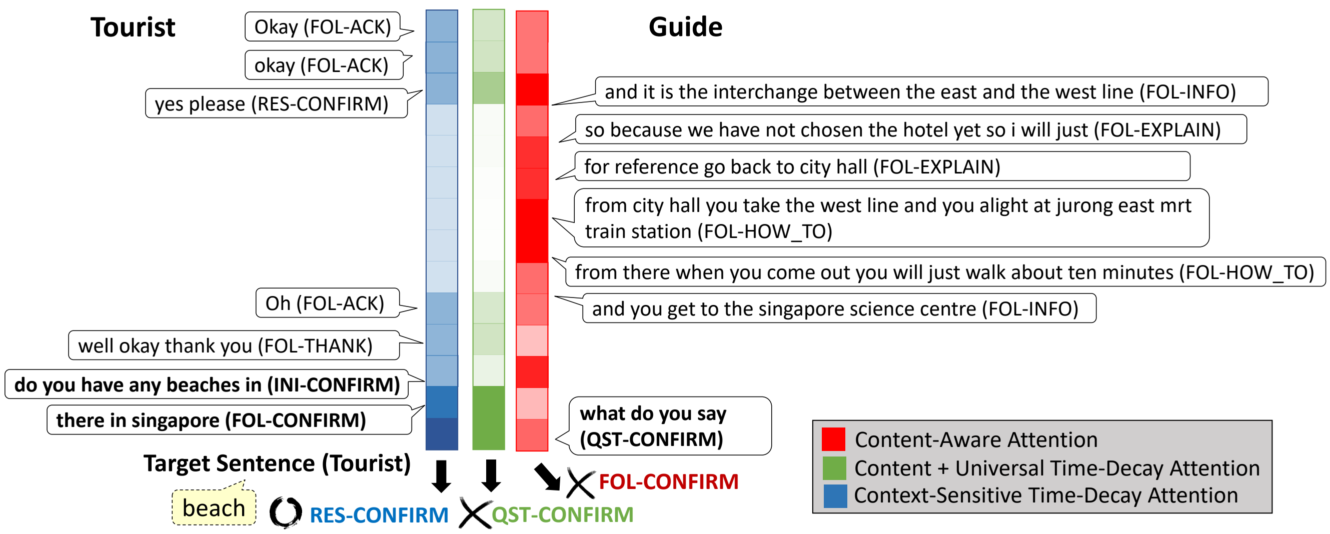
Tourist
Guide
Okay (FOL-ACK)
okay (FOL-ACK)
yes please (RES-CONFIRM)
and it is the interchange between the east and the west line (FOL-INFO)
so because we have not chosen the hotel yet so i will just (FOL-EXPLAIN)
for reference go back to city hall (FOL-EXPLAIN)
from city hall you take the west line and you alight at jurong east mrt train station (FOL-HOW_TO)
from there when you come out you will just walk about ten minutes (FOL-HOW_TO)
Oh (FOL-ACK)
and you get to the singapore science centre (FOL-INFO)
well okay thank you (FOL-THANK)
do you have any beaches in (INI-CONFIRM)
there in singapore (FOL-CONFIRM)
what do you say (QST-CONFIRM)
Target Sentence (Tourist)
beach
RES-CONFIRM
QST-CONFIRM
FOL-CONFIRM
🟥 Content-Aware Attention
🟩 Content + Universal Time-Decay Attention
🟦 Context-Sensitive Time-Decay Attention

Fig. 2. The visualization of the attention weights enhanced by the proposed time-decay function compared with the weights learned by the content-aware attention model.

formation lies in few recent utterances and once again demonstrates our claim.

As mentioned above, one can control the level of flexibility in the time-decay attention at will; for example, the proposed context-sensitive universal time-decay attention possesses the best flexibility. The more freedom means less controllable, so integrating different time-decay perspectives does not guarantee to improve the efficacy of the model. It is possible that the combination may interfere attention model learning. Surprising, experiments show that the universal models outperform the models with a single time-decay attention type (rows (d)-(g)); furthermore, the most flexible model (row (h)) achieves the best results.

To justly examine the effectiveness of the universal design, we set the context-sensitive models with convex, linear, and concave time-decay attention to have larger models than the universal model. The experimental results (row (h)) reveal that our proposed context-sensitive universal attention model based on a smaller model can still outperform the ones using only a single type of time-decay attention, further demonstrating the positive interaction between attention functions and efficacy of our design.

4.7. Qualitative Analysis

From the 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 2. 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. The learned content-aware attention (red, right; row (c)) focuses on the wrong part of the dialogue and hence predicts the wrong label. By introducing the jointly-learned time-decay attention (green, middle; row (g)), the model can focus on the more relevant part, however, it miss the critical utterance ("*do you have any **beaches** in*" (`INI-CONFIRM`)) and fails predicting label of the target sentence. The proposed context-sensitive attention mechanism (blue, left; row (h)) can adapt to the complex dialogue contexts and successfully predict the corresponding intent ("*beach*") (`RES-CONFIRM`)).

5. 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

[19] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al., "End-to-end memory networks," in *Proceedings of NIPS*, 2015, pp. 2431–2439.

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

[21] Stephen Butterworth, "On the theory of filter amplifiers," *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930.

[22] 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.

[23] 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.

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