On the Power of Pre-Trained Text Representations: Models and Applications in Text Mining

Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han
{yumeng5,jiaxinh3,yuz9,hanj}@illinois.edu
Department of Computer Science, University of Illinois Urbana-Champaign
USA

ABSTRACT

Recent years have witnessed the enormous success of text representation learning in a wide range of text mining tasks. Earlier word embedding learning approaches represent words as fixed low-dimensional vectors to capture their semantics. The word embeddings so learned are used as the input features of task-specific models. Recently, pre-trained language models (PLMs), which learn universal language representations via pre-training Transformer-based neural models on large-scale text corpora, have revolutionized the natural language processing (NLP) field. Such pre-trained representations encode generic linguistic features that can be transferred to almost any text-related applications. PLMs outperform previous task-specific models in many applications as they only need to be fine-tuned on the target corpus instead of being trained from scratch.

In this tutorial, we introduce recent advances in pre-trained text embeddings and language models, as well as their applications to a wide range of text mining tasks. Specifically, we first overview a set of recently developed self-supervised and weakly-supervised text embedding methods and pre-trained language models that serve as the fundamentals for downstream tasks. We then present several new methods based on pre-trained text embeddings and language models for various text mining applications such as topic discovery and text classification. We focus on methods that are weakly-supervised, domain-independent, language-agnostic, effective and scalable for mining and discovering structured knowledge from large-scale text corpora. Finally, we demonstrate with real-world datasets how pre-trained text representations help mitigate the human annotation burden and facilitate automatic, accurate and efficient text analyses[1].

KEYWORDS

Text Embedding, Language Models, Topic Discovery, Text Mining

ACM Reference Format:
Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han. 2021. On the Power of Pre-Trained Text Representations: Models and Applications in Text Mining. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447548.3470810

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.

TUTORS AND PAST TUTORIAL EXPERIENCES

We have four tutors. All are contributors and in-person presenters of the tutorial.

- **Yu Meng**, Ph.D. student, 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 and KDD’20.
- **Jiaxin Huang**, Ph.D. student, 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 and KDD’20.
- **Yu Zhang**, Ph.D. student, Computer Science, UIUC. His research focuses on weakly supervised text mining with structural information. He received WWW’18 Best Poster Award Honorable Mention. He has delivered a tutorial in IEEE BigData’19.
- **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 2020 tutorial and CIKM 2019 keynote).

TUTORIAL OUTLINE

- **Introduction**
  - Overview of Recent Pre-Trained Text Representation Models

[1] Tutorial website can be found at https://yumeng5.github.io/kdd21-tutorial/

KDD ’21, August 14–18, 2021, Virtual Event, Singapore
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3470810

  - Overview of the Applications of Pre-Trained Text Representations in Text Mining
- **Text Embedding and Language Models**
  - Euclidean Context-Free Embeddings [3, 17, 20]
  - Non-Euclidean Context-Free Embeddings [12, 19, 24]
  - Contextualized Language Models [5, 6, 10, 21, 27]
  - Weakly-Supervised Embeddings [11, 16]
- **Topic Discovery with Embeddings**
  - Traditional Topic Models [1, 2, 18]
  - Topic Discovery via Clustering Pre-Trained Embeddings [23]
  - Embedding-Based Discriminative Topic Mining [11, 16]
- **Weakly-Supervised Text Classification**
  - Flat Text Classification [4, 13, 15, 25]
  - Text Classification with Taxonomy Information [14, 22]
  - Text Classification with Metadata Information [29, 30]
- **Other Text Mining Applications Empowered by Pre-Trained Language Models**
  - Phrase/Entity Mining [7]
  - Named Entity Recognition [26]
  - Taxonomy Construction [9]
  - Aspect-Based Sentiment Analysis [8]
  - Text Summarization [28]
- **Summary and Future Directions**

ACKNOWLEDGMENTS

Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and SocialSim Program No. W911NF-17-C-0099, 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. 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, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *NIPS*.

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

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2016), 135–146.

[4] Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of Semantic Representation: Dataless Classification.. In *AAAI*.

[5] 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*.

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

[7] 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*.

[8] Jiaxin Huang, Yu Meng, Fang Guo, Heng Ji, and Jiawei Han. 2020. Weakly-Supervised Aspect-Based Sentiment Analysis via Joint Aspect-Sentiment Topic Embedding. In *EMNLP*.

[9] 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*.

[10] 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).

[11] 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*.

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

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

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

[15] 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*.

[16] 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*.

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

[18] David M. Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with Pachinko allocation. In *ICML*.

[19] Maximilian Nickel and Douwe Kiela. 2017. Poincaré Embeddings for Learning Hierarchical Representations. In *NIPS*.

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

[21] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*.

[22] 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*.

[23] 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*.

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

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

[26] Sam Wiseman and Karl Stratos. 2019. Label-Agnostic Sequence Labeling by Copying Nearest Neighbors. In *ACL*.

[27] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*.

[28] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *ICML*.

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

[30] Yu Zhang, Zhihong Shen, Yuxiao Dong, Kuansan Wang, and Jiawei Han. 2021. MATCH: Metadata-Aware Text Classification in A Large Hierarchy. In *WWW*.