Language Resources and Evaluation (2025) 59:1747–1773
https://doi.org/10.1007/s10579-024-09744-w

SURVEY

Cantonese natural language processing in the transformers era: a survey and current challenges

Rong Xiang[2] · Emmanuele Chersoni[1] · Yixia Li[3] · Jing Li[2] · Chu-Ren Huang[1] · Yushan Pan[4] · Yushi Li[5]

Accepted: 19 April 2024 / Published online: 8 June 2024

Abstract

Despite being spoken by a large population of speakers worldwide, Cantonese is under-resourced in terms of the data scale and diversity compared to other major languages. This limitation has excluded it from the current "pre-training and fine-tuning" paradigm that is dominated by Transformer architectures. In this paper, we provide a comprehensive review on the existing resources and methodologies for Cantonese Natural Language Processing, covering the recent progress in language understanding, text generation and development of language models. We finally discuss two aspects of the Cantonese language that could make it potentially challenging even for state-of-the-art architectures: *colloquialism* and *multilinguality*

Keywords Cantonese · NLP for social media · Multilingualism · Code-switching · Evaluation resources

Rong Xiang and Emmanuele Chersoni have contributed equally to this work.

✉ Emmanuele Chersoni
emmanuele.chersoni@polyu.edu.hk

[1] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China

[2] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

[3] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

[4] Department of Computing, Xi-an Jiaotong-Liverpool University, Suzhou, China

[5] Department of Intelligent Science, Xi-an Jiaotong-Liverpool University, Suzhou, China

1 Introduction

Cantonese, or Yue Chinese, is a diaspora language with over 85 million speakers all over the world (Lai, 2004; García & Fishman, 2011; Yu, 2013; Eberhard et al., 2022).[1] It is commonly used in colloquial scenarios (e.g., daily conversation and social media) but also in formal and written contexts, such as in the Legislative Council of the Hong Kong Special Administrative Region, or in sections of special local interests in the newspapers, such social and entertainment, or in horse racing and betting information. Otherwise Standard Chinese (SCN),[2] sometimes called Putonghua (普通话) or Guoyu (國語), is generally favored in formal and written contexts (Luke, 1995; Lee, 2016; Li, 2017; Wong & Lee, 2018).

In terms of digital language support, Mandarin Chinese thrives with a mature Natural Language Processing (NLP) environment. Chinese NLP has a versatile and growing literature from major conferences, such as ACL and COLING. In contrast, as for digital language support Cantonese is at the vital level, one level lower than thriving (cf. Ethnologue). In fact, Cantonese is an rare exception as a main diaspora language, as most diaspora languages -including but not limited to Arabic, Chinese, English, French, Hindi, Japanese, Korean, Portuguese, Spanish, etc.- have both a thriving digital language support and a strong NLP community, while Cantonese does not.

More specifically, while current NLP paradigms have been deeply changed by large-scale pre-training models based on Transformer architectures, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2020), GPT-3 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023), which have achieved state-of-the-art (SOTA) level of performance on several tasks. Compared to the previous generation systems, the progress was particularly remarkable in task requiring fine-grained semantic understanding, such as textual entailment, question answering and causal reasoning (Wang et al., 2018, 2019). On the other hand, language technologies for Cantonese have not yet benefited from this revolution (Xiang et al., 2022). From this point of view, the number of publications in the ACL Anthology is emblematic (see Fig. 1): only 61 papers are related to "Cantonese", compared to 9756 papers for English, and 5312 (4919 + 393) for SCN/Mandarin.

The history of publications in Cantonese NLP, as in Fig. 1, shows that the numbers of papers published yearly remains in single digit, although there is a moderate increasing trend (Fig. 2). However, as an emergent language in NLP, it is surprising that only a small portion (17/61, 27.9%) introduces language resources, as shown by Table 1. This explains why Cantonese NLP has a problem in terms of scarcity of resources and lack of alignment to state-of-the-art practices.

In light of these concerns, this paper presents a first overview of Cantonese NLP, going through essential issues regarding this language's uniqueness, data scarcity, research progress, and major challenges. As a pilot study, we also present some

[1] https://www.ethnologue.com/language/yue.

[2] Notice that the written form of SCN includes both simplified and traditional orthographies for writing in a specific Chinese dialect or topolect.

preliminary analysis on Cantonese data from social media and discuss the possible challenges. We found that, given the prominence of *colloquial language* and *code-switching* in the data, it is desirable that future models will be developed to properly deal with such phenomena. Finally, we conclude our contribution by indicating some possible directions for future research.

The remainder of this article, summarized in Fig. 3, is organized as follows: Chapter 2 provides background and characteristics of Cantonese as a language, as well as the differences between Cantonese and Standard Chinese. Chapter 3 summarizes the studies on Cantonese corpora, benchmarks, linguistic resources, natural language understanding, natural language generation and language models. Chapter 4 demonstrates the main challenges of Cantonese study which are colloguialism and multilinguality. Chapter 5 presents the possible future research directions of Cantonese NLP. Chapter 6 summarizes this survey, its current limitations and significance.

2 Uniqueness of Cantonese

Cantonese is the second for number of native speakers among all Sinitic languages/dialects of Chinese (Matthews & Yip, 2011). As a diaspora language, the native speakers of Cantonese reside originally in South China, including Guangdong, Hong Kong, Macao, and part of Guangxi. In addition, it is perhaps the most common diaspora language for overseas Chinese communities in South-East Asia, North America, and Western Europe (Sachs & Li, 2007; Yu, 2013).

The word Cantonese comes from Canton, the former English name of Guangzhou, capital of Guangdong, which was once considered the home of the most prestigious form of Cantonese. On the other hand, through years of mass media and pop culture influence, Hong Kong can now be considered as the most influential cultural centre of Cantonese.

Similar to many Sinitic languages that are traditionally called dialects of Chinese, Cantonese has both vernacular and formal strata that correspond roughly to spoken and written forms, but not always. That is, one could either speak in a formal and literal style or write colloquially, but both would be considered as marked. Cantonese diverges substantially from SCN in phonology, lexicon and grammar, with the difference increasing even more in informal communicative situations. Unlike other Chinese varieties, Cantonese developed its own input tools, such as *Yuepin* (Jyutping)[3] Cantonese is not fully supported for NLP as existing Chinese NLP tool-kits and packages are typically designed based on SCN (either in simplified or traditional form). Given the fact that formal strata of all Sinitic languages tend to largely coincide with SCN, colloquial Cantonese processing remains a challenge. As a diaspora language, spoken Cantonese also has several varieties that are spoken at different

[3] https://jyutping.org/en/. It is worth mentioning that, on the basis of the Jyutping romanization scheme, Lau et al. (2022b) recently proposed Rime-Cantonese, a multi-purpose lexicon that can be used to build Cantonese keyboards.

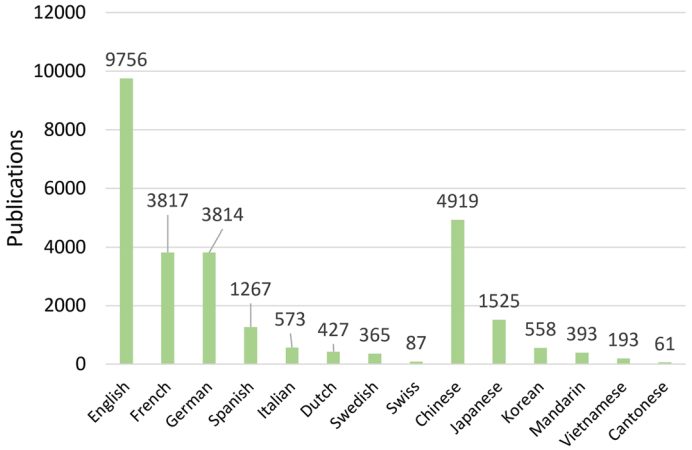

Fig. 1 Number of publications in the ACL Anthology indexed by languages as of Mar 2024. The publications were retrieved via searching the language name in either the title or the abstract

Table 1 Papers on Cantonese by research topic (statistics checked on March 2024)

Research topics	# of Papers
Phonetics &Phonology& Speech Recognition	22
Lexicography &Syntax& Semantics &Morphology	10
NLP Resources	17
NLP Tasks	12
Total	61

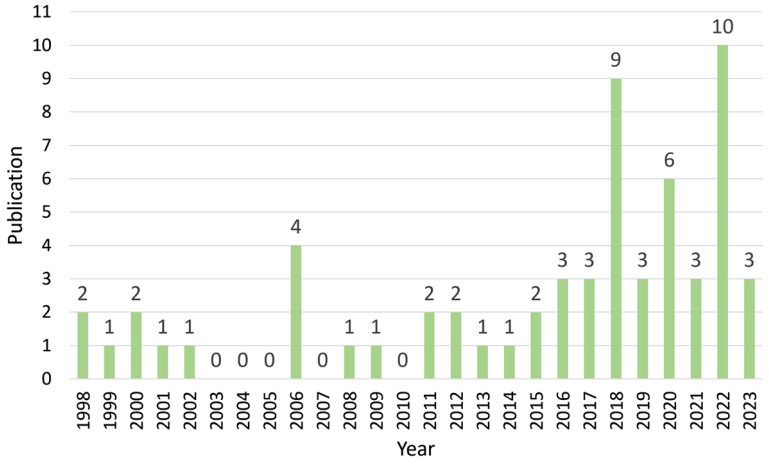

Fig. 2 Yearly publications of the 61 papers for Cantonese NLP in the ACL Anthology from 1998 to 2024

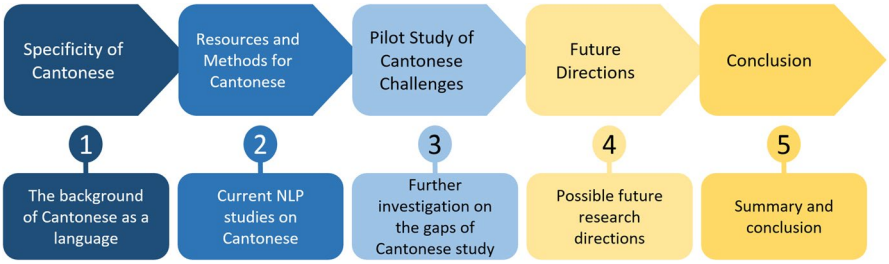

Fig. 3 Outline of the survey

parts of the world. This is why this paper focuses on the study of *colloquial Cantonese* (we just refer to it as *Cantonese* henceforth).

