Character-level Annotation for Chinese Surface-Syntactic Universal Dependencies

Chuanming Dong, Yixuan Li, and Kim Gerdes

Inalco, Paris

Sorbonne Nouvelle
Lattice (CNRS)

Sorbonne Nouvelle
LPP (CNRS)
Almanach (Inria)

Plan

1. Chinese Wordhood
2. Syntactic Parsing for Chinese
3. Enriching Chinese treebanks with word-internal structures
4. Training and parsing on the character level

Chinese Wordhood

- Scriptura Continua
- Chinese Word Segmentation (CWS)
  - Often recognised as the first step for different Chinese NLP tasks
  - Confusing notion of word in modern Chinese

	咖啡 *ka-fei*	**一个 *yi-ge***	**小朋友们 *xiao-peng-you-men***
gloss	(transliterated)	one -quantifier	little -friend-friend -plural
meaning	coffee	one; a/an	children
GB standards	咖啡	一 个	小朋友 们
UD treebanks	咖啡	一 个	小朋友们
segmenters	咖啡	一个 / 一 个	小朋友们 / 小 朋友 们 / …

Syntactic Parsing for Chinese

- Chinese has commonly significantly lower f-scores for parsing than European languages (Dozat & Manning 2017)

		English PTB-SD 3.3.0		Chinese PTB 5.1	
Type	**Model**	**UAS**	**LAS**	**UAS**	**LAS**
Transition	Ballesteros et al. (2016)	93.56	91.42	87.65	86.21
	Andor et al. (2016)	94.61	92.79	–	–
	Kuncoro et al. (2016)	**95.8**	**94.6**	–	–
Graph	Kiperwasser & Goldberg (2016)	93.9	91.9	87.6	86.1
	Cheng et al. (2016)	94.10	91.49	88.1	85.7
	Hashimoto et al. (2016)	94.67	92.90	–	–
	Deep Biaffine	95.74	94.08	**89.30**	**88.23**

Table 4: Results on the English PTB and Chinese PTB parsing datasets

Syntactic Parsing for Chinese

- Chinese has commonly significantly lower f-scores for parsing than European languages (Dozat & Manning 2017)

	Catalan		Chinese		Czech	
Model	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**
Andor et al.	92.67	89.83	84.72	80.85	88.94	84.56
Deep Biaffine	**94.69**	**92.02**	**88.90**	**85.38**	**92.08**	**87.38**

	English		German		Spanish	
Model	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**
Andor et al.	93.22	91.23	90.91	89.15	92.62	89.95
Deep Biaffine	**95.21**	**93.20**	**93.46**	**91.44**	**94.34**	**91.65**

Table 5: Results on the CoNLL '09 shared task datasets

Syntactic Parsing for Chinese

- Previous results on UD 2.0 with character-based segmenter (Shao & al. 2018)

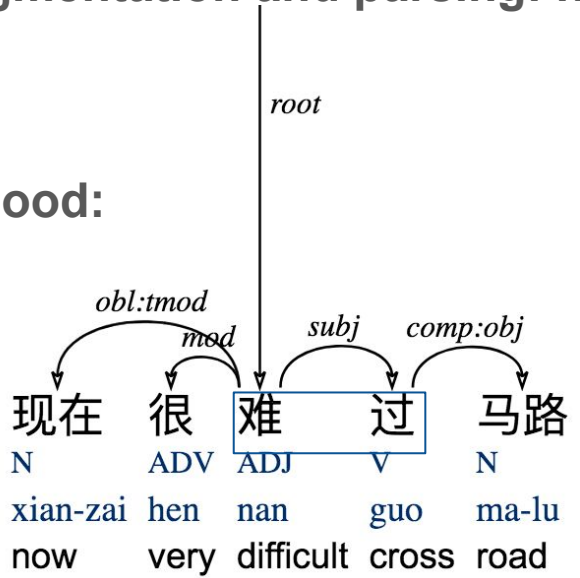
	Segmentation Accuracy		UDPipe parser				Dozat et al. (2017)			
			UAS		LAS		UAS		LAS	
	UDPipe	This Paper	UDPipe	This Paper	UDPipe	This Paper	UDPipe	This Paper	UDPipe	This Paper
Arabic	93.77	97.16	72.34	78.22	66.41	71.79	77.52	83.55	72.89	78.42
Chinese	90.47	93.82	63.20	67.91	59.07	63.31	71.24	76.33	68.20	73.04
Hebrew	85.16	91.01	62.14	71.18	57.82	66.59	67.61	76.39	64.02	72.37
Japanese	92.03	93.77	78.08	81.77	76.73	80.83	80.21	83.79	79.44	82.99
Vietnamese	85.53	87.79	47.72	50.87	43.10	46.03	50.28	53.78	45.54	48.86

