# Must-Read Papers on Pre-trained Language Models (PLMs)

Contributed by Xiaozhi Wang and Zhengyan Zhang.

Introduction

Pre-trained Languge Model (PLM) has achieved great success in NLP since 2018. In this repo, we list some representative work on PLMs and show their relationship with a diagram. Feel free to distribute or use it! Here you can get the source PPT file of the diagram if you want to use it in your presentation.
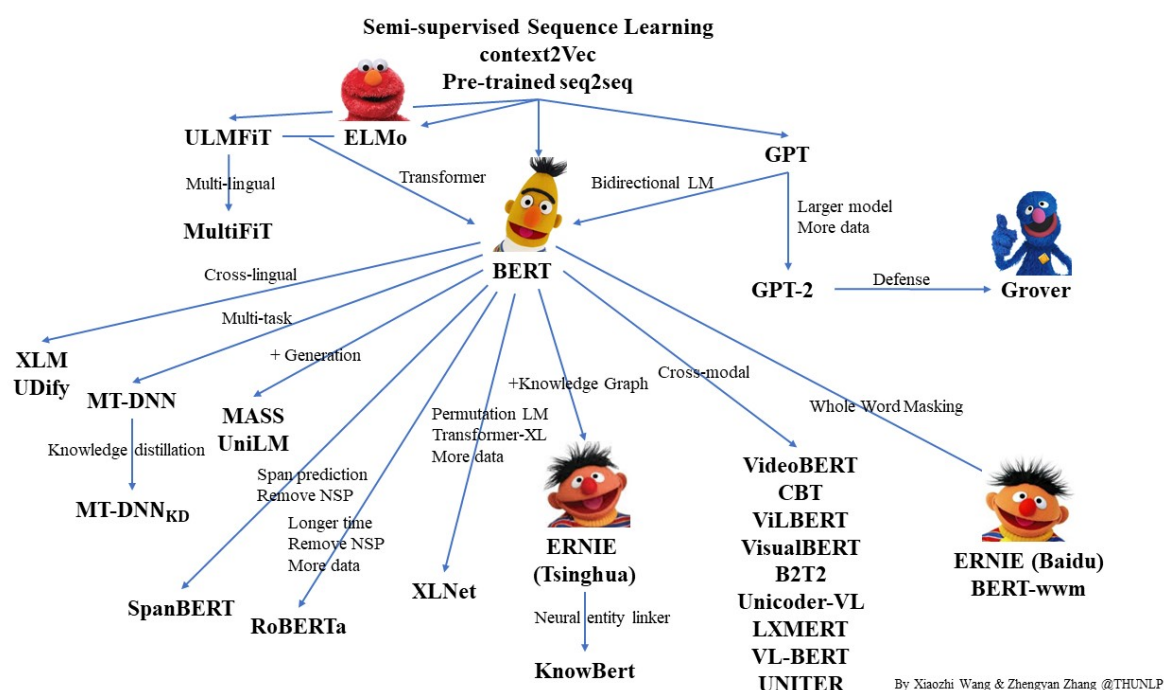
Corrections and suggestions are welcomed.

Open PLMs

We keep training and releasing large-scale PLMs in recent years, which are listed as follows. Welcome to try them.

1. **CPM-2**. Cost-Effective Pre-trained Language Models, 2021. [Model&Code]
2. **CPM-1**. Chinese Pre-trained Language Model, 2020. [Model&Code] [Paper]
3. **OpenCLap**. Open-source Chinese Language Pre-Trained Model Zoo, 2019. [Link]

Survey

Pre-Trained Models: Past, Present and Future. *Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, Jun Zhu*. arXiv:2106.07139 2021. [pdf]