As Hong Kong has one of the biggest and most active online communities using Cantonese, we use Hong Kong Cantonese as a general representative of Cantonese.[4] Hong Kong Cantonese was deeply influenced by a unique congeries of social, economic, political, cultural, environmental, historical, and linguistic factors intrinsically linked to this city (Luke, 1995; Li, 2017; Bauer, 2018). Its multi-cultural background leads to frequent borrowing and code-mixing, hence it nicely mirrors the language landscape of diaspora Cantonese varieties spoken at overseas Cantonese communities. While Cantonese is rarely used in mainland Chinese media (Snow et al., 2004), the political segregation of Hong Kong from mainland China for over 150 years since the Opium Wars allowed Cantonese to develop as the dominant Sinitic language in Hong Kong, as attested by the fact that 'Chinese- stands for Cantonese in Hong Kong and not Mandarin. The sections of Hong Kong Chinese newspapers and magazines have seen a proliferation of articles in Cantonese, with the increasing popularity of a writing style known as 三及第 (*saam1 kap6 dai6*), which is characterized by the hybridization of Cantonese with other languages (classical and modern Chinese, English etc.) as an expressive device (Li, 2017).

The extraordinary status of Cantonese results, for example, in many distinctive, newly-coined Chinese characters (e.g., *gui6* 攰 'exhausted'), directional verbs (*heoi3* 去 'go' and *loi4* 来 'come'), aspect markers (*gan2* 緊 '-ing', *zo2* 咗 '-ed'), and constructions that are specific of the Cantonese syntax, such as the double object construction (*bei2 zo2 jat1 bun2 syu1 ngo5* 畀咗一本書我 'He gave a book to me' vs. *bei2 zo2 ngo5 jat1 bun2 syu1* 畀咗我一本書 'He gave me a book', notice the perfective aspect marker 咗 *zo2*, which is Cantonese-specific).

When dealing with Cantonese text, the frequent uses of *colloquial language* and *multilingualism* via various forms of code mixing and code switching pose fundamental challenges. Compared to Mandarin Chinese, Cantonese is much less conventionalized and often 'improvises- to represent or mimic actual pronunciation, as it is

[4] It should be noticed, however, that Hong Kong social media text contains a mixture of different languages, with a high-level of code-mixing. We do not intend to identify the language variety spoken in Hong Kong social media with colloquial Cantonese, we just mean that such platforms are potentially an important source of data for our target variety. Therefore, one would need an efficient way to identify the posts and the texts that are entirely in Cantonese.

evident from some clearly identifiable examples, e.g. *ham6 baang6 laang6* 冚唪唥 'entire/all'. The issue becomes even more relevant when we consider that the main textual sources for the computational processing of Cantonese are social media data, where typical features of social media language such as non-standard spelling, local slang, neologisms and emojis frequently appear. However, even in formal writings (cf. the Yue Wikipedia), Cantonese significantly differs from Mandarin, and the two varieties are not mutually intelligible in either written or spoken forms.

On the other hand, Cantonese differs from Mandarin in vocabulary by 30-50% (Snow et al., 2004), showing systematic differences from other Chinese varieties in several linguistic aspects (Lee et al., 2011). Such a deep difference has both historical and geographical roots, given the multi-cultural and multi-lingual environments in which Cantonese historically evolved. This is especially true nowadays for Hong Kong Cantonese, where English loanwords are particularly frequent in informal text genres. English abbreviations and/or grammatical elements such as suffixes can be used by mixing Chinese characters and alphabetic writing (e.g. *tung1 deng2* 通頂, 'working overnight'; *nei5 ho2 m4 ho2 ji5* DM *di1 sai3 zit3 bei2 ngo5 aa3* 你可唔可以DM啲細節俾我呀?, 'Can you send over the details via direct message?'[5]),[6] but they can also be transliterated (e.g., *si6 do1* 士多 'store'), including transliterations that have no written representation in Chinese characters.[7] Such irregularity can make the identification of the loanwords particularly difficult.

It has been shown that Hong Kong Cantonese speakers can also adopt a wide variety of strategies to render morpho-syllables that have no written representation, and the most common is the so-called *phonetic borrowing*: a linguistic element can be borrowed not for its semantic content, but because it sounds similar to the target morpheme to be represented (Li, 2000). The phenomenon commonly involves standard Chinese characters that happen to be homophonic in Cantonese with the novel syllable without character. This typically results in assigning a new meaning to the particular character in Cantonese that is not recognized by speakers of SCN or other sinitic languages, e.g. *gau6* 舊, 'old' in SCN, the classifier 'a lump of' in Cantonese (*jat1 gau6 gai1* 一舊雞 → 'old chicken' in SCN, but 'a lump of chicken' in Cantonese) (Li, 2017). This also happens to borrowings from English, further increasing the literacy problems for non-Cantonese readers.

In short, Cantonese possesses a vast array of unique linguistic features that can make it particularly challenging for models developed primarily for SCN.

[5] Examples taken from https://www.cantoneseclass101.com/blog/2019/07/23/cantonese-text-slang/.

[6] Notice there are also examples in SCN, cf. Ding et al. (2017); Xiang et al., (2020b), although they are less frequent.

[7] Notice that individual 'characters' (or sinograms) are typically morpho-syllabic, that is, they are pronounced as a syllable and typically morphemic.

3 Resources and methods for Cantonese

Unlike other diaspora languages such as Arabic, SCN, English or Spanish, which benefit from abundant well-annotated textual resources, there is a general lack of digitized resources for Cantonese data. In this section, the existing resources are divided into three main categories: *Corpora*, *Benchmarks*, and *Expert Resources* (Sects. 3.1–3.3).

In general, using existing Cantonese resources may be difficult for two reasons: (1) the data scale is relatively small (especially compared to SCN); (2) the domain is usually specific and lacks diversity and generality. To make things worse, the open-source situation for Cantonese resources is a concern, as many datasets are not publicly available.

What are the reasons of this scarcity? A main factor could be that the use of social media data, which are potentially one of the main sources for extracting Cantonese text in natural contexts and building benchmarks for NLP tasks, faces a lot of legal obstacles. The use of those data must comply with the requirements of the Personal Data (Privacy) Ordinance [Cap. 486 of the Laws of Hong Kong], which does not allow the collection and use of personal data without the express consent of the data subjects. Moreover, the provisions of the Copyright Ordinance [Cap. 528 of the Laws of Hong Kong, sections 22 and 23] prohibit the copying and adaption of any copyrightable work (it might be the case for some of the contents on social media platforms). Finally, the use of data is regulated by the contractual terms of use that govern the use of all social media platforms. Therefore, the development of open benchmarks for Cantonese is problematic from a legal point of view, and this might be a reason why many evaluation datasets do not get published.[8]

In our overview, following the description of the resources, we illustrate the current progress of the resources and the methodologies for Natural Language Processing for Cantonese, focusing on semantic tasks: first we present the available corpora, NLP benchmarks and expert resources (Sect. 3.1, 3.2 and 3.3); then we describe the advances in Natural Language Understanding (Sect. 3.4), and in Natural Language Generation (Sect. 3.5). We also introduce the publicly available language models that pretrained on Cantonese data (Sect. 3.6).

3.1 Corpora

Cantonese was perhaps the most documented Sinitic languages in early bilingual dictionaries compiled by western missionaries (Huang et al., 2016). Some Cantonese words were included in the first ‘modern- bilingual Chinese dictionary compiled by Matteo Ricci at the end of the 16th century. The majority of the bilingual dictionaries published throughout the 19th century were, indeed, dedicated to Cantonese. Given the important role of Cantonese in the context of the encounter between

[8] Even in many other highly industrialized countries, it is not totally clear what is the general regulation for the usage of social media texts for training language models (cf. the discussion about the situation in EU countries in Eckart de Castilho et al. (2018)), as in the common case of restrictively-licensed platforms this risks to be considered as an illegal form of permanent reproduction of the data.

China and the West, it is perhaps no surprising that the first Cantonese corpus was a bilingual one. Wu (1994) introduced the work on the HKUST Chinese-English Bilingual Parallel Corpus, based on the transcriptions from the Hong Kong legislative Council. the first monolingual Cantonese corpus was most likely the CANCOR-PLee and Wong (1998), consisting of one million characters from Cantonese-speaking children in Hong Kong. Another important corpus for child language acquisition is the CHILDES Cantonese-English Corpus by Yip and Matthews (2007), containing both audio and visual data of children conversation and the related transcripts.

