2022 European Conference on Natural Language Processing and Information Retrieval (ECNLPIR) | 978-1-6654-7382-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ECNLPIR57021.2022.00021

A Hybrid scheme for Parsing Cantonese Text Based on PyCantonese Plus and PyLTP

Chunxiao Huang
School of Chinese Language and Literature
Yunnan University
Kunming, China
hcxfans@yeah.net

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

Shaowen Yao
National Pilot School of software
Yunnan University
Kunming, China
yaosw@ynu.edu.cn

Ye Ding
School of Cyberspace Security
Dongguan University of Technology
Dongguan, China
dingye@dgut.edu.cn

Mingsuo Bao
School of Chinese Language and Literature
Yunnan University
Kunming, China
baoms@ynu.edu.cn

Kun She
College of Computer Sci&Engineering
Univ. of Electronic Sci&Tech of China
Chengdu, China
kun@uestc.edu.cn

***Abstract*—A novel scheme for the Cantonese text parsing based on the method of dynamically expanding corpus is presented. Scheme need prepare an additional Chat format file, i.e., export a Cantonese phase file from Cantonese Dictionary firstly, which is widely used in Chinese Mainland, then annotate Jyutping romanization online and annotate them with part-of-speech (commonly known as POS) tags, finally generate the additional Chat format file based on CHILDES database and the above-mentioned Cantonese phase file , Jyutping file , POS file, etc. When parsing Cantonese text, we execute iterative operations such as word segmentation with PyCantonese, expanding corpus until the segment result is ideal. Then we do other tasks including POS tagging with our POS dictionary and PyCantonese's pos_tagging module, dependency parsing with PyLTP, and visual display all the above parsing results, etc. Test results show the superior performance of the scheme and potential for the parsing Cantonese text.**

Keywords—Cantonese phase file, POS dictionary, PyCantonese Plus, PyLTP, Parsing Cantonese Text

I. INTRODUCTION

Recently, the need for natural language processing (NLP) has a dramatic increase in many downstream applications. Compared with NLP in English, Chinese NLP faces a unique challenge since it usually needs word segmentation. Nevertheless, the imperfect segmentation performed by the CWS (Chinese Word Segmentation) system will misidentify slot boundaries and predict wrong slot categories, therefore suffers from the error propagation. To address this issue, Liuet al. in [1] proposed a character-based method to perform Chinese NLP in a joint model at the character level, achieving state-of-the-art performance.

As one of the most well-known Chinese varieties other than Mandarin Chinese, Cantonese's language data handling and natural language processing (NLP) tasks are important for us.

Lee, Jackson L. et al. in [2] introduce the Python NLP toolkit PyCantonese for Cantonese language analyzing. PyCantonese's corpora is CHAT format file, can readily access Cantonese corpora from CHILDES database[3],the Hong Kong Cantonese Corpus[4](commonly known as HKCanCor), furthermore, about 170,000 rime-cantonese[5] word-romanization pairs have been incorporated into PyCantonese. As of this writing, the current version of PyCantonese has the following functionality: stop words, handling Jyutping romanization, word segmentation, part-of-speech tagging, and parsing Cantonese text.

In this paper, we slightly modify PyCantonese(named PyCantonese Plus), and use PyCantonese Plus, PyLTP and graphviz comprehensively, propose a novel scheme for the Cantonese text parsing. Test results show the superior performance of the scheme. The fundamental advantage of our scheme lies in the method used to dynamically expand corpus, POS dictionary and visually display the analysis results. This work is a more practical extension of previous research on Cantonese linguistics and natural language processing.

II. PYCANTONESE PLUS

PyCantonese is built with a high level of usability and transparency in mind. The inputs to and outputs from PyCantonese are intuitive Python data structures, for high interoperability with other Python programs. Due to neural network-based machine learning coupled with the availability of a large amount of data, recent advances in NLP have been largely, but the fact that only a small amount of Cantonese data is legally available to PyCantonese, and it would be unrealistic for PyCantonese to train or include models based on neural networks.

