11-712: NLP Lab Report

William Yang Wang
ww@cmu.edu

February 14, 2014

Abstract

Dependency parsing is a core task in NLP, and it is a widely used by many applications such as information extraction, question answering, and machine translation. In general, the resources for Chinese dependency parsing are less accessible than English, and publicly available Chinese dependency parsers are still very limited. In this project, the goal is to build a Chinese dependency parser that can be used by others.

1 Basic Information about Chinese Dependency Parsing

Chinese dependency parsing has attracted many interests in the past decade. Bikel and Chiang (2000) and Chiang and Bikel (2002) are among the first to use Penn Chinese Tree Bank for dependency parsing, where they adapted Xia (1999)'s head rules. A few years later, the CoNLL shared task opened a track for multilingual dependency parsing, which also included Chinese (Buchholz and Marsi, 2006; Nilsson et al., 2007). These shared tasks soon popularized Chinese dependency parsing by making datasets available, and there has been growing amount of literature since then (Carreras, 2007; Che et al., 2010; Duan et al., 2007; Nivre et al., 2007; Sagae and Tsujii, 2007; Zhang and Clark, 2008). In this work, we aim at building a new publicly available Chinese dependency parsing tool, using new technologies that aim at improving the accuracy of the state-of-the-art.

2 Past Work on the Syntax of Chinese

Chao (1968) is among the first to study the syntax of Chinese. Unlike English, there has been long debate on the wordhood of Chinese (Duanmu, 1998). Chao and others' work investigate the free and bound forms, prosodic aspects (Shen, 1990), semantics Li (1972); Wu (1999), and morphological aspects (Dai, 1992; Sproat and Shih, 2002; Tang, 1989) to define the unit of word in Chinese. In addition, he has also studied the complex compound constructions (Zhang et al., 2000; Zhou et al., 1999) in Chinese, as well as the parts of speech such as nouns and verbs (Krifka, 1995). More recently, Huang et al. (2009) have studied the lexical and functional categories, argument structure, and the verb phrase in Chinese. Moreover, they have discussed the more unique and challenging parts of syntax in Chinese: the passives, the *ba* construction, and the topic & relative constructions. Interestingly, they have also shed light on some advanced Chinese linguistic issues that have not been well studied in the past: questions, nominal expressions, and anaphora.

Even though there has been many interesting linguistics papers on various aspects of syntax in Chinese, the corresponding computational modeling work has been rather limited. One of the most popular computational tasks in Chinese NLP is word segmentation (Sproat and Emerson, 2003; Xue and Shen, 2003). where researchers have previously investigated sequencial models such as hierarchial hidden Markov model (Zhang et al., 2003), maximum entropy Markov model (Xue and Shen,

2003), and conditional random fields (Zhao et al., 2006) for this task. In addition to tokenization and segmentation, standard structure prediction tasks such as named entity recognition (Wu et al., 2005; Xue and Shen, 2003), part-of-speech tagging (Jiang et al., 2008; Ng and Low, 2004), and constituency parsing (Wang et al., 2006; Wu, 1997) have also been studied in the language-specific setups. As mentioned in Section 1, Chinese dependency parsing was first introduced by Bikel and Chiang (2000), and then became popular after the CoNLL multilingual shared tasks (Buchholz and Marsi, 2006; Nilsson et al., 2007).

In the past decade, there have been growing number of publicly availble Chinese language processing tools. ICTCLAS[1] is one of the most popular word segmentation tool in Chinese NLP. The Stanford Chinese NLP constituency parser (Levy and Manning, 2003), and the dependency parser (Chang et al., 2009) also provide insights for many Chinese NLP applications. More recently, more comphrehensive and Chinese-optimized toolkits were also made available (Che et al., 2010; Qiu et al., 2013). To the best of my knowledge, even though systems such as Malt parser (Nivre et al., 2007) provides solutions to multilingual dependency parsing, but they are not optimized for Chinese, and the accuracy on Penn Chinese Treebank is typically around 70% and lower 80%, which falls behind languages like English and German.

3 Available Resources

After some research and hands-on experiments on real data, I decided to use the open-source Stanford Word Segmenter[2] as the segmentation tool. Comparing to other popular Chinese word segmenters, the Stanford segmenter is well-maintained, and well-documented. The open-source Chinese lexicon I plan to use is also attached in the distribution of Stanford Chinese Segmenter: the Penn Tree Bank lexicon and the PKU lexicon. For the Chinese reference grammar, I am currently investigating the Stanford Dependencies[3], but I am also open to other suggestions.

The test data that I am considering using at this stage: A. a subset of Wang Ling's μtopia Chinese microblog dataset[4]; B. a subset of the test data from HIT-SCIR's LTP test program[5]. I chose the above two datasets because they belong to very different genres: Wang Ling's data is from Sina Weibo, while the second one is taken from newswire headlines. It might be interesting to compare the difficulty in the dependency annotation process, as well as parsing results.

4 Survey of Phenomena in Chinese Dependency Parsing

One notable difference between English and Chinese dependency parsing is the Chinese word segmentation issue, while both English and Chinese parser may also suffer from the issue of part-of-speech tagging errors. However, despite these issues, there are still some interesting phenomena that mark the differences of the two languages:

- Function words are more frequently used in English than in Chinese. For example, when examining Wang Ling's parallel English-Chinese for the total counts of the word "the", there are 2,084 occurances in 2,003 sentences. Whereas in Chinese, there are only 52 occurances of the word "the" out of the 2,003 sentences.

