Embedding-Driven Multi-Dimensional Topic Mining and Text Analysis

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

ABSTRACT

People nowadays are immersed in a wealth of text data, ranging from news articles, to social media, academic publications, advertisements, and economic reports. A grand challenge of data mining is to develop effective, scalable and weakly-supervised methods for extracting actionable structures and knowledge from massive text data. Without requiring extensive and corpus-specific human annotations, these methods will satisfy people's diverse applications and needs for comprehending and making good use of large-scale corpora.

In this tutorial, we will introduce recent advances in text embeddings and their applications to a wide range of text mining tasks that facilitate multi-dimensional analysis of massive text corpora. Specifically, we first overview a set of recently developed unsupervised and weakly-supervised text embedding methods including state-of-the-art context-free embeddings and pre-trained language models that serve as the fundamentals for downstream tasks. We then present several embedding-driven text mining techniques that are weakly-supervised, domain-independent, language-agnostic, effective and scalable for mining and discovering structured knowledge, in the form of multi-dimensional topics and multi-faceted taxonomies, from large-scale text corpora. We finally show that the topics and taxonomies so discovered will naturally form a multi-dimensional TextCube structure, which greatly enhances text exploration and analysis for various important applications, including text classification, retrieval and summarization. We will demonstrate on the most recent real-world datasets (including political news articles as well as scientific publications related to the coronavirus) how multi-dimensional analysis of massive text corpora can be conducted with the introduced embedding-driven text mining techniques[1].

KEYWORDS

Text Embedding, Topic Mining, Multi-Faceted Taxonomy, Text Cube, Massive Text Corpora, Multi-Dimensional Analysis

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 three tutors. All are contributors and in-person presenters of the tutorial.

- **Yu Meng**, Ph.D. student, Computer Science, Univ. of Illinois at Urbana-Champaign. His research focuses on mining structured knowledge from massive text corpora with minimum human supervision. He has delivered a tutorial in VLDB'19.
- **Jiaxin Huang**, Ph.D. student, Computer Science, UIUC. Her research focuses on mining structured knowledge from massive text corpora. She is the recipient of Chirag Foundation Graduate Fellowship in Computer Science. She has delivered a tutorial in VLDB'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 800 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.*, SIGKDD 2019 tutorial and CIKM 2019 keynote).

TUTORIAL OUTLINE

One important feature of this tutorial is that we interleave the introduction to principles and methods with system demonstrations to show how the introduced methods work on various kinds of real-world data sets effectively and efficiently. We will introduce the related open-source software packages as well.

The outline of the topics that will be covered in the tutorial is presented as follows.

- **Introduction**
  - Motivation: Why Mining and Structuring Text in a Multi-Dimensional Way?
  - An Overview of Recently Developed Text Embedding Methods
  - An Overview of Multi-Dimensional Text Mining Applications
- **Overview of Text Embedding Methods**
  - Euclidean Context-Free Embeddings [4, 11, 17, 20]
  - Non-Euclidean Context-Free Embeddings [13, 19, 25]
  - Contextualized Language Models [6, 21, 26]

[1]Tutorial website can be found at https://yumeng5.github.io/kdd20-tutorial/

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08.
https://doi.org/10.1145/3394486.3406483

  - Weakly-Supervised Embeddings [12, 16]
- **Multi-Faceted Taxonomy Construction**
  - Coordinated Expansion of Multiple Concepts [7, 22]
  - Hierarchical Concept Expansion and Construction [8, 27]
- **Multi-Dimensional Topic Mining**
  - Unsupervised Topic Modeling [1, 3, 18]
  - Supervised & Seed-Guided Topic Modeling [2, 9]
  - Embedding-Based Discriminative Topic Mining (Demo: TopicMine) [12, 16]
- **Embedding-Driven Multi-Dimensional Text Analysis**
  - Multi-Dimensional TextCube Construction [14, 15, 23, 28]
  - TextCube-Based Online Analytical Processing [24]
  - TextCube-Aware Document Summarization (Demo: TextCube on Hong Kong Demonstration & COVID-19 Open Research Dataset) [5, 10, 29]
- **Summary and Future Directions**

ACKNOWLEDGMENTS

Research was sponsored in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and SocialSim Program No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS 17-41317, and DTRA HDTRA11810026. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation hereon.

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 and Jon D. McAuliffe. 2007. Supervised Topic Models. In *NIPS*.

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

[4] 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.

[5] Xuan Cai and Wenjie Li. 2013. Ranking Through Clustering: An Integrated Approach to Multi-Document Summarization. *IEEE Transactions on Audio, Speech, and Language Processing* 21 (2013), 1424–1433.

[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] Jiaxin Huang, Yiqing Xie, Yu Meng, Jiaming Shen, Yunyi Zhang, and Jiawei Han. 2020. Guiding Corpus-based Set Expansion by Auxiliary Sets Generation and Co-Expansion. In *WWW*.

[8] 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*.

[9] Jagadeesh Jagarlamudi, Hal Daumé, and Raghavendra Udupa. 2012. Incorporating Lexical Priors into Topic Models. In *EACL*.

[10] Chenliang Li, Weiran Xu, Si Li, and Sheng Gao. 2018. Guiding Generation for Abstractive Text Summarization Based on Key Information Guide Network. In *NAACL-HLT*.

[11] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, and Jiawei Han. 2020. Unsupervised Word Embedding Learning by Incorporating Local and Global Contexts. *Frontiers in Big Data* (2020).

[12] 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*.

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

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

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

[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, Zeqiu Wu, Dongming Lei, Jingbo Shang, Xiang Ren, and Jiawei Han. 2017. SetExpan: Corpus-Based Set Expansion via Context Feature Selection and Rank Ensemble. In *ECMLPKDD*.

[23] Fangbo Tao, Chao Zhang, Xiusi Chen, Meng Jiang, Tim Hanratty, Lance M. Kaplan, and Jiawei Han. 2018. Doc2Cube: Allocating Documents to Text Cube Without Labeled Data. *2018 IEEE International Conference on Data Mining (ICDM)* (2018), 1260–1265.

[24] Fangbo Tao, Honglei Zhuang, Chi Wang Yu, Qi Wang, Taylor Cassidy, Lance M. Kaplan, Clare R. Voss, and Jiawei Han. 2016. Multi-Dimensional, Phrase-Based Summarization in Text Cubes. *IEEE Data Eng. Bull.* 39 (2016), 74–84.

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

[26] 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*.

[27] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle T. Vanni, and Jiawei Han. 2018. TaxoGen: Constructing Topical Concept Taxonomy by Adaptive Term Embedding and Clustering. In *KDD*.

[28] Yu Zhang, Yu Meng, Jiaxin Huang, Frank F. Xu, Xuan Wang, and Jiawei Han. 2020. Minimally Supervised Categorization of Text with Metadata. In *SIGIR*.

[29] Yonghui Zhang, Yunqing Xia, Yi Liu, and Wenmin Wang. 2015. Clustering Sentences with Density Peaks for Multi-document Summarization. In *HLT-NAACL*.