PyCantonese's 4 Cantonese processing modules as shown in Table I.

TABLE I. THE 4 MODULES OF PYCANTONESE

Modules	model	corpus
WordSeg	the longest string matching method [6]	-
POSTag	averaged perceptron model[7]	HKCanCor,rime-cantonese
CSW	a list of 104 Cantonese stop words	HKCanCor,CanCLID
Jyutping Romanization	-	HKCanCor,CanCLID

Note:

WordSeg:Word Segmentation;

POSTag:Part-of-Speech Tagging;

CSW:Cantonese Stop Words;

WSD:Word Sense Disambiguation;

"-" represents no corpus is required.

978-1-6654-7382-8/22/$31.00 ©2022 IEEE
DOI 10.1109/ECNLPIR57021.2022.00021

The part-of-speech annotations in the HKCanCor use a tagset of over 100 tags. To facilitate cross-linguistic NLP work, PyCantonese maps a predicted tag to its equivalent from the Universal Dependencies tagset[8](de Marneffe et al., 2021) with a much smaller tagset of 17 tags.

Though PyCantonese can be directly applied for processing the Cantonese texts,it suffers from several limitations. First, it only supports part of Cantonese NLP tasks. For example, it fails to handle dependency parsing analysis, resulting in incomplete analysis in Cantonese NLP. Second, it cannot correctly deal with many allegorical saying(include proverb, some slang, some spoken, some Wisecrack,etc) in Cantonese. Third, the HKCanCor used by PyCantonese contains insufficient vocabulary, especially the vocabulary commonly used in mainland China.

To address the aforementioned issues, we can consider expanding HKCanCor and using PyLTP to do post-parsing. We first export a Cantonese thesaurus file from Hong Kong Cantonese Dictionary, and annotate Jyutping romanization for these phase base on the website (http://hongkongvision.com/tool/cc_py_conv_zh) or CanCLID, and annotate them with part-of-speech(POS) tags, generate the additional Chat format file with CHILDES database and the above-mentioned Cantonese phase file , Jyutping file , POS file. Establish a dictionary for those words or phrases with unique POS, we call it POS dictionary(there are about 6000 words or phrases). Run train_tagger.py to generate new "tagger.pickle" file, and change the constant _MAX_WORD_LENGTH from 5 to 10(In order to deal with the Cantonese Wisecrack, 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 part of speech. Now we can name the improved PyCantonese as PyCantonese Plus.

PyLTP is the python encapsulation of language technology platform [9](Wanxiang Che,2010, commonly known as LTP). LTP is an integrated Chinese processing platform which includes a suite of high performance natural language processing (NLP) modules and relevant corpora.

III. Brief Description Of PyLTP

PyLTP is the python encapsulation of LTP, which uses XML to transfer data through modules and provides all sorts of high performance Chinese processing modules.

There are two key corpora used by LTP: WordMap and CDT. WordMap is a Chinese thesaurus which contains 100,093 words. Each word sense belongs to five-level categories in WordMap. There are 12 top, about 100 second and 1,500 third level,and more fourth and fifth level categories.The CDT corpus, i.e., Chinese Dependency Treebank [10], consists of 10,000 sentences randomly extracted from the first six-month corpus of People's Daily (China) in 1998. CDT's sentences have been annotated with lexical tags, including word segmentation, part-of-speech tagging, and named entity recognition tags. In the LTP's later version, some new models are adopted and Weibo data is also added to some models.

Table II shows the 6 state-of-the-art Chinese processing modules of LTP v3.4.0.

TABLE II. The 6 modules of LTP v3.4.0

Modules	model	corpus
WordSeg	CRF model[11]	PDC corpus
POSTag	SVMTool3[12]	PDC corpus,Weibo
NER	maximum entropy model [13]	PDC corpus
WSD	SVM model[14](Guo et al., 2007)	WordMap
Parser	high order graph-based model[15]	Weibo
SRL	Bi-LSTM, maximum entropy model [13]	-

Note:

WordSeg:Word Segmentation;

POSTag:Part-of-Speech Tagging;

NER:Named Entity Recognition;

WSD:Word Sense Disambiguation;

Parser:Syntactic Parsing;

SRL:Semantic Role Labeling;

Bi-LSTM:Bi-directional Long Short-Term Memory;

PDC:People's Daily (China);

"-" represents no corpus is required.