The Hong Kong Cantonese Adult Language Corpus (HKCAC) focuses instead on adult language and contributes speech recorded from phone-in programs and forums (Leung & Law, 2001). This corpus also presents speech transcriptions for a total of 170k characters. Another resource, the Hong Kong University Cantonese Corpus (HKUCC) (Wong, 2006) was collected from transcribed spontaneous speech in conversations and radio programs and its annotation include word segmentation, Cantonese pronunciation and parts-of-speech, covering approximately 230,000 words.

Lee et al., (2011) introduced a parallel corpus that aligns Cantonese and SCN at the sentence level for machine translation. The annotation materials are the transcriptions of Cantonese speeches from television shows in Hong Kong, and their corresponding Mandarin subtitles. The corpus contains 4,135 pairs of aligned sentences, with a total of 36,775 characters in Mandarin, and 39,192 in Cantonese. Wong et al. (2017) later published a small parallel dependency treebank for Cantonese and Mandarin, based on the same textual materials. The corpus consists, in total, of 569 aligned sentences and it is annotated with the Universal Dependencies scheme (De et al., 2014; Nivre et al., 2016). Another corpus based on the transcripts of Hong Kong Cantonese movies has been presented by Chin (2015), and made accessible to the users via an online interface.[9]

Spoken Cantonese data from television and radio programmes broadcasted in Hong Kong are the source material also for the corpus introduced by Kwong, (2015). The corpus covers different topics, such as politics, affairs, economics/finance, and food/entertainment, and a variety of textual typologies (interviews, phone call transcriptions, reviews etc.). The Hong Kong Cantonese Corpus by Luke and Wong (2015) includes 150,000 words, and it also consists of transcribed Cantonese speech recordings that are annotated with both segmentation and part-of-speech tags. Ng et al. (2017) proposed the first bilingual speech corpus of Cantonese and English, built with the goal of the assessment of correct Cantonese pronunciation. Finally, the most recent introduction is the MYCanCor corpus (Liesenfeld, 2008), which has been built with 20 h of Cantonese speech recorded in Malaysia (plus the videos and the related transcriptions) to support studies on multimodal communication.

Concerning domain-specific resources, the parallel corpus by Ahrens (2015) includes 6 million words from political speeches from China, Hong Kong, Taiwan and USA, and it contains more than one million words of transcribed speeches of Hong Kong- leaders before and after the handover. It consist of more than 400k words in English, and more than 600k words in Chinese/Cantonese. Pan (2019) introduced a Chinese/English Political Corpus for translation and interpretation

[9] https://hkcc.eduhk.hk/.

studies. With over 6 million word tokens, the corpus consists of transcripts of both Cantonese and Mandarin and their English translations. Lee et al. (2020) introduced a Counselling Corpus in Cantonese to research domain-specific dialogues: 436 input questions were solicited from native Cantonese speakers and 150 chatbot replies were harvested from mental health websites. The authors later extended their work by collecting another dataset used for text summarization and question generation (Lee et al., 2021), containing 12,634 post-restatement pairs and 9,036 post-question pairs, all with manual annotations. It also includes 89,000 unlabeled post-reply pairs collected from the online discussion forums in Hong Kong. Finally, the `SpiCE` corpus by Johnson et al. (2020) is an open-access corpus created specifically for translation tasks and contains bilingual speech conversations in Cantonese and English, for a total of 19 h of conversation. The transcripts have been produced with the Google Cloud Speech-to-Text application, followed by manual corrections, orthographic alignment and phonetic transcriptions.

For corpus reading and preprocessing, Lee et al. (2022) recently introduced the PyCantonese package, which includes reader modules for some of the most popular Cantonese corpora (e.g. the CHILDES Cantonese-English Bilingual Corpus, the Hong Kong Cantonese Corpus etc.), stopword lists, modules for carrying out word segmentation and part-of-speech tagging, parsing and common computational tasks involving Jyutping (e.g. romanization of the characters).

3.2 NLP benchmarks

The gap between Cantonese and other diaspora languages in NLP research and digital support is underlined by the scarcity of benchmark datasets specifically targeting Cantonese. A first example was the shared task for Chinese Spelling Check, which was conducted in co-location with the workshop on NLP for Educational Applications in 2017. The organizers published a benchmark dataset with 6,890 sentences for normalizing Cantonese, mapping from the spoken to the written form (Fung et al., 2017).

Xiang et al. (2019) provided a sentiment analysis benchmark collected `Open-Rice`, a Hong Kong catering website, where over 60k comments are labeled with 5-level ratings indicating sentiment scores. The authors anonymized the data, filtered out comments written in other languages (e.g. SCN, English) and limited the length of the examples to 250 words.[10]

Chen et al. (2020) published a rumor detection benchmark collected from Twitter, including 27,328 web-crawled tweets (13,883 rumors and 13,445 non-rumors) written in Traditional Chinese characters, in part in Taiwanese Mandarin and in part in Cantonese.[11] However, the dataset does not provide the information about the language in which a tweet has been written.

[10] https://github.com/Christainx/Dataset_Cantonese_Openrice.

[11] https://github.com/cxyccc/CR-Dataset.

For text genre categorization, a benchmark has been collected by the ToastyNews project.[12]. The dataset consists of more than 11000 texts, divided into 20 different categories. The texts have been extracted from LIHKG, a popular Hong Kong forum with a structure similar to Reddit, and the category labels have been generated from the discussion threads they belong to.

Finally, for the development of dialogue systems, Wang et al. (2020) presented a food-ordering dialogue dataset for Cantonese called KddRES, including dialogues extracted from Facebook and OpenRice for 10 different Hong Kong restaurants. Using this dataset, it is possible to evaluate systems either on the classification of the intention of customer statements, or on sequence labeling tasks to identify the slot of interests of a conversation (e.g. the selected food, the number of people for a reservation, the time for take-out etc.).

3.3 Expert resources

We refer to language resources that have been handcrafted by trained linguists as *expert resources*. Dictionaries, ontology and knowledge bases are traditional types of expert resources.

jyut6 din2 粵典 is an example of a publicly-available crowd-sourced dictionary for Cantonese, covering 55,581 words in 5638 unique characters (Lau et al., 2022a).[13]. A manually-digitalized version of the dictionary of modern Cantonese has been published by Cheung et al. (2018) and contains more than 12,000 entries, while a lexical database of Hong Kong Cantonese has been proposed by Lai et al. (2020), providing definitions, frequency, strokes, and structure for 51,798 Cantonese words. Moreover, a recent study by Winterstein et al. (2023) focused on Cantonese nominal expressions (e.g. bare nouns, bare classifier phrases, numeral phrases etc.), the authors annotated almost 11K of such constructions in the HKCanCor corpus (Luke & Wong, 2015) for several syntactic and semantic features (e.g. classifier of the noun, type of construction abstractness, animacy, mass/count status of the head noun etc.). The annotations have been made publicly available for future studies on nominal expressions in Cantonese.[14].

As for the ontology, Cantonese has its own version of the WordNet lexical network (Sio et al., 2019), including over 3,500 concepts and 12,000 senses, which are structured in a hierarchy of semantic relations.

Finally, a more domain-specific resource is the expert-customized sentiment lexicon by Klyueva et al. (2018), which focuses on food-related Cantonese words and contains 1887 positive and 858 negative words.

[12] https://github.com/toastynews/lihkg-cat-v2

[13] https://words.hk/

[14] https://osf.io/6hw37?view_only=673e8af11bba4ab6b8559ffe29e5d8ac

3.4 Natural language understanding

Natural Language Understanding refers to the tasks that require models to have a grasp of aspects of the semantics of text. NLP has witnessed important advances on such tasks after the introduction of pretrained language models architecture. A full overview of the recent general progress in this field in NLP would be out of the scope of the present article (we refer the reader to Lenci (2023) for an updated state of the current research), and thus we limit ourselves to the work done for the Cantonese language.

3.4.1 Rumor detection

Using their rumor detection dataset, Chen et al. (2020) devised a method called XGA (XLNet-based Bidirectional Gated Recurrent network with Attention mechanism) to identify rumors in social media posts. Their approach makes use of the XLNet Transformer (Yang et al., 2019) to generate both text and sentiment embeddings for the target texts, before feeding them to a BiGRU network with attention. The same group of authors later proposed an improvement of the system (Ke et al., 2020), this time using a pre-trained BERT language model (Devlin et al., 2019) combined with a Bi-LSTM network with attention, which led to further Accuracy improvements. However, it should be pointed out again that their evaluation dataset actually contains a mixture of Cantonese and Taiwanese Mandarin in Traditional Chinese characters and the performance is not analyzed by language, so it is difficult to assess how well the system is actually doing on Cantonese.

3.4.2 Sentiment analysis

To model sentiment in Cantonese, Zhang et al. (2011) proposed to employ Naive Bayes and SVM with handcrafted features to predict the customers' sentiment in a dataset of OpenRice reviews. The authors showed similar performance for the two classifiers and observed that the feature choice had a major impact, with character-based bigrams being the most efficient feature type in capturing Cantonese sentiment orientation.

The works by Chen et al. (2013, 2015) took advantage instead of the advances in Cantonese sentence segmentation and Part-of-Speech tagging based on a Hidden Markov Model. After applying the above-mentioned preprocessing steps to their data, they created a keyword dictionary based on manually-designed sentiment seed words and assign sentiment polarities to the target sentences via a rule-based system.

