Research Statement: Context-Aware Sequence Model

Kyungwoo Song

September 28, 2020

My research focuses on developing a context-aware sequence model. Context modeling helps understand the abstract meaning of data, such as sentence or user behavior. Contextual information captures the important underlying feature, and it helps the model to capture the relationship between data instances or hidden representations. The importance of context modeling becomes larger when we deal with sequential data which has its own order. This is because even the same word might have a completely different meaning depending on the position and order of words. For the sequential data, we need to consider the context change over time and the relationship between sequential input. Furthermore, we extend our research to handle the multi-granularity and hierarchical context of sequential modeling to deal with long and complex sequences by capturing rich contextual representations. My research focuses on three keywords, sequence, context, and hierarchy on diverse datasets and tasks. Figure 1 visualize my primary research works and experiments that I conducted.

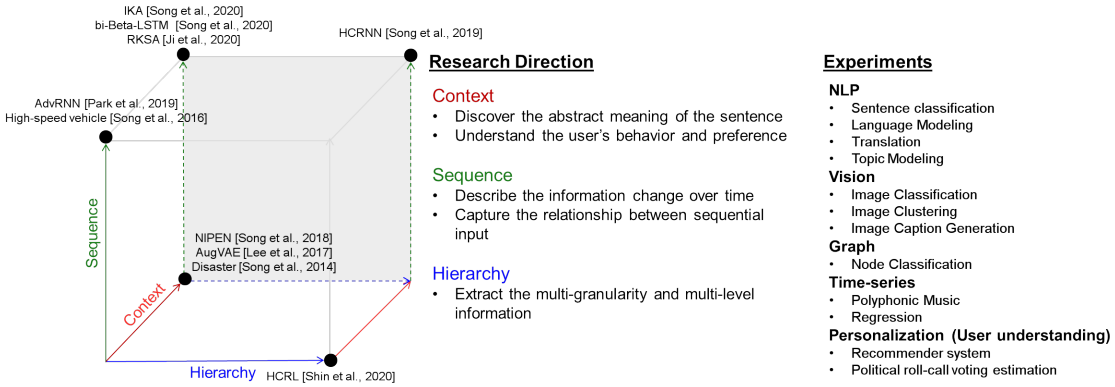


Figure 1: The structure of my past, ongoing, and future research works.

1 Context-Aware Model

The context denotes the generally related thought of the event, and it can be defined on the sentence user's behavior. The context modeling helps us to discover the clear meaning of the sentence or understand the user's behavior.

Neural Ideal Point Estimation Network (NIPEN) I have focused on static context modeling of sentence and user behavior. I investigate the static context on the legislative roll-call data because legislative processes have both contents of bill (sentence) and quantitative record of legislator's voting (user behavior). Under the legislative processes, it is challenging to consider the context of the bill (contents) and the contents of the legislator (ideal point) simultaneously. To solve the issue, we assume that contents and ideal points are composed of several topics and the probability of voting *YEA* increases proportionally to the conformity of the topic of bill and legislator's ideal point for each topic. Under the assumption, we proposed the Neural Ideal Point Model (NIPEN) Song *et al.* [2018], which model each context of sentence and user behavior. With NIPEN, we can understand and interpret the sentence and user behavior itself and the results of legislative voting, which is an interaction between document and user. This research proposes the way of quantifying the characteristic of documents and persons for each topic or agenda (Figure 2).

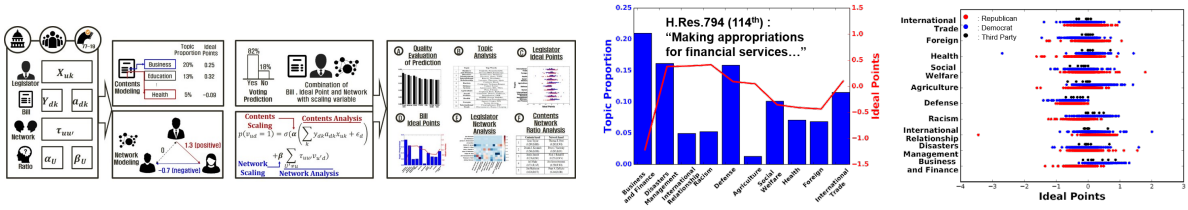
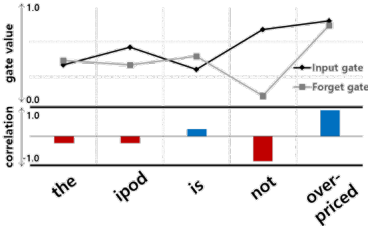