Papers on PLMs Models

1. **Semi-supervised Sequence Learning**. *Andrew M. Dai, Quoc V. Le*. NIPS 2015. [pdf]
2. **context2vec: Learning Generic Context Embedding with Bidirectional LSTM**. *Oren Melamud, Jacob Goldberger, Ido Dagan*. CoNLL 2016. [pdf] [project] (**context2vec**)
3. **Unsupervised Pretraining for Sequence to Sequence Learning**. *Prajit Ramachandran, Peter J. Liu, Quoc V. Le*. EMNLP 2017. [pdf] (**Pre-trained seq2seq**)
4. **Deep contextualized word representations**. *Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer*. NAACL 2018. [pdf] [project] (**ELMo**)
5. **Universal Language Model Fine-tuning for Text Classification**. *Jeremy Howard and Sebastian Ruder*. ACL 2018. [pdf] [project] (**ULMFiT**)
6. **Improving Language Understanding by Generative Pre-Training**. *Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever*. Preprint. [pdf] [project] (**GPT**)
7. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. *Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova*. NAACL 2019. [pdf] [code & model]
8. **Language Models are Unsupervised Multitask Learners**. *Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever*. Preprint. [pdf] [code] (**GPT-2**)
9. **ERNIE: Enhanced Language Representation with Informative Entities**. *Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun and Qun Liu*. ACL 2019. [pdf] [code & model] (**ERNIE (Tsinghua)**)
10. **ERNIE: Enhanced Representation through Knowledge Integration**. *Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian and Hua Wu*. Preprint. [pdf] [code] (**ERNIE (Baidu)**)
11. **Defending Against Neural Fake News**. *Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi*. NeurIPS 2019. [pdf] [project] (**Grover**)
12. **Cross-lingual Language Model Pretraining**. *Guillaume Lample, Alexis Conneau*. NeurIPS 2019. [pdf] [code & model] (**XLM**)
13. **Multi-Task Deep Neural Networks for Natural Language Understanding**. *Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao*. ACL 2019. [pdf] [code & model] (**MT-DNN**)

14. **MASS: Masked Sequence to Sequence Pre-training for Language Generation**. *Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu*. ICML 2019. [pdf] [code & model]

15. **Unified Language Model Pre-training for Natural Language Understanding and Generation**. *Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, Hsiao-Wuen Hon*. Preprint. [pdf] (**UniLM**)

16. **XLNet: Generalized Autoregressive Pretraining for Language Understanding**. *Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le*. NeurIPS 2019. [pdf] [code & model]

17. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. *Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov*. Preprint. [pdf] [code & model]

18. **SpanBERT: Improving Pre-training by Representing and Predicting Spans**. *Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy*. Preprint. [pdf] [code & model]

19. **Knowledge Enhanced Contextual Word Representations**. *Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, Noah A. Smith*. EMNLP 2019. [pdf] (**KnowBert**)

20. **VisualBERT: A Simple and Performant Baseline for Vision and Language**. *Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Kai-Wei Chang*. Preprint. [pdf] [code & model]

21. **ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks**. *Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee*. NeurIPS 2019. [pdf] [code & model]

22. **VideoBERT: A Joint Model for Video and Language Representation Learning**. *Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid*. ICCV 2019. [pdf]

23. **LXMERT: Learning Cross-Modality Encoder Representations from Transformers**. *Hao Tan, Mohit Bansal*. EMNLP 2019. [pdf] [code & model]

24. **VL-BERT: Pre-training of Generic Visual-Linguistic Representations**. *Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai*. Preprint. [pdf]

25. **Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training**. *Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, Ming Zhou*. Preprint. [pdf]

26. **K-BERT: Enabling Language Representation with Knowledge Graph**. *Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, Ping Wang*. Preprint. [pdf]

27. **Fusion of Detected Objects in Text for Visual Question Answering**. *Chris Alberti, Jeffrey Ling, Michael Collins, David Reitter*. EMNLP 2019. [pdf] (**B2T2**)

28. **Contrastive Bidirectional Transformer for Temporal Representation Learning**. *Chen Sun, Fabien Baradel, Kevin Murphy, Cordelia Schmid*. Preprint. [pdf] (**CBT**)

29. **ERNIE 2.0: A Continual Pre-training Framework for Language Understanding**. *Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, Haifeng Wang*. Preprint. [pdf] [code]

30. **75 Languages, 1 Model: Parsing Universal Dependencies Universally**. *Dan Kondratyuk, Milan Straka*. EMNLP 2019. [pdf] [code & model] (**UDify**)

31. **Pre-Training with Whole Word Masking for Chinese BERT**. *Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, Guoping Hu*. Preprint. [pdf] [code & model] (**Chinese-BERT-wwm**)

32. **UNITER: Learning UNiversal Image-TExt Representations**. *Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, Jingjing Liu*. Preprint. [pdf]

33. **MultiFiT: Efficient Multi-lingual Language Model Fine-tuning**. *Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kardas, Sylvain Gugger, Jeremy Howard*. EMNLP 2019. [pdf] [code & model]

34. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. *Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu*. Preprint. [pdf] [code & model] (**T5**)

35. **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**. *Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer*. ACL 2020. [pdf]