[1] http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/English.html
[2] http://nlp.stanford.edu/software/segmenter.shtml
[3] http://nlp.stanford.edu/software/stanford-dependencies.shtml
[4] http://www.cs.cmu.edu/ lingwang/microtopia/
[5] https://github.com/HIT-SCIR/ltp/blob/master/test_data/test_utf8.txt

- The other interesting thing is the position of the head. In English, it seems the head of the tree occurs more frequent on the left-to-middle of the sentence, while the distribution of the head seems to be more complicated in Chinese. This is also verified using the parallel Weibo data.

- Another well-known issue in Chinese is that Chinese is a pro-drop language. This is extremely prominent in the short text. For example, in the Chinese Weibo data, I have observed the sentence "(If you) Want to eat Chicken, (you) Have to bear the sounds of chickens.".

5 Initial Design

My informant, Lingpeng Kong, and I have been labeling Wang Ling's Weibo data in the past week. We have set up an online annotation environment[6] , using the FUDG and GFL annotation tool introduced by Nathan Schneider. After discussing with my informant, we decide to start with the supervised learning approach. This means that we need to annotation additional datasets for training. Since we have more than 2,000 sentences from Weibo, apart from the test sets, I decide to take the rest of the sentences for additional annotation.

6 System Analysis on Corpus A

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

9 Final Revisions

10 Future Work

References

Daniel M Bikel and David Chiang. Two statistical parsing models applied to the chinese treebank. In *Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 12*, pages 1–6. Association for Computational Linguistics, 2000.

Sabine Buchholz and Erwin Marsi. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics, 2006.

Xavier Carreras. Experiments with a higher-order projective dependency parser. In *EMNLP-CoNLL*, pages 957–961, 2007.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D Manning. Discriminative reordering with chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, pages 51–59. Association for Computational Linguistics, 2009.

Yuen Ren Chao. *A grammar of spoken Chinese*. University of California Pr, 1968.

Wanxiang Che, Zhenghua Li, and Ting Liu. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics, 2010.

[6] http://www.ark.cs.cmu.edu:7788/annotate

David Chiang and Daniel M. Bikel. Recovering latent information in treebanks. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1072228.1072354. URL `http://dx.doi.org/10.3115/1072228.1072354`.

Xiang-Ling Dai. *Chinese morphology and its interface with syntax*. PhD thesis, Ohio State University, 1992.

Xiangyu Duan, Jun Zhao, and Bo Xu. Probabilistic parsing action models for multi-lingual dependency parsing. In *EMNLP-CoNLL*, pages 940–946, 2007.

San Duanmu. Wordhood in chinese. *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, pages 135–196, 1998.

Cheng-Teh James Huang, Yen-hui Audrey Li, and Yafei Li. *The syntax of Chinese*. Cambridge University Press Cambridge, 2009.

Wenbin Jiang, Liang Huang, Qun Liu, and Yajuan Lü. A cascaded linear model for joint chinese word segmentation and part-of-speech tagging. In *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Citeseer, 2008.

Manfred Krifka. Common nouns: A contrastive analysis of chinese and english. *The generic book*, pages 398–411, 1995.

Roger Levy and Christopher Manning. Is it harder to parse chinese, or the chinese treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 439–446. Association for Computational Linguistics, 2003.

Charles Na Li. *Semantics and the structure of compounds in Chinese*. PhD thesis, University of California, Berkeley, 1972.

Hwee Tou Ng and Jin Kiat Low. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *EMNLP*, pages 277–284, 2004.

Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 915–932. sn, 2007.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135, 2007.

Xipeng Qiu, Qi Zhang, and Xuanjing Huang. Fudannlp: A toolkit for chinese natural language processing. In *Proceedings of ACL*. Citeseer, 2013.

Kenji Sagae and Jun'ichi Tsujii. Dependency parsing and domain adaptation with lr models and parser ensembles. In *EMNLP-CoNLL*, volume 2007, pages 1044–1050, 2007.

Xiao-nan Susan Shen. *The Prosody of Mandarin Chinese*, volume 118. University of California Pr, 1990.

Richard Sproat and Thomas Emerson. The first international chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 133–143. Association for Computational Linguistics, 2003.

Richard Sproat and Chilin Shih. Corpus-based methods in chinese morphology. *Tutorial at the 19th COLING*, 2002.

Ting-Chi Tang. Studies on chinese morphology and syntax: 2. *Taipei: Student Book Co*, 1989.

Mengqiu Wang, Kenji Sagae, and Teruko Mitamura. A fast, accurate deterministic parser for chinese. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 425–432. Association for Computational Linguistics, 2006.

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403, 1997.

Jianxin Wu. *Syntax and semantics of quantification in Chinese*. PhD thesis, research directed by

Dept. of Linguistics.University of Maryland, College Park, 1999.

Youzheng Wu, Jun Zhao, Bo Xu, and Hao Yu. Chinese named entity recognition based on multiple features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 427–434. Association for Computational Linguistics, 2005.

Fei Xia. Extracting tree adjoining grammars from bracketed corpora. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS-99)*, pages 398–403, 1999.

Nianwen Xue and Libin Shen. Chinese word segmentation as lmr tagging. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 176–179. Association for Computational Linguistics, 2003.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 184–187. Association for Computational Linguistics, 2003.

Jian Zhang, Jianfeng Gao, and Ming Zhou. Extraction of chinese compound words: an experimental study on a very large corpus. In *Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 12*, pages 132–139. Association for Computational Linguistics, 2000.

Yue Zhang and Stephen Clark. A tale of two parsers: investigating and combining graph-based and transition-based dependency parsing using beam-search. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 562–571. Association for Computational Linguistics, 2008.

Hai Zhao, Chang-Ning Huang, and Mu Li. An improved chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, volume 1082117. Sydney: July, 2006.

Xiaolin Zhou, William Marslen-Wilson, Marcus Taft, and Hua Shu. Morphology, orthography, and phonology reading chinese compound words. *Language and Cognitive Processes*, 14(5-6):525–565, 1999.