H-Bert: Enhancing Chinese Pretrained Models with Attention to HowNet

Wei Zhu[1,2]

[1] East China Normal University, Shanghai, China
[2] DataSelect AI Technology, Shanghai, China

Abstract. Pretrained transformers for Chinese show remarkable performances on various natural language processing tasks. However, these models are purely data-driven and fail to incorporate explicit semantic knowledge, like HowNet. In this paper, we propose H-BERT, which enhances the semantic representations of Chinese BERT by incorporating sememe knowledge from HowNet in the pretraining stage via multi-head attention. Our experiments demonstrate that H-BERT can significantly outperform the vanilla BERT on the downstream tasks. Ablation study compares different settings of H-BERT and shows that and case study also shows that knowledge injection is required at both the pretraining and fine-tuning stage.[3]

Keywords: pretrained language models · knowledge graph · knowledge enhanced pretraining.
Type of submission: Poster

1 Introduction

Since the rise of BERT, pretrained language models (PLMs) have dominated state of the art (SOTA) for a comprehensive list of natural language tasks [2, 7, 3]. Despite their powerfulness, PLMs still fall short on a series of tasks that requires entity level and domain level knowledge [13]. As a result, a branch of literature has been dedicated to injecting structured knowledge into PLMs, both in pretraining and fine-tuning stages. One approach is to inject structure information of knowledge graph via entity embedding [15]. Similarly, [9] and [13] pretrains BERT jointly with knowledge embedding training. Another approach is to inject knowledge by adding them into the original sentence. [5] explicitly injects related triples extracted from KG into the sentence to obtain an extended tree-form input for BERT.

However, the literature falls short on three aspects. First, many PLMs that inject knowledge from pretraining suffer from catastrophic forgetting and thus can only perform well on entity-related tasks such as named entity recognition and relation classification, but performs poorly on sentence-level tasks like GLUE [12]. Second, there are few studies on enriching Chinese PLMs with Knowledge. K-BERT studied incorporating HowNet without pretraining by adding the knowledge facts into the input sentence. However, it requires manually select the most important two sememes for each word, which is unsuitable for scale-up and tasks of different domains.

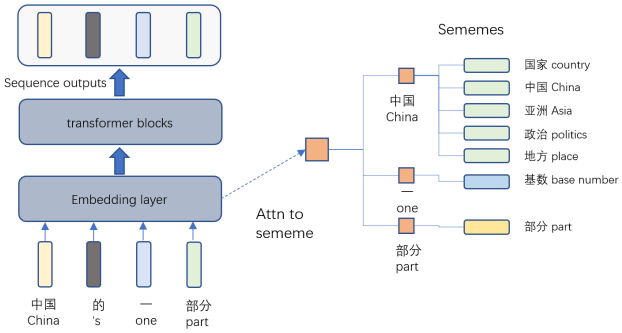

Fig. 1. The model architecture of H-BERT.

In this work, we adopt the core data of HowNet [10] as our knowledge source. HowNet was initially designed and constructed in the 1990s. Furthermore, it has kept frequently updating since it was published in 1999. In HowNet, each word has several sememes, which by linguistic definition, are the minimum semantic units of language and can well represent implicit semantic meanings behind words. For example, as shown in Figure 1, the word 中国(China) has a series of sememes, i.e., 国家(country), 中国(China), 亚洲(Asia), 政治(politics), 地方(place). The sememe set of HowNet is determined by extracting, analyzing, merging, and filtering semantics of thousands of Chinese characters. HowNet is widely applied for knowledge-enhanced word/sentence representations [8, 14] and is shown to be beneficial for a wide range of NLP tasks. However, the previous work does not combine HowNet with language model pretraining.

This work proposes HowNet BERT (H-BERT), a transformer-based model that uses a simple multi-head attention module to incorporate How-Net knowledge. Our H-BERT model is depicted in Figure 1. A sentence is encoded via two modules. First, it is tokenized and embedded via the token embedding layer. Second, we recognize the words which are included in How-Net and obtain their sememes. Tokens in the same word will have the same sememes. We can treat the sememes of a word as sub-word features. Sememes are treated as the minimum units, and the sememes of the tokens will be embedded via the sememe embedding layer. Knowledge from How-Net is injected via a multi-head attention layer from the token representation to the tokens' sememe representation (denoted as attn-to-sememes). Here, this attention module can be implemented right after the token embedding layer or after obtaining the sequential output of the token encoder.

