Character-Level Chinese Dependency Parsing

Meishan Zhang (HIT), Yue Zhang (HIT), Wanxiang Che (HIT), Ting Liu (SUTD)

2014

Kedi LI, Laura DARENNE, Alice WALLARD & Liza FRETEL

M2 TAL Inalco

2023-2024

Que promet l’article ?

Objectif

- contourner l’absence de norme universelle pour la segmentation des mots chinois

- obtenir un arbre de dépendances à la segmentation flexible contenant à la fois une segmentation par mots et une segmentation par caractères

Quelles sont les données utilisées ?

Datasets

Chinese Treebank 5.0, 6.0, 7.0.

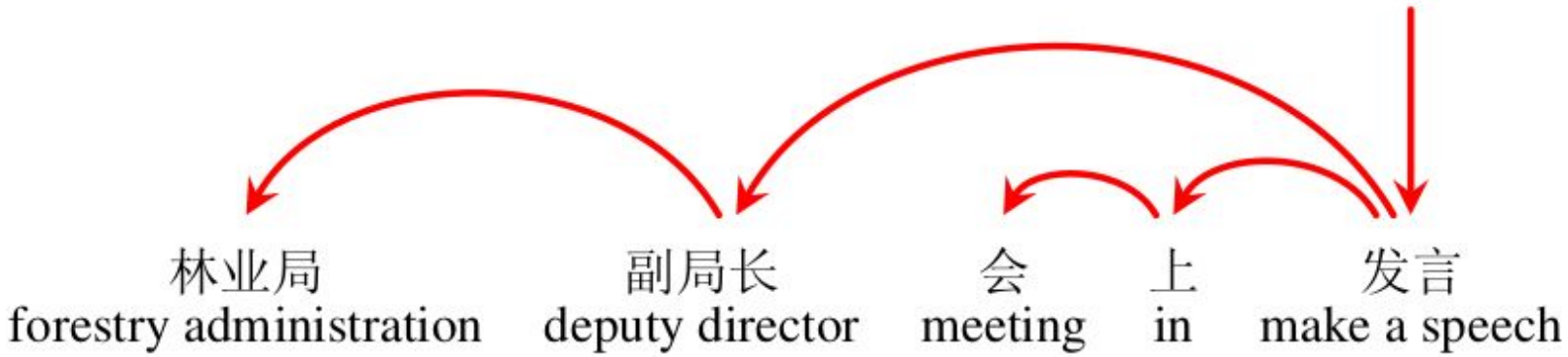
source pour CTB 5.0 et 6.0 : fils de presse (newswire)

source pour CTB 7.0 : fils de presse, blog, magazine d'information ; transcrip--tion oral provenant de la radio

		CTB50	CTB60	CTB70
Training	#sent	18k	23k	31k
	#word	494k	641k	718k
Development	#sent	350	2.1k	10k
	#word	6.8k	60k	237k
	#oov	553	3.3k	13k
Test	#sent	348	2.8k	10k
	#word	8.0k	82k	245k
	#oov	278	4.6k	13k

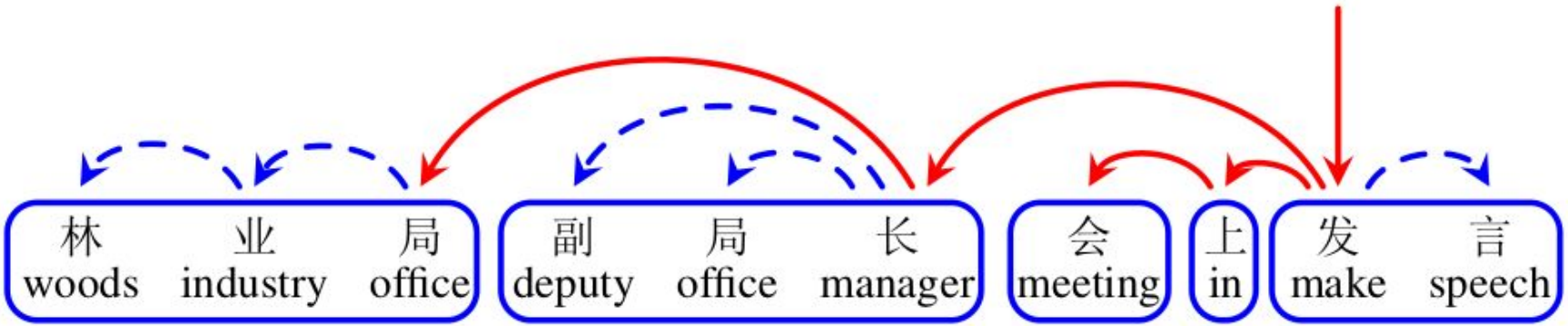
Table 2: Statistics of datasets.

