Decay-Function-Free Time-Aware Attention to Context and Speaker Indicator for Spoken Language Understanding

Jonggu Kim
Computer Science and Engineering,
Pohang University of Science and Technology (POSTECH)
Pohang, Republic of Korea
jgkimi@postech.ac.kr

Jong-Hyeok Lee
Computer Science and Engineering,
Pohang University of Science and Technology (POSTECH)
Pohang, Republic of Korea
jhlee@postech.ac.kr

Abstract

To capture salient contextual information for spoken language understanding (SLU) of a dialogue, we propose time-aware models that automatically learn the latent time-decay function of the history without a manual time-decay function. We also propose a method to identify and label the current speaker to improve the SLU accuracy. In experiments on the benchmark dataset used in Dialog State Tracking Challenge 4, the proposed models achieved significantly higher F1 scores than the state-of-the-art contextual models. Finally, we analyze the effectiveness of the introduced models in detail. The analysis demonstrates that the proposed methods were effective to improve SLU accuracy individually.

1 Introduction

Spoken language understanding (SLU) is a component that understands the user's utterance of a dialogue system. Given an utterance, SLU generates a structured meaning representation of the utterance; i.e., a semantic frame. SLU can be decomposed into several subtasks such as domain identification, intent prediction and slot filling; these subtasks can be jointly assigned using a single model (Hakkani-Tür et al., 2016; Liu and Lane, 2016; Chen et al., 2016b). The accuracy of SLU is important for the dialogue system to generate an appropriate response to a user.

To improve the accuracy of SLU, much work has used contextual information of dialogues to alleviate the ambiguity of recognition of the given utterance. In SLU, selecting important history information is crucial, and it directly influences the improvement of SLU accuracy. To summarize this history, content-aware models (Chen et al., 2016a; Kim et al., 2017) similar to attention models in machine translation (Bahdanau et al., 2014) have been proposed. However, content-aware models are likely to select the wrong history when the histories are similar in content. To alleviate this problem, time-aware models (Chen et al., 2017; Su et al., 2018a,b) which pay attention to recent previous utterances by using the temporal distance between a previous utterance and a current utterance are being considered; the models are based on mathematical formulas, time-decay functions, which are formulated by human, and decomposed into trainable parameters.

However, the previous time-aware models may not be sufficiently accurate. In the models, either a single time-decay function is used or a limited number of time-decay functions are linearly combined; these manual functions may not be sufficiently flexible to learn an optimal time-decay function.

In this paper, we propose flexible and effective time-aware attention models to improve SLU accuracy. The proposed models do not need any manual time-decay function, but learn a time-decay tendency directly by introducing a trainable distance vector, and therefore have good SLU accuracy. The proposed models do not use long short-term memory (LSTM) to summarize histories, and therefore use fewer parameters than previous time-aware models. We also propose current-speaker modeling by using a speaker indicator that identifies the current speaker.

To the best of our knowledge, this is the first method that shows improvement by considering the identity of the current speaker. This information may be helpful for modeling multi-party conversations in addition to human-human conversations.

Prediction of the semantic label of the current utterance even using a conventional time-aware model can be difficult. (Figure 1). The nearest utterance is "Right.", but it is not the most rele-

Proceedings of NAACL-HLT 2019, pages 3718–3726
Minneapolis, Minnesota, June 2 - June 7, 2019.