We conduct experiments on sentence classification (CLS) and sentence pair classification (NLI), which are whole-sentence level tasks, and named entity recognition (NER), which is an entity-level task. Tasks from different domains are selected. Experimental results show that our model consistently outperforms the vanilla ALBERT and K-BERT on a series of tasks, indicating that our model can handle both whole sentence-level and entity-level tasks equally well. Moreover, we find that pretraining with attn-to-sememe but exclude this module during fine-tuning also improves the performance of the vanilla PLM.

2 Methodology

2.1 Sentence Encoder

Figure 1 gives a high-level description of transformer architecture in pretrained models [2]. A sentence is tokenized into tokens and are embedded as $H^0 = (w_1, w_2, ..., w_{T_t})$. After going through BERT encoder, we obtain the contextualized representation $H^L = (h_1, h_2, ..., h_{T_t})$.

2.2 Sememe Embedding

Now we discuss how to incorporate the HowNet knowledge. First, in a sentence, we match all the words (not overlapping) included in the sememe via FlashText [11]. Then sememes of these words are obtained. The tokens will have the same sememe if a word has more than one token after sub-word tokenization. Now we have a sememe sequence $S^0 = (s_1, s_2, ..., s_{T_t})$, in which $s_i = [sem_1, sem_2, ..., sem_{l_i}]$ means the word s_i is in has l_i sememes according to HowNet. For tokens in words that are not in HowNet, $l_i = 1$ and we given it a special padding sememe, denoted as $< s - pad >$. For example, in Figure 1, different words have different sememes, and some have no sememes. The tokens' sememes are embedded to tensor $SE^0 = (se_1, se_2, ..., se_{T_t})$. The sememe embedding layer is randomly initialized and is learned along with pretraining.

2.3 Knowledge Injection

Knowledge is injected via multi-head attention. The token embeddings H^0 are treated as query, and the sememe embeddings SE^0 are treated as key and value. Knowledge enriched representation of the sentence H^S is obtained by multi-head attention from the query to the sememes. We call this knowledge injection module as attn-to-sememes.

2.4 Training and Fine-tuning

We include the attn to sememes in the pretraining stage. During pretraining, for masked language modeling (MLM), the masked tokens will be treated as tokens with no sememes, i.e., it only has a padding sememe $< s - pad >$. We believe that including sememe knowledge during pretraining can help to speed up learning semantic meanings of tokens.

For pretrained H-BERT, we can fine-tune it in two approaches: (a) discarding the attn-to-sememe module and fine-tuning H-BERT in the same way as BERT; (b) taking advantage of HowNet during fine-tuning, that is, extracting the sememe informaitons from the sentence, and using the attn-to-sememe module to inject sememe information to H-BERT's encoder.

Model	#params	chn	lcqmc	xnli	msra_ner	fin_ner	ccks_ner
ALBERT base	9.9M	91.10	84.21	60.97	85.14	78.43	90.63
K-BERT	9.9M	91.53	84.71	61.37	85.78	78.63	90.82
H-BERT v3	10.3M	91.48	84.67	61.46	85.69	78.58	90.78
H-BERT v2	10.3M	91.58	84.84	61.72	85.97	78.76	90.93
H-BERT v0	10.3M	**92.08**	**85.16**	**61.93**	**86.48**	**79.08**	**91.15**
ALBERT large	16.5M	93.54	87.21	64.89	87.53	80.35	92.46
H-BERT v1	16.8M	**94.47**	**88.34**	**65.94**	**89.07**	**81.23**	**93.35**

Table 1. Main results on the 6 benchmark datasets. Average scores of ten runs are reported. We also report each models' number of parameters.

3 Experiments

3.1 Experimental Setup

Pretraining is done on the Chinese Wikipedia corpus. We use the vocabulary of Google Chinese Bert [2] for tokenization. We pretrain three models totally from scratch: (a) ALBERT base [3]; (b) ALBERT large [3]; (c) H-BERT v0, which uses a randomly initialized ALBERT base as the encoder and includes an attn-to-sememe module on the embedding layer; (d) H-BERT v1, which is in the large model setting; (e) H-BERT v2, which is a base model and puts the attn-to-sememe module on the last layer of its Transformer encoder. (f) H-BERT v3, which is H-BERT v0 fine-tuned without attn-to-sememe. For H-BERT v0 and H-BERT v2, the hidden size of transformers is reduced to 640. The hidden size of H-BERT v1 is set to 980. Moreover, the attn-to-sememe module reuses the encoder's parameters. Thus, the number of parameters is comparable to ALBERT base. We apply our pretrained vanilla Albert base on the open-sourced codes of K-Bert [5] to obtain the results.