In a more recent study, Ngai et al. (2018) combined supervised machine learning and unsupervised lexicon-based approaches over multiple-domain sentiment classification. They found that an additional sentiment lexicon can provide extra benefits to machine learning classifiers in both the training and inference stages.

Xiang et al. (2019) first illustrated an unsupervised method to expand a Cantonese sentiment lexicon, and then they incorporate this knowledge into a LSTM with attention, which resulted in an Accuracy score of around 60.8% on a large dataset of restaurant reviews collected from OpenRice.

Beyond the traditional polarity identification, Lee (2019) exploited Mandarin emotion resources and lexical mappings between Cantonese, English, and Mandarin to operate a more fine-grained emotion analysis. In a preliminary evaluation on a 8-class emotion classification task, they obtained 62.5% Accuracy on a small dataset of social media posts.

3.4.3 Cognitive modeling and computational psycholinguistics

Recent NLP research has rediscovered the value of using psycholinguistic data such as human reading times and eye-tracking fixations to build more challenging and cognitively-plausible benchmarks (Hollenstein et al., 2021, 2022). The work by Li et al. (2023) introduces a parallel eye-tracking corpus for Mandarin and Cantonese,[15] based on the textual materials from *Le Petit Prince* by Antoine de Saint-Exupéry and including several fixation metrics for each word. The authors propose a general evaluation on the task of predicting eye fixations using several linguistically-motivated features (e.g. segmentation, POS, syntactic distances, dependency tree depth etc.), plus the contextualized embedding representations of Transformer models. Eye fixations are notoriously related to language processing difficulty (Hale, 2016), and technological improvements in predicting such data might be useful to develop educational applications and/or text simplification systems (Shardlow, 2014) for Cantonese.

3.5 Natural language generation

3.5.1 Dialogue summarization

Lee et al. (2021) explored the generation of questions and restatements for Cantonese dialogues, in the context of counseling chatbots. In both the text summarization and in the question generation task (e.g. the system first has to summarize the main content of the input from the user, then it has to generate appropriate questions), the fine-tuning of the pre-trained BertSum model (Liu & Lapata, 2019) over Cantonese data enabled the largest performance increase.

3.5.2 Machine translation

The earliest attempts in this line are based on heuristic rules, which were in turn handcrafted by human experts (Zhang, 1998), and a bilingual knowledge base for Cantonese-English (Wu et al., 2006).

The more recent studies are based on statistical machine translation techniques. (Huang et al., 2016) adopts a small-scale parallel resource to show the challenge for deep learning models to translate between Cantonese and Mandarin in low-resource scenarios. Following their practice, Wong and Lee (2018) further leveraged lexical

[15] https://github.com/CN-Eyetk/MCFIX.

mappings and syntactic transformations to automatically scale up the parallel data to allow a more efficient model training.

Liu (2022) introduced a large-scale parallel evaluation dataset for Mandarin-Cantonese machine translation.[16] The author extracted parallel sentences from the Cantonese and the Mandarin Wikipedia, using bitext mining to identify semantically similar sentences and then selecting them with a round of manual filtering. The final resource includes more than 35K sentence pairs.

Finally, Dare et al. (2023) experimented with different types of unsupervised Cantonese-Mandarin machine translation systems, exploiting the power of crosslingual word embeddings to produce translations even in absence of a large amount of parallel data. The authors tested several architectures, obtaining their best results with a model combing the Transformer architecture and character-based tokenization. Moreover, they created a new Cantonese corpus, consisting of approximately 1 million sentences.[17]

3.6 Language models for Cantonese

As we stated in the introductory sections, training language models for Cantonese is not easy, given the scarcity of the available data that is not legally restricted. The only exception, at the moment, is represented by the Transformer architectures made available by the ToastyNews. This project, which aims at developing open source NLP tools for Cantonese, introduced a XLNet and an ELECTRA model trained partially on Cantonese data.[18]

The **XLNet** architecture (Yang et al., 2019) is a generalized auto-regressive Transformer using the context word to predict the next word. The autoregressive architecture is constrained to a single direction (either forward or backwards), that is, context representation takes in considerations only the tokens to the left or to the right of i-th position, while BERT representation has access to the contextual information on both sides. To capture bidirectional contexts, XLNet is trained with a permutation method as language modeling objective, where all tokens are predicted but in random order.

ELECTRA (Clark et al., 2020) instead adopts a pre-training approach that reminds the training of Generative Adversarial Networks. The training dynamics of ELECTRA relies on two neural networks, a *generator* and a *discriminator*. During the training phase, the generator network will replace some of the tokens from the sentences of the input corpus with plausible alternatives, and the discriminator network is trained with the objective of identifying which tokens in the input have been replaced (*replaced token detection*).

The training materials include a mixture of blogs and articles in Cantonese, together with the texts of the entire Cantonese Wikipedia. However, it should be

[16] Data and code available at: https://github.com/evelynkyl/yue_nmt.

[17] The Cantonese corpus and the scripts used for scraping the texts are available at: https://github.com/meganndare/cantonese-nlp.

[18] XLNet-HK-Base and the different components of ELECTRA-HK-Base are available for download at https://huggingface.co/toastynews.

pointed out that a big part of the training data is in SCN (around the 60%), and there is a lot of contamination from other languages, including English. We are not aware of any published research on the comparative evaluation between the performance of these models and SCN ones on Cantonese benchmarks, which would be important to assess the impact of language contamination.

4 Closing the gap: resolution of mixed codes due to colloquialism and multilinguality

In the previous sections, we have illustrated the general scarcity of resources in NLP for Cantonese. We also mentioned that Cantonese has a numerous and active social media community, and Cantonese social media language provides an interesting example for analysis, as it can show the main challenges related to the automatic processing of this language.

As we anticipated, *colloquialism* and *multilinguality* are primary obstacles to robust and effective processing. In the next sections, we present an analysis of the two phenomena in Cantonese social media.

4.1 Colloquialism and lexical differences

In the introductory sections, we already discussed how the Cantonese vocabulary deeply diverges from SCN (Ouyang et al., 1993; Snow et al.., 2004), and mentioned the fact that, due to the long tradition of all Sinitic languages sharing a written/formal strata (i.e. written Chinese), the divergence and challenges of Cantonese are in the spoken or informal strata. This include transcriptions of speech, as well as the habit in writing to adopt a colloquial style when dealing with topics of local interest, hence we refer to it as "colloquialism").

In this section, we analyze the colloquial features of Cantonese, with some examples, and present some data from a small-scale study on word surprisal (Hale, 2001, 2016). To start with, we examined the data from three popular Cantonese online forums: DISCUSS, LIHKG, and OpenRice (Hong Kong).[19] The first two are general forums with diverse topics, while OpenRice is s the most popular forum for sharing restaurant and food reviews. Table 2 shows the statistics of the forums, where the three sources altogether contribute 1.1 Gigabytes (G) texts and 0.924 billion (B) tokens. Just to give some figures for comparison, 80 G texts and 16B tokens have been used for pre-training English models on tweets (BERTweet, Nguyen et al. (2020)), and 5.4B tokens have been used for a relatively small size model for SCN (MacBERT, Cui et al. (2021)). This would be, to the best of our knowledge, the largest social media text collection for pre-training a Cantonese model from scratch, although the data size is certainly smaller compared to other languages.

One reason why it is challenging to directly apply or adapt SCN NLP models for Cantonese is the large number of Cantonese specific vocabulary and expressions,

[19] DISCUSS: https://discuss.com.hk; LIHKG: https://lihkg.com; OpenRice:https://www.openrice.com/zh/hongkong.

including words with unknown forms and words with known forms but with novel meanings. These discrepancies made the pre-trained models based on Mandarin ineffective for Cantonese NLP. In addition, due to the low degree of conventionalizing, *spelling mistakes* are prominent in the data, such as the mis-replacement of *fan3 gaau3* 訓覺 instead of *fan3 gaau3* 瞓覺 (*sleep*), together with intentional misspellings in jokes and punning, which are commonly found also in newspapers headlines (Li & Costa, 2009).

As in all social media texts, *slang expressions and idioms* are also frequently found, requiring external knowledge and background for the correct understanding, and most of such expressions are unknown in standard Chinese. Consider the following example: *gam1 ci3 jin2 coeng3 wui2 hou2 naan4 maai5 dou3 fei1 keoi5 dou1 hai6 zap1 sei2 gai1 sin1 zi3 jau5 dak1 tai2 zaa3* 。今次演唱會好難買到飛，佢都係執死雞先至有得睇咋。 (*It's extremely hard to buy tickets for the concert. He would not have a chance to go to the concert if he did not collect a lucky coin*). There are at least two expressions that would be challenging to a SCN trained model. The first is the word 飛 'fare, ticket-, which is a phonetic borrowing as discussed above. A Mandarin trained model would treat it as the verb 'to fly-, with a different PoS and totally different behavior. The second is the expression *zap1 sei2 gai1* 執死雞 is a Cantonese idiom originated from football terminology, literally meaning 'to hold (a) dead chicken-, which is shared by Mandarin and Cantonese. However, in Cantonese, it also has the idiomatic meaning that was originally used in soccer 'scoring a goal with pure luck.- These two meanings in Cantonese cannot be obtained without either a comprehensive Cantonese lexicon of colloquial usages or a large training corpus. Without the prior knowledge of its extended meaning of "to get a great deal", even for humans it would be challenging to make sense of the sentence, not to mention NLP models.

