Cantonese Natural Language Processing in the Transformers Era

Rong Xiang **Ming Liao** **Jing Li**
Department of Computing, Hong Kong Polytechnic University, HKSAR, China
{rongxiang, mliao, jing-amelia.li}@polyu.edu.hk

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*.

1 Introduction

Cantonese, or Yue Chinese, is a diaspora language with over 85 million speakers all over the world (Lai, 2004; García and 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 and 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) (Zhao et al., 2024b; Zhu et al., 2024). 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 (Li et al., 2023; Zhao et al., 2024a).

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; Zhao et al., 2023). 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 Figure 1): only 61 papers are related to "Cantonese", compared to 9,756 papers for English, and 5,312 (4,919 + 393) for SCN/Mandarin.

The history of publications in Cantonese NLP, as in Figure 1, shows that the numbers of papers published yearly remains in single digit, although there is a moderate increasing trend (cf. Figure 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] 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.

Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), pages 69–79
August 16, 2024 ©2024 Association for Computational Linguistics

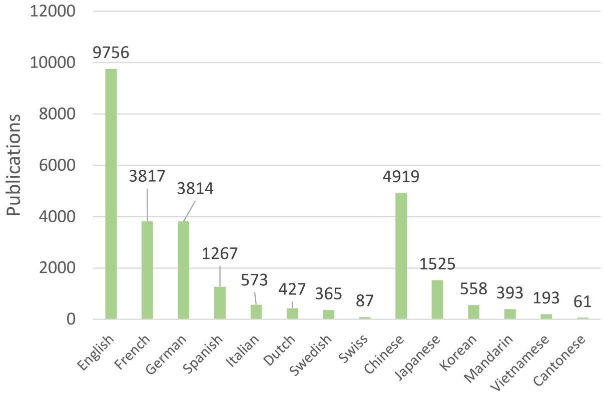

Figure 1: Number of publications in the ACL Anthology indexed by languages as of Mar 2024. The publications were retrieved via searching the language name in either the title or the abstract.

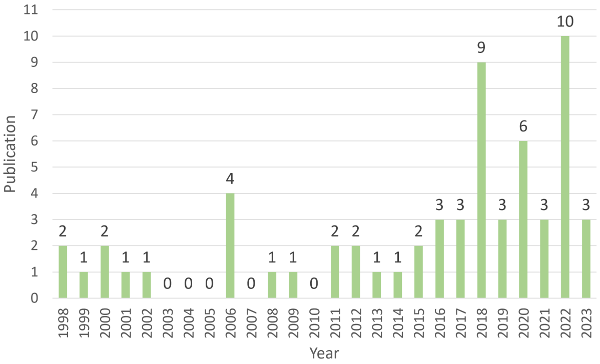

Figure 2: Yearly publications of the 61 papers for Cantonese NLP in the ACL Anthology from 1998 to 2024.

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

Table 1: Papers on Cantonese by research topic (statistics checked on Mar 2024).

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 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.

2 Cantonese NLP Resources

2.1 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 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 CANCORP (Lee 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 and 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 (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 correspond-

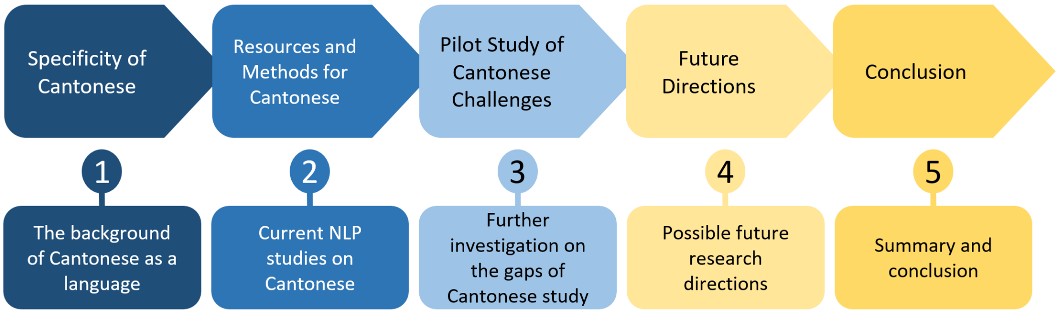
Specificity of Cantonese
Resources and Methods for Cantonese
Pilot Study of Cantonese Challenges
Future Directions
Conclusion
1
2
3
4
5
The background of Cantonese as a language
Current NLP studies on Cantonese
Further investigation on the gaps of Cantonese study
Possible future research directions
Summary and conclusion