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

37. **A Mutual Information Maximization Perspective of Language Representation Learning**. *Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, Dani Yogatama*. ICLR 2020. [pdf]

38. **StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding**. *Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, Luo Si*. ICLR 2020. [pdf]

39. **Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scorings**. *Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, Jason Weston*. ICLR 2020. [pdf]

40. **FreeLB: Enhanced Adversarial Training for Language Understanding**. *Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, Jingjing Liu*. ICLR 2020. [pdf]

41. **Multilingual Alignment of Contextual Word Representations**. *Steven Cao, Nikita Kitaev, Dan Klein*. ICLR 2020. [pdf]

42. **TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data**. *Pengcheng Yin, Graham Neubig, Wen-tau Yih, Sebastian Riedel*. ACL 2020. [pdf] [code]

43. **BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance**. *Timo Schick, Hinrich Schutze*. ACL 2020. [pdf]

44. **TAPAS: Weakly Supervised Table Parsing via Pre-training**. *Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, Julian Martin Eisenschlos*. ACL 2020. [pdf]

45. **On the Sentence Embeddings from Pre-trained Language Models**. *Bohan Li, Hao Zhou, Junx-*

ian He, Mingxuan Wang, Yiming Yang, Lei Li. EMNLP 2020. [pdf]

46. **An Empirical Investigation Towards Efficient Multi-Domain Language Model Pre-training**. *Kristjan Arumae, Qing Sun, Parminder Bhatia*. EMNLP 2020. [pdf]

47. **Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information**. *Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, Lei Li*. EMNLP 2020. [pdf]

48. **Pre-Training Transformers as Energy-Based Cloze Models**. *Kevin Clark, Minh-Thang Luong, Quoc Le, Christopher D. Manning*. EMNLP 2020. [pdf]

49. **PatchBERT: Just-in-Time, Out-of-Vocabulary Patching**. *Sangwhan Moon, Naoaki Okazaki*. EMNLP 2020. [pdf]

50. **Pre-training via Paraphrasing**. *Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, Luke Zettlemoyer*. NeurIPS 2020. [pdf]

51. **ConvBERT: Improving BERT with Span-based Dynamic Convolution**. *Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, Shuicheng Yan*. NeurIPS 2020. [pdf]

Papers on Model Compression & Acceleration

1. **TinyBERT: Distilling BERT for Natural Language Understanding**. *Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, Qun Liu*. Preprint. [pdf] [code & model]

2. **Distilling Task-Specific Knowledge from BERT into Simple Neural Networks**. *Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, Jimmy Lin*. Preprint. [pdf]

3. **Patient Knowledge Distillation for BERT Model Compression**. *Siqi Sun, Yu Cheng, Zhe Gan, Jingjing Liu*. EMNLP 2019. [pdf] [code]

4. **Model Compression with Multi-Task Knowledge Distillation for Web-scale Question Answering System**. *Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, Daxin Jiang*. Preprint. [pdf]

5. **PANLP at MEDIQA 2019: Pre-trained Language Models, Transfer Learning and Knowledge Distillation**. *Wei Zhu, Xiaofeng Zhou, Keqiang Wang, Xun Luo, Xiepeng Li, Yuan Ni, Guotong Xie*. The 18th BioNLP workshop. [pdf]

6. **Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding**. *Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao*. Preprint. [pdf] [code & model]

7. **Well-Read Students Learn Better: The Impact of Student Initialization on Knowledge Distillation**. *Iulia Turc, Ming-Wei Chang, Kenton Lee, Kristina Toutanova*. Preprint. [pdf]

8. **Small and Practical BERT Models for Sequence Labeling**. *Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, Amelia Archer*. EMNLP 2019. [pdf]

9. **Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT**. *Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, Kurt Keutzer*. Preprint. [pdf]

10. **ALBERT: A Lite BERT for Self-supervised Learning of Language Representations**. *Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut*. ICLR 2020. [pdf]

11. **Extreme Language Model Compression with Optimal Subwords and Shared Projections**. *Sanqiang Zhao, Raghav Gupta, Yang Song, Denny Zhou*. Preprint. [pdf]

12. **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter**. *Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf*. Preprint. [pdf]

13. **Reducing Transformer Depth on Demand with Structured Dropout**. *Angela Fan, Edouard Grave, Armand Joulin*. ICLR 2020. [pdf]

14. **Thieves on Sesame Street! Model Extraction of BERT-based APIs**. *Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, Mohit Iyyer*. ICLR 2020. [pdf]