Results on different languages from *Universal Word Segmentation: Implementation and Interpretation* (Shao & al. 2018). The parsing accuracies are reported in unlabelled attachment score (UAS) and labelled attachment score (LAS).

Syntactic Parsing for Chinese

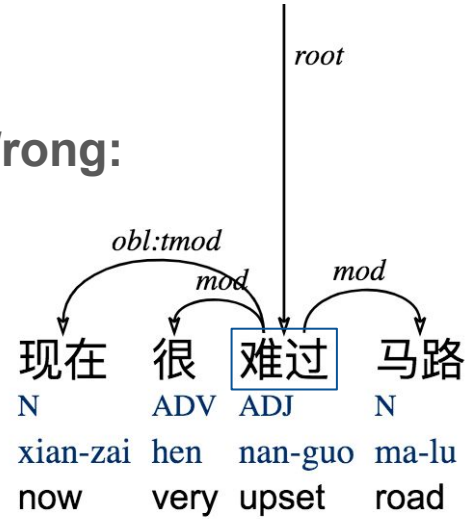
- **Segmentation and parsing: hen-and-egg problem**

Good:



"Now it's difficult to cross the road."

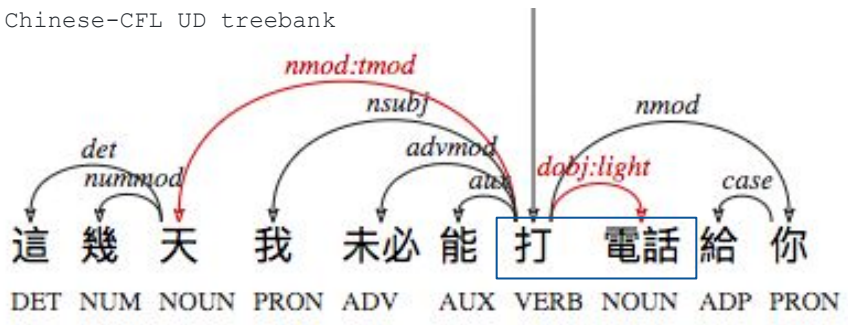
Wrong:



Syntactic Parsing for Chinese

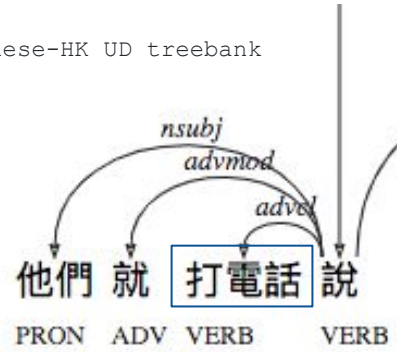
- Segmentation and parsing: hen-and-egg problem
- **Incoherent segmentations in UD corpora**

Chinese-CFL UD treebank



zhe ji tian wo wei-bi neng da dian-hua gei ni
this few day I may_not can hit phone to you
"Maybe I can't call you these days"

Chinese-HK UD treebank



ta-men jiu da-dian-hua shuo
they just call say
"They just called and said..."

