Spa-NLP 2022

The 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge

Proceedings of the Workshop

May 27, 2022

The Spa-NLP organizers gratefully acknowledge the support from the following sponsors.

Gold

Silver

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-955917-50-6

Introduction

We are excited to present the inaugural workshop on semiparametric methods on NLP.

The field of natural language processing (NLP) has undergone a paradigm shift with the dramatic success of large pre-trained language models (LMs) on almost every downstream task. These large parametric models are based on the transformer architecture and are trained on massive collections of data using self-supervised learning, which are then fine-tuned on a relatively smaller set of task-specific supervised examples. The success of this simple recipe of homogeneous architectures and transfer learning has led to its widespread adoption.

Despite these successes, parametric models lack several desirable properties. For example, these models use knowledge stored in their parameters to perform tasks without providing provenance or transparency into the model mechanisms. This is further exacerbated when they make an erroneous prediction as it is challenging to understand what went wrong and how to fix it. Moreover, as new information arrives, existing knowledge becomes obsolete and should be updated. However, it is currently challenging to update the knowledge stored in the parameters of LMs. Amongst other issues, this has implications on personal privacy as we do not have a robust way to execute requests for deletion of personal information which could be stored in the parameters of the model.

Nonparametric instance-based models, on the other hand, offer many of the properties described above by design — a model capacity that naturally grows with data, easy addition and deletion of knowledge, and provenance for predictions based on the nearest neighbors with respect to the input. However, these models often suffer from weaker empirical performance compared to deep parametric models. Semiparametric models are statistical models that consist of a fixed parametric and a flexible nonparametric component. Combining the advantages of both paradigms has the potential to remedy many of the shortcomings described previously. For example, the nonparametric component can provide vast amounts of background knowledge and the parametric component can encode the logic required to solve the problem.

Recently, many recent works have independently proposed approaches that combine a parametric model with a nonparametric model in areas from question answering, language modeling, machine translation, and even protein structure prediction. Given the increasingly promising results on various tasks of such semiparametric models, we believe this area is ripe for targeted investigation on understanding efficiency, generalization, limitations, and to widen its applicability.

This workshop invited previously unpublished work as archival submissions, in addition to a non-archival track of previously-published work, recognising the fast-moving nature of this area, and the large amount of recently introduced work. After withdrawals, We have accepted a total of 5 archival papers, and 21 non-archival papers. Our final program thus includes 26 papers, 5 of which will be included in the proceedings.

We are excited to host six stellar invited speakers, who will each lend their perspective to this exciting and rapidly-evolving area. In the morning session, we will host Anna Potapenko, and in the afternoon session, we will host Danqi Chen, Jason Weston, Andrew McCallum and Hannaneh Hajishirzi. We shall finish with a panel discussion. We thank these speakers, our program committee, the ACL workshop chairs, and our sponsors, Google and Meta, for helping to make this workshop possible.

Non-Archival Papers

The following papers were submitted to our workshop as non-archival submissions.

- *Learning To Retrieve Prompts for In-Context Learning* Ohad Rubin, Jonathan Herzig, Jonathan Berant
- *A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models* Da Yin, Li Dong, Hao Cheng, Xiaodong Liu, Kai-Wei Chang, Furu Wei, Jianfeng Gao
- *KNN-BERT: Fine-Tuning Pre-Trained Models with KNN Classifier* Linyang Li, Demin Song, Ruotian Ma, Xipeng Qiu, Xuanjing Huang
- *Learning to Retrieve Passages without Supervision* Ori Ram,Gal Shachaf,Omer Levy,Jonathan Berant,Amir Globerson
- *Hyperlink-induced Pre-training for Passage Retrieval in Open-domain Question Answering* Jiawei Zhou,Xiaoguang,Lifeng Shang,Lan Luo,Ke Zhan,Enrui Hu,Xinyu Zhang,Hao Jiang,Zhao Cao,Fan Yu,Xin Jiang,Qun Liu,Lei Chen
- *Towards Continual Knowledge Learning of Language Models* Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, Minjoon Seo
- *Learning Cross-Lingual IR from an English Retriever* Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, Avirup Sil
- *C-MORE: Pretraining to Answer Open-Domain Questions by Consulting Millions of References* Xiang Yue, Xiaoman Pan, Wenlin Yao, Dian Yu, Dong Yu, Jianshu Chen
- *Towards Unsupervised Dense Information Retrieval with Contrastive Learning* Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, Edouard Grave
- *Internet-augmented language models through few-shot prompting for open-domain question answering* Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, Nikolai Grigorev
- *Towards Interactive Language Modeling* Maartje ter Hoeve, Evgeny Kharitonov, Dieuwke Hupkes, Emmanuel Dupoux
- *GUD-IR: Generative Retrieval for Semiparametric Models* Aman Madaan, Niket Tandon, Peter Clark, Yiming Yang
- *Less is More: Summary of Long Instructions is Better for Program Synthesis* Kirby Kuznia, Swaroop Mishra, Mihir Parmar, Chitta Baral
- *How Many Data Samples is an Additional Instruction Worth?* Ravsehaj Singh Puri, Swaroop Mishra, Mihir Parmar, Chitta Baral
- *Is Retriever Merely an Approximator of Reader?* Sohee Yang,Minjoon Seo
- *TemporalWiki: A Lifelong Benchmark for Training and Evaluating Ever-Evolving Language Models* Joel Jang, Seonghyeon ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Minjoon Seo
- *Unsupervised Cross-Task Generalization via Retrieval Augmentation* Bill Yuchen Lin, Kangmin Tan, Chris Scott Miller, Beiwen Tian, Xiang Ren

