Building Natural Language Processing Applications with EasyNLP

Chengyu Wang
Alibaba Group
Hangzhou, China
chengyu.wcy@alibaba-inc.com

Minghui Qiu
Alibaba Group
Hangzhou, China
minghui.qmh@alibaba-inc.com

Jun Huang
Alibaba Group
Hangzhou, China
huangjun.hj@alibaba-inc.com

ABSTRACT

The successful application of Pre-Trained Models (PTMs) has revolutionized the development of Natural Language Processing (NLP) by large-scale self-supervised pre-training. However, it is not easy to obtain high-performing models in domain-specific applications and deploy them online with strict QPS (Query Per Second) requirements for industrial practitioners. To solve these issues, the EasyNLP toolkit is designed for building PTM-based NLP applications with ease, which supports a comprehensive suite of NLP algorithms and is suitable for meeting the inference requirements in industry. It features knowledge-enhanced pre-training that captures rich domain knowledge to better support domain-specific applications. In addition, the knowledge distillation and prompt-based few-shot learning functionalities are provided to improve the performance of large-scale PTMs with little training data available, and to distill models to smaller ones that are suitable for online deployment. EasyNLP provides a unified framework of model training, inference and deployment for real-world applications, using simple high-level APIs or command-line tools. Currently, EasyNLP has powered over ten business units within Alibaba Group and is seamlessly integrated to the Platform of AI (PAI) products on Alibaba Cloud. EasyNLP is also beneficial for academia, as it integrates state- of-the-art methods and models to make it easy for researchers to benchmark and develop their own algorithms. We have released EasyNLP to public at GitHub (https: //github.com/alibaba/EasyNLP).

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

natural language processing, Pre-Trained Models, knowledge-enhanced pre-training, knowledge distillation, few-shot learning

ACM Reference Format:
Chengyu Wang, Minghui Qiu, and Jun Huang. 2022. Building Natural Language Processing Applications with EasyNLP. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3511808.3557510

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557510

1 AN OVERVIEW OF EASYNLP

Large-scale Pre-Trained Models (PTMs) have achieved significant improvements in the performance of downstream Natural Language Processing (NLP) tasks [5]. With the explosion of data and computational resources available, the parameter space and the scale of PTMs are becoming larger over time.

Despite the remarkable success, there still exist some burdens on the applications of large-scale PTMs for industrial practitioners, with reasons stated as follows:

- Large-scale PTMs do not always bring better performance in domain-specific applications of users on the cloud, due to the lack of domain knowledge. Therefore, it is highly necessary to make the underlying PTMs truly understand knowledge facts by knowledge-enhanced pre-training.
- Although ultra-large PTMs such as GPT-3 [1] have achieved good few-shot/zero-shot performance, the huge size still restricts the usage of these models in real-world applications. Hence, few-shot fine-tuning and knowledge distillation are fundamental functionalities for industrial practitioners.
- There have been a series of widely-used open-source NLP libraries such as transformers[1] and AllenNLP[2]. Yet, real-world application needs call for a unified NLP framework of model training, inference and deployment on the cloud, with advanced features provided for users.

Based on our industrial experiences on Alibaba Cloud[3], we have built the PyTorch-based EasyNLP toolkit that is designed to make the applications of large-scale PTMs for online applications efficiently and effectively.[4] The framework of EasyNLP is presented in Fig. 1. Briefly speaking, features of our EasyNLP toolkit include:

- **Rich APIs and Easy-to-use Functionalities.** EasyNLP provides easy-to-use APIs and command-line tools to call cutting-edge NLP models, include ModelZoo (for building PTMs), AppZoo (for building downstream NLP applications), etc. It also has DataHub, providing users with simple interfaces to load and process NLP datasets.
- **Knowledge-enhanced PTMs.** EasyNLP is equipped with knowledge-enhanced pre-training techniques based on our proposed DKPLM framework [8], which decomposes the processes of knowledge-enhanced pre-training and task-specific fine-tuning/inference. Based on DKPLM, we have provided the knowledge-enhanced PTMs of various domains so that users can tune the models and deploy them online in the same way as BERT. Users can also train PTMs using their own knowledge bases using our pre-training APIs.

[1] https://github.com/huggingface/transformers
[2] https://github.com/allenai/allennlp
[3] https://www.alibabacloud.com
[4] A detailed introduction to the EasyNLP toolkit can be found in [6]. Readers are more than welcomed to check out the source codes at https://github.com/alibaba/EasyNLP.

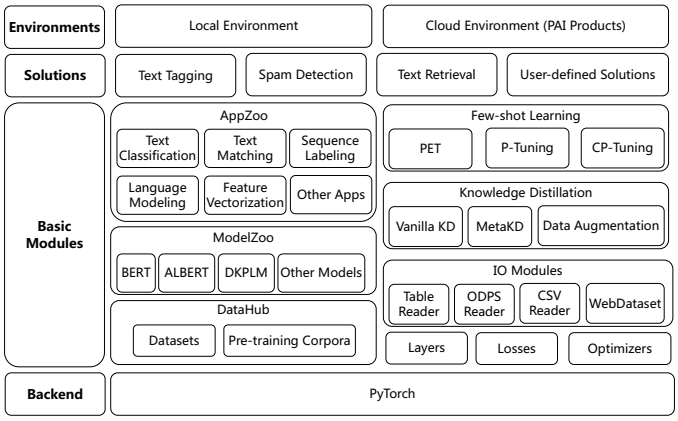

Figure 1: The overall framework of EasyNLP.

