# CSE/LIN 467/567: Extra assignment report

Wenfeng Pan, Nannan Zhai

**Abstract**

In this paper, a report has been made mainly on the approaches and methods to covert the Penn Treebank to Universal dependencies in Chinese. The tool we use is the Stanford parser. Stanford parser is basically a lexicalized probability context-free parser, and also uses dependency analysis. Different analysis results can be output according to different grammatical points. Therefore, it can be considered as a parser using a hybrid analysis method. Through syntactic structure analysis, we can analyze the stem of the statement and the relationship between the components. For complex sentences, the correct sentence component relationship cannot be obtained only by part of speech analysis. For example, the subject of the verb predicate ”propose”, we can know that it is ”Li Keqiang”, not the ”outer high bridge” which is more similar to ”propose”.
All codes are available at `https://github.com/pwfee/CSE-567-Extra-Assignment`

## 1 Introduction

For a long time, the study of Chinese natural language processing (NLP) has encountered bottlenecks. One of the important reasons is that Chinese linguistics cannot be transferred to the existing mature deep learning model. This is one of the important reasons why Chinese NLP is more difficult than English. In the study of natural language processing, we can combines deep learning, linguistics and psychology to make up for the shortcomings of traditional Chinese NLP in language understanding through NLU, and achieved good results. Syntactic analysis is also the basic work in natural language processing. It analyzes the syntactic structure of the sentence (the subject-predicate structure) and the interdependence of vocabulary (parallel, dependent, etc.). Through syntactic analysis, it can lay a solid foundation for NLP application scenarios such as semantic analysis, sentiment orientation, and viewpoint extraction. With the use of deep learning in NLP, especially the application of the LSTM model with its own syntactic relationship, syntactic analysis has become less necessary. However, syntactic analysis can still play a large role in long sentences with very complicated syntactic structures and fewer sample labels. Therefore, it is still necessary to study syntactic analysis.

The figure above show the steps of the transform from a regular sentence to Universal dependencies. Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on an evolution of (universal) Stanford dependencies (de Marneffe et al., 2006, 2008, 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). The general philosophy is to provide a

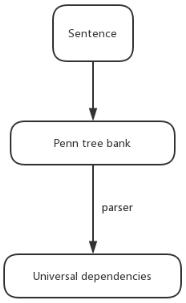


Figure 1: Steps of transform

universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. This is illustrated in the following parallel examples from English, Bulgarian, Czech and Swedish, where the main grammatical relations involving a passive verb, a nominal subject and an oblique agent are the same, but where the concrete grammatical realization varies.

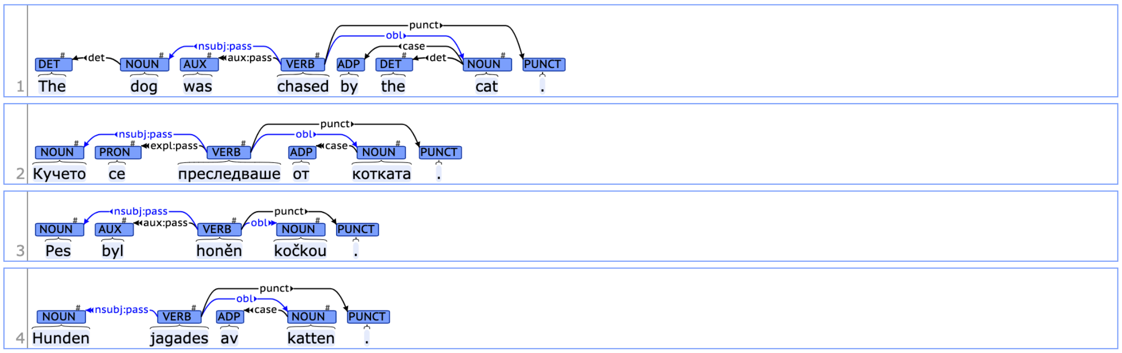

1 The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT
det nsubj:pass aux:pass obl case det punct
2 Кучето се преследваше от котката .
NOUN PRON VERB ADP NOUN PUNCT
nsubj:pass expl:pass obl case punct
3 Pes byl honěn kočkou .
NOUN AUX VERB NOUN PUNCT
nsubj:pass aux:pass obl punct
4 Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT
nsubj:pass obl case punct

Figure 2: UD examples of English,Bulgarian,Czech,Swedish

## 2 Converting rules

After we study UniversalChineseGrammaticalRelations.java [1], we found some rules of different grammatical relations.