Figure 2: (Left) The overall structure of our proposed model, NIPEN. (Middle) The bill, H.Res.794 (114th), considers the appropriations for financial services and general government, and the major topic is *Business and Finance*, and the bill's ideal point in *Business and Finance* is -1.217. (Right) The vote casting will be determined by the legislator's view on *Business and Finance*, and this topic shows the greatest disagreement between the Republicans and the Democrats. In the real world, the voting results were the same as expected: 1) the voting was very partisan, 92.2% Republican voted *YEA*, and the 90.3% Democrat voted *NAY*.

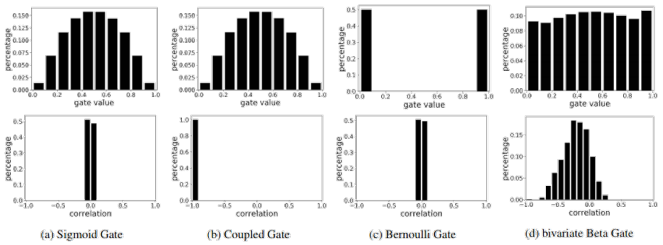
2 Context-Aware Sequence Model

The importance of sequential modeling has increased, and it is necessary to handle the context of sequential data. For sentence modeling, there are many kinds of research to incorporate the inductive bias of sequential order, such as RNN and Transformer. To understand the sequential data deeply, capturing the relationship between sequential input is important. First, we propose an explicit correlation gate structure to handle the dynamics of context. Second, we propose implicit kernel attention, a generalization of attention in Transformer and GAT, to capture the complex relationship in the dataset adaptively.

Bivariate Beta-LSTM (bBLSTM) Second, we focused on the sequential context of the sentence. The sentence is composed of words, and some words are correlated positively or negatively. We determine the meaning of the sentence by composing appropriate words with proper weight, which represents the level of correlation. To capture the context of the given sentence, we need to consider two properties; 1) correlation modeling between input words, 2) flexible valued gate structures to remove unnecessary information, and preserve important information, as shown in Figure 3a. It is challenging to handle the correlation on the sequential data, and most previous models, such as LSTM Hochreiter and Schmidhuber [1997], lack of explicit correlation modeling. The traditional gate structure handles the correlation implicitly, and their sigmoid function-based gate functions might not represent the value between 0 and 1 flexibly. We improve the traditional model by incorporating the correlated input and forget gate based on the bi-variate Beta distribution, which represents the values between 0 and 1 flexibly and correlation Song *et al.* [2020a], as shown in Figure 3b. Under the flexible correlated gate structure, our proposed model, Bivariate Beta-LSTM, determines the level of composition between words, understand the meaning of sentences efficiently. This work envisions how to incorporate the neural network models with probabilistic components to improve its flexibility and capture the rich contextual information.



(a) An illustrative example of the input gate, the forget gate, and their correlation for part of a given sentence in sentiment classification datasets. Blue and Red bars denote the positive and negative correlations, respectively.



(a) Sigmoid Gate (b) Coupled Gate (c) Bernoulli Gate (d) bivariate Beta Gate

(b) Analysis of input gate value (first row) and the correlation between input and forget gate (second row). Our proposed model, bBLSTM(5G) and bBLSTM(5G+p) are based on the bivariate Beta distribution, and it represents the value between 0 and 1 flexibly with correlation.

Figure 3: To capture the sentiment of the sentence, "the ipad is not over-priced", the positive correlation is necessary to aggregate the meaning of "not" and "over-priced" at $t = 4$, with flexible gate value (left). We explicitly formulate the gate structure, which represents correlation and flexible gate value (right).

Implicit Kernel Attention (IKA) Attention is widely utilized to improve model performance as well as to explain the model mechanisms. Transformer adopts the scaled dot-product attention, and the graph attention network (GAT) utilizes concatenation-based attention. We provide a new interpretation of attention, and this interpretation leads to generalized attention by formalizing the attention weight into a multiplication of two terms: similarity and magnitude. We derive the explicit separation, so the derivation reveals that the attention in Transformer and GAT is a product of 1) the Radial Basis Function (RBF) kernel between instances (similarity) and 2) the exponential of L^2 norm for each instance (magnitude). The similarity measures the relative importance, and the magnitude computes the individual importance. Figure 4 indicates that each kernel has its own inductive bias, which leads each kernel to capture different contexts of given data. Hence, the appropriate kernel might be different for each dataset or task.