- **Few-shot Learning.** EasyNLP integrates a variety of popular prompt-based few-shot learning algorithms as introduced in [2], allowing users to fine-tune PTMs with few training samples. Furthermore, we provide a new few-shot learning algorithm named CP-Tuing [7] that is free of manual verbalizer construction based on contrastive learning.
- **Knowledge Distillation.** EasyNLP provides knowledge distillation algorithms that quickly distill large PTMs to small and efficient models for online deployment. For scenarios where training data of the target task is scarce, our MetaKD algorithm [3] is beneficial for improving the accuracy of student models based on cross-domain datasets.
- **Compatible with Open-source Community.** EasyNLP has APIs to support model training from other open-source libraries such as transformers, enhanced with our in-house distributed learning framework. It is also compatible with PTMs in EasyTransfer ModelZoo [4] based on TensorFlow.
- **Product-ready Platform.** EasyNLP is closely connected to the Machine Learning Platform of AI (PAI) products on Alibaba Cloud[5], including PAI-DSW for model development, PAI-DLC for cloud-native training, PAI-Designer for zero-code training and PAI-EAS for online serving.

2 FUTURE ROADMAP

It is obvious that EasyNLP is far from prefect. Based on our experiences of using EasyNLP to support business units inside Alibaba Group, we have identified several technical challenges that guide us to draw the roadmap for future development, listed as follows: i) the support for multi-modality PTMs; ii) better knowledge-enhanced PTMs for closed-domains to satisfy the industrial needs; iii) richer APIs to support complicated NLP and multi-modality tasks; and iv) advanced techniques for distributed model training and inference acceleration of large-scale PTMs.

3 RELEVANCY TO CIKM

EasyNLP is a comprehensive toolkit for building various NLP applications to support real-world, industrial scenarios. It has been seamlessly integrated into the PAI products, and has also been released to the open-source community. We present the design principles and unique features to show its values. Several industrial use cases will also be given in the talk. In addition, EasyNLP is highly beneficial for academia, as it integrates state-of-the-art methods and models to make it easy for researchers to benchmark and develop their own algorithms.

4 SPEAKER BIOGRAPHIES

Chengyu Wang received his Ph.D degree from East China Normal University (ECNU) in 2020. Currently, he works on deep learning algorithms on various topics for Alibaba Cloud Machine Learning Platform of AI (PAI), including natural language processing, human speech understanding, transfer learning and few-shot learning. He has published 60+ research papers in international conferences and journals, such as ACL, KDD, SIGIR, WWW, AAAI, TKDE, CIKM, WSDM, etc.

Minghui Qiu holds a Ph.D. degree at the School of Information Systems, Singapore Management University, in 2015. He was a visiting scholar in Language Technologies Institute, Carnegie Mellon University from 2013 to 2014. He is currently a senior algorithm expert in Platform of AI (PAI) in Alibaba Cloud, working on deep learning and transfer learning for many NLP and IR tasks.

Jun Huang received a Ph.D. degree of Modern Physics from University of Science and Technology of China in 2008. He was an associate research fellow of China Academy of Engineering Physics. Now he leads a team for developing AI algorithms on the Platform of AI of Alibaba Group, responsible for developing innovative algorithms and platforms to serve important internal and external businesses of Alibaba. His interests include high performance distributed implementation of AI algorithms and applying them to real applications.

REFERENCES

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

[2] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *CoRR* abs/2107.13586 (2021).

[3] Haojie Pan, Chengyu Wang, Minghui Qiu, Yichang Zhang, Yaliang Li, and Jun Huang. 2021. Meta-KD: A Meta Knowledge Distillation Framework for Language Model Compression across Domains. In *ACL/IJCNLP*. 3026–3036.

[4] Minghui Qiu, Peng Li, Chengyu Wang, Haojie Pan, Ang Wang, Cen Chen, Xianyan Jia, Yaliang Li, Jun Huang, Deng Cai, and Wei Lin. 2021. EasyTransfer: A Simple and Scalable Deep Transfer Learning Platform for NLP Applications. In *CIKM*. 4075–4084.

[5] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained Models for Natural Language Processing: A Survey. *CoRR* abs/2003.08271 (2020).

[6] Chengyu Wang, Minghui Qiu, Taolin Zhang, Tingting Liu, Lei Li, Jianing Wang, Ming Wang, Jun Huang, and Wei Lin. 2022. EasyNLP: A Comprehensive and Easy-to-use Toolkit for Natural Language Processing. *CoRR* abs/2205.00258 (2022).

[7] Ziyun Xu, Chengyu Wang, Minghui Qiu, Fuli Luo, Runxin Xu, Songfang Huang, and Jun Huang. 2022. Making Pre-trained Language Models End-to-end Few-shot Learners with Contrastive Prompt Tuning. *CoRR* abs/2204.00166 (2022).

[8] Taolin Zhang, Chengyu Wang, Nan Hu, Minghui Qiu, Chengguang Tang, Xiaofeng He, and Jun Huang. 2022. DKPLM: Decomposable Knowledge-enhanced Pre-trained Language Model for Natural Language Understanding. In *AAAI*.

[5] https://www.alibabacloud.com/product/machine-learning