We studied the bigram distributions of DISCUSS, containing forum threads in 20 different topics, and compare it with the Gigaword corpus, which is composed of text from news outlets in Chinese (Huang, 2009; Parker et al., 2011). Both datasets concern contemporary and widely-discussed events in diverse news topics and are written in traditional Chinese. For both datasets, we sampled 260 megabytes of textual data and computed the average frequency of the union of the top 1000 most frequent bigrams in the two datasets. The relative frequencies of the bigrams are shown in Fig. 4. We can observe, at a glance, that the distribution of DISCUSS exhibits a high spike on the left, and then it has a long tail of low-frequency bigrams. Notice that, given the bigger size and the more standardized nature of GigaWord, the relative frequencies of many of the shared bigrams in the long tail are comparably higher.

To explore the predictability of Cantonese text by SCN models, we utilized two representative models to extract and compare surprisal scores for Cantonese sentences and the corresponding translations in Simplified and Traditional Chinese. We chose to use the *BERT-CKIP* model,[20] which was trained on Traditional Chinese on a concatenation of a 2020 dump of the Chinese Wikipedia and the Chinese

[20] https://github.com/ckiplab/ckip-transformers.

Table 2 Scales of textual data from 3 different Cantonese forums (0.924 billion tokens and 1.1 Gigabytes size in total)

Data Source	Token Count (M)	Text Size (MB)
DISCUSS	118.7	258.8
LIHKG	632.7	651.9
OpenRice	172.1	226.1

Gigaword Corpus (Huang, 2009; Parker et al., 2011); and the *RoBERTa-HFL* model,[21] an implementation of RoBERTa by Cui et al. (2021). It has been trained on both Simplified and Traditional characters on a 2019 dump of the Chinese Wikipedia and various news and question answering websites.

The surprisal of a word w (Hale, 2001; Levy,2008) is generally defined as the negative log probability of the word conditioned on the sentence context, according to the following:

$$Surprisal\,(w) = -logP(w|context) \tag{1}$$

The higher the surprisal for a given linguistic expression, the more unpredictable that expression is for a given computational model. If a model instead is able to provide confident estimates of words occurring in a corpus, the surprisal will be low.

To run our small experiment, we adopted the implementation of the minicons library (Misra, 2022), which provides handy functions to estimate probability and surprisal scores of a sentence. We randomly sampled 50 sentences from the Cantonese forums in Section 4.1, and for each of them we generate the translation in both Traditional and Simplified Chinese using the Baidu translation interface.[22] Then we computed the surprisal score for each sentence using the two SCN models, and took the average across sentences. The sampling was repeated 10 times (Table 3 reports the average across different samples). Notice that, since both BERT-CKIP and RoBERTa-HFL are bidirectional models trained, the surprisal scores for each word are computed by masking the words in the sentence one-by-one, computing their probabilities in context and then applying the formula in (1). Once the scores for single words are obtained, the minicons library outputs their average as the surprisal score for the sentence,[23]

We tested both Cantonese sentences and Taiwan Mandarin sentences from the Academia Sinica Corpus (Huang & Chen, 1992). Note that both Hong Kong and Taiwan use traditional characters with variations in lexical choices. Thus, our study was carried out in three different writing systems to ensure that the differences in writing systems do not contribute to the surprisal scores. Thus each set of data are tested in (1) original writing forms, (2) converted writing forms with each other (i.e. Hong Kong vs. Taiwan), and 3) converted to simplified Chinese.

[21] https://huggingface.co/hfl/chinese-roberta-wwm-ext.

[22] https://fanyi.baidu.com/.

[23] This method for estimating probabilities/surprisals for sentences with bidirectional language models is known as *pseudo log-likelihood* and it has been introduced by Salazar et al. (2020). This method has a standard implementation in the minicons library.

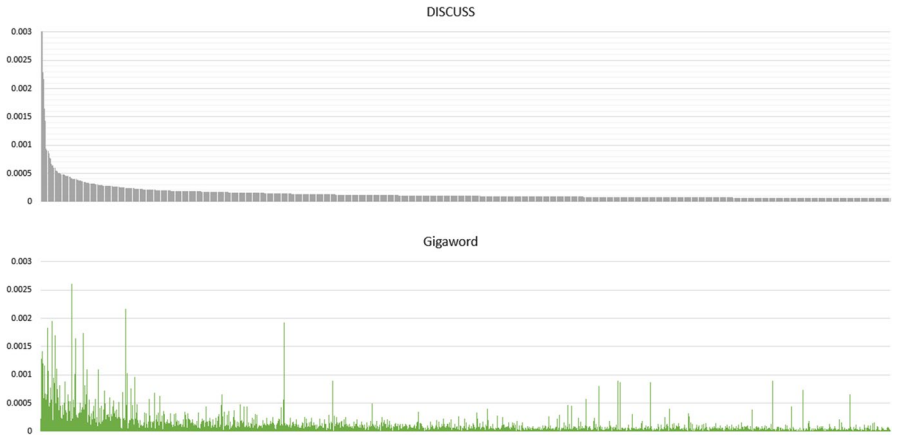

Fig. 4 Distribution of bigrams from DISCUSS and Gigaword datasets. The x-axis shows the union dataset of the top 1,000 bigrams from each dataset ordered by the average relative frequency in the two datasets. The top curve refers to DISCUSS, the bottom one to Gigaword

Table 3 Surprisal analysis on 50 Cantonese and Traditional Chinese sentences

	BERT-CKIP	RoBERTa-HFL
Can_Orig	4.30	4.39
Trad_Translated	3.17	2.89
Simp_Translated_Can	5.84	2.79
Trad_Orig	0.61	1.09
Can_Translated	1.71	2.20
Simp_Translated_Trad	5.38	1.15

The average surprisal scores are shown in the table. Can_Orig: 50 Cantonese sentences. Trad_Translated: 50 Traditional Chinese sentences translated from Can_Orig. Simp_Translated_Can: 50 Simplified sentences translated from Can_Orig

The results in Table 3 show that for both models and for three possible writing system settings (i.e. original, switched, simplified), the Cantonese sentences tend to have higher surprisal scores. The experiment establishes that it is more difficult for SCN trained models to predict Cantonese sentences. One of the reasons of the additional difficulties may be the usage of different words in Cantonese: we computed that, compared to the translated sentences, there is an overlap of characters of 69.1% for the Traditional Chinese translation and 65.5% for the Simplified Chinese one (i.e. more than 30% of the Cantonese characters do not appear in the translations). Still, given the relatively high overlap degree, it is likely that Cantonese-specific words play a role together with other factors, such as regional usages of the same words/characters and differences in grammar.

The two models behave very differently when the Cantonese text is translated into Simplified Chinese: RoBERTa-HFL, which is trained on both Traditional and Simplified characters, reports lower surprisal scores than on the original Cantonese sentences, and has a slightly higher score for the translation from Traditional to Simplified (which might be due to the ambiguity of the conversion, as for a traditional character there might be multiple corresponding characters in Simplified Chinese); BERT-CKIP has instead extremely high surprisal scores when either Cantonese or Traditional Chinese are translated into Simplified Chinese, as it was not exposed to Simplified characters during pretraining. In any case, we can notice that predicting words in Cantonese is much more challenging for SCN models, and that extra difficulties may come in when there is a conversion from Traditional to Simplified characters.

4.2 Multilinguality

To better understand the nature of multilingualism, we examine the contribution of different languages to Hong Kong social media data. The open-source toolkit *fastlangid* is employed to analyze the language usage ratio of the datasets.[24] More specifically, we used *fastlangid* with the default settings and the parameter $k = 1$, meaning that only the most likely language shall be detected. The percentages are shown in Table 4, where the statistics have been computed as an aggregation of sentence-level results. As it can be seen, the code-switching behavior across Cantonese and SCN is frequent; English is also very often attested in our data,[25] and we can even observe code-mixing with other languages. This is because Cantonese-speaking areas happen to integrate speakers of multiple nationalities (Yue-Hashimoto, 1991; Li, 2006).

To exemplify the multilingualism phenomenon in Cantonese, we present some typical code-switching cases of Cantonese and English. The original texts are followed by the English translations in brackets. The switched scripts are underlined in both the original texts and the translations.

E1: *sau1 dou3* offer, *gam1 nin4 gau2 jyut6 zung6 heoi3 m4 heoi3 dou3 ngoi6 gwok3 duk6 syu1 hou2?* 收到offer,今年9月仲去唔去到外國讀書好? (*Got the* offer. *Will it be better or not to go for overseas study in September this year?*)

E2: *hai6 ge3 zau6 wai4 jau5 hai2 hoeng1 gong2 maai5 liu5, tung4 maai4 dim2 gaai2 hoeng1 gong2 di1 din6 hei3 dim3 m4 gaau2 haa6 di1 si3 sik6 wut6 dung6*。係嘅就唯有喺香港買了, 同埋點解香港啲電器店唔搞下啲試食活 動。 (*I can only buy it in Hong Kong. And why don't the electrical appliance stores* of *Hong Kong do* some *trial promotion campaigns.*)

E3: *zaa3 zoeng3 bei2 gaau3 taam5, bat1 gwo3 min6 hou2* Q, *zan1 hai6 hou2 zeng3* 。炸醬比較淡, 不過麵好Q, 真係好正。(*The fried sauce is bland, but the noodles are very* chewy. *it's really tasty.*)

[24] https://github.com/currentsapi/fastlangid.

[25] It should be kept in mind that English is still one of the primary languages in Hong Kong education.

