# When Cantonese NLP Meets Pre-training: Progress and Challenges

**Rong Xiang**[1] **Hanzhuo Tan**[1] **Jing Li**[1] **Mingyu Wan**[2] **Kam-Fai Wong**[3]
[1] Department of Computing, Hong Kong Polytechnic University, HKSAR, China
[2] School of Continuing Education, Hong Kong Baptist University, HKSAR, China
[3] Department of SEEM, The Chinese University of Hong Kong, HKSAR, China
[1]{rong.xiang, hanzhuo.tan}@connect.polyu.hk
[1]jing-amelia.li@polyu.edu.hk [2]claramywan629@hkbu.edu.hk
[3]kfwong@se.cuhk.edu.hk

## Abstract

Cantonese is an influential Chinese variant with a large population of speakers worldwide. However, it is under-resourced in terms of the data scale and diversity, excluding Cantonese Natural Language Processing (NLP) from the state-of-the-art (SOTA) "pre-training and fine-tuning" paradigm. This tutorial will start with a substantially review of the linguistics and NLP progress for shaping language specificity, resources, and methodologies. It will be followed by an introduction to the trendy transformer-based pre-training methods, which have been largely advancing the SOTA performance of a wide range of downstream NLP tasks in numerous majority languages (e.g., English and Chinese). Based on the above, we will present the main challenges for Cantonese NLP in relation to Cantonese language idiosyncrasies of *colloquialism* and *multilingualism*, followed by the future directions to line NLP for Cantonese and other low-resource languages up to the cutting-edge pre-training practice.

## 1 Tutorial Description

In our tutorial, there will be five parts (shown in Figure 1), each presented by a tutorial presenter. The first part will be the overview of Cantonese NLP research (PART I) and the second exhibits the progress in language specify, resources, and methodologies (PART II). Then, we will introduce the roles played by language pre-training in the SOTA NLP practice (PART III), based on which we further discuss the challenges to benefit Cantonese NLP from the trendy "pre-training and fine-tuning" fashion (PART IV). Lastly, the potential solutions will be pointed out to shed light on the promising future direction of Cantonese NLP (PART V).

The detailed content is shown in the following.

*PART I: Cantonese NLP Overview (in 30 min).* At the beginning, we will briefly introduce Cantonese language and its related research in NLP.

Cantonese is a language from the Chinese family with over 73 million speakers in the world (García and Fishman, 2011; Yu, 2013). It is mostly used in colloquial scenarios (e.g., daily conversation and social media) and exhibits different vocabulary, grammar, and pronunciation compared to standard Chinese (SCN)[1], which is mainly designed for formal writing (Wong and Lee, 2018).

Despite the substantial efforts in Chinese Natural Language Processing (NLP), most previous studies center around SCN, where limited work attempts to explore how to process Cantonese with the cutting-edge NLP techniques (Xiang et al., 2019; Lee et al., 2021). Modern NLP paradigms have been deeply revolutionized by large-scale pre-training models, e.g., BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), which have achieved SOTA performance on many NLP tasks via fine-tuning.

Although the general NLP field is thriving, Cantonese NLP has drawn limited attention so far, as demonstrated by the recent publications in the ACL Anthology — only 47 papers are related to "Cantonese", compared to 7,018 papers for English, 2,355 for common Chinese, and 323 for Mandarin.

This tutorial will present a roadmap lining Cantonese NLP up to the SOTA practice based on pre-training. We will start with the previous progress made by linguistics and NLP researchers, followed by the major challenges caused by the language specificity, and end with the promising future directions to allow Cantonese and other low resource languages to benefit from the advanced NLP techniques. The details will be covered in PART II-V.

*PART II: Progress in Language Specificity, Resources, and Methodologies (in 40 min).* Cantonese (or Yue) is the second most popular dialect among all Chinese variants (Matthews and Yip, 2011). For-

[1] Standard Chinese is known as Standard Northern Mandarin, which is emerged as the lingua franca among the speakers of various Mandarin and other varieties of Chinese (Hokkien, Cantonese, and beyond).

*Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 16–21
November 20-21, 2022. ©2022 Association for Computational Linguistics

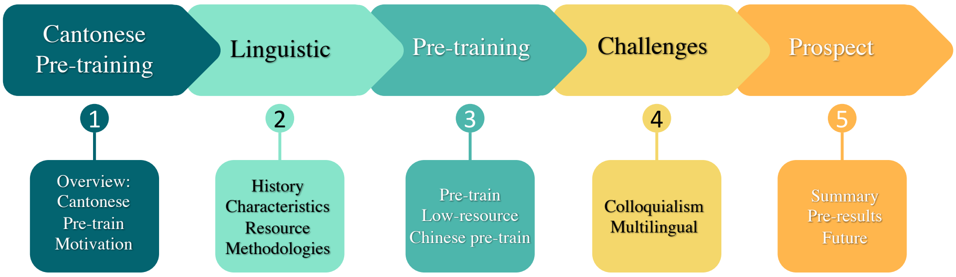
Cantonese Pre-training
Linguistic
Pre-training
Challenges
Prospect
1
2
3
4
5
Overview: Cantonese Pre-train Motivation
History Characteristics Resource Methodologies
Pre-train Low-resource Chinese pre-train
Colloquialism Multilingual
Summary Pre-results Future