- *Controllable Semantic Parsing via Retrieval Augmentation* Panupong Pasupat, Yuan Zhang, Kelvin Guu
- *StreamingQA: A Benchmark for Adaptation to New Knowledge over Timein Question Answering Models* Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, Angeliki Lazaridou
- *On the Effect of Pretraining Corpora on In-context Few-shot Learning by a Large-scale Language Model* Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, Nako Sung
- *Exploring Dual Encoder Architectures for Question Answering* Zhe Dong, Jianmo Ni, Daniel M. Bikel, Enrique Alfonseca, Yuan Wang, Chen Qu, Imed Zitouni

Organizing Committee

Program Chairs

Rajarshi Das, University of Massachusetts Amherst
Patrick Lewis, University College London and Facebook AI Research
Sewon Min, University of Washington
June Thai, University of Massachusetts Amherst
Manzil Zaheer, Google DeepMind

Program Committee

Program Committee Members

Akari Asai, Paul G. Allen School of Computer Science and Engineering, University of Washington
Arthur Mensch, DeepMind
Ameya Godbole, University of Southern California
Ashwin Paranjape, Stanford University
Ahsaas Bajaj, University of Massachusetts, Amherst
Binh Vu, University of Southern California
Jean Maillard, Facebook AI
Jonathan Herzig, Tel Aviv University
Kevin Lin, University of California Berkeley
Kai Sun, Meta
Mor Geva, Allen Institute for Artificial Intelligence
Ni Lao, mosaix.ai
Peng Qi, JD AI Research
Devendra Singh Sachan, Mila - Quebec AI Institute
Shehzaad Zuzar Dhuliawala, Swiss Federal Institute of Technology
Shayne Longpre, Massachusetts Institute of Technology
Sohee Yang, Korea Advanced Institute of Science and Technology
Tong Wang, Microsoft
Urvashi Khandelwal, Google
Wenhan Xiong, Facebook
Yichen Jiang, Department of Computer Science, University of North Carolina, Chapel Hill
Yizhong Wang, Department of Computer Science, University of Washington
Yuxiang Wu, University College London

Invited Speakers

Danqi Chen, Princeton University
Hannaneh Hajishirzi, University of Washington and Allen Institute for AI
Andrew McCallum, University of Massachusetts, Amherst
Anna Potapenko, Google Deepmind
Jason Weston, Facebook AI Research

Table of Contents

Program

Friday, May 27, 2022

09:20 - 09:30 *Opening Remarks*

09:30 - 10:10 *Invited Talk 1: Anna Potapenko*

10:10 - 10:20 *Archival Track Contributed Talk: Efficient Machine Translation Domain Adaptation: Pedro Henrique Martins, Zita Marinho, Andre Martins*

10:20 - 10:30 *Non-Archival Track Contributed Talks 1: Internet-augmented language models through few-shot prompting for open-domain question answering: Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, Nikolai Grigorev*

10:30 - 10:40 *Non-Archival Track Contributed Talks 2: Towards Unsupervised Dense Information Retrieval with Contrastive Learning: Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, Edouard Grave*

10:40 - 10:50 *Non-Archival Track Contributed Talks 3: Towards Continual Knowledge Learning of Language Models: Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, Minjoon Seo*

10:50 - 11:00 *Coffee Break*

11:00 - 12:00 *Poster Session I*

12:00 - 13:30 *Lunch Break*

13:30 - 14:10 *Invited Talk 2: Danqi Chen*

14:10 - 14:50 *Invited Talk 3: Jason Weston*

14:50 - 15:00 *Coffee Break*

15:00 - 16:00 *Poster Session II*

16:00 - 16:40 *Invited Talk 4: Andrew McCallum*

16:40 - 17:20 *Invited Talk 5: Hannah Hajishirzi*

17:20 - 17:50 *Panel Discussion*

17:50 - 18:00 *Closing Remarks*