Figure 3: Outline of the survey.

ing 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 Marneffe 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. [3]

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, 2018), which has been built with 20 hours 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 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 hours 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] https://hkcc.eduhk.hk/.

2.2 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 OpenRice, 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. [4]

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 [5]. However, the dataset does not provide the information about the language in which a tweet has been written.

For text genre categorization, a benchmark has been collected by the ToastyNews project [6]. 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 Pilot Study for Cantonese

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.

3.1 Colloquialism and Lexical Differences

In the introductory sections, we already discussed how the Cantonese vocabulary deeply diverges from SCN (Ouyang, 1993; Snow, 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”).

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

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

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).[7] 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

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

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

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

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

(B) tokens. Just to give some figures for comparison, 80G 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, 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 and 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 Figure 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 [8], which was trained on Traditional Chinese on a concatenation of a 2020 dump of the Chinese Wikipedia and the Chinese Gigaword Corpus (Huang, 2009; Parker et al., 2011); and the *RoBERTa-HFL* model [9], 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:

$$Surprisal(w) = -logP(w|context) \quad (1)$$

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

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

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

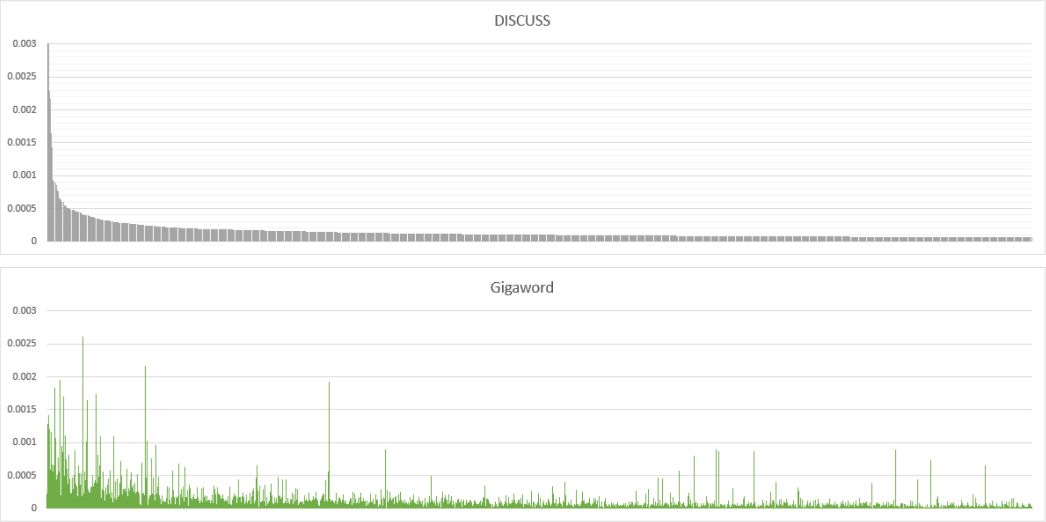

Figure 4: Distribution of bigrams from DISCUSS and Gigaword datasets. The x-axis shows the union dataset of the top 1,000 bigrams from each dataset ordered by the average relative frequency in the two datasets. The top curve refers to DISCUSS, the bottom one to Gigaword.

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 [10]. 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. [11]

We tested both Cantonese sentences and Taiwan Mandarin sentences from the Academia Sinica Corpus (Huang and 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. 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

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

[11] 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.

	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

Table 3: Surprisal analysis on 50 Cantonese and Traditional Chinese sentences. 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.

