

Character-level Annotation for Chinese Surface-Syntactic Universal Dependencies

Chuanming Dong (Inalco), Yixuan Li (CNRS & Paris 3), Kim Gerdes (CNRS
& Paris 3)
2019

Kedi LI, Laura DARENNE, Alice WALLARD
& Liza FRETTEL

M2 TAL Inalco
2023-2024

Problématique

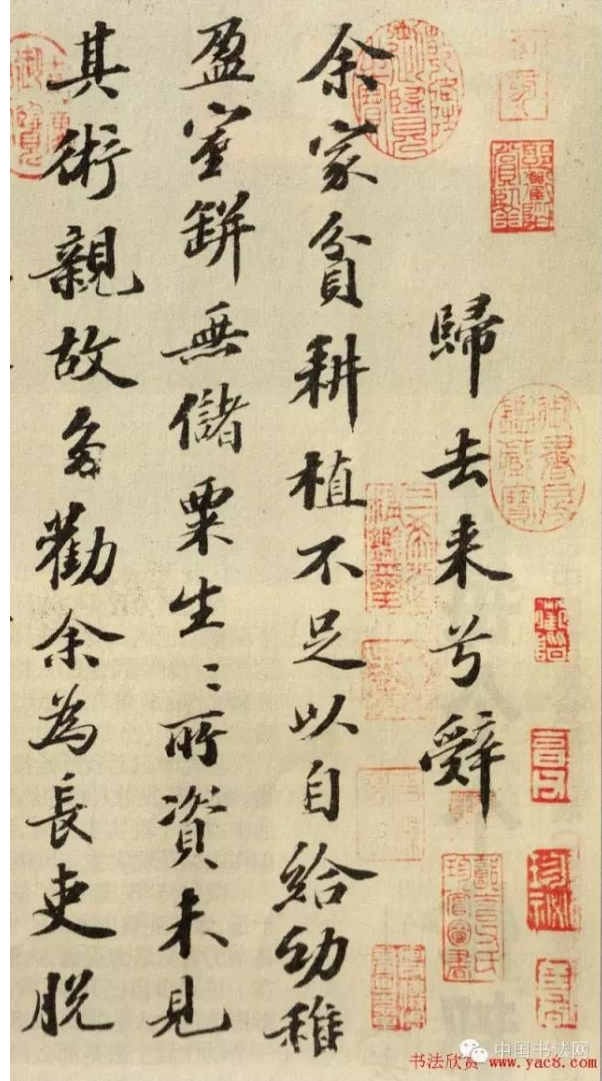
Le chinois est une *écriture continue*.

La *tokenisation* est le premier obstacle du traitement automatique du chinois.

中文太难了，学习中文的人经常分不清单词之间的分界。

中文太难了，学习中文的人经常分不清单词之间的分界。

Si la tokenisation se trompe, cela créera beaucoup d'erreurs,
et ce **jusqu'à la fin de la chaîne de traitement !**



Solution

Se débarrasser de cette étape \Rightarrow considérons chaque caractère comme un token, et passons à l'étape suivante !



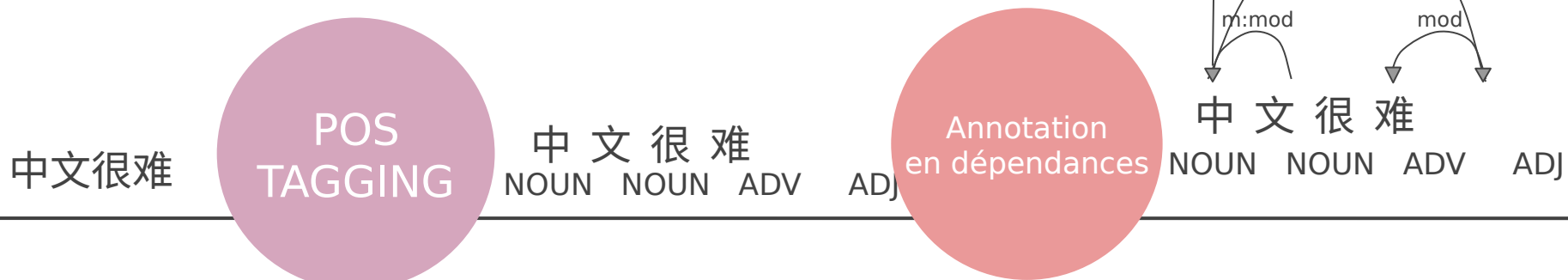
Ce qu'on faisait avant...



Ce qu'on faisait avant...



Et maintenant



Chinese words

Mots simples

monosyllabique : 花 hua (fleur)

polysyllabique : 巧克力 qiaokeli (chocolat)

Mots compliqués polysyllabique : 桃花 taohua (fleur de pêcher)
(plupart de mots chinois)

桃花



桃 花

Schéma d'annotations

relations de dépendance

— — —

m:mod Quand le sens du gouverneur est modifié par le gouverné.

ex: 中 < m:mod 国 (Centre < m:mod pays = Chine)

m:conj Relation de conjonction: quasiment le même sens.

ex: 自 > m:conj 己 (Soi > m:conj soi = Soi-même)

m:arg Relation de sujet-prédicat (souvent entre V et N).

ex: 毕 > m:arg 业 (Finir > m:arg études = être diplômé)

m:flat Constructions sans tête, relations inconnues, translittérations

ex: 巴 > m:flat 黎 (Ba > m:flat Li = Paris)

Les étiquettes morphe (m) permettent de reconstituer les mots. Mais on utilise plutôt la colonne XPOS pour les reconstituer.