Arbre de dépendance basique



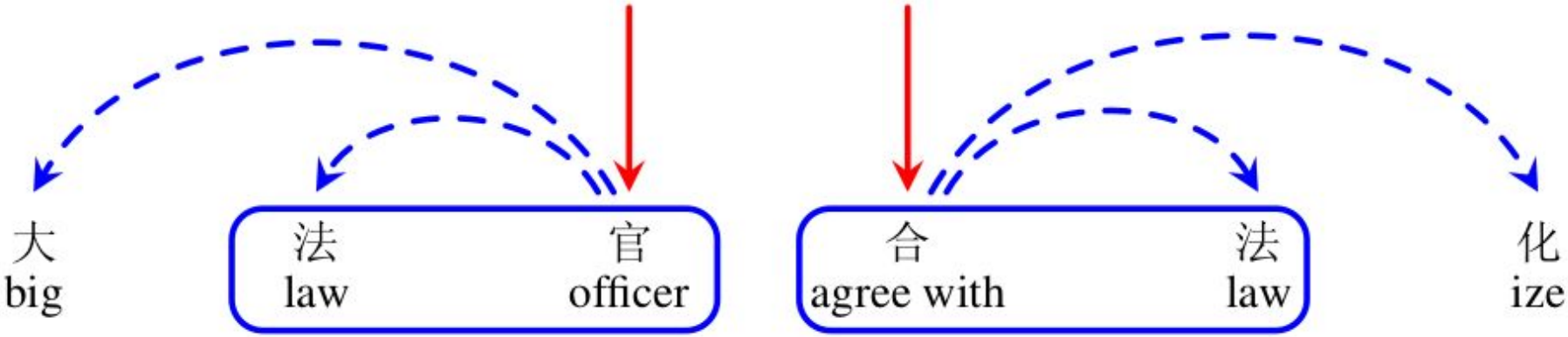
(a) a word-based dependency tree

Arbre de dépendance avec leurs nouveaux modèles



(c) a character-level dependency tree investigated in this paper with both real intra- and inter-word dependencies

Choix du découpage des mots

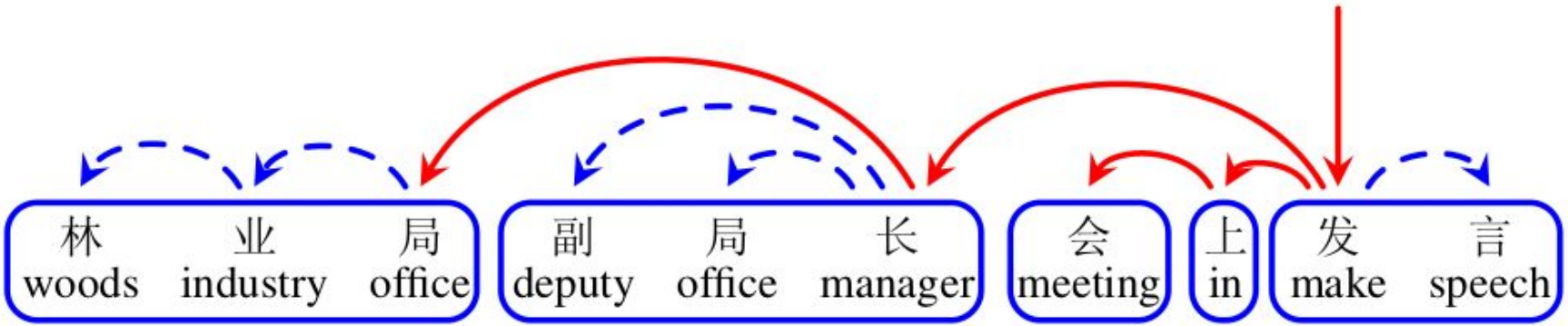


Travaux précédents

- Hai Zhao. 2009. *Character-level dependencies in chinese: Usefulness and learning*. In Proceedings of the EACL, pages 879-887, Athens, Greece, March.
- Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. *Chinese parsing exploiting characters*. In Proceedings of the 51st ACL, pages 125-134, Soa, Bulgaria, August.
- Yue Zhang and Stephen Clark. 2011. *Syntactic processing using the generalized perceptron and beam search*. Computational Linguistics, 37(1):105-151.