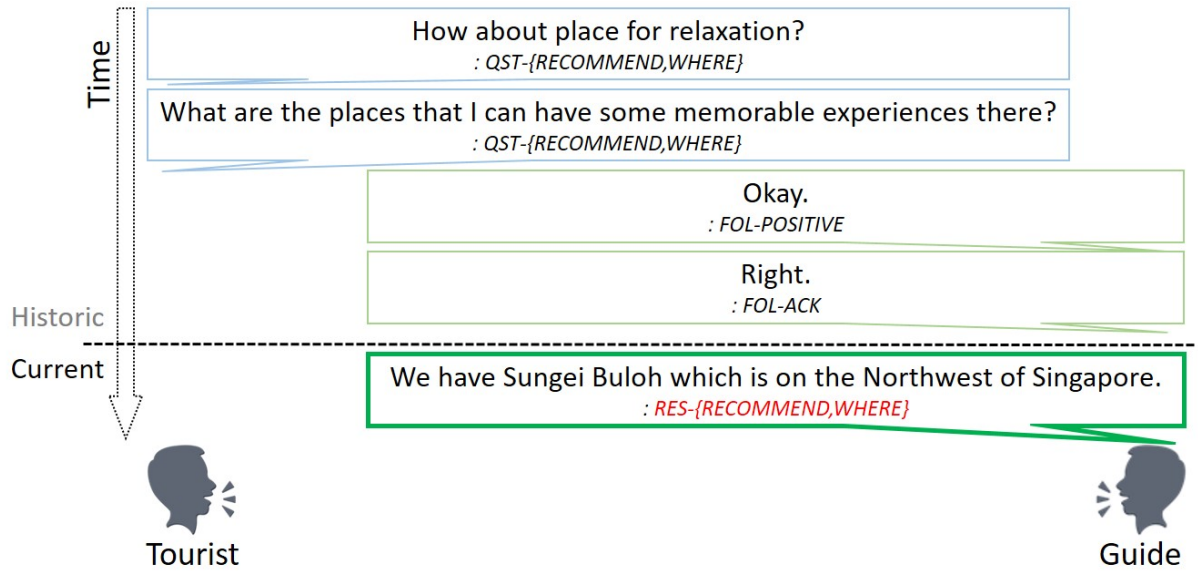

Figure 1: An example of utterances with their semantic labels (speech acts combined with associated attributes) from DSTC 4. The semantic labels are italicized.

vant utterance to the current utterance; the most relevant utterance is "What are the places that I can have some memorable experiences there?". If we do not know the current speaker is *Guide*, we cannot easily assess the relative importance of the nearest histories of the two speakers. We believe that the proposed 'speaker indicator' can help our model to identify such information.

In experiments on the Dialog State Tracking Challenge 4 (DSTC 4) dataset, the proposed models achieved significantly higher accuracy than the state-of-the-art contextual models for SLU. Also, we examine how the proposed methods affect the SLU accuracy in detail. This result shows that the proposed methods were effective to improve SLU accuracy individually. Our contributions are as follows:

- We propose a decay-function-free time-aware attention model that automatically learn the latent time-decay function of the history without a manual time-decay function. The proposed model achieves a new state-of-the-art F1 score.
- We propose a current-speaker modeling method that uses a speaker indicator to identify the current speaker. We present how to incorporate speaker indicator in the proposed attention model for further improvement of SLU accuracy.
- We propose a model that is aware of content as well as time, which also achieved a higher F1 score than the state-of-the-art contextual models.
- We analyze the effectiveness of proposed methods in detail.

Our source code to reproduce the experimental results is available at https://github.com/jgkimi/Decay-Function-Free-Time-Aware.

2 Related Work

Joint semantic frame parsing has the goal of learning intent prediction and slot filling jointly. By joint learning, the model learns their shared features, and this ability is expected to improve the accuracy on both tasks. A model based on bidirectional LSTM for joint semantic frame parsing (Hakkani-Tür et al., 2016) is trained on the two tasks in sequence, by adding an intent label to the output of the final time-step of LSTM. Similarly, an attention-based LSTM predicts slot tags for each time-step, then feeds the hidden vectors and their soft-aligned vectors to a fully-connected layer for intent prediction (Liu and Lane, 2016). Knowledge-guided joint semantic frame parsing (Chen et al., 2016b) incorporates syntax or semantics-level parsing information into a model by using a recurrent neural network (RNN) for joint semantic frame parsing.