(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.

3.2 Multilinguality

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%

Table 4: Ratio of language usage. Cantonese and Standard Chinese are dominant in all the datasets under consideration.

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 [12]. 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 [13], 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.*)

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

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

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

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.

4 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 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.

Limitations

The main limitation of this work is that we only conduct our pilot study on limited number of domains since the textual data demands more efforts to clean. In future work, we plan to extend our study in more domains and more specifically focus on multi/cross lingual scenarios.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

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

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*.

Xinyu Chen, Liang Ke, Zhipeng Lu, Hanjian Su, and Haizhou Wang. 2020. A Novel Hybrid Model for Cantonese Rumor Detection on Twitter. *Applied Sciences*, 10(20):7093.

Andy Chin. 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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the 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/ACM Transactions on Audio, Speech and Language Processing*, 29:3504–3514.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford Dependencies: A Cross-linguistic Typology. In *Proceedings of LREC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

David M Eberhard, Gary F Simons, and Charles D Fennig. 2022. *Ethnologue: Languages of the World*. Dallas: SIL International.

Richard Eckart de Castilho, Giulia Dore, Thomas Margoni, Penny Labropoulou, and Iryna Gurevych. 2018. A Legal Perspective on Training Models for Natural Language Processing. In *Proceedings of LREC*.

Gabriel Fung, Maxime Debosschere, Dingmin Wang, Bo Li, Jia Zhu, and Kam-Fai Wong. 2017. NLPTEA 2017 Shared Task–Chinese Spelling Check. In *Proceedings of the IJCNLP Workshop on Natural Language Processing Techniques for Educational Applications*.

Ofelia García and Joshua A Fishman. 2011. *The Multilingual Apple: Languages in New York City*. Walter de Gruyter.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL-HLT*.

John Hale. 2016. Information-Theoretical Complexity Metrics. *Language and Linguistics Compass*, 10(9):397–412.

Chu-Ren Huang. 2009. Tagged Chinese Gigaword Version 2.0. *Linguistic Data Consortium*.

Chu-Ren Huang and Keh-jiann Chen. 1992. A Chinese Corpus for Linguistic Research. In *Proceedings of COLING*.

Guangpu Huang, Arseniy Gorin, Jean-Luc Gauvain, and Lori Lamel. 2016. Machine Translation Based Data Augmentation for Cantonese Keyword Spotting. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6020–6024. IEEE.

Khia A Johnson, Molly Babel, Ivan Fong, and Nancy Yiu. 2020. SpiCE: A New Open-access Corpus of Conversational Bilingual Speech in Cantonese and English. In *Proceedings of LREC*.

Olivia OY Kwong. 2015. Toward a Corpus of Cantonese Verbal Comments and their Classification by Multi-dimensional Analysis. In *Proceedings of PACLIC*.

Him Mark Lai. 2004. *Becoming Chinese American: A History of Communities and Institutions*, volume 13. Rowman Altamira.

Carmen Lee. 2016. *Multilingualism Online*. Routledge.

Jackson Lee, Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022. PyCantonese: Cantonese Linguistics and NLP in Python. In *Proceedings of LREC*.

John Lee, Tianyuan Cai, Wenxiu Xie, and Lam Xing. 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*.

John SY Lee. 2011. Toward a Parallel Corpus of Spoken Cantonese and Written Chinese. In *Proceedings of IJCNLP*.

John SY Lee, Baikun Liang, and Haley Fong. 2021. Restatement and Question Generation for Counsellor Chatbot. In *Proceedings of the Workshop on NLP for Positive Impact*.

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

Man-Tak Leung and Sam-Po Law. 2001. HKCAC: The Hong Kong Cantonese Adult Language Corpus. *International Journal of Corpus Linguistics*, 6(2):305–325.

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

David CS Li. 2017. *Multilingual Hong Kong: Languages, Literacies and Identities*. Springer.