Figure 1: Outline for the tutorial

mal writing exhibits no essential difference among the SCN in various regions (Kataoka and Lee, 2008; Yu, 2013), whereas in informal situations, colloquial Cantonese diverges substantially from SCN in phonology, orthography, lexicon, and grammar.

In dealing with Cantonese data, *Colloquialism* and *Multilingualism* are two fundamental challenges. Unlike standard Chinese, which character standardization can be dated back to the Qin dynasty (221 BC), Cantonese is mainly derived from pronunciation and may contain many colloquial features, such as non-standard spelling, local slang, neologisms. On the other hand, Cantonese language historically evolved in multi-lingual environments and this is especially true for HK Cantonese, as shown by the large inventory of English loanwords borrowed through phonetic transliteration.

Unlike SCN, which benefits from abundant well-annotated textual resources, there is a chronic lack of digital resources for Cantonese data. The existing resources are summarized into three categories: *Corpora*, *Benchmarks*, and *Expert Resources*. In general, using existing Cantonese resources may be difficult for three reasons: (1) the data scale is relatively small (especially compared to SCN); (2) the domain is usually specific and lacks diversity and generality; (3) many resources mix Cantonese and SCN which might confuse NLP models and hinder them from mastering 'authentic' Cantonese.

Cantonese specific NLP methods are relatively less explored. The following presents a detailed review of Cantonese NLP methods in (1) Natural Language Understanding and (2) Natural Language Generation. Several language understanding tasks will be introduced, including word segmentation, spell checking, rumor detection, sentiment analysis, and dialogue slot filling. As for language generation, we will summarize previous studies on dialogue summarization, machine translation, etc.

*PART III: Pre-training in SOTA NLP (in 40 min).* The cutting-edge NLP takes advantages of the promising results achieved by the pre-training of language representations. A typical pre-trained and fine-tune scheme refers to pre-train a large model on massive unlabelled corpora by self-supervised objectives, and fine-tune the model on downstream tasks with task-specific loss. Such self-supervised objectives, e.g. Masked Langauge Modeling (MLM) and Next Sentene Prediction (NSP) (Devlin et al., 2019) enable the model to gain generalized language representations without human supervision. During fine-tuning stage, the pre-trained representations can be further used to learn a specific Natural Language Understanding (NLU) task with small-scale annotations via incremental training.

Transformer (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020) is the most widely employed pre-trained architecture in NLP. The transformer encoder consumes the input text and project it into high-dimension vectors, which are feed into the transformer decoder to generate the output sequences. Inspired by the transformer architecture, researchers explore the transformer encoder for NLU tasks and transformer decoder for Natural Language Generation (NLG) tasks.

Since the transformer-based pre-training (Devlin et al., 2019; Liu et al., 2019) was introduced to the world, championing the leaderboards of many NLP benchmarks, the "pre-training and fine-tuning" paradigm has profoundly revolutionized the way we research NLP for most of the majority languages, such as English and Chinese. Nevertheless, the success of language pre-training is built upon the availability of rich language resources and large-scale textual corpora, hindering Cantonese and other low-resource languages from gaining the benefit of pre-training. The following presents the challenging low-resource issue in Cantonese.

*PART IV: Challenges from Colloquialism and*

*Multilingualism (40 min).* In Cantonese NLP, *Colloquialism* and *Multilingualism* jointly present the low-resource challenges. Cantonese is by nature colloquial, where using casual language would sparsify the context and require more data for contextual model training, making the low-resource issue serious. Like other low-resource language processing (Li et al., 2018), it is possible to gather large-scale data from social media. However, models might also compromise their performance using just noisy social media data. It not only requires effective data cleaning and augmentation, but also non-trivial model capabilities of capturing salient information in context with diverse quality.

Cantonese exhibits a code-switching convention with multiple languages. Substantially richer context is hence required for NLP models to gain the multilingual understanding capabilities, despite the limitation to make it happen in the low-resource scenarios. Although it is possible to transfer the knowledge gained in a similar language with rich resources (henceforth *cross-lingual learning*) (Friedrich and Gateva, 2017; Khalil et al., 2019; Zhang et al., 2019), Cantonese, as a vibrant language, absorbs the knowledge from numerous languages beyond SCN and English.

