

A Novel Method for Semantic Analysis of Cantonese Text Based on Can_Man Dictionary

Huang Chunxiao

School of Chinese Lang&Literature
Yunnan University
Kunming, China
hcxfans@yeah.net

Li Chunyu

School of Math&Information Science
Guangzhou University
Guangzhou, China
Lcyfans@yeah.net

Yao Shaowen

Engineering Research Center of
Cyberspace
Yunnan University
Kunming, China
yaosw@ynu.edu.cn

Bao Mingsuo

School of Chinese Lang&Literature
Yunnan University
Kunming, China
baoms@ynu.edu.cn

Abstract

A novel scheme to address the lack of a corpus for semantic analysis of Cantonese sentences is presented. Scheme need prepare a POS Dictionary, some additional CHAT format files and a Can_Man Dic-tionary. For those words or phrases with unique part-of-speech (commonly known as POS), we established a dictionary which we call POS Dictionary and saved as a JSON file. We export a Cantonese phase file from the example sentences of the Great Dictionary of Hong Kong Cantonese and other corpus sources, annotate Jyutping Romanization and annotate them with POS tags, finally generate the additional CHAT format file. Furthermore, we make 60 CHAT format files from HKCC corpus. And we established a dictionary according to the Hong Kong Cantonese Dic-tionary and saved as a JSON file, which we call Can_Man Dictionary. When parsing Cantonese text, we first execute word segmentation with PyCantonese, then do other tasks including POS tagging with the POS Dictionary and PyCantonese's pos_tagging module, converting the Cantonese text's words into Mandarin words with the Can_Man Dictionary to reconstruct a Mandarin words List and do some semantic parsing task such as Cantonese Abstract Meaning Representation(CanAMR) and Semantic Dependency Parsing(SDP) on the Mandarin word List with Hanlp, visual display all the above parsing results, etc. Test results show the superior performance of the scheme and potential for the parsing Cantonese text.

1 Introduction

©2023 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

The imperfect segmentation on Chinese sentences will misidentify slot boundaries and predict wrong slot categories. To address this issue, Yijin Liu et al. (2019) proposed a character-based method in a joint model at the character level to perform Chinese NLP, achieving state-of-the-art performance.

As one of the most well-known Chinese varieties other than Mandarin, Cantonese’s language data handling and NLP tasks are important for us. Unfortunately, there is only one open source software dedicated to Cantonese currently, namely PyCantonese (Lee, Jackson L. et al., 2022). Based on Pycantonese, Chunxiao Huang et al. (2022) introduce a scheme for the Cantonese text parsing used the method of dynamically expanding corpus, which iterative operations such as word segmentation with PyCantonese, expanding corpus until the segment result is ideal.

While recent advances in NLP have been largely due to neural network-based machine learning coupled with the availability of a large amount of data, the fact that only a small amount of Cantonese data is legally available to Cantonese parsing means that it would be unrealistic for Cantonese parsing to train or include models based on neural networks; for instance, in Universal Dependencies (UD) Treebank, the tree count of Cantonese-HK is 1004, however, the tree count of Chinese-GSDSimp is 4997.

Considering those words in Cantonese vocabulary which are different from Mandarin also have a certain correspondence with Mandarin vocabulary, in order to overcome the shortage of corpus in CanAMR and SDP, in this paper, we propose a new idea, that is, make a JSON file according to The Great Dictionary of Hong Kong Cantonese (Liyan Zhang et al., 2018) and other corpus sources, use the file to convert the corresponding Cantonese words in Cantonese sentences into Mandarin words and reconstruct a Mandarin words List, so that the CanAMR and SDP tasks of Cantonese sentences becomes Chinese Abstract Meaning Representation (Bin Li et al., 2016) and SDP tasks of Mandarin words List. It should be pointed out that although the grammar differences between Cantonese and Mandarin are not significant, there are also some grammar structures in Cantonese that are significantly different from Mandarin. Therefore, the method discussed in this paper is only suitable for semantic analysis of Cantonese sentences and not for syntactic analysis.

2 Preparations

2.1 The POS Dictionary file

For those words or phrases with unique POS, we established a dictionary which we call POS Dictionary(there are about 13000 words or phrases) and saved as a JSON file. Run `train_tagger.py` of PyCantonese to generate new "tagger.pickle" file, and change the constant `_MAX_WORD_LENGTH` from 5 to 18(In order to deal with the Cantonese allegorical sayings, which may contain more than 5 words.). In the phase of POS tagging, first search for a word or phrase in the POS Dictionary, if the word is found, the word’s POS is subject to the POS Dictionary, else use PyCantonese’s `pos_tagging` module to tag the POS. Now we can name the improved PyCantonese as PyCantonese Plus.