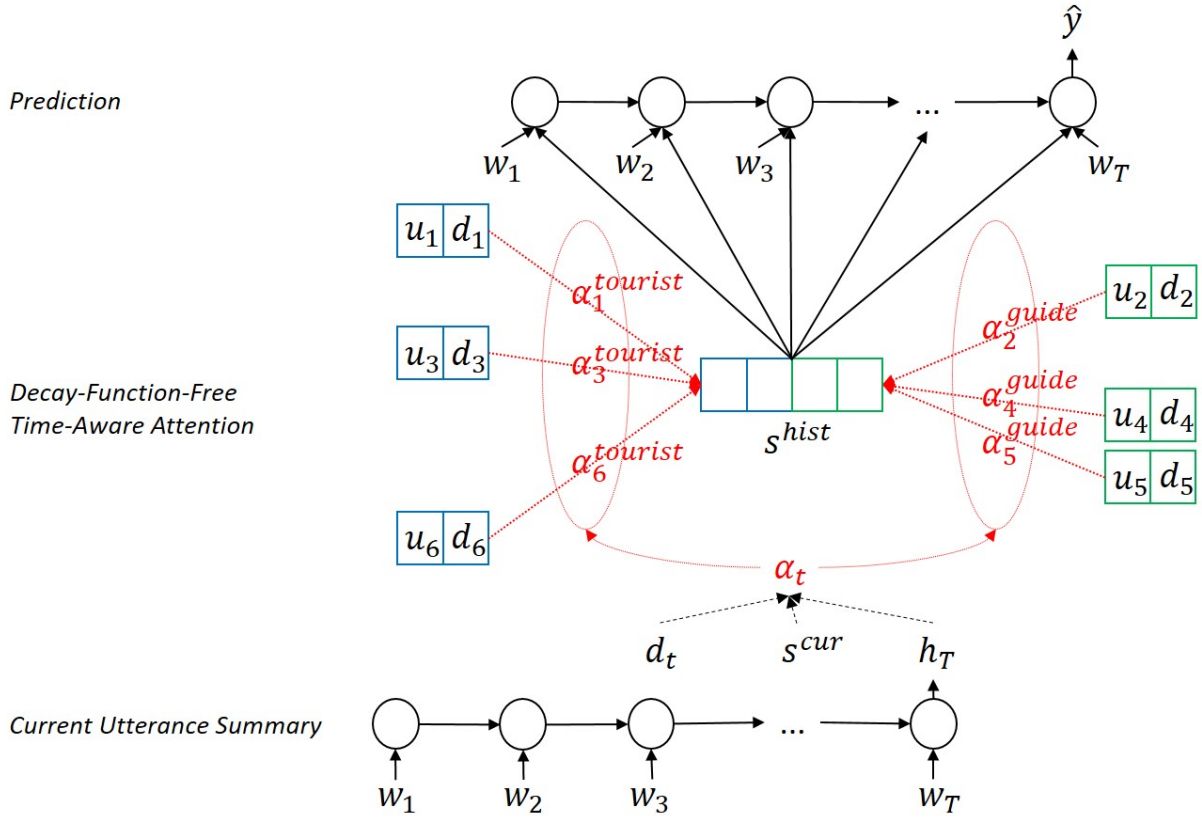

Figure 2: Overall architecture of the decay-function-free time-aware attention with speaker indicator (role-level). $w_1, ..., w_T$ are word vectors of the current utterance, d_t is the t^{th} distance vector, u_t is the t^{th} utterance vector, s^{cur} is the current speaker indicator, h_T is the current utterance summary vector and α_t is the importance of the t^{th} historic utterance. For simplicity, we represent bidirectional LSTM layers as unidirectional LSTM layers.

Other research on SLU uses context information. A model based on support vector machine and a hidden Markov model uses contextual information to show the importance of contextual information in SLU tasks, intent prediction and slot detection (Bhargava et al., 2013). RNN-based models can exploit context to classify domains (Xu and Sarikaya, 2014), and have been combined with previously-estimated intent and slot labels to predict domain and intent (Shi et al., 2015). A memory network that contains historic utterance vectors encoded by RNN has been used to select the most relevant history vector by multiplicative soft-attention (Chen et al., 2016a); the selected vector is fed to an RNN-based slot tagger as context information.

A memory network can be regarded as use of content-based similarity between the current utterance and previous utterances. A memory network can be separated to capture historic utterances for each speaker independently (Kim et al., 2017), and a contextual model can use different LSTM layers to separately encode a history summary for each speaker (Chi et al., 2017). For another task, addressee and response selection in multi-party conversations, a distinct RNN-based encoder for each speaker-role (sender, addressee, or observer) has been used to generate distinct history summaries (Zhang et al., 2018).

Recent work on contextual SLU has introduced time information of contexts into models because content-based attention may cause a wrong choice that introduce noises. The reciprocal of temporal distance between a current utterance and contexts can be used as a time-decay function, and the function can be decomposed into trainable parameters (Chen et al., 2017). Similarly, a universal time-aware attention model (Su et al., 2018a) has been proposed; it is a trainable linear combination of three distinct (convex, linear and concave) time-decay functions . An extension of this model is a context-sensitive time-decay attention (Su et al., 2018b) that generates its parameters from the current utterance by using a fully-connected layer so that the content information of the current utterance is also considered in the attention.