David CS Li and Virginia Costa. 2009. Punning in Hong Kong Chinese Media: Forms and Functions. *Journal of Chinese Linguistics*, 37(1):77–107.

Jiazheng Li, Runcong Zhao, Yongxin Yang, Yulan He, and Lin Gui. 2023. Overprompt: Enhancing chatgpt through efficient in-context learning. *arXiv preprint arXiv:2305.14973*.

Qingxin Li. 2006. *Maritime Silk Road*. China Intercontinental Press.

Andreas Maria Liesenfeld. 2018. MYCanCor: A Video Corpus of Spoken Malaysian Cantonese. In *Proceedings of LREC*.

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. *arXiv preprint arXiv:1907.11692*.

Kang Kwong Luke and May LY Wong. 2015. The Hong Kong Cantonese Corpus: Design and Uses. *Journal of Chinese Linguistics*, 25(2015):309–330.

KK Luke. 1995. Between Big Words and Small Talk: The Writing System in Cantonese Paperbacks in Hong Kong. 香港文化與社會.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Raymond WM Ng, Alvin CM Kwan, Tan Lee, and Thomas Hain. 2017. Shefce: A Cantonese-English Bilingual Speech Corpus for Pronunciation Assessment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5825–5829. IEEE.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A Pre-trained Language Model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, and Natalia Silveira. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of LREC*.

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

Jun Pan. 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*, pages 82–88.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Maeda Kazuaki. 2011. Chinese Gigaword. In *Web Download. Philadelphia: Linguistic Data Consortium*.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of ACL*.

Don Snow. 2004. *Cantonese as Written Language: The Growth of a Written Chinese Vernacular*, volume 1. Hong Kong University Press.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-purpose Language Understanding Systems. *Advances in Neural Information Processing Systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv preprint arXiv:1804.07461*.

Hongru Wang, Min Li, Zimo Zhou, Gabriel Pui Cheong Fung, and Kam-Fai Wong. 2020. KddRES: A Multi-level Knowledge-driven Dialogue Dataset for Restaurant Towards Customized Dialogue System. *arXiv preprint arXiv:2011.08772*.

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

Tak-sum Wong, Kim Gerdes, Herman Leung, and John SY Lee. 2017. Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank. In *Proceedings of Depling*.

Tak-sum Wong and John SY Lee. 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*, pages 89–96.

Dekai Wu. 1994. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. *arXiv preprint cmp-lg/9406007*.

Rong Xiang, Ying Jiao, and Qin Lu. 2019. Sentiment-augmented Attention Network for Cantonese Restaurant Review Analysis. In *Proceedings of KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*.

Rong Xiang, Hanzhuo Tan, Jing Li, Mingyu Wan, and Kam-Fai Wong. 2022. When Cantonese NLP Meets Pre-training: Progress and Challenges. In *Proceedings of AACL-IJCNLP: Tutorials*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

Virginia Yip and Stephen Matthews. 2007. *The Bilingual Child. Early Development and Language Contact*. Cambridge University Press.

Henry Yu. 2013. Mountains of Gold: Canada, North America, and the Cantonese Pacific. In *Routledge Handbook of the Chinese Diaspora*, pages 124–137. Routledge.

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

Runcong Zhao, Lin Gui, and Yulan He. 2023. Cone: Unsupervised contrastive opinion extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1066–1075.

Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. 2024a. Narrativeplay: An automated system for crafting visual worlds in novels for role-playing. In *Proceedings of* *the AAAI Conference on Artificial Intelligence*, volume 38, pages 23859–23861.

Runcong Zhao, Qinglin Zhu, Hainiu Xu, Jiazheng Li, Yuxiang Zhou, Yulan He, and Lin Gui. 2024b. Large language models fall short: Understanding complex relationships in detective narratives. *arXiv preprint arXiv:2402.11051*.

Qinglin Zhu, Runcong Zhao, Jinhua Du, Lin Gui, and Yulan He. 2024. Player*: Enhancing llm-based multi-agent communication and interaction in murder mystery games. *arXiv preprint arXiv:2404.17662*.