2.2 The Additional CHAT Format Files

PyCantonese is built with a high level of usability and transparency in mind. For Cantonese, a focal point of corpora available for research purposes is the CHILDES database (MacWhinney, B, 2000). Given CHAT has been widely used in existing Cantonese corpora for academic research, it is natural for PyCantonese to adopt it as the corpus format.

The Hong Kong Cantonese Corpus (Luke, K. K. and Wong, M. L. Y, 2015), commonly known as HKCanCor, is a Cantonese corpus based on spontaneous speech and radio programs in Hong Kong from the late 1990's. The source data of HKCanCor was in an XML format, and has been converted into the CHAT format. Its transcribed and released version contains about 200,000 Chinese characters. It is included in PyCantonese for the purposes of researching Cantonese. To parse the CHAT format, PyCantonese uses the PyLangAcq package (Lee, J. L. et al., 2016), a Python library for handling CHAT conversational data.

The Corpus of Mid-20th Century Hong Kong Cantonese (Chin, Andy Chion, 2019), known as HKCC, including the first and second phase, developed by The Education University of Hong Kong, the second phase of HKCC (2019) altogether 60 movies were transcribed, with about 770,000 character tokens. These movies are balanced in terms of genre and speakers (including gender) so that the corpus data can represent the Cantonese language spoken in the mid-20th century).

The HKCanCor used by PyCantonese contains insufficient vocabulary, especially the vocabulary commonly used in mainland China, we follow Chunxiao Huang et al. (2022) to consider expanding HKCanCor. We make 60 CHAT format files (Corresponding to the 60 movies) from HKCC V2. Furthermore, we export a Cantonese thesaurus file from the Great Dictionary of Hong Kong Cantonese and other corpus sources, annotate Jyutping romanization for these phase base on CanCLID (2021), and annotate them with POS tags, generate the additional CHAT format file with CHILDES database and the above-mentioned Cantonese phase file, Jyutping file, POS file.

2.3 The Can_Man Dictionary File

Can_Man Dictionary is established according to the Hong Kong Cantonese Dictionary and other corpus sources, are saved as JSON file, which can be called the Can_Man Dictionary file. The Can_Man Dictionary can be used to convert the corresponding Cantonese words(or phrases) into Mandarin words(or phrases).

3 Our Scheme

When parsing a Cantonese sentence, we load the additional CHAT format files to execute operations such as word segmentation, Jyutping and etc. Then we do other task includes POS tagging, CanAMR, SDP, etc. The whole process is shown in Fig.1, and it can be divided into the following steps:

Step 1: Loads our additional CHAT files and input a Cantonese sentence S^C to do word segmentation and Jyutping with PyCantonese Plus tool. We write the results of

word segmentation as $\{ s^C_i \}$. Then loads the POS Dictionary file to do POS tagging for S^C with PyCantonese Plus tool, the result of POS tagging can be written as $\{ p^C_i \}$.

Step 2: Loads the Can_Man Dictionary file whose key is Cantonese words or phrases and value is the corresponding Mandarin words or phrases, etc. The JSON file also can be seen as a Cantonese Dictionary to convert Cantonese word to Mandarin words. When translating polysemy or polysyllabic words in Cantonese, it is necessary to use Jyutping and POS information. Uses these Mandarin words to reconstruct a Mandarin words List $\{ s^M_i \}$.

Step 3: Uses Hanlp (Han He and Jinho D. Choi , 2021) tool with $\{ s^M_i \}$ to get the “penman.Graph” or “CoNLLSentence” results of the CAMR(Chinese Abstract Meaning Representation) and SDP.

Step 4: Converts the CAMR and SDP results with $\{ s^M_i \} \rightarrow \{ s^C_i \}$. Now the result of CAMR and SDP can be written as $\{ G^M_i \}$ and $\{ C^M_i \}$.

Step 5: Visual displays all the above parsing results with graphviz’s Digraph.