Deux systèmes d’analyse des dépendances basées sur les transitions

Arbre de dépendance avec leurs nouveaux modèles



(c) a character-level dependency tree investigated in this paper with both real intra- and inter-word dependencies

The arc-standard parser

step	action	stack	queue	dependencies
0	-	ϕ	林 业 ···	ϕ
1	$SH_w(NR)$	林/NR	业 局 ···	ϕ
2	SH_c	林/NR 业/NR	局 副 ···	ϕ
3	AL_c	业/NR	局 副 ···	$A_1 = \{林 \frown 业\}$
4	SH_c	业/NR 局/NR	副 局 ···	A_1
5	AL_c	局/NR	副 局 ···	$A_2 = A_1 \bigcup \{业 \frown 局\}$
6	PW	林业局/NR	副 局 ···	A_2
7	$SH_w(NN)$	林业局/NR 副/NN	局 长 ···	A_2
...	...	...	...	...
12	PW	林业局/NR 副局长/NN	会 上 ···	A_i
13	AL_w	副局长/NN	会 上 ···	$A_{i+1} = A_i \bigcup \{林业局/NR \frown 副局长/NN\}$
...	...	...	...	...

(a) character-level dependency parsing using the arc-standard algorithm

Construction des dépendances

état de transition

=

une pile (stack) + une file d'attente (queue)

- la pile : arbre de dépendance séquencé partiellement analysé
- la file d'attente : mots non traités.

Actions définissant les changements d’état entre les mots

- ALw arc-left
- ARw arc-right
- PR pop-root
- SHw last shift

Actions définissant les changements d’état au sein du mot

- ALc intra-word arc-left
- ARc intra-word arc-right
- PW pop-word
- SHc inter-word shift

The arc-eager parser

step	action	stack	deque	queue	dependencies
0	-	ϕ		林 业 $\cdots$	
1	SH_c(NR)	ϕ	林/NR	业 局 $\cdots$	ϕ
2	AL_c	ϕ	ϕ	业/NR 局 $\cdots$	$A_1 = \{\text{林}^\frown\text{业}\}$
3	SH_c	ϕ	业/NR	局 副 $\cdots$	A_1
4	AL_c	ϕ	ϕ	局/NR 副 $\cdots$	$A_2 = A_1 \bigcup \{\text{业}^\frown\text{局}\}$
5	SH_c	ϕ	局/NR	副 局 $\cdots$	A_2
6	PW	ϕ	林业局/NR	副 局 $\cdots$	A_2
7	SH_w	林业局/NR	ϕ	副 局 $\cdots$	A_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
13	PW	林业局/NR	副局长/NN	会 上 $\cdots$	A_i
14	AL_w	ϕ	副局长/NN	会 上 $\cdots$	$A_{i+1} = A_i \bigcup \{\text{林业局/NR}^\frown\text{副局长/NN}\}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

(b) character-level dependency parsing using the arc-eager algorithm, $t = 1$

Figure 3: Character-level dependency parsing of the sentence in Figure 1(c).

Construction des dépendances

état de transition

=

une pile (stack) + deux files d'attente (queue + deque)

Actions définissant les changements d’état entre les mots

- ALw arc-left
- ARw arc-right
- PR pop-root
- SHw last shift

Actions définissant les changements d’état au sein du mot

- ALc intra-word arc-left
- ARc intra-word arc-right
- PW pop-word
- SHc inter-word shift

L’expérience

Objectif : annotation en dépendance

à partir :

- analyseur syntaxique basé sur les transitions et règles de détermination de la tête des structures de mots : Zhang et Clark, 2009

- annotations des caractères intra-mots et : Zhang et al., 2013

corpus de référence : Chinese Penn tree Bank (CTB)

- annoté manuellement

métrique :

- précision
- rappel
- F1

-> segmentation, étiquetage des POS, analyse des liens de dépendance

-> analyse des liens de dépendance à l’intérieur des mots

Les modèles

Modèles de base (baseline)

2 chaînes de traitement : segmentation liées + POS + analyse syntaxique en dépendance des mots