During fine-tuning, we mainly follow the hyper-params of [3]. Each model runs 10 times to ensure reproducibility.

3.2 Datasets

We experiment across a diverse set of 6 benchmark NLP tasks and demonstrate the effectiveness of our model. For text classification, we select ChnSentiCorp (**chn**)[4]. For NLI, we select LCQMC [6] (**lcqmc**) and XNLI [1] (**xnli**). For named entity recognition task, three datasets from different domains are selected. MSRA NER (**msra_ner**) [4] is from the open domain, Finance NER[5] (**fin_ner**) is from the financial domain, and CCKS NER[6] (**ccks_ner**) is collected from the medical records.

3.3 Results and Analysis

The experimental results are reported in Table 1. The main takeaways are:

- H-BERT v0 consistently outperforms ALBERT base and K-BERT with comparable parameters, demonstrating H-BERT's effectiveness.
- H-BERT v1 outperforms ALBERT large, demonstrating that our method can also work for large pretrained models.
- H-BERT v3 performs worse than H-BERT v0, but it is better than ALBERT base, showing that attn-to-sememe helps improve the generalization ability of pretrained models. In addition, adopting attn-to-sememe during fine-tuning is beneficial for downstream tasks.
- Comparing H-BERT v2 with H-BERT v0, we can see that it is better to apply attn-to-sememes at the embedding layer of the encoder model.

[4] https://github.com/pengming617/bert classification
[5] https://embedding.github.io/evaluation/#extrinsic
[6] https://biendata.com/competition/CCKS2017 2/

4 Conclusion

This article proposes to enhance the Chinese pre-trained language models with simple attention to sememes module. Experiments on 6 benchmark datasets shows that: From the architecture point of view, attn-to-sememes should be applied at the embedding layer; (2) attn-to-sememes are required at both pretraining and fine-tuning stages for better downstream performances. Our model can beat the vanilla ALBERT significantly across 6 datasets with roughly the same amount of parameters, showing that our model can effectively inject knowledge into PLMs.

References

1. Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., Stoyanov, V.: XNLI: Evaluating cross-lingual sentence representations. In: EMNLP 2018. pp. 2475–2485. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1269, https://www.aclweb.org/anthology/D18-1269
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv e-prints arXiv:1909.11942 (Sep 2019)
4. Levow, G.A.: The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. pp. 108–117. Association for Computational Linguistics, Sydney, Australia (Jul 2006), https://www.aclweb.org/anthology/W06-0115
5. Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., Wang, P.: K-BERT: Enabling Language Representation with Knowledge Graph. arXiv e-prints arXiv:1909.07606 (Sep 2019)
6. Liu, X., Chen, Q., Deng, C., Zeng, H., Chen, J., Li, D., Tang, B.: LCQMC:a large-scale Chinese question matching corpus. In: CL. pp. 1952–1962. ACL, Santa Fe, New Mexico, USA (Aug 2018)
7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv e-prints arXiv:1907.11692 (Jul 2019)
8. Niu, Y., Xie, R., Liu, Z., Sun, M.: Improved word representation learning with sememes. In: ACL (2017)
9. Peters, M.E., Neumann, M., Logan, Robert L., I., Schwartz, R., Joshi, V., Singh, S., Smith, N.A.: Knowledge Enhanced Contextual Word Representations. arXiv e-prints arXiv:1909.04164 (Sep 2019)
10. Qi, F., Yang, C., Liu, Z., Dong, Q., Sun, M., Dong, Z.: Openhownet: An open sememe-based lexical knowledge base. arXiv preprint arXiv:1901.09957 (2019)
11. Singh, V.: Replace or Retrieve Keywords In Documents at Scale. arXiv e-prints arXiv:1711.00046 (Oct 2017)
12. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)
13. Wang, X., Gao, T., Zhu, Z., Liu, Z., Li, J., Tang, J.: Kepler: A unified model for knowledge embedding and pre-trained language representation. arXiv preprint arXiv:1911.06136 (2019)
14. Zhang, Y., Yang, C., Zhou, Z., Liu, Z.: Enhancing transformer with sememe knowledge. In: RepL4NLP@ACL (2020)
15. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: Enhanced Language Representation with Informative Entities. arXiv e-prints arXiv:1905.07129 (May 2019)