[9] B. Wang, D. Zhao, C. Lioma, Q. Li, P. Zhang, and J. Grue Simonsen,

‘‘Encoding word order in complex embeddings,’’ 2019, arXiv:1912.12333.

[10] D. Xu, C. Ruan, S. Kumar, E. Korpeoglu, and K. Achan, ‘‘Self-attention

with functional time representation learning,’’ in Proc. 33rd Int. Conf.

Neural Inf. Process. Syst. Red Hook, NY, USA: Curran Associates Inc.,

Dec. 2019, no. 1426, pp. 15915–15925.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou,

W. Li, and P. J. Liu, ‘‘Exploring the limits of transfer learning with a unified

text-to-text transformer,’’ 2019, arXiv:1910.10683.

[12] T.-C. Chi, T.-H. Fan, P. J. Ramadge, and A. I. Rudnicky, ‘‘KERPLE:

Kernelized relative positional embedding for length extrapolation,’’ 2022,

arXiv:2205.09921.

[13] Z. Huang, D. Liang, P. Xu, and B. Xiang, ‘‘Multiplicative position-

aware transformer models for language understanding,’’ 2021,

arXiv:2109.12788.

[14] G. Ke, D. He, and T.-Y. Liu, ‘‘Rethinking positional encoding in language

pre-training,’’ 2020, arXiv:2006.15595.

[15] O. Press, N. A. Smith, and M. Lewis, ‘‘Train short, test long: Attention with

linear biases enables input length extrapolation,’’ 2021, arXiv:2108.12409.

[16] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, ‘‘RoFormer:

Enhanced transformer with rotary position embedding,’’ 2021,

arXiv:2104.09864.

[17] X. Wang, Z. Tu, L. Wang, and S. Shi, ‘‘Self-attention with structural

position representations,’’ 2019, arXiv:1909.00383.

[18] A. Haviv, O. Ram, O. Press, P. Izsak, and O. Levy, ‘‘Transformer language

models without positional encodings still learn positional information,’’

2022, arXiv:2203.16634.

[19] C. B. Gemirter and D. Goularas, ‘‘A Turkish question answering system

based on deep learning neural networks,’’ J. Intell. Syst., Theory Appl.,

vol. 4, no. 2, pp. 65–75, Sep. 2021.

[20] C. Zeng and S. Li, ‘‘Analyzing the effect of masking length dis-

tribution of MLM: An evaluation framework and case study on

Chinese MRC datasets,’’ Wireless Commun. Mobile Comput., vol. 2021,

pp. 1–17, Nov. 2021. [Online]. Available: https://www.hindawi.com/

journals/wcmc/2021/5375334/

[21] C. Liu, C. Zhu, and W. Zhu, ‘‘Chinese named entity recognition based

on BERT with whole word masking,’’ in Proc. 6th Int. Conf. Comput.

Artif. Intell. New York, NY, USA: Association for Computing Machinery,

Aug. 2020, pp. 311–316, doi: 10.1145/3404555.3404563.

[22] B. Chan, S. Schweter, and T. Müller, ‘‘German’s next language model,’’

in Proc. 28th Int. Conf. Comput. Linguistics, Barcelona, Spain, Dec. 2020,

pp. 6788–6796. [Online]. Available: https://aclanthology.org/2020.coling-

main.598

[23] Y. Dai, L. Li, C. Zhou, Z. Feng, E. Zhao, X. Qiu, P. Li, and D. Tang,

‘‘‘Is whole word masking always better for Chinese BERT?’: Probing on

Chinese grammatical error correction,’’ 2022, arXiv:2203.00286.

[24] Y. Chen, Z. Liang, Z. Tan, and D. Lin, ‘‘Named entity recognition in

power marketing domain based on whole word masking and dual feature

extraction,’’ Appl. Sci., vol. 13, no. 16, p. 9338, Aug. 2023. [Online].

Available: https://www.mdpi.com/2076-3417/13/16/9338

[25] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, ‘‘Pre-training with whole word

masking for Chinese BERT,’’ IEEE/ACM Trans. Audio, Speech, Language

Process., vol. 29, pp. 3504–3514, 2021.