Syntactic Parsing for Chinese

- Segmentation and parsing: hen-and-egg problem
- Incoherent segmentations in UD corpora
- **Out-Of-Vocabulary (OOV): worse results on texts with a great quantity of out-of-vocabulary terms (patent texts)**

Synt. Parsing Experiment	LAS	UAS
CoNLL training, CoNLL test	88.21	90.75
CoNLL training, patent test	79.61	84.87

English patent texts
(Burga & al. 2013)

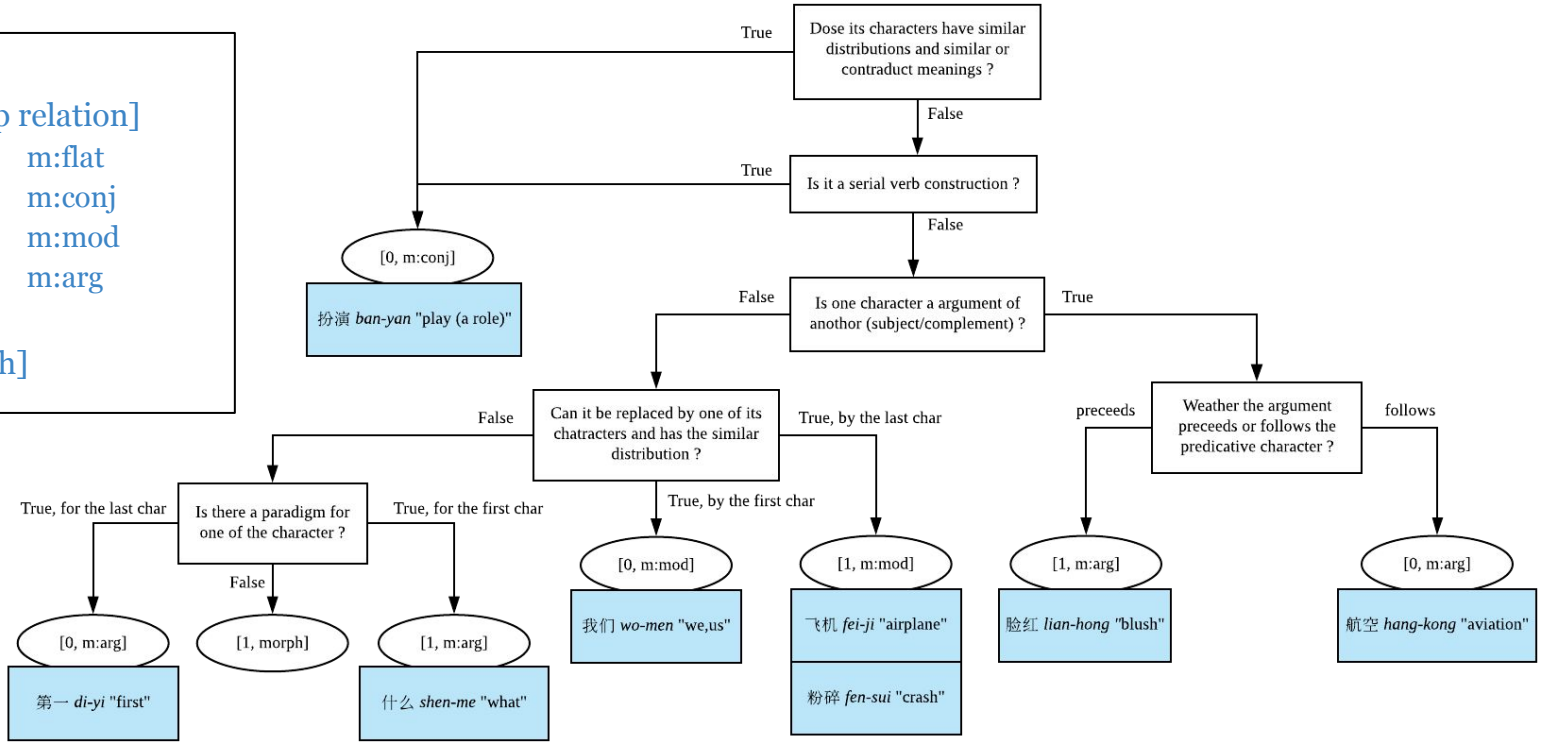
Enriching Chinese treebanks with word-internal structures - Previous works

- Character-level dependencies parsing on Chinese corpus (Zhao 2009; Li & Zhou 2012; Zhang & al. 2014; Li & al. 2018)
    - large-scale annotation on Penn Treebank (PTB) and constituent Chinese Treebank (CTB)
    - usefulness of the word-internal structures in Chinese syntactic parsing