3 Proposed Model

We propose a time-aware model that includes a speaker indicator (Figure 2). In addition, we propose a content-and-time-aware model that includes a speaker indicator. The models are trained in an end-to-end way, in which every model parameter is automatically learned based on a downstream SLU task. The objective of the proposed models are to optimize the conditional probability of labels of SLU, given the current utterance $p(\hat{y}|x)$, by minimizing the cross-entropy loss.

The following description of the proposed model considers three steps: current utterance summary, context summary, and prediction.

3.1 Current Utterance Summary

To select salient parts of contextual histories, the current utterance is used. To summarize a current utterance matrix U that consists of words w_i as a vector (i.e., $U = \{w_1, w_2, ..., w_T\}$), U is fed to bidirectional LSTMs, and the final hidden vector $h_T \in \mathbb{R}^{dim}$ is taken as a current utterance summary.

3.2 Decay-Function-Free Time-Aware Attention

In this subsection, we introduce a decay-function-free time-aware attention model. To summarize contexts, we use a time difference (distance) between a historic utterance and the current utterance; this distance represents the interval between the historic utterance and the current utterance. We use the distance of the t^{th} history from the current utterance as an index to select a dense vector from a distance-embedding matrix $D \in \mathbb{R}^{dim \times |D|}$, then use the vector as the t^{th} distance vector d_t.

To compute the importance α_t of the t^{th} history, both in the sentence-level attention and in the role-level attention, our time-aware attention uses the current utterance summary h_T and the history distance d_t simultaneously and additively:

$$\alpha_t = w_{att}^T \sigma\big(h_T + d_t + b_{att}\big), \tag{1}$$

where w_{att}^T is the transpose of a trainable weight vector for the attention, b_{att} is a trainable bias vector for the attention, and σ is the hyperbolic tangent function.

Computing a time-aware context summary vector s_{time}^{hist} depends on whether the role-level or sentence-level attention is considered. For the role-level attention, we use the softmax operation applied to all α_t of the same speaker, either a guide or a tourist, to obtain a role-level probabilistic importance α_t^{role} of t^{th} history. We then multiply α_t^{role} by t^{th} history vector, which is a concatenation of the corresponding intent-dense vector u_t and the distance vector d_t. We use the element-wise sum of the vectors of the same speaker to construct two summary vectors s_{time}^{guide} and $s_{time}^{tourist}$. Finally, s_{time}^{guide} and $s_{time}^{tourist}$ are concatenated to form a time-aware history summary vector s_{time}^{hist} as:

$$\alpha_t^{role} = softmax_{role}(\alpha_t), \tag{2}$$

$$s_{time}^{role} = \sum_t \alpha_t^{role}(u_t \oplus d_t), \tag{3}$$

$$s_{time}^{hist} = s_{time}^{guide} \oplus s_{time}^{tourist}, \tag{4}$$

where $\oplus$ represents a concatenation operation.

For the sentence-level attention to obtain a sentence-level probabilistic importance α_t^{sent} of t^{th} history, we use the softmax operation applied to all α_t regardless of the speaker, then multiply α_t^{sent} by the t^{th} history vector, which is a concatenation of the corresponding intent dense vector u_t and the distance vector d_t. We use the element-wise sum of the vectors to construct a time-aware summary vector s_{time}^{hist} as:

$$\alpha_t^{sent} = softmax_{sent}(\alpha_t), \tag{5}$$

$$s_{time}^{hist} = \sum_t \alpha_t^{sent}(u_t \oplus d_t). \tag{6}$$

Then, s_{time}^{hist} is used as a context summary s^{hist} in the prediction step.

3.3 Decay-Function-Free Content-and-Time-Aware Attention

Although a time-aware attention model is powerful by itself, content can be considered at the same time to improve accuracy. We propose another contextual attention model that is aware of content, in addition to time. This model is called content-and-time-aware attention. The model uses an importance value β_t for the t^{th} history. To compute β_t, we uses the trainable parameters w_{att} and b_{att} of the time attention as:

$$\beta_t = w_{att}^T \sigma\big(h_T + u_t + b_{att}\big), \tag{7}$$

where u_t is the intent dense vector of t^{th} history, and σ is the hyperbolic tangent function.