Table 4 Ratio of language usage

Language	Cantonese (%)	SCN (%)	English (%)	Others (%)
DISCUSS	31.49	52.00	9.19	7.32
LIHKG	40.57	33.40	11.85	14.18
OpenRice	73.65	18.91	4.93	2.55

Cantonese and Standard Chinese are dominant in all the datasets under consideration

The code-switching phenomenon in E1 is commonly observed in the data: the English nouns "offer" is directly taken and inserted in a Cantonese context. E2 uses "D" in the alphabet as an alternative to Cantonese tokens *di1* "啲" (*of*) and *dim2* "點" (*some*) because of their similar pronunciations. For E3, "Q" is borrowed from Hokkien, another Chinese variety of the Southern Min group that is widely used in Fujian and Taiwan, and it means "chewy". The borrowing can be explained by the geographical proximity of the Cantonese and Hokkien speaking areas and by the constant migratory flows between the two regions.

In sum, our analysis shows how colloquialism and code-switching with multiple languages are pervasive in Cantonese social media data, and thus models for Cantonese NLP will have to be robust to such phenomena. For example, future Cantonese language understanding systems could be integrated with spelling correction and dialect identification components, in order to mitigate the irregularity of the input data.

5 Future directions

Given the current situation of Cantonese NLP, an obvious strategy to improve the performance of Natural Language Understanding systems for this language would be to train new Transformer-based language models specifically for Cantonese. From this perspective, however, we mentioned that the usage of one of the potential main sources of Cantonese text -social media- may be legally problematic.

Two other promising directions for future studies on Cantonese are *data augmentation* and *cross-lingual learning*, which could help to cope with the lack of resources for this language.

5.1 Data augmentation

To deal with low-resource scenarios, strategies for augmenting the training data are commonly used in modern NLP. Generally speaking, data augmentation strategies can be grouped in two families: *label-invariant*, which create new training instances from a given instance by preserving the original label; and *sibyl-variant* ones, which create new instances by changing the label of the original sample in a predictable way (Gulzar et al., 2022).

In the first case, one could think about increasing the size of the training datasets for Cantonese by using simple heuristics, either at the lexical or at the syntactic

level. At the lexical level, new examples could be easily created via word replacement, deletion or swap (Wei & Zou, 2019), and the process has been shown to lead to the generation of high-quality textual data, especially when the manipulation can rely on ontology information (e.g. the Cantonese WordNet) for better semantic accuracy (Xiang et al., 2020a, 2021). At the syntactic level, transformations such as verb argument swaps or the replacement of syntactic sub-trees have been shown in the literature to increase the robustness of machine learning models (Şahin & Steedman, 2018; Min et al., 2020; Shi et al., 2021). Additionally, the recent methods of *supervised contrastive learning* allow for the augmentation of the number of examples by modifying directly the neural representations of a text sequence, and optimizing the representations for the target task at the same time (Gao et al., 2021). For example, multiple views can be created from a single instance by passing it multiple times through a Transformer encoder and applying every time a different dropout mask to the generated embedding representation (Sedghamiz et al., 2021).

In the second case, a promising trend of studies makes use of the recent advances in the technologies for text style transfer to generate new examples by modifying the relevant semantic attributes of the available data (Jin et al., 2019; Dai et al., 2019). This could be applied, for example, to the creation of novel instances of the opposite polarity in sentiment analysis, or with an increased/decreased emotional load for tasks aiming at classifying the degree of subjectivity of a text.

5.2 Cross-lingual learning

It is likely that, even for models trained only on SCN with no further adaptation, a some amount of knowledge can be transferred to carry out tasks in Cantonese. From this point of view, it is important to mention that cross-language transfer is another area of NLP that recently reported impressive advances. One of the reason of this success is the publication of Transformer language models that have simultaneously been trained on multiple languages, e.g. Multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), Multilingual BART (Liu et al., 2020) and mGPT (Shliazhko et al., 2022). Significantly, those models have proved to have zero-shot learning capabilities (i.e. they can be trained on a high-resource language and tackle the same task on an unseen, low-resource one) (Choi et al., 2011), they can generalize across different scripts and, to same extent, across languages with very different typological features (Pires et al., 2019), and finally, they can predict eye movements in reading in multiple languages (Hollenstein et al., 2021b, 2022). On the other hand, such models suffer from the so-called *curse of multilinguality*, that is, the progressive deterioration of per-language performance as more languages are covered by the model (Conneau et al., 2020; Pfeiffer et al., 2022). In this line of research, a very interesting model for the purposes of Cantonese NLP has been recently introduced by Yang et al. (2022), who proposed CINO,[26] a model specialized for Sinitic languages. CINO is a version of the XLM-R Multilingual Transformer (Conneau et al., 2020) that has been trained on texts from several Chinese varieties, including

[26] https://github.com/iflytek/cino.

Cantonese. The vocabulary of the model merges those of the tokenizers for Chinese, Tibetan, Uyghur Arabic, Mongolian and Hangul, in a way that it covers more than 135K tokens and can handle language data in any of the above-mentioned scripts. The objective function is multilingual masked language modeling, and the sampling rate of the languages during training has been calibrated in order to avoid that the higher-resource language (SCN) is over-represented in the internal representations of the model. We believe attempts like CINO are extremely promising for solving tasks in Cantonese NLP, given the richness of textual data and resources for Mandarin Chinese and the possibilities of transfer learning between the two varieties.

Finally, some important developments for Cantonese NLP could come from the research of the newly-introduced Large Language Models (LLMs), systems trained on massive amounts of text and with a vast increase of parameter size (Brown et al., 2020; Scao et al., 2022; Black et al., 2022; Achiam et al., 2023; Touvron et al., 2023a, b; Jiang et al., 20203; Almazrouei et al., 2023; Bai et al., 2023; Ren et al., 2023). Compared to the language models of the previous generation, LLMs have been reported to show the so-called *emergent abilities*, that is, the capacity of solving tasks on which they were not explicitly trained on (Wei et al., 2022; Dettmers et al., 2022; Zhao et al., 2023; Chang et al., 2023). This is generally done via textual instructions called *prompts*, in a zero shot or in a few shot learning scenario. Many of the most popular LLMs (e.g. ChatGPT) are mainly trained on Western languages and their performance was proved to be weaker for languages using non-Latin scripts, especially for tasks involving text generation (Bang et al., 2023).

However, the research on LLMs for Chinese has been progressing quickly in the last year. For example, the work of Cui et al. (2023) aimed at adapting the LLaMa and Alpaca architectures (Touvron et al., 2023a; Taori et al., 2023) to Chinese.[27] The authors introduced several optimizations, including the expansion of the Chinese vocabulary of the original model, secondary pretraining using Chinese data and fine-tuning with Chinese instructions. Moreover, new Chinese LLMs have been made publicly available thanks to the recent efforts of companies like Alibaba (the Qwen LLMs family, Bai et al. (2023)[28]) and Huawei (the PanGu LLMs family, Ren et al. (2023); Wang et al. (2023).[29]).

Although, to our knowledge, no evaluations of Chinese LLMs have been carried out on Cantonese benchmarks yet, we hope that future research will quickly close this gap and experiment with new ways of transferring the knowledge learned from large amounts of SCN textual data to the low-resource Chinese varieties.

6 Conclusions

In this paper, our goal is to present the status of the research on Cantonese NLP, to describe the uniqueness of this language and to suggest possible solutions for addressing the current shortcoming, due to the lack of resources. Indeed, most

[27] https://github.com/ymcui/Chinese-LLaMA-Alpaca.

[28] https://github.com/QwenLM/Qwen.

[29] https://github.com/huawei-noah/Pretrained-Language-Model.

research on Cantonese NLP has not translated into the release of useful models, corpora and benchmark datasets, which are often not publicly available or not up to date. A possible reason of this difficulty is the limited number of online sources of Cantonese text with non-restrictive licenses (Eckart de Castilho et al., 2018), which does not leave too many options to researchers for putting together new benchmarks and for training large-scale models that are Cantonese-specific.

After reviewing the existing resources and methods, we analyzed the two main challenges that such data pose to automatic systems: the pervasive colloquialism and the multilinguality of Cantonese text, which often leads to the simultaneous presence of multiple languages in the same message or post. As strategies to tackle the challenges of Cantonese NLP, we could safely indicate data augmentation and crosslingual learning as two possible ways to go, in case the collection and balancing of large-scale Cantonese corpora turn out to be too problematic.

Cantonese is one of the most pervasive diaspora languages with native speaking communities spread around the world and has a vibrant and multicultural online community, and unique features that deserve a special attention for computational modeling. With our contribution, we hope we will manage to stimulate a new interest around this language in the NLP community, and to encourage future studies that will be devoted to resource sharing and to the reproducibility of the research results on public benchmarks.

Acknowledgements The authors would like to thank Tracy Xin Luo, Kate Wong, Dr. Tak-sum Wong and Prof. David Chor Shing Li for their help in revising the Cantonese examples, and Prof. Eliza Mik for her advice on the legal aspects of using social media data for language model pretraining. We would also like to thank the anonymous reviewers for their insightful feedback, which greatly helped to improve the quality of the paper.

Author contributions R.X. was responsible for the conceptualization of the study (together with E.C.), for writing the code and part of the evaluation section. E.C. was responsible for the conceptualization of the study and most of the writing. Yi.L. supported R.X. in writing the code and took care of data curation for the experiments in Section 4. J. L., C.R.H., Y.P. and Yu.L. contributed to the general revision of the manuscript.

