Creating a Topic Model of Language Models

Clara Meister, Franz Knobel

December 3, 2021

Student: Franz Knobel (knobelf@student.ethz.ch)
Supervisor: Prof. Ryan Cotterell, Clara Meister
Duration: 6 months

1 Overview

Pre-trained language models serve as the building blocks for many tasks in natural language processing (NLP). They have not only yielded state-of-the-art performance on myriad task, but have also reduced the computational resources often required to train NLP models by providing an advanced starting point for practitioners. Yet, due to the sheer size of these models, the learned probability distribution over natural language strings is difficult to analyze. The support of the distribution alone—the set of all possible strings that can be built using a specific vocabulary—is countably infinite. While a number of techniques have recently been proposed for analyzing the attributes of natural language that these models learn, it is still unclear what portions of the semantic space they learn (or fail) to represent.

Here we propose bringing back a standard model from natural language processing, the topic model, in order to gain a better understanding of this subject. By sampling strings from a pre-trained language model, we generate a pseudo-corpus that should provide an unbiased representation of the information learned by the model. We then learn topic models—using techniques such as LDA—to understand the distribution over topics that the model captures. Using this analysis technique, we hope to gain a better understanding of the variation in downstream performance across pre-trained models. We next propose a method for using these topic models to aid in an important downstream application of pretrained language models: natural language generation.

Recently, pre-trained language models have been applied to a number of text generation tasks, such as Abstractive Summarization or Story Generation. These tasks typically require fine-tuning the model on some task-related dataset—otherwise, sampling (unconditionally) from the language model would generate random text, likely irrelevant text. We propose trying to control text generation using the topic model learned for the pre-trained model: given a chosen topic, we use the distribution over words learned by the LDA model, interpolating it with the distribution from our language model, in order to steer generation towards that topic. Such a method would avoid the used of computational resources required to fine-tune such models, and will hopefully make controlled generation techniques produce more natural text.

2 Goal of the Thesis

In this thesis, the student will employ a combination of techniques from "*Topic Modeling*", "*Pre-trained Language Models*" and "*Topic-guided Text Generation*". The first part of this thesis will consist of learning and analyzing the different topic distribution that pre-trained language models capture. The second part of this thesis will involve analyzing these distributions and the effects they have on various model attributes and performance. Lastly, the resulting topic distributions will be used to try to guide text generation from pre-trained language models towards a specific topic.

3 Tasks

1. Proposal writing + Literature review (incl. short summary of each paper read)
   (a) Proposal writing

(b) Topic modeling [5, 10, 12]

(c) Language modeling

(d) Language generation strategies

(e) Pre-trained language models [7, 4, 8, 2, 1]

(f) Topic-guided text generation [9, 11, 6]

(g) On interpolating between probability [3] distributions

2. Code infrastructure setup

   (a) Familiarize with pre-trained language model libraries in python, including how to generate text from them

   (b) Implement LDA models in python

   (c) Implement optimization algorithms for learning LDA models

   (d) Modify language generation code base to incorporate LDA distributions

3. Experimentation

   (a) Learn LDA models for various pretrained models

   (b) Design and conduct experiments on the effectiveness of proposed decoding strategy

4. Analysis and thesis writing

   (a) Analyze and compare the LDA distributions learned by different models resulting distributions

   (b) Analyze the impact of guiding generation and how different parameters affect the resulting text

   (c) Thesis writing

4 Expected Project Timeline

Start: November 2021
End: April 2022

- November 2021 - November 2021: Literature review
- December 2021 - January 2022: Code infrastructure
- February 2022 - February 2022: Experiment management
- March 2022 - April 2022: Analysis and thesis writing

5 Grading

The Master's Thesis (MT) is a graded semester performance. In order to successfully complete the MT, a grade of 4.0 or higher must be obtained. The supervisor establishes the assessment criteria in a written report, which can include a presentation. In principle, the following evaluation scale is applied:

Grade	Requirements
6.00	Work and results are publishable for international workshops
5.50	Thesis quality significantly exceeds expectations
5.00	Thesis meets expectations
4.50	Thesis partially meets expectations and has minor deficits
4.00	Thesis meets minimum quality requirements; but has major deficits and is clearly below expectations

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models, 2021.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[3] F. H. Bursal. On interpolating between probability distributions. *Applied Mathematics and Computation*, 77(2):213–244, 1996. ISSN 0096-3003. doi: https://doi.org/10.1016/S0096-3003(95)00216-2. URL `https://www.sciencedirect.com/science/article/pii/S0096300395002162`.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, 2018.

[6] J. H. Lau, T. Baldwin, and T. Cohn. Topically driven neural language model, 2017.

[7] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, Sep 2020. ISSN 1869-1900. doi: 10.1007/s11431-020-1647-3. URL `http://dx.doi.org/10.1007/s11431-020-1647-3`.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2018. URL `https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf`.

[9] h. tang, M. Li, and B. Jin. A topic augmented text generation model: Joint learning of semantics and structural features. pages 5093–5102, 01 2019. doi: 10.18653/v1/D19-1513.

[10] W. Wang, Z. Gan, W. Wang, D. Shen, J. Huang, W. Ping, S. Satheesh, and L. Carin. Topic compositional neural language model, 2018.

[11] W. Wang, Z. Gan, H. Xu, R. Zhang, G. Wang, D. Shen, C. Chen, and L. Carin. Topic-guided variational autoencoders for text generation, 2019.

[12] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, and W. Buntine. Topic modelling meets deep neural networks: A survey, 2021.