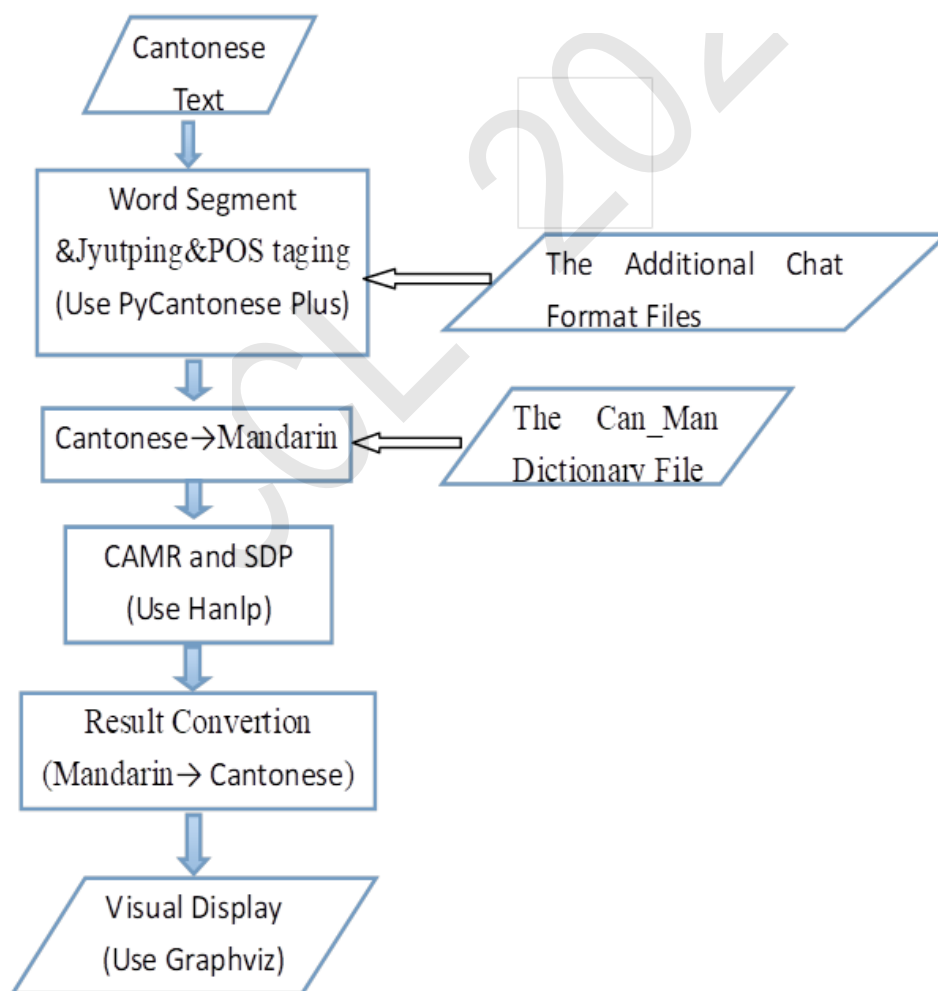


Fig.1. The process of the proposed scheme

4 Test Results

4.1 Words Converting Tests

Our scheme can convert polysemy or polysyllabic words in Cantonese with Jyutping and POS information. Some sample test results are shown in Table 1.

Cantonese	Jyutping	POS	Mandarin	English
蝕底	sit6dai2	verb	吃虧	to suffer losses
蝕底	sit6dai2	adj	虧大啦	to be taken advantage of
伯父	baak8fu2*	noun	老頭兒	old man
伯父	baak8fu4	noun	伯伯	uncle

Table 1: The test results of converting polysemy or polysyllabic words.

4.2 Including Two-Part Allegorical Saying Sentence Tests

When parsing sentences that contains two-part allegorical saying, our scheme treats two-part allegorical saying as a whole. For example, the Cantonese sentence “呢個部門就得我哋三個人，單眼佬睇榜——一眼睇曬。(This department is clear to the three of us at a glance.)” contains two-part allegorical saying “單眼佬睇榜——一眼睇曬 (finish at a glance)”, Hanlp’s CAMR results are as follows:

```
(x1 / 我
  :arg0-of (m1 / mean
            :arg0 1)
  :arg0-of (m2 / mean
            :arg0 (1)
            :arg1 (x4 / 人
                  :arg1-of (i1 / include-91)
                  :quant (x2 / 3)
                  :cunit (x3 / 个))))
```

And our scheme 's CAMR results are as follows:

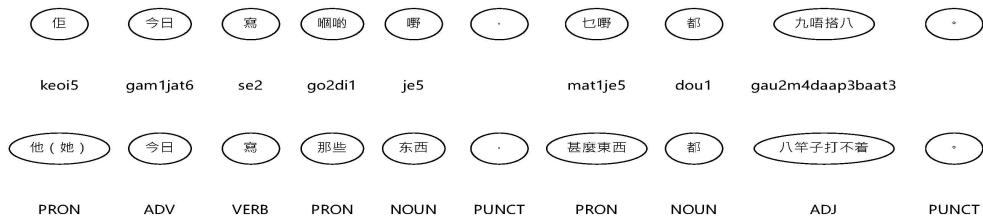
```
(x4 / 得-01
  :domain-of (x3 / 就)
  :arg0 (x2 / 部門
        :domain-of (x1 / 呢個))
  :arg1 (x5 / 我哋
        :arg0-of (m2 / mean)
        :arg0-of (m1 / mean
```

```

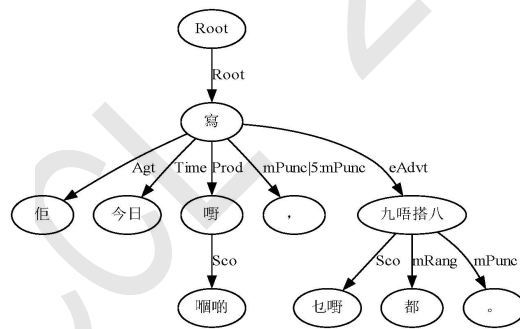
:arg1 (x9 / 單眼佬睇榜——一眼睇曬)
:arg1 (x8 / 人
      :quant (x6 / 3)
      :cunit (x7 / 個)
      :arg1-of m2))))
    
```