IV. Our Scheme

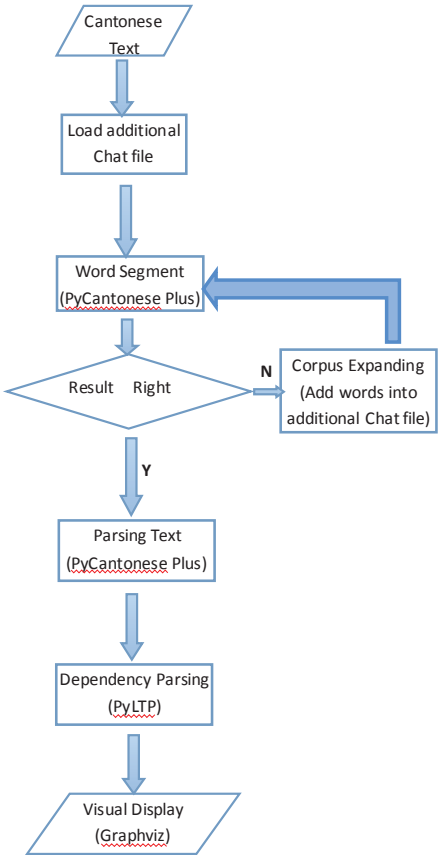
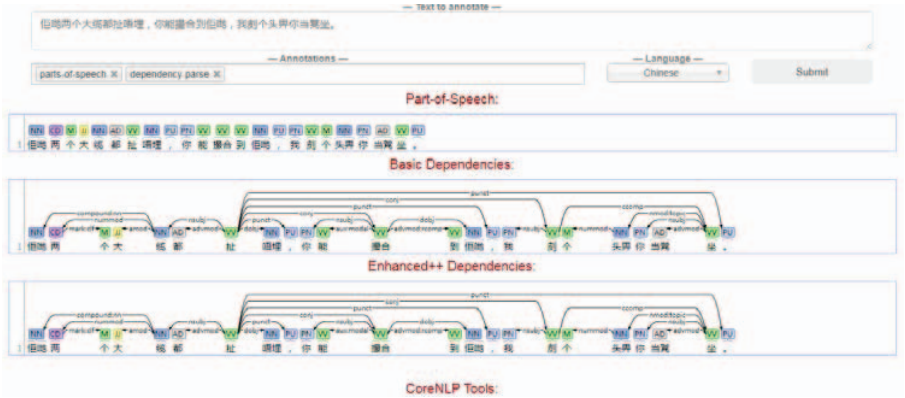
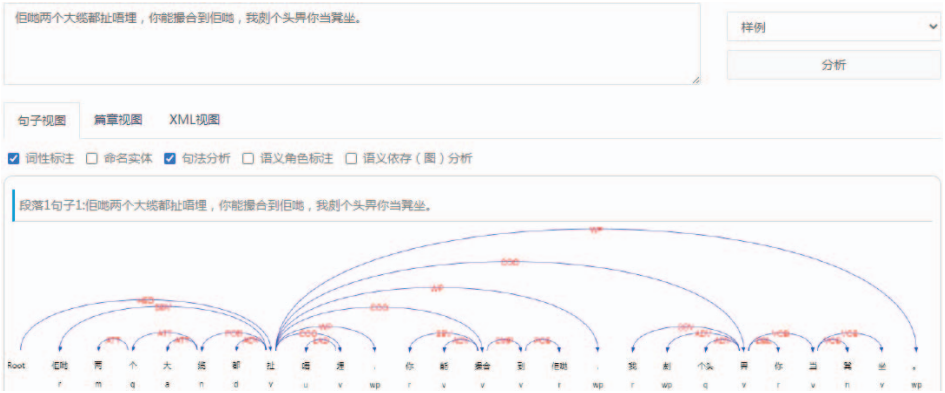


