The Penn Chinese Treebank

- Part of Chinese Treebank project(Version 5.0),

  the latest is Version 8.0

- 500,000 words (824,000 Chinese characters) Corpus of Chinese news
- Fully segmented, POS-tagged and syntactically bracketed
- Guidelines for the above are available online
- Available for download, but payable
- Produced by Linguistic Data Consortium (LDC): Martha Palmer, et al

Creation

- Newswire Sources:
  - 698 articles — Xinhua News website(1994-1998)
  - 55 articles — Information Services Department of HKSAR (1997)
  - 132 articles — Sinorama magazine, Taiwan (1996-1998 & 2000-2001)
- Contains 507,222 words, 824,983 Hanzi (Chinese characters),

  18,782 sentences, and 890 data files

Meta-data/Annotation

- Meta-data:
  - Language : Mandarin Chinese
  - Language ID(s): cmn
  - Format: Constituent format
  - Document ID, Date, Created source, e

```
<DOC>
<DOCID>DNPIN.19960212.0364</DOCID>
<HEADER>
<DATE>1996-02-12</DATE>
</HEADER>
<BODY>
<headline>
( (IP-HLN (NP-TPC (NP-PN (NR 中国))
                  (QP (CD 十四)
                      (CLP (M 个)))
                  (NP (NN 边境)
```

- Annotation:
  - All files have been annotated at least twice. The first pass was done by one annotator, and the resulting files were checked by a second annotator. Some files were also double-blind annotated and then adjudicated to create gold standard files.
  - 4versions of files: bracketed, raw, segmented and postagged. bracketed files are sequentially named as : chtb_nnnn.fid (nnnn is sequential file number)

Usage

- Serves as a dictionary for a class-based selection preference sub-model to incorporate external semantic knowledge, by Xiong et al.
- Uses as a corpus in Chinese Syntactic Reordering for Statistical Machine Translation, by Wang et al.
- Helps automatic semantic role labeling for Chinese verbs, by Xue et al.
- Works as a base for a fast, accurate deterministic parser for Chinese, by M Wang et al.

References

- Xue, Naiwen, et al. "The Penn Chinese TreeBank: Phrase structure annotation of a large corpus." *Natural language engineering* 11.02 (2005): 207-238.
- Xiong, Deyi, et al. "Parsing the penn chinese treebank with semantic knowledge." *Natural Language Processing–IJCNLP 2005*. Springer Berlin Heidelberg, 2005. 70-81.
- Xue, Nianwen, and Martha Palmer. "Automatic semantic role labeling for Chinese verbs." *IJCAI*. Vol. 5. 2005.
- Wang, Chao, Michael Collins, and Philipp Koehn. "Chinese Syntactic Reordering for Statistical Machine Translation." *EMNLP-CoNLL*. 2007.
- Wang, Mengqiu, Kenji Sagae, and Teruko Mitamura. "A fast, accurate deterministic parser for Chinese." *Proceedings of the 21st ICCLand the 44th annual meeting of the ACL*. Association for Computational Linguistics, 2006.

ISLRN

Reference	Chinese Treebank 5.0
Date of Submission	Jan. 24, 2014, 4:28 p.m.
Status	accepted
ISLRN	426-628-131-806-1
Resource Type	Primary Text
Media Type	Text
Language	Mandarin Chinese
Access Medium	Distribution: 1 CD
Description	OLAC identifier: oai:www.ldc.upenn.edu:LDC2005T01 Release type: General Non-member fee: 300.00 USD Reduced-license fee: 150.00 USD Extra-copy fee: 150.00 USD Online documentation: http://catalog.ldc.upenn.edu/docs/LDC2005T01 Application: tagging Application: parsing Application: natural language processing Related research project: TIDES Related research project: GALE Membership year: 2005 Data source: newswire
Version	1.0
Creator	Martha Palmer, Fu-Dong Chiou, Nianwen Xue, and Tsan-Kuang Lee
Distributor	Linguistic Data Consortium