Then, β_t is used in the same way as α_t, but independently. s^{hist}_{time} is computed as in the previous subsection, β_t is used to compute s^{hist}_{cont} for the role-level attention as:

$$\beta^{role}_t = softmax_{role}(\beta_t), \quad (8)$$

$$s^{role}_{cont} = \sum_t \beta^{role}_t (u_t \oplus d_t), \quad (9)$$

$$s^{hist}_{cont} = s^{guide}_{cont} \oplus s^{tourist}_{cont}. \quad (10)$$

To compute s^{hist}_{cont} for the sentence-level attention, β_t is used as:

$$\beta^{sent}_t = softmax_{sent}(\beta_t), \quad (11)$$

$$s^{hist}_{cont} = \sum_t \beta^{sent}_t (u_t \oplus d_t). \quad (12)$$

Finally, the time-aware history summary s^{hist}_{time} and the content-aware history summary s^{hist}_{cont} are concatenated to generate a history summary s^{hist} regardless of the attention level:

$$s^{hist} = s^{hist}_{time} \oplus s^{hist}_{cont}. \quad (13)$$

3.4 Speaker Indicator

Speaker indicator is a trainable vector $s^{cur} \in \mathbb{R}^{dim}$ which indicates the identity of the current speaker; i.e., either a tourist or a guide in DSTC 4. An embedding lookup method is used after a speaker embedding matrix $S \in \mathbb{R}^{dim \times |S|}$ is defined. The speaker embedding matrix is randomly initialized before the model is trained.

To use speaker indicator s^{cur} in the proposed attentions, Eq. 1 is rewritten as:

$$\alpha_t = w^T_{att} \sigma\big(h_T + d_t + s^{cur} + b_{att}\big), \quad (14)$$

and Eq. 7 is rewritten as:

$$\beta_t = w^T_{att} \sigma\big(h_T + u_t + s^{cur} + b_{att}\big). \quad (15)$$

3.5 Prediction

To predict the true label in spoken language understanding, our model consumes the current utterance U again. We use another bidirectional LSTM layer which is distinct from that of the current utterance summary. To prepare for t^{th} input v_t of the LSTM layer, we concatenate t^{th} word vector w_t of the current utterance U with the history summary vector s^{hist}:

$$v_t = w_t \oplus s^{hist}. \quad (16)$$

Then, we feed each v_t to the LSTM layer sequentially, and the final hidden vector of the LSTM layer is used as an input of a feed-forward layer to predict the true label $\hat{y}$.

4 Experiments

To test the proposed models, we conducted language-understanding experiments on a dataset of human-human conversations.

4.1 Dataset and Settings

We conducted experiments on the DSTC 4 dataset which consists of 35 dialogue sessions on touristic information for Singapore; they were collected from Skype calls of three tour guides with 35 tourists. The 35 dialogue sessions total 21 h, and include 31,034 utterances and 273,580 words (Kim et al., 2016). DSTC 4 is a suitable benchmark dataset for evaluation, because all of the dialogues have been manually transcribed and annotated with speech acts and semantic labels at each turn level. a semantic label consists of a speech act and associated attribute(s). The speaker information (guide and tourist) is also provided. Human-human dialogues contain rich and complex human behaviors and bring much difficulty to all tasks that are involved in SLU. We used the same training dataset, the same test dataset and the same validation set as in the DSTC 4 competition: 14 dialogues as the training dataset, 6 dialogues as the validation dataset, and 9 dialogues as the test dataset.

We used Adam (Kingma and Ba, 2015) as the optimizer in training the model. We set the batch size to 256, and used pretrained 200-dimensional word embeddings GloVe (Pennington et al., 2014). We applied 30 training epochs with early stopping. We set the size dim of every hidden layer to 128, and the context length to 7. We used the ground truth intents (semantic labels) to form an intent-dense vector like previous work. To evaluate SLU accuracy, we used the F1 score, which is the harmonic mean of precision and recall. To validate the significance of improvements, we used a one-tailed t-test. We ran each model ten times, and report their average scores.

