Research Statement: Context Aware Sequential Model

Kyungwoo Song

June 8, 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.

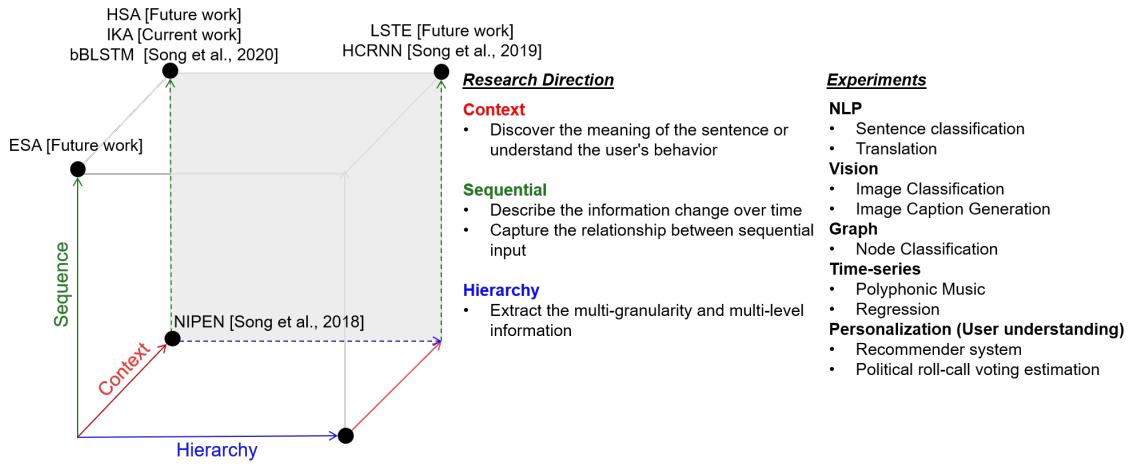


Figure 1: The structure of my past, ongoing, and future research works.

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).

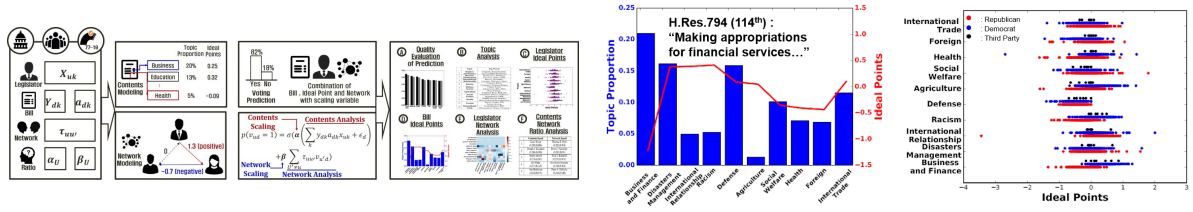


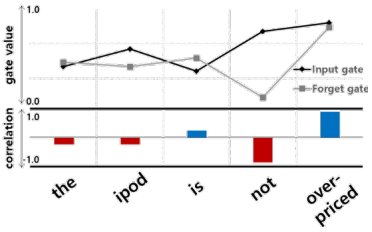
Figure 2: (Left) The overall structure of our proposed model, NIPEN. (Middle) The bill, H.Res.794 (114th), considers the appropriations for financial services and general government, and the major topic is *Business and Finance*, and the bill's ideal point in *Business and Finance* is -1.217. (Right) 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 same as expected: 1) the voting was very partisan, 92.2% Republican voted *YEA* and the 90.3% Democrat voted *NAY*.

2 Context-aware Sequential Model

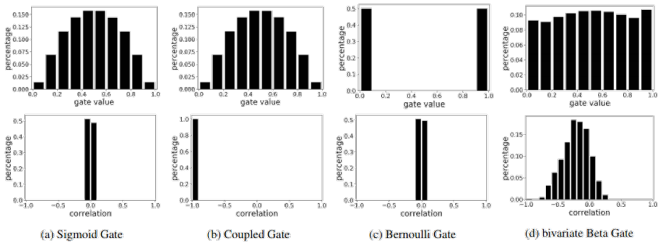
Recently, the importance of sequential modeling has increased, and it is necessary to handle the context of sequential data. For sentence modeling, there are many kinds of research to incorporate the inductive bias of sequential order such as RNN, and Transformer. However, static context modeling is not enough to reflect the context changes in the sentence or user's sequential behavior. 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 scaled-dot product attention, 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. First, correlation modeling between input words is important. Second, flexible valued gate structures are necessary 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.* [2020], 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.

Implicit Kernel Attention (IKA) In neural networks, *Attention* has become an essential structure. Attention captures the important features and allows the model to focus on the essential features. The scaled dot-product attention compute the dot-product between query and key, which is a linear projection of hidden feature. It is well known that the dot-product of two vectors is a product of two terms, 1) cosine of the angle between two vectors, which denotes the similarity, 2) norm of each vector which measures individual scale. In this paper, we further analyze the meaning of scaled dot-product attention. Because the scaled dot-product attention uses an un-normalized dot-product between query and key, the attention weights are influenced by 1) similarity between query and key, relative importance, and 2) the magnitude



(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.



(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).

of each query and key, individual importance. This opens a question on how to separate the scaled dot-product attention as the similarity and magnitude term explicitly, which have different meanings, and how to generalize it. This paper formalizes generalized scaled dot-product attention by translating the attention weight into a multiplication of two terms: similarity and magnitude. We derive the explicit separation in Eq. 1 that the scaled dot-product attention is a product of 1) the Radial Basis Function (RBF) kernel function between query and key with fixed hyper-parameters, and 2) exponential of L^2 norm for each representation vectors.