[1] From CoreNLP code `https://github.com/stanfordnlp/CoreNLP/blob/master/src/edu/stanford/nlp/trees/international/pennchinese/UniversalChineseGrammaticalRelations.java`

## 2.1 Nominal subject (nsubj)

A nominal subject is a subject which is an noun phrase.

```
Input:
(ROOT
 (IP
   (NP
     (NP (NR 上海) (NR 浦东))
     (NP (NN 开发)
       (CC 与)
       (NN 法制) (NN 建设)))
   (VP (VV 同步))))
Output:
nsubj(同步，建设)
```

regexCompiler

```
"IP <( ( NP|QP=target!< NT ) $+ ( /^VP|VCD|IP/ !< VE !<VC !<SB !<LB !<:NP !<:PP )) !$- BA",
// Handle the case where the subject and object is separated by a comma
"IP <( ( NP|QP=target!< NT ) $+ (PU (<: " + COMMA_PATTERN + " $+ ( /^VP|VCD|IP/ !< VE !<VC !<SB
     !<LB !<:NP !<:PP )))) !$- BA",
// Handle the case where the subject and object is separated by a LCP
"IP <( ( NP|QP=target!< NT ) $+ (LCP ($+ ( /^VP|VCD|IP/ !< VE !<VC !<SB !<LB !<:NP !<:PP )))) !$-
      BA",
// There are a number of cases of NP-SBJ not under IP, and we should try to get some of them as
     this
// pattern does. There are others under CP, especially CP-CND
"NP !$+ VP < ( ( NP|DP|QP=target !< NT ) $+ ( /^VP|VCD/ !<VE !< VC !<SB !<LB))",
"IP < (/^NP/=target $+ (VP < VC))" // Go over copula
```

## 2.2 Nominal passive subject (nsubjpass)

The noun is the subject of a passive sentence.

```
Input:
(IP
 (NP (NN 镍))
 (VP (SB 被)
   (VP (VV 称作)
     (NP (PU “)
       (DNP
         (NP
           (ADJP (JJ 现代))
           (NP (NN 工业)))
         (DEG 的))
       (NP (NN 维生素))
       (PU ” )))))
Output:
nsubjpass(称作-3，镍-1)
```

regexCompiler

```
"IP < (NP=target $+ (VP|IP < SB|LB))");
```

## 2.3 Direct object (dobj)

```
Input:
(IP
 (NP (NR 上海) (NR 浦东))
 (VP
   (VCD (VV 颁布) (VV 实行))
        (AS 了)
        (QP (CD 七十一)
            (CLP (M 件)))
        (NP (NN 法规性) (NN 文件))))

In recent years Shanghai 's Pudong has promulgated and implemented
some regulatory documents.
Output:
dobj(颁布, 文件)
```

regexCompiler

```
"VP < ( /^V*/ $+ NP|DP=target ) !< VC ",
"VP < ( /^V*/ $+ (AS $+ NP|DP=target) ) !< VC ",
" VP < ( /^V*/ $+ NP|DP=target ! $+ NP|DP) !< VC ",
"CP < (IP $++ NP=target ) !<< VC");
```

## 2.4 Adjective modifier (amod)

```
Input:
(NP
 (ADJP (JJ 新))
 (NP (NN 情况)))
Output:
amod(情况-34, 新-33)
```

regexCompiler

```
"NP|CLP|QP < (ADJP=target $++ NP|CLP|QP ) ",
"NP $++ (CP=target << VA !<< VV) > NP ",
"NP < ( CP=target $++ NP << VA !<< VV)",
"NP|QP < ( DNP=target < JJ|ADJP !< NP|QP $++ NP|QP )");
```