Enriching Chinese treebanks with word-internal structures - Annotation

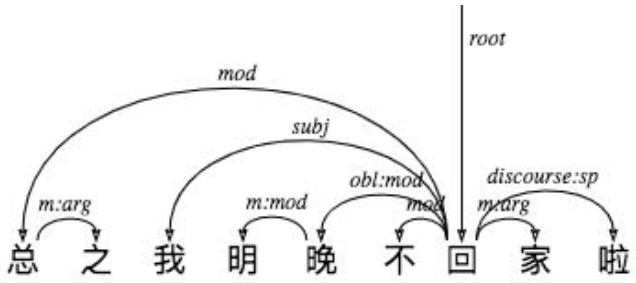
Label:
[head position, dep relation]
- m:flat
- m:conj
- m:mod
- m:arg

Pseudo: [1, morph]



Enriching Chinese treebanks with word-internal structures - Annotation

We annotated the 500 most frequent words
Corpus: all Chinese UD/SUD treebanks (CFL, GSD, HK, PUD)
Annotators: 2 (inter-annotator agreement of 88%)

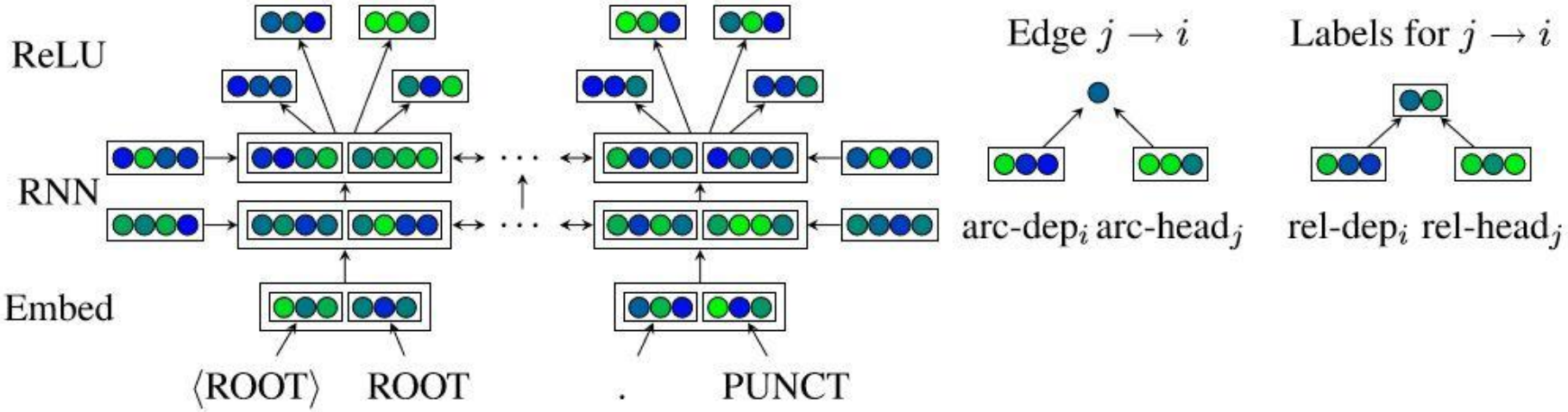


Enriching Chinese treebanks with word-internal structures - Annotation

Tricky examples & Problems:

一般,64,1,m:arg???*
一起, 一直, 一定, 一 样

Training and parsing on the character level



Dozat, (2017)

Training and parsing on the character level - Tagger

Category	Precision	Recall	F-score
ADJ	65.69%	50.00%	56.78%
ADP	63.48%	69.75%	66.47%
ADV	80.08%	76.40%	78.20%
AUX	59.84%	81.56%	69.03%
CCONJ	92.68%	58.46%	71.70%
DET	96.81%	68.94%	80.53%
INTJ	100.00%	0.00%	0.00%
NOUN	88.17%	82.27%	85.12%
NUM	63.92%	98.41%	77.50%
PART	84.03%	91.74%	87.72%
PRON	94.06%	93.14%	93.60%
PROPN	38.17%	89.29%	53.48%
PUNCT	99.84%	99.84%	99.84%
SCONJ	100.00%	0.00%	0.00%
SYM	100.00%	0.00%	0.00%
VERB	76.29%	77.56%	76.92%
TOTAL	**81.85%**	**81.62%**	**81.74%**

