Pretrained Language Representations for Text Understanding: A Weakly-Supervised Perspective

Yu Meng
yumeng5@illinois.edu
University of Illinois
Urbana-Champaign, IL, USA

Jiaxin Huang
jiaxinh3@illinois.edu
University of Illinois
Urbana-Champaign, IL, USA

Yu Zhang
yuz9@illinois.edu
University of Illinois
Urbana-Champaign, IL, USA

Yunyi Zhang
yzhan238@illinois.edu
University of Illinois
Urbana-Champaign, IL, USA

Jiawei Han
hanj@illinois.edu
University of Illinois
Urbana-Champaign, IL, USA

ABSTRACT

Language representations pretrained on general-domain corpora and adapted to downstream task data have achieved enormous success in building natural language understanding (NLU) systems. While the standard supervised fine-tuning of pretrained language models (PLMs) has proven an effective approach for superior NLU performance, it often necessitates a large quantity of costly human-annotated training data. For example, the enormous success of ChatGPT and GPT-4 can be largely credited to their supervised fine-tuning with massive manually-labeled prompt-response training pairs. Unfortunately, obtaining large-scale human annotations is in general infeasible for most practitioners. To broaden the applicability of PLMs to various tasks and settings, *weakly-supervised learning* offers a promising direction to minimize the annotation requirements for PLM adaptions.

In this tutorial, we cover the recent advancements in pretraining language models and adaptation methods for a wide range of NLU tasks. Our tutorial has a particular focus on *weakly-supervised* approaches that do not require massive human annotations. We will introduce the following topics in this tutorial: (1) pretraining language representation models that serve as the fundamentals for various NLU tasks, (2) extracting entities and hierarchical relations from unlabeled texts, (3) discovering topical structures from massive text corpora for text organization, and (4) understanding documents and sentences with weakly-supervised techniques.

ACM Reference Format:
Yu Meng, Jiaxin Huang, Yu Zhang, Yunyi Zhang, and Jiawei Han. 2023. Pretrained Language Representations for Text Understanding: A Weakly-Supervised Perspective. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3580305.3599569

TUTORS AND PAST TUTORIAL EXPERIENCES

We have five tutors. All are contributors and in-person presenters.

- **Yu Meng**, Ph.D. candidate, Computer Science, UIUC. His research focuses on mining structured knowledge from massive text corpora with minimum human supervision. He received the Google PhD Fellowship (2021) in Structured Data and Database Management. He has delivered tutorials in VLDB'19, KDD'20, KDD'21, AAAI'22, KDD'22 and WWW'23.
- **Jiaxin Huang**, Ph.D. candidate, Computer Science, UIUC. Her research focuses on mining structured knowledge from massive text corpora. She received the Microsoft Research PhD Fellowship (2021) and the Chirag Foundation Graduate Fellowship (2018) in Computer Science, UIUC. She has delivered tutorials in VLDB'19, KDD'20, KDD'21, AAAI'22, KDD'22 and WWW'23.
- **Yu Zhang**, Ph.D. candidate, Computer Science, UIUC. His research focuses on weakly-supervised text mining with structural information. He received the Yunni & Maxine Pao Memorial Fellowship (2022) and WWW Best Poster Award Honorable Mention (2018). He has delivered tutorials in IEEE BigData'19, KDD'21, AAAI'22, KDD'22, EDBT'23 and WWW'23.
- **Yunyi Zhang**, Ph.D. candidate, Computer Science, UIUC. His research focuses on weakly supervised text mining, text classification, and taxonomy construction. He has numerous research publications at KDD, WWW, WSDM, ACL, and EMNLP and has delivered tutorials in EDBT'23.
- **Jiawei Han**, Michael Aiken Chair Professor, Computer Science, UIUC. His research areas encompass data mining, text mining, data warehousing and information network analysis, with over 900 research publications. He is Fellow of ACM, Fellow of IEEE, and received numerous prominent awards, including ACM SIGKDD Innovation Award (2004) and IEEE Computer Society W. Wallace McDowell Award (2009). He delivered 50+ conference tutorials or keynote speeches (*e.g.*, KDD'22 tutorial and CIKM'19 keynote).

TUTORIAL OUTLINE

- **Pretrained Language Models (PLMs)**
  - Categorization of PLMs [2–4, 11, 13, 18, 24]
  - Adaptations of PLMs to Downstream Tasks [5, 12, 23]
- **Text Mining Fundamentals**
  - Phrase Mining [6]
  - Named Entity Recognition [8, 9, 19]
  - Taxonomy Construction [10]
- **Text Representation Enhanced Topic Discovery**
  - Traditional Topic Models [1]
  - Embedding-Based Discriminative Topic Mining [15, 22]

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599569

  - Topic Discovery with PLMs [21, 26, 30, 34]
- **Weakly-Supervised NLU**
  - Text Classification with Weak Supervision [14, 20, 27, 29]
  - Structure-Enhanced Text Classification [25, 31–33]
  - Advanced NLU Tasks [7, 16, 17, 28]

ACKNOWLEDGMENTS

Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. In *J. Mach. Learn. Res.*

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

[3] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[5] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *ACL*.

[6] Xiaotao Gu, Zihan Wang, Zhenyu Bi, Yu Meng, Liyuan Liu, Jiawei Han, and Jingbo Shang. 2021. UCPhrase: Unsupervised Context-aware Quality Phrase Tagging. In *KDD*.

[7] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv* (2022).

[8] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2020. Few-Shot Named Entity Recognition: A Comprehensive Study. *arXiv* (2020).

[9] Jiaxin Huang, Yu Meng, and Jiawei Han. 2022. Few-Shot Fine-Grained Entity Typing with Automatic Label Interpretation and Instance Generation. In *KDD*.

[10] Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring. In *KDD*.

[11] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*.

[12] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *ACL*.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv* (2019).

[14] Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In *ACL*.

[15] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative Topic Mining via Category-Name Guided Text Embedding. In *WWW*.

[16] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating Training Data with Language Models: Towards Zero-Shot Language Understanding. In *NeurIPS*.

[17] Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning Language Models as Training Data Generators for Augmentation-Enhanced Few-Shot Learning. In *ICML*.

[18] Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining. In *NeurIPS*.

[19] Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training. In *EMNLP*.

[20] Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text Classification Using Label Names Only: A Language Model Self-Training Approach. In *EMNLP*.

[21] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations. In *WWW*.

[22] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding. In *KDD*.

[23] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *EMNLP*.

[24] OpenAI. 2023. GPT-4 Technical Report.

[25] Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names. In *NAACL-HLT*.

[26] Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!. In *EMNLP*.

[27] Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-Class: Text Classification with Extremely Weak Supervision. In *NAACL-HLT*.

[28] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In *EMNLP*.

[29] Lu Zhang, Jiandong Ding, Yi Xu, Yingyao Liu, and Shuigeng Zhou. 2021. Weakly-supervised Text Classification Based on Keyword Graph. In *EMNLP*.

[30] Yunyi Zhang, Fang Guo, Jiaming Shen, and Jiawei Han. 2022. Unsupervised Key Event Detection from Massive Text Corpora. In *KDD*.

[31] Yu Zhang, Bowen Jin, Qi Zhu, Yu Meng, and Jiawei Han. 2023. The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study. In *WWW*.

[32] Yu Zhang, Yu Meng, Jiaxin Huang, Frank F Xu, Xuan Wang, and Jiawei Han. 2020. Minimally supervised categorization of text with metadata. In *SIGIR*.

[33] Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. 2022. Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification. In *WWW*.

[34] Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. 2023. Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. In *WSDM*.