From this decomposition, we generalize the attention in three ways. First, we propose implicit kernel attention with an implicit kernel function, instead of manual kernel selection ($\text{IKA}_{\text{S}}, \text{IKA}_{\text{NS}}$). Second, we generalize L^2 norm as the L^p norm (IKAN, IKAN-direct). Furthermore, we analyze the relationship between the magnitude and the sparsity of attention weights. Third, we extend our attention to structured multi-head attention (MIKAN). We validate our models on classification, translation, and regression tasks. Our generalized implicit attention is exchangeable with the attention in Transformer and GAT without increasing the algorithmic complexity in order Song *et al.* [2020b].

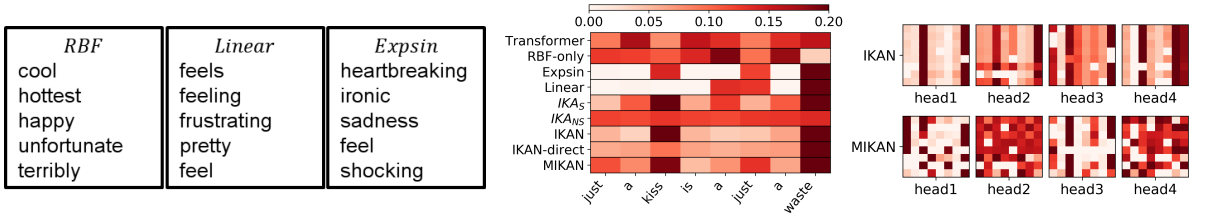


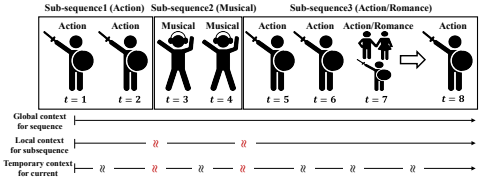
Figure 4: (Left): The five most similar words for given query words, *queen* and *sad* under the Glove word vectors. RBF, Linear, Expsin kernel captures the interpretable and meaningful similar words, and their words are different depending on the kernel. Because each kernel captures the different context, the appropriate kernel might be different on the dataset or task. (Middle): Attention weights for the given sentence on MR, whose task is the sentiment classification of a movie review. The sentiment of the given sentence "just a kiss is a just a waste" is negative. To identify the sentiment, the model should focus on the *just*, *kiss*, *just*, and *waste*, and MIKAN captures all important words. (Right): Attention weights matrix of first four heads among eight heads for IKAN and MIKAN. MIKAN shows relatively diverse attention weights across the multi-head because of the structured model with the copula.

3 Hierarchical Context-Aware Sequence Model

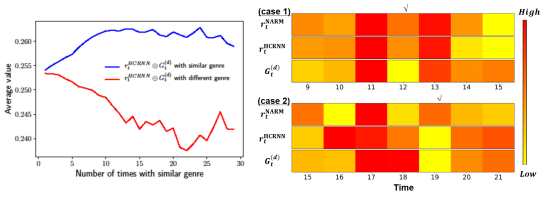
Many sequences, such as text and music, have hierarchical structure naturally Nevill-Manning and Witten [1997]. To understand the sequence deeply, we need to capture the multi-granularity context, hierarchical context. The importance of hierarchical sequential context modeling increases for a complex and long sequence, such as user log history.

Hierarchical Context enabled Recurrent Neural Network (HCRNN) I focus on the context of user behavior history with hierarchical context modeling. User history is a sequence of user's actions such as clicks or skips. With user history, music, and video streaming services want to recommend appropriate items to the user. To recommend an item that the user wants, we need to reflect the user's context (interest), and the user's long history might have a more diverse context than that of the sentence. We divide these user's context into a global context for the entire sequence of user's action, the local context for sub-sequence, and temporary context for the current time, as shown in Figure 5a. In short, we need to model hierarchical user's context modeling and its dynamics to consider the user's long history well. To address the issue, we proposed the Hierarchical Context enabled Recurrent Neural Network (HCRNN), which handles the sequential hierarchical context Song *et al.* [2019], different from LSTM. HCRNN incorporate the topic modeling and memory network for global context and utilize attention mechanism to attend related global context to the current sub-sequence. Besides, we model the temporary context, current interest, with the local context and the recent user behavior. Additionally,

we introduce an interest drift gate, which controls the sequential changes in each context. Our proposed model captures the interest change points, as shown in Figure 5b. This paper proposes a three-level hierarchical context modeling that handles the long and complex sequence, such as user log history.



