Machine Reading Comprehension based on Language Model with Knowledge Graph

Seonghyun Kim* et al.

AIR Team, AI Labs, Saltlux Inc., Seoul, Korea **seonghyunkim@saltlux.com*

Introduction

- Question answering (QA) has grown substantially in the past decade, with applications in the life sciences, financial services, insurance, healthcare, financial software, government, and public accounting.
- Traditional QA system provided accurate and concise answers by understanding the natural language questions and mapping precisely to structured queries over the knowledge base (KB) [1, 2].
- In contrast of KB based QA, an information retrieval (IR) based QA system as also known machine reading comprehension (MRC) system that can give a result on unstructured natural language data has been recently proposed [3-5].
- Especially, BERT (Bidirectional Encoder Representations from Transformers) has shown remarkable result in MRC over human performance [6-9].
- However, state of the art models of MRC such as BERT rarely consider incorporating knowledge graphs (KGs), which can provide rich structured knowledges for better language understanding.
- Therefore, we suggest pre-trained language model that learn language with knowledge information from KG to enhance the understanding of natural language.

Methods

- In order to represent the entity information to language model, we define 707,667 entities that contain 40 types of predicate from open KG API as candidate key entities (in detail, see our paper) [10].
- The key entities are used to preprocess given sentence as shown in Table 1.
- Our model is based on BERT model, and the training process and hyper parameters are set according to published BERT model [11].

Preprocessing step	Sentence example
Original sentence	이순신은 조선 중기의 무신이다.
Extracting candidate key entities	이순신, 조선, 조선 중기
Extracting key entities	이순신, 조선
Tagging key entities	[ENT] 이순신 [/ENT] 은 [ENT] 조선 [/ENT] 중기의 무신이다.
Morphological analysis	이순신/NNP 은/JX 조선/NNP 중기/NNG 의/JKG 무신/NNG 이/VCP 다/EP ./SF
Result	[ENT] 이순신/NNP [/ENT] 은/JX [ENT] 조선/NNP [/ENT] 중기/NNG 의/JKG 무신/NNG 이/VCP 다/EP ./SF

Table 1. A preprocessing example of given sentence

- We use Korean Wikipedia data for pre-training that was continued until 30,000 steps in 128 batch sizes.
- To validate the model, we experiment MRC task using KorQuAD v1.0 dataset and all of source codes are shared at github [12].

Results & Conclusion

- First, we analyze the training loss and masked token prediction accuracy after pre-training as shown in below table.
- The training loss of proposed model ($BERT_{Entity}$) is higher than that of the baseline BERT model ($BERT_{Base}$).
- Interestingly, the pre-training performance was not good for the proposed model, but the fine-turning performance for MRC was better for our model (Table 3).

Model	Loss	Masked token prediction accuracy
$BERT_{Base}$	0.83	0.79
$BERT_{Entity}$	1.12	0.75

Table 2. Pre-training results of proposed language model

Model	F1	Exact matching
$BERT_{Base}$	64.51	83.76
$BERT_{Entity}$	78.13	87.25

Table 3. MRC result using KorQuAD v1.0 dataset

- As a result, when the entity information from KG is used for training language, the language understanding performance is further improved because the language model has more meaningful features to learn.
- In the other hand, the knowledge information plays a major role in understanding a language.

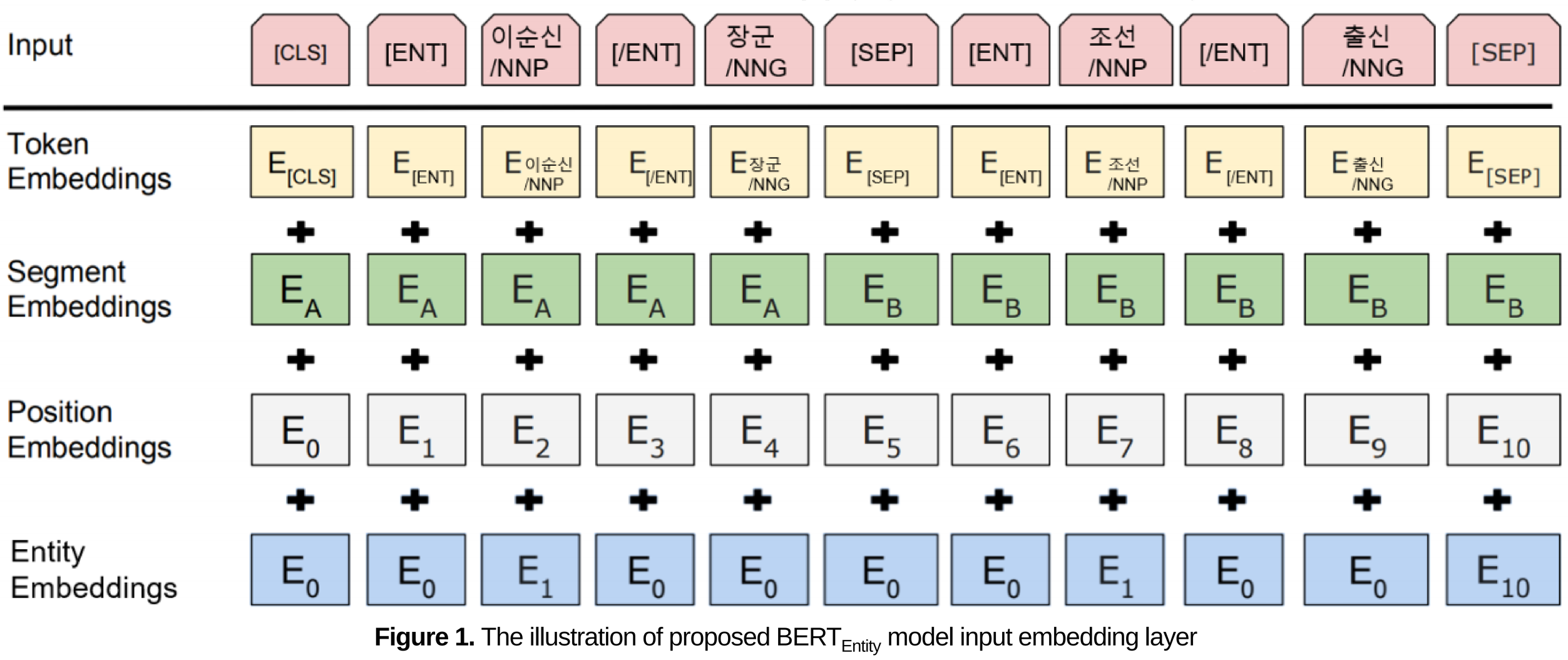

Figure 1. The illustration of proposed $BERT_{Entity}$ model input embedding layer

Acknowledgement

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform).

References

[1] Cui et al., 2017, VLDB [2] Yih et al., 2016, ACM [3] Chen et al., 2017, arXiv [4] Joshi et al., 2017, arXiv [5] Ng et al., 2000, ACL [6] Devlin et al., 2018, arXiv [7] Rajpurkar et al., 2016, arXiv [8] Lim et al., 2016, arXiv [9] Park et al., KIISE [10] adams.ai [11] https://github.com/google-research/bert [12] https://github.com/MrBananaHuman/BertEntity