*PART V: Future NLP Directions for Cantonese and other Low-Resource Languages (30 min).* Data scarcity and limited methodology exploration are top issues for Cantonese in benefiting deep semantics and general NLP tasks. To mitigate low-resource problems, data augmentation is an alternative to scale up the Cantonese dataset for NLP model training. For example, we might employ heuristic rules (Ratner et al., 2017; Lison et al., 2020), machine learning (Şahin and Steedman, 2018) and information retrieval (Riedel et al., 2010; Hedderich et al., 2021) to automatically boost the data scales. In the augmentation process, we might need to learn how to distinguish SCN from Cantonese. Though both are encoded in the Chinese language system, the former dominates the Chinese resources while the latter is a minority (Wu and Dredze, 2020; Cui et al., 2021).

Cross-lingual learning might provide another be a promising alternative for the pre-training in low resource, which borrows knowledge from other languages (Wisniewski et al., 2014; Zhang et al., 2019; Khalil et al., 2019). We may take advantage of SOTA pre-trained transformers to capture the general and specific language features for transfer learning (Devlin et al., 2019; Clark et al., 2020). In addition, based on Cantonese's phonological history, future work may consider injecting phonetic knowledge into language learning or developing multi-modal understanding across text and speech.

## 2 Type of the Tutorial

This is an *introductory* tutorial of *Cantonese NLP*, where we draw NLP community's attention to look at the research of Cantonese — a language with over 73 million speakers in the world (García and Fishman, 2011; Yu, 2013) while only has 47 papers in ACL Anthology related to it. The tutorial will present a roadmap going through the essential issues regarding language specificity, data scarcity, research progress, and major challenges for Cantonese NLP to be benefited from the cutting-edge NLP paradigms based on language pre-training.

## 3 Target Audience

Our tutorial is designed for the attendees of premier computational linguistics conferences, who preferably have interests and working experience in the processing of Asian languages and low-resource languages. The audiences would better have the following prerequisites.

- **Language Representation Learning**. Familiar with the basic concepts and main ideas of language pre-training, e.g., word embeddings (Mikolov et al., 2013), BERT (Devlin et al., 2019), and how the learned representations are employed to train various NLP tasks.

- **Linguistics**. Have the basic knowledge of the fundamental linguistic concepts (Jurafsky, 2000), e.g., *semantics*, *syntax*, *lexicography*, *morphology*, *phonetics*, etc.

- **Machine Learning.** Understand the traditional machine learning paradigm using handcrafted features (Svensén and Bishop, 2007) and the trendy deep learning-based methods (Goodfellow et al., 2016) allowing automatic feature learning in neural architectures.

## 4 Tutorial Outline (3 hours)

- PART I: Cantonese NLP overview (30 min).
  - Background of Cantonese.
  - Brief review of Cantonese NLP.
  - Brief introduction of language pre-training.
  - Problem definition and motivation.

- The outline of tutorial.

- PART II: Progress in language specificity, resources, and methodologies (40 min).
  - Brief history of Cantonese.
  - Linguistic characteristics of Cantonese.
  - Summary of Cantonese NLP resources.
  - Summary of Cantonese NLP methodologies.
- PART III: Pre-training in SOTA NLP (40 min).
  - Language pre-training methods.
  - Pre-training in low resource.
  - Chinese pre-training.
- PART IV: Challenges from colloquialism and multilingualism (40 min).
  - How colloquialism challenges pre-training.
  - How multilingualism challenges pre-training.
- PART V: Future NLP directions for Cantonese and other low-resource languages (30 min).
  - Summary of the tutorial.
  - Future work for Cantonese NLP and beyond.

## 5 Reading list

For trainees interested in reading important studies before the tutorial, we recommend the following: Ouyang (1993); Snow (2004); Sachs and Li (2007). Vaswani et al. (2017); Devlin et al. (2019); Liu et al. (2019); Brown et al. (2020); Sun et al. (2019); Liu et al. (2019); Nguyen et al. (2020).

## 6 Tutorial Presenters

Our tutorial will contain 5 parts and here we introduce the presenter for each of them.

• **Kam-Fai Wong** (*PART I*). Kam-Fai Wong a full professor in the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (CUHK). His research interest focuses on Chinese natural language processing and database. He is the fellow of ACL and has published over 260 technical papers in different international journals, conferences, and books. Also, He was the founding Editor-In-Chief of ACM Transactions on Asian Language Processing (TALIP) and the president of Asian Federation of Natural Language Processing (AFNLP).