## 2.5 Adverbial modifier (advmod)

```
Input:
  (VP
    (ADVP (AD 基本))
    (VP (VV 做到) (AS 了)
Output:
  advmod(做到-74, 基本-73)
```

regexCompiler

```
"VP|ADJP|IP|CP|PP|NP < (ADVP=target !< (AD < /^(\\u4e0d|\\u6CA1|\\u6CA1\\u6709)$/))",
"VP|ADJP < AD|CS=target",
"QP < (ADVP=target $+ QP)",
"QP < ( QP $+ ADVP=target)");
```

# 3 All the result

## 3.1 Constituent parsing results

First, we have to decide the data split scheme.
Refer to Improved Inference for Unlexicalized Parsing[2], we choose this scheme.

| - | TrainSet | DevSet | TestSet |
|---|---|---|---|
| ArticleID | 1-270, 400-1151 | 301-325 | 271-300 |

Then we use Python to extract the from the SGML files ctb5.1/bracketed/chtb$_*$.fid.utf8

Then, we train the model with Chinese TreeBank 5.1.

```
$ java -cp berkeleyParser.jar edu.berkeley.nlp.PCFGLA.GrammarTrainer -path revise/train.txt -out
    chinese.gr --treebank SINGLEFILE
```

After training, we use berkeleyParser test the model with test data.

```
$ java -jar berkeleyParser.jar -gr chinese.gr -inputFile revise/test.txt -outputFile revise/parsed.
    txt
```

Finally, we use evalb to evaluate the performance of the model.

```
$ evalb -p COLLINS.prm revise/train.txt revise/parsed.txt
```

## 3.2 Dependency parsing results

We use Stanford Parser to convert the treebank to universal dependency.

https://nlp.stanford.edu/software/lex-parser.html

```
java -cp "*" -Xmx1g edu.stanford.nlp.trees.international.pennchinese.
    UniversalChineseGrammaticalStructure -checkConnected -basic -keepPunct -conllx -treeFile
    treebank.txt
```

After using parse.py, it automatically separate the data and call the Stanford Parser to convert the treebank.

```
$ python parse.py

[Start]
[Convert][Train Dataset]Filename:chtb_073.fid.utf8 (1/712)
[Convert][Train Dataset]Filename:chtb_215.fid.utf8 (2/712)
[Convert][Train Dataset]Filename:chtb_214.fid.utf8 (3/712)
[Convert][Train Dataset]Filename:chtb_072.fid.utf8 (4/712)
[Convert][Train Dataset]Filename:chtb_741.fid.utf8 (5/712)
```

[2] http://www.coli.uni-saarland.de/~yzhang/rapt-ws1112/papers/petrov_2007.pdf

```
[Convert][Train Dataset]Filename:chtb_527.fid.utf8 (6/712)
[Convert][Train Dataset]Filename:chtb_526.fid.utf8 (7/712)
...

[Convert][Test Dataset]Filename:chtb_665.fid.utf8 (175/178)
[Convert][Test Dataset]Filename:chtb_403.fid.utf8 (176/178)
[Convert][Test Dataset]Filename:chtb_1100.fid.utf8 (177/178)
[Convert][Test Dataset]Filename:chtb_1101.fid.utf8 (178/178)
[End]Data successfully separate
```

```
$ ./udpipe --train model/ctb_ud.model data/train.conllu
```

```
$ ./udpipe --accuracy --parse model/ctb_ud.model data/test.conllu

Parsing from gold tokenization with gold tags - forms: 109239,
UAS: 79.78%, LAS: 76.73%
```