15. **DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference**. *Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, Jimmy Lin*. ACL 2020. [pdf]

16. **Contrastive Distillation on Intermediate Representations for Language Model Compression**. *Siqi Sun, Zhe Gan, Yuwei Fang, Yu Cheng, Shuohang Wang, Jingjing Liu*. EMNLP 2020. [pdf]

17. **BERT-of-Theseus: Compressing BERT by Progressive Module Replacing**. *Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, Ming Zhou*. EMNLP 2020. [pdf]

18. **TernaryBERT: Distillation-aware Ultra-low Bit BERT**. *Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, Qun Liu*. EMNLP 2020. [pdf]

19. **When BERT Plays the Lottery, All Tickets Are Winning**. *Sai Prasanna, Anna Rogers, Anna Rumshisky*. EMNLP 2020. [pdf]

20. **Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing**. *Zihang Dai, Guokun Lai, Yiming Yang, Quoc Le*. NeurIPS 2020. [pdf]

21. **DynaBERT: Dynamic BERT with Adaptive Width and Depth**. *Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, Qun Liu*. NeurIPS 2020. [pdf]

22. **BERT Loses Patience: Fast and Robust Inference with Early Exit**. *Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, Furu Wei*. NeurIPS 2020. [pdf]

Papers on Model Analysis

1. **Revealing the Dark Secrets of BERT**. *Olga Kovaleva, Alexey Romanov, Anna Rogers, Anna Rumshisky*. EMNLP 2019. [pdf]

2. **How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations**. *Betty van Aken, Benjamin Winter, Alexander Löser, Felix A. Gers*. CIKM 2019. [pdf]

3. **Are Sixteen Heads Really Better than One?**. *Paul Michel, Omer Levy, Graham Neubig*. Preprint. [pdf] [code]

4. **Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment**. *Di Jin, Zhijing Jin, Joey Tianyi Zhou, Peter Szolovits*. Preprint. [pdf] [code]

5. **BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model**. *Alex Wang, Kyunghyun Cho*. NeuralGen 2019. [pdf] [code]

6. **Linguistic Knowledge and Transferability of Contextual Representations**. *Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, Noah A. Smith*. NAACL 2019. [pdf]

7. **What Does BERT Look At? An Analysis of BERT's Attention**. *Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher D. Manning*. BlackBoxNLP 2019. [pdf] [code]

8. **Open Sesame: Getting Inside BERT's Linguistic Knowledge**. *Yongjie Lin, Yi Chern Tan, Robert Frank*. BlackBoxNLP 2019. [pdf] [code]

9. **Analyzing the Structure of Attention in a Transformer Language Model**. *Jesse Vig, Yonatan Belinkov*. BlackBoxNLP 2019. [pdf]

10. **Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains**. *Samira Abnar, Lisa Beinborn, Rochelle Choenni, Willem Zuidema*. BlackBoxNLP 2019. [pdf]

11. **BERT Rediscovers the Classical NLP Pipeline**. *Ian Tenney, Dipanjan Das, Ellie Pavlick*. ACL 2019. [pdf]

12. **How multilingual is Multilingual BERT?**. *Telmo Pires, Eva Schlinger, Dan Garrette*. ACL 2019. [pdf]

13. **What Does BERT Learn about the Structure of Language?**. *Ganesh Jawahar, Benoît Sagot, Djamé Seddah*. ACL 2019. [pdf]

14. **Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT**. *Shijie Wu, Mark Dredze*. EMNLP 2019. [pdf]

15. **How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings**. *Kawin Ethayarajh*. EMNLP 2019. [pdf]

16. **Probing Neural Network Comprehension of Natural Language Arguments**. *Timothy Niven, Hung-Yu Kao*. ACL 2019. [pdf] [code]

17. **Universal Adversarial Triggers for Attacking and Analyzing NLP**. *Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, Sameer Singh*. EMNLP 2019. [pdf] [code]

18. **The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives**. *Elena Voita, Rico Sennrich, Ivan Titov*. EMNLP 2019. [pdf]

19. **Do NLP Models Know Numbers? Probing Numeracy in Embeddings**. *Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, Matt Gardner*. EMNLP 2019. [pdf]

20. **Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs**. *Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretič, Samuel R. Bowman*. EMNLP 2019. [pdf] [code]

21. **Visualizing and Understanding the Effectiveness of BERT**. *Yaru Hao, Li Dong, Furu Wei, Ke Xu*. EMNLP 2019. [pdf]