F-score of word level POS (UPOS) for our word-based tagger

Category	Precision	Recall	F-score
ADJ	65.52%	42.54%	51.58%
ADP	60.11%	87.90%	71.40%
ADV	75.00%	70.80%	72.84%
AUX	64.71%	86.03%	73.86%
CCONJ	92.68%	58.46%	71.70%
DET	91.22%	86.45%	88.77%
INTJ	100.00%	20.00%	33.33%
NOUN	77.87%	85.56%	81.54%
NUM	65.14%	93.65%	76.84%
PART	91.56%	94.50%	93.00%
PRON	92.47%	88.24%	90.30%
PROPN	54.05%	71.43%	61.54%
PUNCT	99.84%	100.00%	99.92%
SCONJ	20.00%	4.35%	7.14%
SYM	100.00%	100.00%	100.00%
VERB	83.31%	76.41%	79.71%
TOTAL	**88.85%**	**88.70%**	**88.78%**

F-score of word level POS for our character-based tagger after the recombination

Training and parsing on the character level - comparison

	WB	CB
UAS	78.96%	81.72%
OLS	81.29%	85.93%
LAS	66.65%	72.99%

This paper

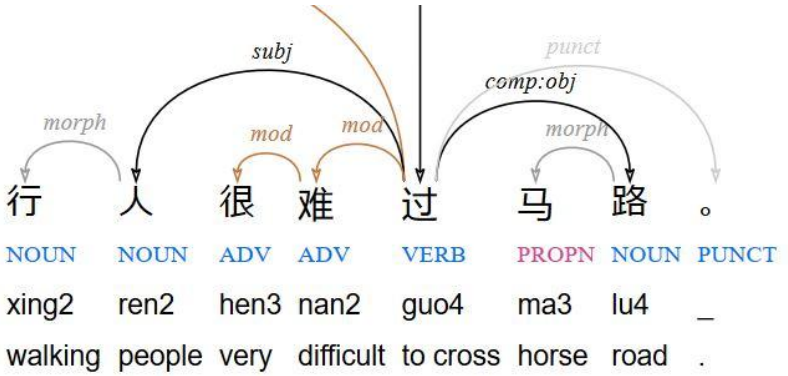
	UPOS	XPOS	UAS	LAS	CLAS
zh	85.26	85.07	68.95	65.88	62.03

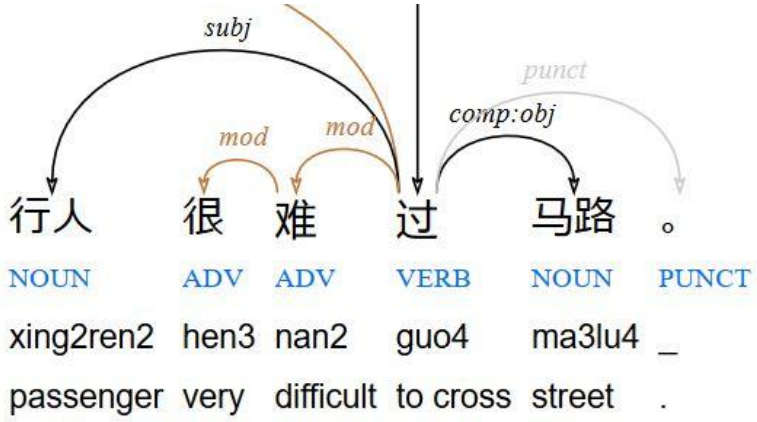
Parsing result on Chinese UD treebanks in Dozat, 2017

Training and parsing on the character level - segmentation

	Morph (Gold)	Deprel (Gold)	TOTAL
Morph	2099	2	2101
Deprel	0	3128	3128
Wrong Head	4	1092	1096
TOTAL	2103	4222	6325

Word segmentation accuracy after recombination: 99.8%





Conclusion

- Possibility to skip the word segmentation preprocessing
- Improvements on parsing using word-internal structures
  - Head position
  - Dependency relation
- High accuracy of detecting internal and external dependency relations
- Future work: regularization of different treebanks with new Chinese SUD annotation guidelines

Thank you for your attention