Fig. 1. The process of the proposed scheme

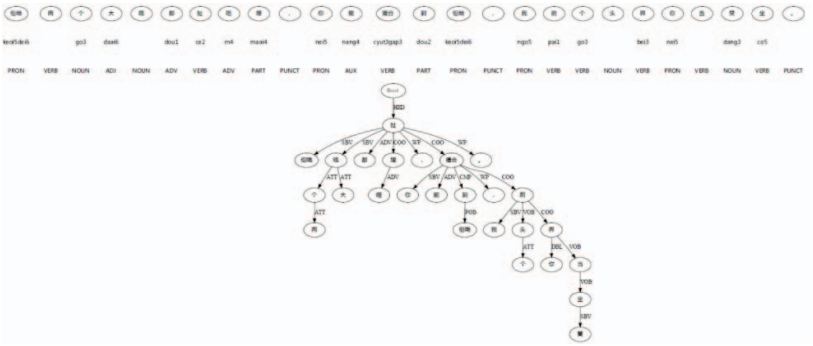
When parsing a Cantonese sentence, we load an additional Chat format file which is made from Cantonese Dictionary phase to execute iterative operations such as word



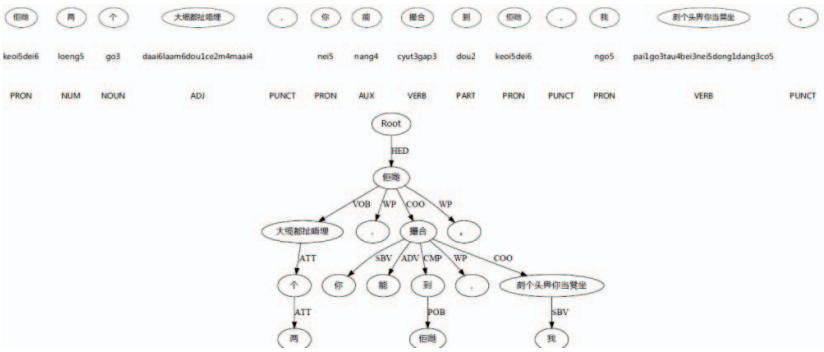
(a). Result of CoreNLP(v 4.4.0)



(b). Result of N-LTP(LTP v4.0)



(c). Result of PyCantonese+PyLTP+graphviz



(d). Result of Our scheme

Fig.2. Test Results

segmentation, expanding corpus until the segment result right. Then we do other task includes part-of-speech tagging, dependency parsing, etc. The whole process is shown in Fig.1,and it can be divided into the following steps:

Step 1: Loads our additional Chat file and input a Cantonese sentence to do word segmentation with PyCantonese Plus tool.

Review the segment result, add some phase into our additional Chat file if the result is not ideal, then do word segmentation with PyCantonese Plus tool again until the segment result is right.

Step 2: Loads our POS dictionary and gets the Cantonese text's token list with PyCantonese Plus tool to access the corresponding phrase's Jyutping and POS results for dependency parsing later.

Step 3: Uses PyLTP tool to get the vector of the dependency parsing result. Extract the dependent parent node and dependency relationship, match the dependent parent node words based on the VectorOfParseResult.

Step 4: Visual display all the above parsing results with graphviz's Digraph.

V. Test Result

To evaluate the ability of our scheme, we conduct experiments on four Cantonese tasks with PyCantonnese, N-LTP[16],CoreNLP[17] and our scheme. The results are shown in Table III, we have the following observations:

- CoreNLP and N-LTP have no ability to mark Cantonese Jyutping, while PyCantonnese and our scheme can.
- When parsing sentences that contains Wisecrack, all of them did not treat Wisecrack as a whole except our scheme, which shows the superiority of our scheme.
- Neither PyCantonnese nor N-LTP can handle some Cantonese words used in Chinese Mainland correctly, while CoreNLP and our scheme can. This is because that CoreNLP and our scheme can consider the shared knowledge with the independently training paradigm.