- arc-standard (STD)
- arc-eager (EAG)

<u>associées</u> à des types d'annotations :

- extraites
- obtenues du modèle d'annotation Hatori et al. (2012)

-> Modèles étudiés

chaîne de traitement	annotation intra-mot	annotation inter-mots
STD	real	pseudo
STD	pseudo	real
STD	real	real
EAG	real	pseudo
EAG	pseudo	real
EAG	real	real

STD (real, real)	SEG	POS	DEP	WS
$\alpha = 1$	95.85	91.60	76.96	95.14
$\alpha = 2$	96.09	91.89	77.28	**95.29**
$\alpha = 3$	96.02	91.84	77.22	95.23
$\alpha = \mathbf{4}$	**96.10**	**91.96**	**77.49**	**95.29**
$\alpha = 5$	96.07	91.90	77.31	95.21

Table 3: Development test results of the character-level arc-standard model on CTB60.

EAG (real, real)		SEG	POS	DEP	WS
$\alpha = 1$	$t = 1$	**96.00**	91.66	74.63	**95.49**
	$t = 2$	95.93	91.75	76.60	95.37
	$\mathbf{t = 3}$	95.93	**91.74**	**76.94**	95.36
	$t = 4$	95.91	91.71	76.82	95.33
	$t = 5$	95.95	91.73	76.84	95.40
$t = 3$	$\alpha = 1$	95.93	91.74	76.94	95.36
	$\alpha = 2$	96.11	91.99	77.17	95.56
	$\alpha = \mathbf{3}$	**96.16**	**92.01**	**77.48**	**95.62**
	$\alpha = 4$	96.11	91.93	77.40	95.53
	$\alpha = 5$	96.00	91.84	77.10	95.43

Table 4: Development test results of the character-level arc-eager model on CTB60.

	SEG	POS	DEP	WS
STD (real, real)	96.10	91.96	77.49	95.29
STD (real, real)/wo	95.99	91.79	77.19	95.35
Δ	-0.11	-0.17	-0.30	+0.06
EAG (real, real)	96.16	92.01	77.48	95.62
EAG (real, real)/wo	96.09	91.82	77.12	95.56
Δ	-0.07	-0.19	-0.36	-0.06

Table 5: Feature ablation tests for the novel word-structure features, where "/wo" denotes the corresponding models without the novel intra-word dependency features.

Résultats

Model	CTB50				CTB60				CTB70			
	SEG	POS	DEP	WS	SEG	POS	DEP	WS	SEG	POS	DEP	WS
The arc-standard models												
STD (pipe)	97.53	93.28	79.72	–	95.32	90.65	75.35	–	95.23	89.92	73.93	–
STD (real, pseudo)	97.78	93.74	–	**97.40**	**95.77**‡	91.24‡	–	**95.08**	**95.59**‡	90.49‡	–	**94.97**
STD (pseudo, real)	97.67	94.28‡	81.63‡	–	95.63‡	**91.40**‡	76.75‡	–	95.53‡	90.75‡	75.63‡	–
STD (real, real)	**97.84**	**94.62**‡	**82.14**‡	97.30	95.56‡	91.39‡	**77.09**‡	94.80	95.51‡	**90.76**‡	**75.70**‡	94.78
Hatori+ '12	97.75	94.33	81.56	–	95.26	91.06	75.93	–	95.27	90.53	74.73	–
The arc-eager models												
EAG (pipe)	97.53	93.28	79.59	–	95.32	90.65	74.98	–	95.23	89.92	73.46	–
EAG (real, pseudo)	97.75	93.88	–	97.45	95.63‡	91.07‡	–	95.06	**95.50**‡	90.36‡	–	**95.00**
EAG (pseudo, real)	97.76	**94.36**‡	81.70‡	–	95.63‡	91.34‡	76.87‡	–	95.39‡	90.56‡	75.56‡	–
EAG (real, real)	**97.84**	**94.36**‡	**82.07**‡	**97.49**	**95.71**‡	**91.51**‡	**76.99**‡	**95.16**	95.47‡	**90.72**‡	**75.76**‡	94.94

Table 6: Main results, where the results marked with ‡ denote that the p-value is less than 0.001 compared with the pipeline word-based models using pairwise t-test.

OOV

---

rappel :

- STD **67,98%**
- EAG **69,01%**

précision :

- STD **87,64%**
- EAG **89,07%**