Funding Open access funding provided by The Hong Kong Polytechnic University. The authors did not receive any external funding for this research.

Declarations

Conflict of interest The authors declare that they do not have any competing financial and non-financial interests.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, FL., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al. (2023) GPT-4 Technical Report. arXiv preprint arXiv:2303.08774

Ahrens, K. (2015) Corpus of Political Speeches. Hong Kong Baptist University Library, URL https://digital.lib.hkbu.edu.hk/corpus/

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., & Malartic, Q. et al (2023) The Falcon Series of Open Language Models. arXiv preprint arXiv:2311.16867

Bai J., Bai S., Chu Y., Cui Z., Dang K., Deng X., Fan Y., Ge W., Han Y., Huang F. et al (2023) Qwen Technical Teport. arXiv preprint arXiv:2309.16609

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., & Fung, P. (2023) A Multitask., Multilingual., Multimodal Evaluation of ChatGPT on Reasoning., Hallucination., and Interactivity. arXiv preprint arXiv:2302.04023

Bauer, R. S. (2018). Cantonese as written language in Hong Kong. *Global Chinese, 4*(1), 103–142.

Black S., Biderman S., Hallahan E., Anthony Q., Gao L., Golding L., He H., Leahy C., McDonell K., Phang J. et al (2022) GPT-NeoX-20B: An open-source autoregressive language model. arXiv preprint arXiv:2204.06745

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020) Language models are few-shot learners. In: Larochelle H., Ranzato M., Hadsell R., Balcan M., Lin H (eds) Advances in neural information processing systems

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2023). A survey on evaluation of large language models. *ACM Trans Intel Syst Technol, 15*, 1–45.

Chen J., Liu Y., Zhang G., Cai Y., Wang T., Min H (2013) Sentiment analysis for Cantonese opinion mining. In: International Conference on Emerging Intelligent Data and Web Technologies., IEEE

Chen, J., Huang, D. P., Hu, S., Liu, Y., Cai, Y., & Min, H. (2015). An opinion mining framework for Cantonese reviews. *Journal of Ambient Intelligence and Humanized Computing, 6*(5), 541–547.

Chen, X., Ke, L., Lu, Z., Su, H., & Wang, H. (2020). A novel hybrid model for Cantonese rumor detection on Twitter. *Applied Sciences, 10*(20), 7093.

Cheung, L. Y., Ngai, L. W., & Poon, L. M. (2018). *The dictionary of Hong Kong Cantonese*. Cosmo Books.

Chin A (2015) A Linguistics Corpus of Mid-20th Century Hong Kong Cantonese. Department of Linguistics and Modern Language Studies., The Hong Kong Institute of Education., Retrieved 23(3):2015

Choi, H., Kim, J., Joe, S., Min, S., & Gwon, Y. (2021) Analyzing Zero-shot cross-lingual transfer in supervised NLP tasks. In: International Conference on Pattern Recognition., IEEE., pp 9608–9613

Clark K., Luong MT., Le QV., Manning CD (2020) ELECTRA: Pre-training text encoders as discriminators rather than generators. In: Proceedings of the International Conference on Learning Representations

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., & Stoyanov, V. (2020) Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL

Cui Y., Yang Z., Yao X (2023) Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. arXiv preprint arXiv:2304.08177

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio Speech and Language Processing, 29*, 3504–3514.

Dai, N., Liang, J., Qiu, X., & Huang, X. (2019). Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. arXiv preprint arXiv:1905.05621

Dare, M., Fajardo Diaz, V., So, AHZ., Wang, Y., Zhang, S. (2023) Unsupervised Mandarin-Cantonese Machine Translation . arXiv preprint arXiv:2301.03971

De Marneffe, M.C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.D. (2014) Universal Stanford Dependencies: A Cross-linguistic Typology. In: Proceedings of LREC

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). GPT3. int8: 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems, 35*, 318–332.

Devlin J., Chang MW., Lee K., Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL

Ding, H., Zhang, Y., Liu, H., Huang, C.R. (2017) A preliminary phonetic investigation of alphabetic words in Mandarin Chinese. In: Interspeech., pp 3028–3032

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2022). *Ethnologue: Languages of the World*. SIL International.

Eckart de Castilho R., Dore G., Margoni T., Labropoulou P., Gurevych I (2018) A legal perspective on training models for natural language processing. In: Proceedings of LREC

Fung, G., Debosschere, M., Wang, D., Li, B., Zhu, J., & Wong, K.F. (2017) NLPTEA 2017 shared task–Chinese spelling check. In: Proceedings of the IJCNLP Workshop on Natural Language Processing Techniques for Educational Applications

Gao T., Yao X., Chen D (2021) SIMCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of EMNLP

García, O., & Fishman, J. A. (2011). *The multilingual apple: Languages in New York City*. Walter de Gruyter.

Gulzar, M.A., Peng, N., Kim, M. et al. (2022). Sibylvariant transformations for robust text classification. In: Findings of ACL.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In: Proceedings of NAACL-HLT

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass, 10*(9), 397–412.

Hollenstein N., Chersoni E., Jacobs CL., Oseki Y., Prévot L., Santus E. (2022). CMCL 2022 shared task on multilingual and Crosslingual Prediction of Human Reading Behavior. In: Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics

Hollenstein N., Pirovano F., Zhang C., Jäger L., Beinborn L. (2021b). Multilingual language models predict human reading behavior. In: Proceedings of NAACL

Hollenstein, N., Chersoni E., Jacobs CL., Oseki Y., Prévot L., Santus E. (2021a). CMCL 2021 shared task on eye-tracking prediction. In: Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics

Huang, C.R. (2009). Tagged Chinese Gigaword Version 2.0. Linguistic Data Consortium

Huang, C.R., & Chen, K.j. (1992). A Chinese corpus for linguistic research. In: Proceedings of COLING

Huang, G., Gorin, A., Gauvain, J.L., & Lamel, L. (2016). Machine translation based data augmentation for Cantonese keyword spotting. In: IEEE International Conference on Acoustics., Speech and Signal Processing., IEEE., pp 6020–6024

Jiang, AQ., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, DS., Casas, Ddl., Bressand, F., Lengyel, G., Lample, G., & Saulnier, L. et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825

Jin, Z., Jin, D., Mueller, J., Matthews, N., Santus, E. (2019). IMaT: Unsupervised text attribute transfer via iterative matching and translation. In: Proceedings of EMNLP

Johnson, K.A., Babel, M., Fong, I., Yiu, N. (2020) SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English. In: Proceedings of LREC

Ke, L., Chen, X., Lu, Z., Su, H., Wang, H. (2020). A novel approach for cantonese rumor detection based on deep neural network. In: 2020 IEEE International Conference on Systems., Man., and Cybernetics (SMC)., IEEE., pp 1610–1615

Klyueva, N., Long, Y., Huang, CR., & Lu, Q. (2018) Food-related sentiment analysis for Cantonese. In: Proceedings of the PACLIC Joint Workshop on Linguistics and Language Processing

Kwong, O.O .(2015) Toward a corpus of Cantonese Verbal Comments and their classification by multidimensional analysis. In: Proceedings of PACLIC

Lai, H.M. (2004) Becoming Chinese American: A History of Communities and Institutions., vol 13. Rowman Altamira

Lai, R., & Winterstein, G. (2020). Cifu: A frequency Lexicon of Hong Kong Cantonese. In: Proceedings of LREC

Lau, C.M., Chan, G.W.y., Tse RKw., Chan LSy. (2022a). Words.hk: A comprehensive cantonese dictionary dataset with definitions., Translations and transliterated examples. In: Proceedings of the LREC Workshop on Dataset Creation for Lower-Resourced Languages

Lau, M., Zhong, M., Lau, C.M., Su, J., Chan, H., Cheung, B. (2022b). Rime-Cantonese: A Normalized Cantonese Jyutping Lexicon. LDC2022L01. Web Download. Philadelphia: Linguistic Data Consortium.

Lee J., Chen L., Lam C., Lau CM., Tsui, TH., (2022). PyCantonese: Cantonese Linguistics and NLP in Python. In: Proceedings of LREC

Lee, J.S. (2011). Toward a parallel corpus of spoken cantonese and written Chinese. In: Proceedings of IJCNLP

Lee, J., (2019) An emotion detection system for Cantonese. In: Proceedings of FLAIRS

Lee, J., Cai, T., Xie, W., & Xing, L. (2020). A counselling corpus in Cantonese. In: Proceedings of the LREC Joint Workshop on Spoken Language Technologies for Under-resourced languages and Collaboration and Computing for Under-Resourced Languages

Lee, J.S., Liang, B., & Fong, H. (2021). Restatement and question generation for counsellor chatbot. In: Proceedings of the Workshop on NLP for Positive Impact

Lee, C. (2016). *Multilingualism online*. Routledge.

Lee, T., & Wong, C. (1998). CANCORP: The Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique Asie Orientale, 27*(2), 211–228.

Lenci, A. (2023). Understanding natural language understanding systems. A critical analysis. *Sistemi Intelligenti, 2*, 277–302.

Leung, M. T., & Law, S. P. (2001). HKCAC: The Hong Kong Cantonese Adult Language Corpus. *International Journal of Corpus Linguistics, 6*(2), 305–325.

Levy, R. (2008). Expectation-Based Syntactic Comprehension. *Cognition, 106*(3), 1126–1177.

