Adapting Pretrained Representations for Text Mining

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

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

ABSTRACT

Pretrained text representations, evolving from context-free word embeddings to contextualized language models, have brought text mining into a new era: By pretraining neural models on large-scale text corpora and then adapting them to task-specific data, generic linguistic features and knowledge can be effectively transferred to the target applications and remarkable performance has been achieved on many text mining tasks. Unfortunately, a formidable challenge exists in such a prominent pretrain-finetune paradigm: Large pretrained language models (PLMs) usually require a massive amount of training data for stable fine-tuning on downstream tasks, while human annotations in abundance can be costly to acquire.

In this tutorial, we introduce recent advances in pretrained text representations, as well as their applications to a wide range of text mining tasks. We focus on *minimally-supervised* approaches that do not require massive human annotations, including (1) self-supervised text embeddings and pretrained language models that serve as the fundamentals for downstream tasks, (2) unsupervised and distantly-supervised methods for fundamental text mining applications, (3) unsupervised and seed-guided methods for topic discovery from massive text corpora and (4) weakly-supervised methods for text classification and advanced text mining tasks.

ACM Reference Format:
Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Adapting Pretrained Representations for Text Mining. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3534678.3542607

TARGET AUDIENCE AND PREREQUISITES

Researchers and practitioners in the fields of data mining, text mining, natural language processing, information retrieval, database systems, and machine learning. While the audience with a good background in these areas would benefit most from this tutorial, we believe the material to be presented would give both general audience and newcomers an introductory pointer to the current work and important research topics in this field, and inspire them to learn more. Our tutorial is designed as self-contained, so only preliminary knowledge about basic concepts in data mining, text mining, machine learning, and their applications are needed.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542607

TUTORS AND PAST TUTORIAL EXPERIENCES

We have four 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 and AAAI'22.
- **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 and AAAI'22.
- **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 and AAAI'22.
- **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'21 tutorial and CIKM'19 keynote).

TUTORIAL OUTLINE

- **Pretrained Text Representations**
  - Context-Free Embeddings [14, 22, 23, 28]
  - Contextualized Language Models [2–4, 11, 17, 25]
  - Adaptations of PLMs to Downstream Tasks [5, 7, 24]
- **Text Mining Fundamentals**
  - Phrase Mining [6]
  - Named Entity Recognition [8, 9, 18]
  - Taxonomy Construction [10]
- **Text Representation Enhanced Topic Discovery**
  - Traditional Topic Models [1]
  - Embedding-Based Discriminative Topic Mining [13, 21]
  - Topic Discovery with PLMs [20, 27]
- **Weakly-Supervised Text Classification**

- Flat Text Classification [12, 15, 19, 29]
- Hierarchical Text Classification [16, 26]
- Metadata-Aware Text Classifcation [30–33]

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 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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff 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*.

[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] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *ICML*.

[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* abs/2012.14978 (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] 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 preprint arXiv:1907.11692* (2019).

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

[13] 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*.

[14] Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance M. Kaplan, and Jiawei Han. 2019. Spherical Text Embedding. In *NeurIPS*.

[15] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-Supervised Neural Text Classification. In *CIKM*.

[16] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-Supervised Hierarchical Text Classification. In *AAAI*.

[17] 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*.

[18] 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*.

[19] 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*.

[20] 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*.

[21] 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*.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*.

[23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*.

[24] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language Models as Knowledge Bases?. In *EMNLP*.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. (2019).

[26] 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*.

[27] 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*.

[28] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2019. Poincaré GloVe: Hyperbolic Word Embeddings. In *ICLR*.

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

[30] Yu Zhang, Xiusi Chen, Yu Meng, and Jiawei Han. 2021. Hierarchical Metadata-Aware Document Categorization under Weak Supervision. In *WSDM*.

[31] Yu Zhang, Shweta Garg, Yu Meng, Xiusi Chen, and Jiawei Han. 2022. Motifclass: Weakly supervised text classification with higher-order metadata information. In *WSDM*.

[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*.