As baseline models, we used the state-of-the-art contextual models, and most accurate participant of DSTC 4 (DSTC 4 - Best) (Kim et al., 2016). For comparison with our models, we used the scores

Model	F1 score	
	Sent.-Level	Role-Level
DSTC 4 - Best	*61.4*	
No Context	65.06	
LSTM-Used Context Summary without Attention	72.15	
LSTM-Used Content-Aware Attention	71.27	71.84
Speaker Role Modeling (Chi et al., 2017)	*66.8*	*70.1*
Convex Time-Aware Attention (Chen et al., 2017)	*74.6*	*74.2*
Universal Time-Aware Attention (Su et al., 2018a)	74.22	74.12
Universal Content + Time Attention (Su et al., 2018a)	74.40	74.33
Context-Sensitive Time Attention (Su et al., 2018b)	74.20	73.53
Decay-Function-Free Time-Aware Attention	75.58**	75.58**
with Speaker Indicator	75.95**	**76.56****
Decay-Function-Free Content-and-Time-Aware Attention	75.59**	75.30**
with Speaker Indicator	76.11**	76.14**

Table 1: SLU accuracy on DSTC 4. *: $p < 0.05$; **, $p < 0.01$ compared to all the baseline models. Italicized scores are reported in the references. Model names are described in the text.

reported in the papers[1]. We ran three additional baseline models in which the prediction stage is the same: (1) 'No Context' uses no context summary; (2) 'LSTM-Used Context Summary without Attention' uses the context summary of bidirectional LSTM without an attention mechanism, and (3) 'LSTM-Used Content-Aware Attention' uses context summary of bidirectional LSTM after content-aware attention is applied to histories, as in previous approaches.

4.2 Results

We conducted an experiment to compare the proposed models with the baseline models in the SLU accuracy (Table 1). All of the proposed models achieved significant improvements compared to all the baseline models.

We conducted an experiment to identify details of how possible combinations of the proposed methods affect the SLU accuracy (Table 2). In addition to the combinations of the proposed methods, we tested another content-and-time-aware attention method (Content x Time) which computes attention values using both intent and distance at a time, and shares the values to compare with the proposed content-and-time-aware attention.

[1] Su et al. (2018a) and Su et al. (2018b) specified that they used different training/valid/test datasets that had been randomly selected from the whole DSTC 4 data with different rates for the experiments. Therefore, we do not use the reported score in our comparison, but produced the results under the same conditions by using the open-source code.

5 Discussion

In the first subsection, we analyze the effectiveness of the decay-function-free time-aware attention and decay-function-free content-and-time-aware attention by comparison with others. In the next subsection, we analyze the effectiveness of the proposed methods in their possible combinations. We also analyze the effectiveness of the use of a distance vector in the history representation under the various conditions. Finally, we analyze attention weights of the proposed models in a qualitative way to convince of the effectiveness of them.

We also conducted an experiment to show the effectiveness of the use of a distance vector in the history representation under the same condition as in the role-level attention (Table 3). Although we propose to use both intent and distance by concatenating them as a history representation, intent can be used alone; this approach is more intuitive than using both intent and distance.

5.1 Comparison with Baseline Models

In Table 1, Decay-Function-Free Time-Aware Attention and Decay-Function-Free Content-and-Time-Aware Attention achieved significantly higher F1 scores that all baseline models. Especially, the role-level Decay-Function-Free Time-Aware Attention with speaker indicator achieved an F1 score of 76.56 % (row 11), which is a state-of-the-art SLU accuracy.

Attention Type	F1 score Sent.-Level	Role-Level
no attention	70.49	70.43
Content	73.03	72.87
Content	73.05	72.68
Time	75.58	75.58
Time	75.95**	**76.56****
Content + Time	75.59	75.30
Content + Time	75.83	75.96**
Content + **Time**	75.94*	75.97**
Content + Time	76.11**	76.14**
Content x Time	75.59	75.50
Content x Time	75.63	75.64

Table 2: SLU accuracy of possible combinations of the proposed methods. "no attention": sum of all history vectors without calculating α, "Content" is content-aware attention (Decay-Function-Free Content-Aware Attention), "Time" is the proposed time-aware attention (Decay-Function-Free Time-Aware Attention), "Content + Time" is the proposed content-and-time-aware attention (Decay-Function-Free Content-and-Time-Aware Attention), "Content x Time" is variant content-and-time-aware attention (Decay-Function-Free Inseparate Content-and-Time-Aware Attention). *: $p < 0.05$; **, $p < 0.01$ compared to the same attention without speaker indicator. An attention type in bold is the speaker-involved part.