4.3 Visual Display Tests

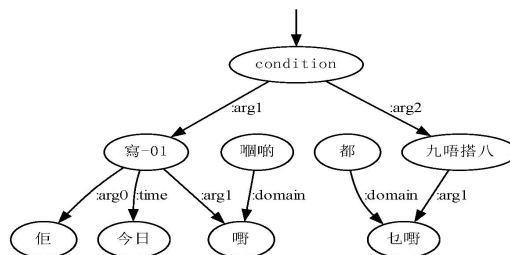
To evaluate the ability of our scheme, we conduct experiments on three Cantonese tasks, a sample Cantonese sentence “佢今日寫啲嘢，乜嘢都九唔搭八。(The things he wrote today are just wrong.)” is parsed with our scheme respectively, the test results are shown in Fig.2.



(a) Results of PyCantonese Plus for the sample Cantonese sentence



(b) SDP Results of the sample Cantonese sentence



(c) CanAMR Results of the sample Cantonese sentence

Fig.2. Test Results of a sample Cantonese sentence

5 Conclusions and Future Work

In this paper, we first generate a POS Dictionary, some additional CHAT format files and a Can_Man Dictionary, then gave a mixed scheme, which includes gloss function in Mandarin for the Cantonese text parsing based on PyCantonese Plus, Hanlp, graphviz etc. Experimental results demonstrated that the scheme has superior performance and potential for the Cantonese linguistics. However, our scheme still has many shortcomings and deficiencies. The possible further work includes providing gloss function in English, embedding CanAMR, SDP, SRL(semantic role labeling) and other semantic analysis functions into PyCantonese Plus.

Acknowledgements

This paper was supported by the Scientific and Technological Plan in Key Fields of Yunnan Province under Grant No. 202202AD080002.

References

- Bin Li, YuanWen, Lijun Bu, Weiguang Qu, Nianwen Xue. 2016. Annotating the Little Prince with Chinese AMRs. *The 10th Linguistic Annotation Workshop*, pages 7-15, Berlin, Germany, August 11, 2016.
- CanCLID. 2021. CanCLID's rime-cantonese Homepage. <https://github.com/rime/rime-cantonese>, last accessed 2023/6/15.
- Chin, Andy Chion. 2019. Initiatives of digital humanities in Cantonese studies: A corpus of mid-twentieth-century Hong Kong Cantonese. In *Anna Tso Wing Bo (Ed.), Digital Humanities and New Ways of Teaching (pp. 71-88)*. Singapore: Springer.
- Chunxiao Huang, Chunyu Li, Shaowen Yao, Ye Ding, Mingsuo Bao, and Kun She. 2022. A Hybrid scheme for Parsing Cantonese Text Based on PyCantonese and PyLTP. *2022 European Conference on Natural Language Processing and Information Retrieval*. IEEE CPS.
- Han He and Jinho D. Choi. 2021. The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. pp. 5555-5577. Association for Computational Linguistics.
- HKCC. 2019. HKCC(Phase 2) Homepage. <https://hkcc.edu.hk/v2/>, last accessed 2023/6/15.
- Lee, Jackson L., Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022. PyCantonese: Cantonese Linguistics and NLP in Python. *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 6607-6611.
- Lee, J. L., Burkholder, R., Flinn, G. B., and Coppess, E. R. 2016. Working with CHAT transcripts in Python. *Technical Report TR-2016-02*. Department of Computer Science, University of Chicago.
- Liyan Zhang, Limei Pan, Liehuai Li. 2018. The Great Dictionary of Hong Kong Cantonese. *Hong Kong Cosmosbooks*, (張勵妍, 潘禮美, 倪列懷. 香港粵語大詞典. 天地圖書. 2018).
- Luke, K. K. and Wong, M. L. Y. 2015. The Hong Kong Cantonese Corpus: Design and Uses. *Journal of Chinese Linguistics Monograph Series*, 25:312-333.
- MacWhinney, B. 2000. The CHILDES Project: Tools for Analyzing Talk. *Lawrence Erlbaum Associates*, Mahwah, NJ, 3rd edition.
- Yijin Liu, Fandong Meng, Jinchao Zhang, Jie Zhou, Yufeng Chen, and Jinan Xu. 2019. Cm-net: A novel collaborative memory network for spoken language understanding. *Proc. EMNLP 2019*, pp. 1050-1059.