(a) The long user history contains multiple hierarchical contexts; a global context for the entire sequence, the local context for sub-sequence, and a temporary context for the current time. To handle the user's interest drift, the temporary context must change at every point (black wave) but should change more when the new sub-sequence starts (red wave).



(b) Average our gate value after appearing items with similar genre consecutively. Our gate represents a large value if the current input of the item has a similar genre with the previous input of the item. In the opposite case, our gate grasps the user's interest drift and become smaller.

Figure 5: Hierarchical context modeling is effective in handling a complex sequence such as user log history (Left). Our proposed model captures the interest drift point with hierarchical context modeling (right).

4 Future Research Plan

I have focused on sequential contextual modeling, which handles the diverse and complex context of the sequence. My research will continue to understand the sequence and user's behavior deeply. The attention is a core of many state-of-the-art for sequence modeling, and it handles the sequence without recurrence modeling. I will investigate the attention and its application to social sciences.

Efficient Sparse Attention (ESA) Sparse attention has recently emerged for sequence, image, and graph-related tasks Martins and Astudillo [2016]; Child *et al.* [2019]; Tay *et al.* [2020]. Sparse attention encourages performance and interpretability. At the same time, the importance of handling longer and more complex sequence have increased. The user's activity log gets longer as time goes by, and the necessity of generating longer sentences or speech datasets has increased. Linear time sparse attention is effective in capturing the long-range sequence with high performance and interpretation. However, there has been no research that proposes linear time sparse attention. I have focused on the relationship between kernel and attention. From the relation, I will propose a linear time sparse attention.

Higher-order Context-Aware Attention (HCA) Traditional attention computes the relationship between the instances. They only compute the pair-wise relationship, second-order interaction. However, the relationship between instances might be different depends on the contexts. If we incorporate the contextual information by formulating the higher-order attention, we can capture the more diverse and complex relationship between instances. There are two research, area attention Li *et al.* [2019] and context-aware attention Yang *et al.* [2019]. However, area attention aggregate only restricted area, and context-aware attention lacks explicit higher-order interaction between instances. To formulate the higher-order attention, I will connect the relationship between factorization machine Rendle [2010] and attention. The factorization machine basically computes second-order interaction between instances, but it is possible to compute higher-order interaction efficiently Rendle [2010]; Blondel *et al.* [2016].

Network Analysis with Repulsive Graph Modeling The user-user relationship and user-item relationship are important to capture and understand the user's preference. The relationship might denote the positive relationship and negative relationship. However, most graph-related works focus on a positive relationship by assuming the network homophily. To understand the user's behavior, negative relationship modeling is necessary. It is natural that the users who are in a negative relationship have different hidden representations. I will propose negative attention to incorporate the repulsive modeling into the graph-related models Kipf and Welling [2016]; Veličković *et al.* [2017].

References

Mathieu Blondel, Akinori Fujino, Naonori Ueda, and Masakazu Ishihata. Higher-order factorization machines. In *Advances in Neural Information Processing Systems*, pages 3351–3359, 2016.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

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

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Yang Li, Lukasz Kaiser, Samy Bengio, and Si Si. Area attention. In *International Conference on Machine Learning*, pages 3846–3855, 2019.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623, 2016.

Craig G Nevill-Manning and Ian H Witten. Identifying hierarchical structure in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, 7:67–82, 1997.

Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000. IEEE, 2010.

Kyungwoo Song, Wonsung Lee, and Il-Chul Moon. Neural ideal point estimation network. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Kyungwoo Song, Mingi Ji, Sungrae Park, and Il-Chul Moon. Hierarchical context enabled recurrent neural network for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4983–4991, 2019.

Kyungwoo Song, JoonHo Jang, Il-Chul Moon, et al. Bivariate beta lstm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

Kyungwoo Song, Yohan Jung, Dongjun Kim, and Il-Chul Moon. Implicit kernel attention. *arXiv preprint arXiv:2006.06147*, 2020.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse sinkhorn attention. *arXiv preprint arXiv:2002.11296*, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. Context-aware self-attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 387–394, 2019.