• **Mingyu Wan** (*PART II*). Mingyu Wan is a postdoctoral fellow at the Department of Chinese and Bilingual Studies of Hong Kong Polytechnic University. Her research interest includes Financial NLP, CSR Modelling, Misinformation Detection, Sentiment/Emotion Analysis, Machine Learning, Language Resource Construction etc. She has 6 journal publications and more than 10 international conference proceedings in the NLP venue. She organizes the first Computing Social Responsibility Workshop cohosted at LREC 2022 conference.

• **Hanzhuo Tan** (*PART III*). Hanzhuo Tan is a Ph.D. student at the Department of Computing of Hong Kong Polytechnic University. His research interest includes self-supervised pre-training, NLP for social media, etc. He has 2 journal paper published in IEEE Transactions. He did six-month intership at Baidu PaddleNLP group on pre-training social transformer.

• **Rong Xiang** (*PART IV*). Rong Xiang is a postdoctoral fellow at the Department of Computing, Hong Kong Polytechnic University (PolyU). His research interests are acquisition and the application of human intelligence into machine learning networks. He has done substantial work in sentiment analysis, social media analysis and lexical semantics. He has published over 20 research papers in premier NLP venues. He co-organized CogALex 2020 and PACLIC 33.

• **Jing Li** (*PART V*). Jing Li is an assistant professor at the Department of Computing, Hong Kong Polytechnic University (PolyU). Before joining PolyU, she was a senior researcher in Tencent AI Lab. Her research interests are topic modeling, language representation learning, and NLP for colloquial and social media languages. She has published over 30 research papers in the top NLP venues and was invited to serve as the action editor for ACL rolling review (ARR) and the area chair for ACL 2021.

## 7 Other Information

*Inclusion of Others' Work.* This tutorial will survey the progress of Cantonese NLP and language pretraining, which substantially contain others' work.

*Divergency considerations.* Audiences who cannot speak Cantonese or Chinese will also be able to understand our tutorial. It will be conducted in English, where Cantonese cases will be presented with their English translations. Background knowledge will be provided to lower prerequisites (only those in Section 3 are needed). In the tutorial, we will discuss how the findings from Cantonese NLP can be generalized to other low-resource languages to benefit audiences in diverse streams.

*Estimation of audience size.* 100-200.

## References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE Transactions on Audio, Speech and Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Annemarie Friedrich and Damyana Gateva. 2017. Classification of telicity using cross-linguistic annotation projection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2559–2565.

Ofelia García and Joshua A Fishman. 2011. *The multilingual apple: languages in New York City*. Walter de Gruyter.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Michael A Hedderich, Lukas Lange, and Dietrich Klakow. 2021. Anea: distant supervision for low-resource named entity recognition. *arXiv preprint arXiv:2102.13129*.

Dan Jurafsky. 2000. *Speech & language processing*. Pearson Education India.

Shin Kataoka and Cream Lee. 2008. A system without a system: Cantonese romanization used in hong kong place and personal names. *Hong Kong Journal of Applied Linguistics*, 11(1):79–98.

Talaat Khalil, Kornel Kiełczewski, Georgios Christos Chouliaras, Amina Keldibek, and Maarten Versteegh. 2019. Cross-lingual intent classification in a low resource industrial setting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6419–6424.

John Lee, Baikun Liang, and Haley Fong. 2021. Restatement and question generation for counsellor chatbot. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 1–7, Online. Association for Computational Linguistics.

Jing Li, Yan Song, Zhongyu Wei, and Kam-Fai Wong. 2018. A joint model of conversational discourse latent topics on microblogs. *Computational Linguistics*, 44(4):719–754.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1518–1533, Online. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

S. Matthews and V. Yip. 2011. Cantonese: A comprehensive grammar (2nd ed.). *Routledge Grammars*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh-Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Jueya Ouyang. 1993. Putonghua guangzhouhua de bijiao yu xuexi (the comparison and learning of mandarin and cantonese).

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, 3, page 269. NIH Public Access.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Gertrude Tinker Sachs and David CS Li. 2007. Cantonese as an additional language in hong kong. *Multilingua*, 26(95):130.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009.

Don Snow. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*, volume 1. Hong Kong University Press.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Markus Svensén and Christopher M Bishop. 2007. Pattern recognition and machine learning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1779–1785, Doha, Qatar. Association for Computational Linguistics.

Tak-sum Wong and John SY Lee. 2018. Register-sensitive translation: A case study of mandarin and cantonese (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 89–96.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Rong Xiang, Ying Jiao, and Qin Lu. 2019. Sentiment augmented attention network for cantonese restaurant review analysis. In *Proceedings of the 8th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, pages 1–9. KDD WISDOM.

Henry Yu. 2013. Mountains of gold: Canada, north america, and the cantonese pacific. In *Routledge Handbook of the Chinese Diaspora*, pages 124–137. Routledge.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2019. Cross-lingual dependency parsing using code-mixed TreeBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 997–1006, Hong Kong, China. Association for Computational Linguistics.