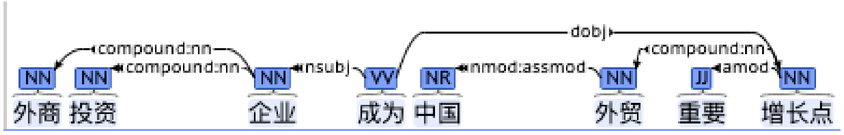
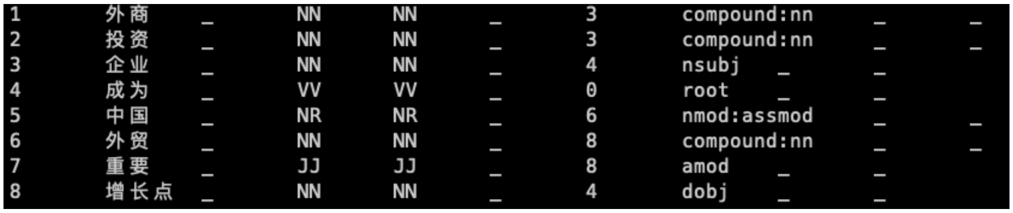


Figure 3: Demo of Chinese Universal Dependencies



```
1    外商    _    NN    NN    _    3    compound:nn    _    _
2    投资    _    NN    NN    _    3    compound:nn    _    _
3    企业    _    NN    NN    _    4    nsubj    _    _
4    成为    _    VV    VV    _    0    root    _    _
5    中国    _    NR    NR    _    6    nmod:assmod    _    _
6    外贸    _    NN    NN    _    8    compound:nn    _    _
7    重要    _    JJ    JJ    _    8    amod    _    _
8    增长点  _    NN    NN    _    4    dobj    _    _
```

Figure 4: Demo of Chinese in CoNLL-U Format

# 4 Discussion and conclusion

In this project, we figure out how to convert the penn treebank to universal dependency in Chinese.

However, we face some problem when using the UDPipe[3]. In the final step of UDPipe parser, it may check the root of each sentence.[4]

After converting the penn treebank to the CoNLL-U Format, we get ”ROOT” as deprel in each sentence, but the UDPipe only accept the root write as ”root”.

[3] https://github.com/ufal/udpipe
[4] https://github.com/ufal/udpipe/blob/master/src/parsito/parser/parser_nn_trainer.cpp

Therefore, it failed everytime when training the model.

## References

[1] Jacqueline Aguilar, Charley Beller, Paul McNamee, and Benjamin Van Durme. *A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards.*. Nicholas Andrews, Jason Eisner, and Mark Dredze.2014.

[2] Pi-Chuan Chang, Michel Galley, and Christopher D Manning *Robust entity clustering via phylogenetic inference. In Association for Computational Linguistics. Optimizing chinese word segmentation for machine translation performance. In Third Workshop on Statistical Machine Translation.* In Third Workshop on Statistical Machine Translation.2014

[3] Annotation guidelines: Events v1.1. Linguistic Data Consortium.2013

[4] Francis Ferraro, Max Thomas, Matthew R. Gormley, Travis Wolfe, Craig Harman, and Benjamin Van Durme. *Concretely Annotated Corpora.* In AKBC Workshop at NIPS.2014.

[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. *Incorporating non-local information into information extraction systems by gibbs sampling.*. In ACL.2005

[6] Dan Klein and Christopher D Manning. *Accurate unlexicalized parsing.* In ACL.2003.

[7] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. *The Stanford CoreNLP natural language processing toolkit.* In ACL: Demos..2014.

[8] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. *Annotated gigaword.* In AKBCWEKEX Workshop at NAACL 2012.

[9] Pontus Stenetorp, Sampo Pyysalo, Goran Topi´c, Sophia Ananiadou, and Akiko Aizawa. *Normalisation with the brat rapid annotation tool.* . In International Symposium on Semantic Mining in Biomedicine.2012

[10] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. *Feature-rich part-of-speech tagging with a cyclic dependency network.*. Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006.

[11] Jacqueline Aguilar, Charley Beller, Paul McNamee, and Benjamin Van Durme. *A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards.*. Nicholas Andrews, Jason Eisner, and Mark Dredze.2014.

[12] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. *Ace 2005 multilingual training corpus ldc2006t06.*.

[13] Mo Yu, Matthew Gormley, and Mark Dredze. *Factor-based compositional embedding models.*. In NIPS Workshop on Learning Semantics.2014.

[14] Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. *The penn chinese treebank: Phrase structure annotation of a large corpus.* Natural language engineering, 11(02):207–238..2005