From the factorization, we propose a new attention method which formulates an implicit kernel function and utilizes a generalized L^p norm. First, we propose implicit kernel attention (IKA), and IKA learns an appropriate kernel shape and its hyper-parameters by a data-driven approach instead of kernel manual selection. Kernel function in attention is a similarity measure to capture the dependency, and the dependency might be different with respect to the measure as shown in Figure 4a. Second, we interpret p in L^p norm as a hyper-parameter, so we define IKA with norm (IKAN), which improves IKA by changing L^p norm to control the magnitude of independent importance and the sparsity of attention weights as shown in Figure 4b. Third, we propose multi-head IKAN (MIKAN), which adopts a copula-augmented inference to estimate a structured spectral density of MHAs jointly.

$$\alpha_{ij} = \underbrace{\exp\left(\frac{-\|q_i - k_j\|_2^2}{2\sqrt{d_k}}\right)}_{\text{similarity}} \times \underbrace{\exp\left(\frac{\|q_i\|_2^2 + \|k_j\|_2^2}{2\sqrt{d_k}}\right)}_{\text{magnitude}} \Big/ \sum_l \left(\exp\left(\frac{-\|q_i - k_l\|_2^2}{2\sqrt{d_k}}\right) \times \exp\left(\frac{\|q_i\|_2^2 + \|k_l\|_2^2}{2\sqrt{d_k}}\right)\right) \quad (1)$$

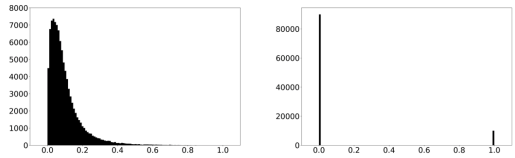
3 Hierarchical Context-aware Sequential Model

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

RBF	*Linear*	*Expsin*
cool	feels	heartbreaking
hottest	feeling	ironic
happy	frustrating	sadness
unfortunate	pretty	feel
terribly	feel	shocking

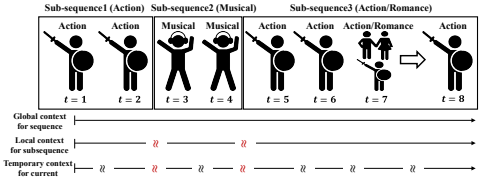
(a) The five most similar words for given query word, *sad* on Glove Pennington *et al.* [2014]. Every kernel captures the meaningful words, and their words are different depending on the kernel. Because each kernel has its own inductive bias, the appropriate kernel might be different on the dataset or task.



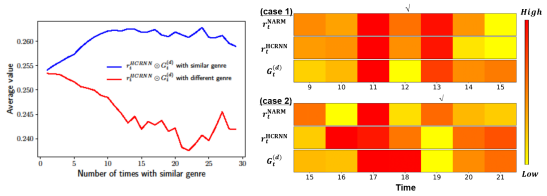
(b) Visualization of attention weight on different L^p norm for synthetic dataset. As $p \to 0$, the sparsity level of attention weights increases, and the most value become zero or one.

Figure 4: We explicitly derive that the scaled dot product attention can be expressed as a product of RBF kernel and exponential of L^2 norm. We generalize the scaled-dot product attention in two ways, kernel and norm. Kernel forms a topology of input space by introducing a metric of two given input instances, and the optimal kernel might be different depending on the dataset and task (left). The norm in attention compute the importance of query and key independently, and it affects the sparsity of attention weights (right)

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 represent large value if the current input of item has similar genre with previous input of item. In the opposite case, our gate grasps the user's interest drift and become smaller.

Figure 5: Hierarchical context modeling is important to handle 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 sentence deeply. The transformer is a core of many state-of-the-art for sequence modeling, and it handles the sequence without recurrence modeling. The parallel characteristic of the transformer depends on the two components, attention and positional encoding. I will investigate the attention and positional encoding in the near future.

Efficient Self-Attention (ESA) 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. The transformer is a core model of treating sequence datasets, but it still suffers from handling long sequences. The transformer depends on the scaled-dot product attention between instances, and it requires $O(L^2)$ time complexity for handling length L sequence. Its high computation complexity makes the transformer hard to model a longer sequence. For capturing the context of the sequence, sequential modeling should be performed in advance, and we will introduce a new attention method which has linear time complexity.

Higher-order Self-Attention (HSA) 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.

Locally Stationary Time Encoding (LSTE) Positional encoding helps the model to capture the order of instances without a recurrence structure. For discrete-time dataset such as sentence, absolute positional encoding Vaswani *et al.* [2017], and relative positional encoding Shaw *et al.* [2018]; Dai *et al.* [2019] captures the word order well. We will extend positional encoding to time encoding for handling the continuous-time dataset, such as user behavior history. The attention layer computes the dot product between positional encoding or time encoding itself to find the relationship between instances. We can interpret the results of the inner product as a kernel, and positional encoding or time encoding as random Fourier features. With LSTE, we can incorporate the user's behavior into the time encoding. The LSTE will helpful for the model to understand the user's sequential behavior without a recurrence structure.

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.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019.

Marc G Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2(Dec):299–312, 2001.

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

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

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543, 2014.

Michael Reed and Barry Simon. *II: Fourier Analysis, Self-Adjointness*, volume 2. Elsevier, 1975.

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

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL-HLT (2)*, 2018.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 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.