Li, J., Peng, B., Hsu, Y.Y., & Chersoni, E. (2023). Comparing and predicting eye-tracking data of mandarin and Cantonese. In: Proceedings of the EACL Workshop for Similar Languages., Varieties and Dialects

Li, D. C. (2000). Cantonese-English Code-Switching Research in Hong Kong: A Y2K Review. *World Englishes, 19*(3), 305–322.

Li, Q. (2006). *Maritime silk road*. Intercontinental Press.

Li, D. C. (2017). *Multilingual Hong Kong: Languages Literacies and Identities*. Springer.

Li, D. C., & Costa, V. (2009). Punning in Hong Kong Chinese media: Forms and functions. *Journal of Chinese Linguistics, 37*(1), 77–107.

Liesenfeld, A.M. (2018). MYCanCor: A Video Corpus of Spoken Malaysian Cantonese. In: Proceedings of LREC

Liu ,Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. arXiv preprint arXiv:1908.08345

Liu, E.K.Y. (2022) Low-resource neural machine translation: A Case Study of Cantonese. In: Proceedings of the COLING Workshop on NLP for Similar Languages., Varieties and Dialects

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics, 8*, 726–742.

Luke, K. (1995). Between big words and small talk: The Writing System in Cantonese Paperbacks in Hong Kong.

Luke, K. K., & Wong, M. L. (2015). The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics, 25*(2015), 309–330.

Matthews, S., & Yip, V. (2011). *Cantonese: A comprehensive grammar*. Routledge Grammars.

Min, J., McCoy, R.T., Das, D., Pitler, E., & Linzen, T. (2020). Syntactic data augmentation increases robustness to inference heuristics. In: Proceedings of ACL

Misra, K. (2022) minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. arXiv preprint arXiv:2203.13112

Ng RW., Kwan AC., Lee T., Hain., T. (2017). Shefce: A cantonese-english bilingual speech corpus for pronunciation assessment. In: IEEE International Conference on Acoustics., Speech and Signal Processing (ICASSP)., IEEE., pp 5825–5829

Ngai E., Lee M., Choi Y., Chai, P., (2018) Multiple-domain sentiment classification for cantonese using a combined approach. In: Proceedings of PACIS

Nguyen DQ., Vu T., Nguyen, A.T. (2020). BERTweet: a pre-trained language model for English tweets. arXiv preprint arXiv:2005.10200

Nivre J., De Marneffe MC., Ginter F., Goldberg Y., Hajic J., Manning CD., McDonald R., Petrov S., Pyysalo S., Silveira, N. (2016). Universal dependencies v1: A multilingual treebank collection. In: Proceedings of LREC

Ouyang, J. (1993). Putonghua Guangzhouhua De Bijiao Yu Xuexi (The Comparison and Learning of Mandarin and Cantonese)

Pan, J. (2019). The Chinese/English Political Interpreting Corpus (CEPIC): A New Electronic Resource for Translators and Interpreters. In: Proceedings of the Human-Informed Translation and Interpreting Technology Workshop., pp 82–88

Parker R., Graff D., Chen K., Kong J., Kazuaki, M. (2011). Chinese Gigaword. In: Web Download. Philadelphia: Linguistic Data Consortium

Pfeiffer J., Goyal N., Lin XV., Li X., Cross J., Riedel S., Artetxe, M. (2022). Lifting the curse of Multilinguality by pre-training modular transformers. In: Proceedings of NAACL

Pires, T., Schlinger, E., Garrette, D. (2019). How multilingual is multilingual BERT? In: Proceedings of ACL

Ren X., Zhou P., Meng X., Huang X., Wang Y., Wang W., Li P., Zhang X., Podolskiy A., Arshinov G. et al. (2023). PanGu-Σ: Towards Trillion Parameter Language Model with Sparse Heterogeneous Computing. arXiv preprint arXiv:2303.10845

Sachs, G. T., & Li, D. C. (2007). Cantonese as an additional language in Hong Kong. *Multilingua, 26*(95), 130.

Şahin GG., Steedman M (2018) Data augmentation via dependency tree morphing for low-resource languages. In: Proceedings of EMNLP

Salazar, J., Liang, D., Nguyen, TQ., & Kirchhoff, K. (2020) Masked language model scoring. In: Proceedings of ACL

Scao TL., Fan A., Akiki C., Pavlick E., Ilić S., Hesslow D., Castagné R., Luccioni AS., Yvon F., Gallé M. et al. (2022). BLOOM: A 176B-parameter Open-access Multilingual Language Model. arXiv preprint arXiv:2211.05100

Sedghamiz H., Raval S., Santus E., Alhanai T., Ghassemi, M. (2021). SupCL-Seq: Supervised contrastive learning for downstream optimized sequence representations. In: Findings of EMNLP

Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications, 4*(1), 58–70.

Shi, H., Livescu, K., & Gimpel, K. (2021). Substructure substitution: Structured data augmentation for NLP. arXiv preprint arXiv:2101.00411

Shliazhko O., Fenogenova A., Tikhonova M., Mikhailov V., Kozlova A., Shavrina, T. (2022) mGPT: Few-shot learners go multilingual. arXiv preprint arXiv:2204.07580

Sio, J.U.S., Da Costa, L.M. (2019) Building the Cantonese Wordnet. In: Proceedings of the Global WordNet Conference

Snow, D. (2004). *Cantonese as written language: The growth of a written Chinese vernacular*. Hong Kong University Press.

Taori R., Gulrajani I., Zhang T., Dubois Y., Li X., Guestrin C., Liang P., Hashimoto, T.B. (2023). Stanford alpaca: An instruction-following LLaMA model

Touvron H., Lavril T., Izacard G., Martinet X., Lachaux MA., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Edouard G., Lample, G., (2023a). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov ,N., Batra, S., Bhargava, P., & Bhosale, S. et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288

Wang Y., Chen H., Tang Y., Guo T., Han K., Nie Y., Wang X., Hu H., Bai Z., Wang Y. et al (2023) PanGu-π: Enhancing Language Model Architectures via Nonlinearity Compensation. arXiv preprint arXiv:2312.17276

Wang H., Li M., Zhou Z., Fung GPC., Wong, K.F. (2020). KddRES: A multi-level Knowledge-driven dialogue dataset for restaurant towards customized dialogue system. arXiv preprint arXiv:2011.08772

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S.R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1905.00537

Wei J., Zou, K. (2019) EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of EMNLP-IJCNLP

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou ,D., & Metzler, D. et al. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682

Winterstein, G., Vergnaud, D., Lupien, J., Laperle, S., Yu, H., Davis, C., Luk, P.S.Z. (2023). An Empirical., Corpus-based., Approach to Cantonese Nominal Expressions. In: Proceedings of PACLIC

Wong Ts., Gerdes K., Leung H., Lee, J.S. (2017) Quantitative comparative syntax on the cantonese-mandarin parallel dependency treebank. In: Proceedings of Depling

Wong Ts., Lee, J.S. (2018). Register-sensitive translation: A case study of mandarin and cantonese. In: Proceedings of the Conference of the Association for Machine Translation in the Americas., pp 89–96

Wong, P. W. (2006). The Specification of POS Tagging of the Hong Kong University Cantonese Corpus. *International Journal of Technology and Human Interaction, 2*(1), 21–38.

Wu Y., Li X., Lun, S.C. (2006). A structural-based approach to Cantonese-English machine translation. In: International Journal of Computational Linguistics & Chinese Language Processing

Wu, D. (1994). Aligning a parallel English-Chinese corpus statistically with lexical ccriteria. arXiv preprint cmp-lg/9406007

Xiang R., Chersoni E., Long Y., Lu Q., Huang, C.R. (2020a). Lexical data augmentation for text classification in deep learning. In: Canadian Conference on Artificial Intelligence., Springer

Xiang R., Jiao Y., Lu, Q. (2019). Sentiment-augmented attention Network for Cantonese restaurant review analysis. In: Proceedings of KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)

Xiang R., Tan H., Li J., Wan M., Wong, K.F. (2022). When Cantonese NLP Meets Pre-training: Progress and Challenges. In: Proceedings of AACL-IJCNLP: Tutorials

Xiang R., Wan M., Su Q., Huang CR., Lu, Q. (2020b). Sina Mandarin Alphabetical Words: A Web-driven Code-mixing Lexical Resource. In: Proceedings of AACL-IJCNLP

Xiang, R., Chersoni, E., Lu, Q., Huang, C. R., Li, W., & Long, Y. (2021). Lexical data augmentation for sentiment analysis. *Journal of the Association for Information Science and Technology, 72*(11), 1432–1447.

Yang Z., Xu Z., Cui Y., Wang B., Lin M., Wu D., Chen, Z. (2022). CINO: A Chinese minority pre-trained language model. In: Proceedings of COLING

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237

Yip, V., & Matthews, S. (2007). *The bilingual child. Early development and language contact*. Cambridge University Press.

Yu, H. (2013). Mountains of Gold: Canada, North America, and the Cantonese Pacific. *Routledge Handbook of the Chinese Diaspora* (pp. 124–137). Routledge.

Yue-Hashimoto, A. (1991). The Yue dialect. *Journal of Chinese Linguistics Monograph Series, 1*(3), 292–322.

Zhang, X. (1998). Dialect MT: A Case Study between Cantonese and Mandarin. In: Proceedings of COLING

Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of internet restaurant reviews written in Cantonese. *Expert Systems with Applications, 38*(6), 7674–7682.

Zhao WX., Zhou K., Li J., Tang T., Wang X., Hou Y., Min Y., Zhang B., Zhang J., Dong Z. et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223