Title of Your Ph.D. Thesis

Your Name

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

in

Systems Engineering and Engineering Management

The Chinese University of Hong Kong

January 2021

Thesis Assessment Committee

Professor AAA (Chair)

Professor BBB (Thesis Supervisor)

Professor CCC (Committee Member)

Professor DDD (External Examiner)

Abstract

摘要

Contents

List of Tables

List of Figures

Acknowledgments

Acknowledgement here.

Chapter 1

Introduction

Introduction, motivation and big picture of your phd thesis

1.1 Contributions

The contributions of this dissertation are summarized as follows:

- **Description of your first paper**

- **Description of your second paper**

- ...

1.2 Produced Publications

The research work of this dissertation has produced some direct and indirect publications as listed below:

- ...

1.3 Thesis Outline

Chapter 2

Literature Review

2.1 Unsupervised Language Model Pre-training

The advent of unsupervised Language Model (LM) pre-training has led to significant performance gains on a variety of language understanding (Radford et al., 2018; Devlin et al., 2019; Yang et al., 2019b; Clark et al., 2020) and language generation (Radford et al., 2019; Dong et al., 2019) tasks. Typically, a deep Long Shot-Term Memory networks (LSTM) (Hochreiter and Schmidhuber, 1997; Peters et al., 2018) or Transformer (Vaswani et al., 2017) is first pre-trained on large-scale corpus and then the contextualized embeddings from the pre-trained language model are provided for the downstream tasks in the manner of feature extraction or fine-tuning.

Several pre-training techniques have been developed for generating general-purpose contextualized embeddings. Peters et al. (2017, 2018) employ two unidirectional LSTM LMs, namely, a forward left-to-right LM and a backward right-to-left LM, to encode bi-directional contexts and the pre-training is conducted via auto-regressive language modeling, as shown in the left part of Figure 2.1. GPT (Radford et al., 2018) and GPT-2 (Radford et al., 2019) adopt the same auto-regressive pre-training objective but change to model sequential word flow with unidirectional Transformer. BERT (Devlin et al., 2019), as well as its variants (Conneau and Lample, 2019; Liu et al., 2019b; Wang et al., 2019; Joshi et al., 2020), propose to learn