Attention Type	F1 score Intent only	Int. & Dist.
no attention	70.20	70.43
Content	71.09	72.87**
Content	71.26	72.68**
Time	75.17	75.58**
Time	75.11	**76.56****
Content + Time	75.04	75.30**
Content + Time	75.62	75.96*
Content + **Time**	75.13	75.97**
Content + Time	75.67	76.14**
Content x Time	75.08	75.50**
Content x Time	75.03	75.64**

Table 3: SLU accuracy of possible combinations of the proposed methods in role-level attention with different history representations. Int. used intent vector; Dist. used distance vector. *: $p < 0.05$; **, $p < 0.01$ compared to using intent only. Other codes are as in Table 2.

5.2 Detailed Analysis on Proposed Methods

The proposed methods had good SLU accuracy (Table 2). Every time-aware attention with and without speaker indicator (rows 4 to 11) improved the F1 score compared to the content-aware attention with and without speaker indicator (rows 2 and 3) and to no attention (row 1). This result means that the proposed time-aware attention was effective to improve the SLU accuracy. Any of the content-and-time-aware attention with or without speaker indicator (rows 6 to 11) did not improve the F1 score compared to the time-aware attention with and without speaker indicator (rows 4 and 5). This result means that incorporating content could not make further improvement of the accuracy. Also, without speaker indicator, all the time-aware attention (rows 4, 6 and 10) achieved similar F1 scores.

Use of speaker indicator also showed tendencies. It did not significantly improve the SLU accuracy of Decay-Function-Free Content-Aware Attention (rows 2 and 3) or Decay-Function-Free Inseparate Content-and-Time-Aware Attention (rows 10 and 11), but did improve the accuracy of the proposed models, Decay-Function-Free Time-Aware Attention (rows 4 and 5) and Decay-Function-Free Content-and-Time-Aware Attention (rows 6 to 9). Decay-Function-Free Content-and-Time-Aware Attention with speaker indicator (rows 7 to 9) were more accurate than Decay-Function-Free Inseparate Content-and-Time-Aware Attention with speaker indicator (row 11). This result means that using speaker indicator, separation of content and time improved the accuracy. The improvement in the role-level tended to be greater than that in the sentence-level. The improvement was greatest when speaker indicator was involved in the proposed role-level Decay-Function-Free Time-Aware Attention (row 5).

5.3 Effectiveness of Use of Distance in History Representation

In all models, the use of both intent and distance vectors significantly achieved higher F1 than the use of an intent vector only (Table 3). The results indicate that distance embeddings are helpful both for attention and for the history representation. Decay-Function-Free Time-Aware Attention achieved the biggest improvement (row 5) among all the models.