TABLE III. The results of our scheme's ability comparison to other NLP toolkit

System	Jyutping task	Wisecrack task	Dependency parsing task	Visual display
PyCantonnese	√			
N-LTP(LTP v4.0)			√	√
CoreNLP(v 4.4.0)				√
Our scheme	√	√	√	√

A sample Cantonese sentence "佢哋两个大缆都扯唔埋，你能撮合到佢哋，我剁个头畀你当凳坐。(They are completely mismatched. If you can match them, I will cut off my head and use it as a stool for you.)" is parsed with CoreNLP, N-LTP, PyCantonnese and our scheme respectively, the test results are shown in Fig.2.

VI. Conclusions and Future Work

In this paper, we first generate an additional Chat format file and a POS dictionary, then gave a hybrid scheme for the Cantonese text parsing based on PyCantonese Plus, PyLTP, graphviz etc. Experimental results demonstrated that the scheme has superior performance and potential for the Cantonese linguistics and natural language processing. However, our scheme still has many shortcomings and deficiencies. The possible further work includes providing gloss function in English and Mandarin, embedding NER, WSD,SRL and Syntactic Parsing into PyCantonese.

Acknowledgment

This paper was supported by the Scientific and Technological Plan in Key Fields of Yunnan Province under Grant No. 202202AD080002.And this work is supported in part by the National Natural Science Foundation of China under grant No. 61976051.The authors also wish to acknowledge the assistance and support of the National Natural Science Foundation of China under grant No. 61976051 and No.U19A2067.

References

[1] Yijin Liu, Fandong Meng, Jinchao Zhang, Jie Zhou, Yufeng Chen, and Jinan Xu, "Cm-net: A novel collaborative memory network for spoken language understanding," in Proc. EMNLP, 2019, pp. 1050–1059.

[2] Lee, Jackson L., Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui.PyCantonese: Cantonese Linguistics and NLP in Python.Proceedings of the 13th Language Resources and Evaluation Conference. 2022.

[3] MacWhinney, B.The CHILDES Project:Tools for Analyzing Talk. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition.2000.

[4] 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

[5] CanCLID. (2021). rime-cantonese, https://github.com/rime /rime-cantonese.

[6] Fung, R. and Bigi, B. (2015). Automatic word segmentation for spoken cantonese. In 2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE),pages 196–201. IEEE.

[7] Bird, S., Loper, E., and Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

[8] de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2):255–308.

[9] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08, pp13-16, Beijing, China MacWhinney, B. (2000). The CHILDES Project:Tools for Analyzing Talk. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition.

[10] Liu, Ting, Jinshan Ma, and Sheng Li. 2006. Building a dependency treebank for improving Chinese parser. Journal of Chinese Language and Computing, 16(4):207–224.

[11] Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML 2001, pages 282–289. Morgan Kaufmann, San Francisco, CA.

[12] Wang, Lijie, Wanxiang Che, and Ting Liu. 2009. An SVMTool-based Chinese POS Tagger. Journal of Chinese Information Processing, 23(4):16–22.

[13] Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. Comput.Linguist., 22(1):39–71.

[14] Guo, Yuhang, Wanxiang Che, Yuxuan Hu, Wei Zhang,and Ting Liu. 2007. Hit-ir-wsd: A wsd system for english lexical sample task. In SemEval-2007,pages 165–168.

[15] Che, Wanxiang, Zhenghua Li, Yongqiang Li, Yuhang Guo, Bing Qin, and Ting Liu. 2009. Multilingual dependency-based syntactic and semantic parsing.In CoNLL 2009, pages 49–54, Boulder, Colorado,June.

[16] Che, W., Feng, Y., Qin, L. , & Liu, T. . (2021). N-LTP: An Open-source Neural Language Technology Platform for Chinese. Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

[17] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.