22. **Visualizing and Measuring the Geometry of BERT**. *Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, Martin Wattenberg*. NeurIPS 2019. [pdf]

23. **On the Validity of Self-Attention as Explanation in Transformer Models**. *Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Roger Wattenhofer*. Preprint. [pdf]

24. **Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel**. *Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, Ruslan Salakhutdinov*. EMNLP 2019. [pdf]

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

26. **To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks**. *Matthew E. Peters, Sebastian Ruder, Noah A. Smith*. RepL4NLP 2019, [pdf]

27. **On the Cross-lingual Transferability of Monolingual Representations**. *Mikel Artetxe, Sebastian Ruder, Dani Yogatama*. Preprint, [pdf] [dataset]

28. **A Structural Probe for Finding Syntax in Word Representations**. *John Hewitt, Christopher D. Manning*. NAACL 2019. [pdf]

29. **Assessing BERT's Syntactic Abilities**. *Yoav Goldberg*. Technical Report. [pdf]

30. **What do you learn from context? Probing for sentence structure in contextualized word representations**. *Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick*. ICLR 2019. [pdf]

31. **Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling**. *Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, Samuel R. Bowman*. ACL 2019. [pdf]

32. **BERT is Not an Interlingua and the Bias of Tokenization**. *Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong*. DeepLo 2019. [pdf] [dataset]

33. **What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models**. *Allyson Ettinger*. Preprint. [pdf] [code]

34. **How Language-Neutral is Multilingual BERT?**. *Jindřich Libovický, Rudolf Rosa, and Alexander Fraser*. Preprint. [pdf]

35. **Cross-Lingual Ability of Multilingual BERT: An Empirical Study**. *Karthikeyan K, Zihan Wang, Stephen Mayhew, Dan Roth*. ICLR 2020. [pdf]

36. **Finding Universal Grammatical Relations in Multilingual BERT**. *Ethan A. Chi, John Hewitt, Christopher D. Manning*. ACL 2020. [pdf]

37. **Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly**. *Nora Kassner, Hinrich Schütze*. ACL 2020. [pdf]

38. **Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT**. *Zhiyong Wu, Yun Chen, Ben Kao, Qun Liu*. ACL 2020. [pdf]

39. **Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models**. *Bill Yuchen Lin, Seyeon Lee, Rahul Khanna and Xiang Ren*. EMNLP 2020. [pdf]

40. **Identifying Elements Essential for BERT's Multilinguality**. *Philipp Dufter, Hinrich Schütze*. EMNLP 2020. [pdf]

41. **AUTOPROMPT: Eliciting Knowledge from Language Models with Automatically Generated Prompts**. *Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, Sameer Singh*. EMNLP 2020. [pdf]

42. **The Lottery Ticket Hypothesis for Pre-trained BERT Networks**. *Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, Michael Carbin*. NeurIPS 2020. [pdf]

Papers on Finetuning or Adaptation

1. **SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization**. *Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, Tuo Zhao*. ACL 2020. [pdf]

2. **Do you have the right scissors? Tailoring Pre-trained Language Models via Monte-Carlo Methods**. *Ning Miao, Yuxuan Song, Hao Zhou, Lei Li*. ACL 2020. [pdf]

3. **ExpBERT: Representation Engineering with Natural Language Explanations**. *Shikhar Murty, Pang Wei Koh, Percy Liang*. ACL 2020. [pdf]

4. **Don't Stop Pretraining: Adapt Language Models to Domains and Tasks**. *Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith*. ACL 2020. [pdf]

5. **Recall and Learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting**. *Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, Xiangzhan Yu*. EMNLP 2020. [pdf]

6. **Masking as an Efficient Alternative to Finetuning for Pretrained Language Models**. *Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, Hinrich Schütze*. EMNLP 2020. [pdf]

7. **CogLTX: Applying BERT to Long Texts**. *Ming Ding, Chang Zhou, Hongxia Yang, Jie Tang*. NeurIPS 2020. [pdf]

Papers on Prompt-based Tuning

Here is our new paper list on prompt-based tuning for pre-trained language models. [repo]

Tutorial & Resource

1. **Transfer Learning in Natural Language Processing**. *Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, Thomas Wolf*. NAACL 2019. [slides]
2. **Transformers: State-of-the-art Natural Language Processing**. *Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Jamie Brew*. EMNLP 2020. [pdf] [code]