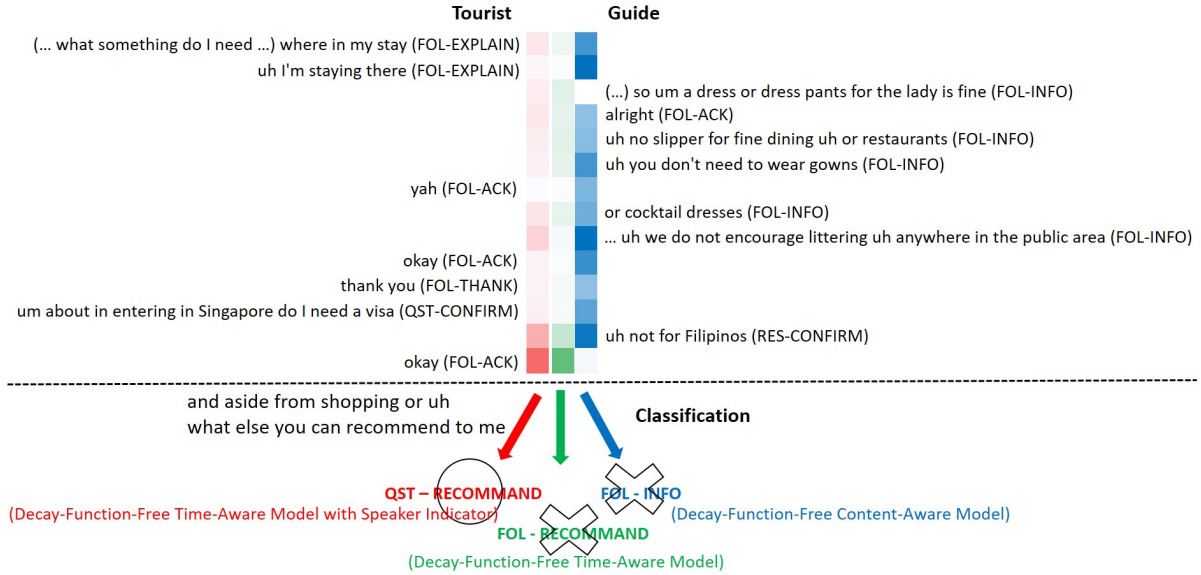

Figure 3: The visualization of the attention weights of the proposed models and the baseline content-aware model. Color gradient indicates intensity given a single datum after training. The color gradient at the left side indicates attention intensities of Decay-Function-Free Time-Aware Model with Speaker Indicator, the color gradient in the center indicates attention intensities of Decay-Function-Free Time-Aware Model without Speaker Indicator, and the color gradient at the right side indicates attention intensities of Decay-Function-Free Content-Aware Model.

5.4 Qualitative Analysis

To assess whether the proposed time-aware attention and speaker indicator can learn a time-decay tendency of the history effectively, we inspected the weights trained in Decay-Function-Free Time-Aware Attention with and without the speaker indicator. We also inspected Decay-Function-Free Content-Aware Attention to compare with them. We observed (Figure 3) that the weights of the proposed models were trained well compared to Decay-Function-Free Content-Aware Attention. The proposed time-aware attention with/without speaker indicator tended to pay attention to recent histories, whereas the content-aware attention does not. As a result, Decay-Function-Free Time-Aware Attention with speaker indicator could generate the true label, *QST-RECOMMEND*, by avoiding noisy contextual information like "uh I'm staying there (FOL-EXPLAIN)" or "... uh we do not encourage littering uh anywhere in the public area (FOL-INFO)".

6 Conclusion

In this paper, we propose decay-function-free time-aware attention models for SLU. These models summarize contextual information by taking advantage of temporal information without a manual time-decay function. We also propose a current-speaker detector that identifies the current speaker. In experiments on the DSTC 4 benchmark dataset, the proposed models achieved a state-of-the-art SLU accuracy. Detailed analysis of effectiveness of the proposed methods demonstrated that the proposed methods increase the accuracy of SLU individually.

Acknowledgments

We would like to thank the reviewers for their insightful and constructive comments on this paper.

References

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

A. Bhargava, A. Celikyilmaz, D. Hakkani-Tür, and R. Sarikaya. 2013. Easy contextual intent prediction and slot detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8337–8341.

P. Chen, T. Chi, S. Su, and Y. Chen. 2017. Dynamic time-aware attention to speaker roles and contexts for spoken language understanding. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 554–560.

Yun-Nung Chen, Dilek Hakkani-Tür, Gokhan Tur, Jianfeng Gao, and Li Deng. 2016a. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Interspeech 2016*, pages 3245–3249.

Yun-Nung (Vivian) Chen, Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Jianfeng Gao, and Li Deng. 2016b. Syntax or semantics? knowledge-guided joint semantic frame parsing. IEEE Workshop on Spoken Language Technology (SLT 2016).

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 the The 8th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 163–168.

Dilek Hakkani-Tür, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech 2016*, pages 715–719.

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

Young-Bum Kim, Sungjin Lee, and Ruhi Sarikaya. 2017. Speaker-sensitive dual memory networks for multi-turn slot tagging. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 541–546.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of 3rd International Conference on Learning Representations (ICLR)*.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech 2016*, pages 685–689.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Y. Shi, K. Yao, H. Chen, Y. Pan, M. Hwang, and B. Peng. 2015. Contextual spoken language understanding using recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5271–5275.

Shang-Yu Su, Pei-Chieh Yuan, and Yun-Nung Chen. 2018a. How time matters: Learning time-decay attention for contextual spoken language understanding in dialogues. In *Proceedings of The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Shang-Yu Su, Pei-Chieh Yuan, and Yun-Nung Chen. 2018b. Learning context-sensitive time-decay attention for role-based dialogue modeling. *arXiv preprint arXiv:1809.01557*.

P. Xu and R. Sarikaya. 2014. Contextual domain classification in spoken language understanding systems using recurrent neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (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 *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.