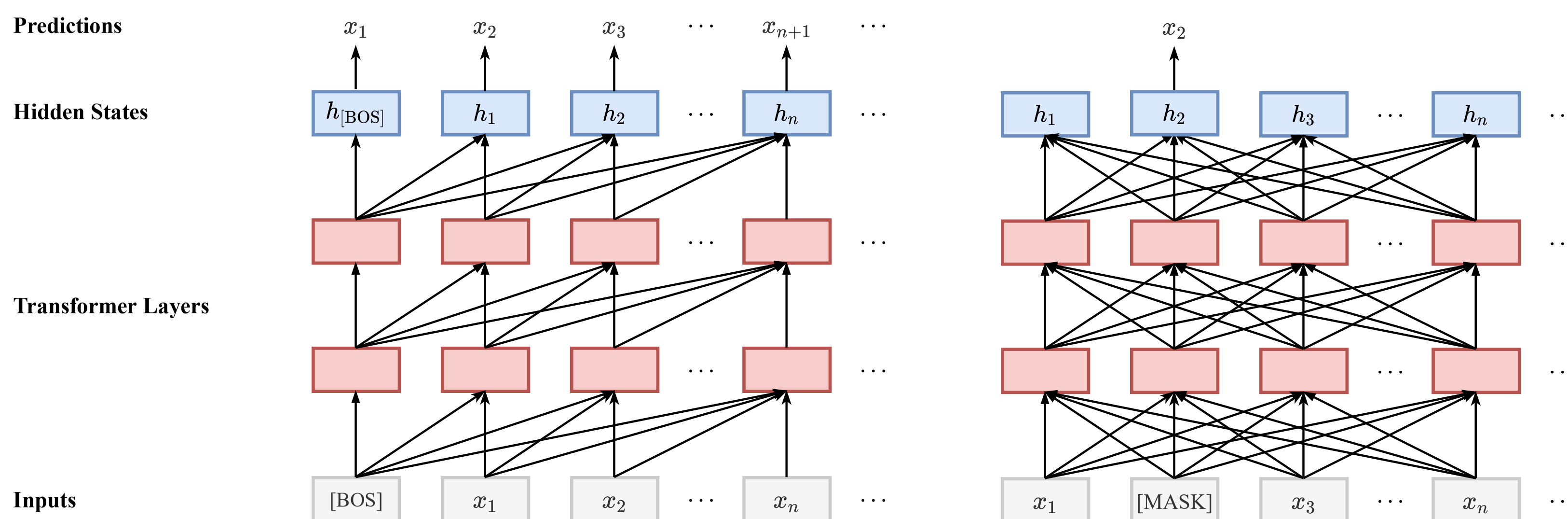

Figure 2.1: *Auto-regressive language modeling (left) and masked language modeling (right).*

contextualized embeddings based on masked language modeling—reconstruct the masked input with the special [MASK] token and the surrounding context (see the right part of Figure 2.1). In order to unify the auto-regressive pre-training and auto-encoding based pre-training, UniLM (Dong et al., 2019) jointly optimizes the pre-training of Transformer with auto-regressive language modeling objective and masked language modeling objective. Similarly, XLNet (Yang et al., 2019b) designs permutation language modeling objective to capture the context information from all positions while preserving the temporal dependency among the predictions. MASS (Song et al., 2019), BART (Lewis et al., 2020) and T5 (Raffel et al., 2019) are the initial attempts to pre-train sequence-to-sequence architecture and achieve promising results on machine translation and abstractive summarization.

2.2 Multilingual Language Model Pre-Training

Multilingual language model pre-training is a multilingual extension of language model pre-training, where the deep neural architecture (Vaswani et al., 2017; Peters et al., 2018) is pre-trained on large-scale multilingual corpus, e.g., a collection of wikipedia documents in different languages. The yielding multilingual pre-trained language models (mPTLMs) (Che et al., 2018; Devlin et al., 2019; Conneau and Lample, 2019; Mulcaire et al., 2019; Conneau et al., 2020) have been regarded as the gold standard for a variety of unsupervised cross-lingual natural language understanding tasks (Prettenhofer and Stein, 2010; Schwenk and Li, 2018; Zeman et al., 2018; Liu et al., 2019a). The most impressive features of mPTLMs is that even performing cross-lingual transfer in a zero-shot manner—only fine-tune the mPTLMs on the source-language training data—it can still significantly outperform the existing works based on cross-lingual embeddings (Mikolov et al., 2013; Faruqui and Dyer, 2014; Smith et al., 2018) or machine translation (Banea et al., 2008; Duh et al., 2011), as observed in Pires et al. (2019); Wu and Dredze (2019); Keung et al. (2019); Artetxe et al. (2020). After being additionally trained on machine-translated data from source language, the cross-lingual performances of the cross-lingual models exploiting mPTLMs are further improved (Huang et al., 2019; Conneau et al., 2020; Cao et al., 2020), especially on sequence classification (Schwenk and Li, 2018) and sequence pair classification tasks (Conneau et al., 2018; Yang et al., 2019a).

Chapter 3

Chapter 3

Chapter 4

Chapter 4

Chapter 5

Chapter 5

Chapter 6

...

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of ACL*, pages 4623–4637.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP*, pages 127–135.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proceedings of ICLR*.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of CoNLL*, pages 55–64.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, pages 7059–7069.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Proceedings of NeurIPS*, pages 13063–13075.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of ACL*, pages 429–433.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*, pages 462–471.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of EMNLP*, pages 2485–2494.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of EMNLP*, pages 1355–1360.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*, pages 7871–7880.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019a. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of ACL*, pages 2358–2368.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *Proceedings of NAACL-HLT*, pages 3912–3918.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of ACL*, pages 1756–1765.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*, pages 4996–5001.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of ACL*, pages 1118–1127.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of LREC*.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2018. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *Proceedings of ICML*, pages 5926–5936.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP*, pages 833–844.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. PAWS-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP*, pages 3687–3692.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of NeurIPS*, pages 5753–5763.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of CoNLL*, pages 1–21.