Schéma d'annotations

relations tête-fille

Trois positions de la tête:

- Gaucher
- Droitier
- Coordination

Tester les relations:

1. Le caractère ajouté modifie-t-il l'ensemble de la distribution ?
2. Les caractères individuels ont-ils le même POS que le mot entier ?
3. Peut-on grammaticalement inverser les caractères d'un mot (pour tester la relation de coordination) ?
4. ...

我们

wo men
je, moi pluriel
□ nous

Head-modifier

现代化

xiandai hua
moderne -iser
□ moderniser

Modifier-head

Schéma d'annotations

Annotation POS des caractères

Comment annoter le POS de chaque caractère ?

FORM	POS:char1	POS:char2	POS:char3	POS:char4	...	POS:word	Frequency
电影 dian-ying 'film'	NOUN	NOUN	-	-	...	(NOUN)	96
发展 fa-zhan 'development'	VERB	VERB	-	-	...	(VERB)	95
平方公里 ping-fang-gong-li 'square kilometer'	NOUN	NOUN	NOUN	NOUN	...	(NOUN)	90

Table 1 Character POS Dictionary



In short, I will not go home tomorrow night.

Figure 1 word-based treebank

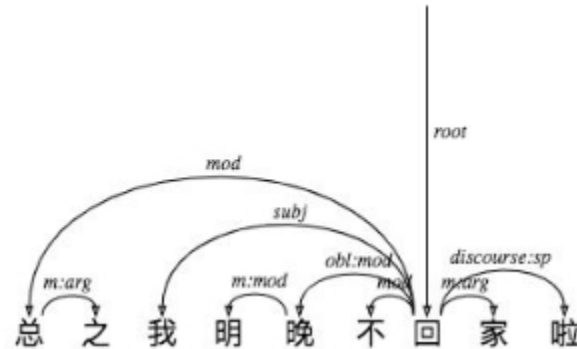


Figure 2 character--based treebank

Différence entre les arbres de dépendance par caractère et par mot

Annotation en POS : Expériences

Entraînement du POS Tagger

— — —

Corpus : SUD Chinese

1. Division du corpus
 - Train (151 954 mots)
 - Dev (4 469 mots)
 - Test (4 232 words)
2. Conversion mots → caractères
3. Entraînement avec un algo de deep learning (comme LSTM)

Expérience réalisée:

Entraînement avec un [Dozat parser](#) (2016) : Parser en dépendances Neurologique basé sur des graphes; + représentation LSTM **caractère par caractère** des mots.

Résultats:

91% de précision sur la tâche de POS tagging.


Corpus utilisé

— — —

Chinese SUD treebanks (composé de 4 treebanks UD du chinois):

- Traditional Chinese Universal Dependencies Treebank annotated by Google (GSD)
- Parallel Universal Dependencies, créé pour CoNLL 2017 (PUD)
- Treebank de chinois traditionnel de sous-titres de films et de procédures législatives de Hong Kong (HK)
- Essais rédigés par des apprenants du Chinois en tant que Foreign Language (CFL),

Pré-traitements effectués:

- Fichier de caractères vectorisés entraîné par BERT sur un modèle pré-entraîné caractère par caractère (source des corpus pour l'entraînement et nom du  èle :)
- Découpage du corpus Chinese SUD treebanks en caractères
- Conversion automatique en dépendances *Morphe*

Entraînement

— — —

Entraînement du POS tagger avec le *Dozat Parser*. Mais dans *Dozat*, POS / dépendances = entraînements séparés
=> il faut faire 2 entraînements pour cette expérience.

4 entraînements:

CB tagger	WB tagger
CB parser	WB parser

Chinese SUD treebank: 10% test ; 90% train

Afin de comparer les entraînements, utilisation de la colonne XPOS pour reconstituer les POS des mots et comparer les résultats mot par mot (*recombination*)

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%

Table 3 F-score of word level POS (UPOS) for our word-based tagger

Category	Precision	Recall	F-score
ADJ	89.37%	87.98%	88.67%
ADP	88.55%	81.38%	84.81%
ADV	89.33%	90.17%	89.75%
AUX	75.46%	89.96%	82.07%
CCONJ	95.92%	63.51%	76.42%
DET	89.36%	77.78%	83.17%
INTJ	66.67%	66.67%	66.67%
NOUN	93.20%	94.10%	93.65%
NUM	93.53%	100.00%	96.65%
PART	96.43%	96.72%	96.57%
PRON	96.06%	97.99%	97.01%
PROPN	73.37%	82.12%	77.50%
PUNCT	100.00%	100.00%	100.00%
SCONJ	0.00%	0.00%	0.00%
SYM	100.00%	100.00%	100.00%
VERB	92.22%	89.89%	91.04%
TOTAL	91.99%	91.87%	91.93%

Table 2 F-score of character level POS for our character-based 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%

Table 3 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%

Table 4 F-score of word level POS (XPOS) for our character-based tagger after the recombination

Category	Precision	Recall	F-score	Category	Precision	Recall	F-score
ADJ	65.69%	50.00%	56.78%	ADJ	65.52%	42.54%	51.58%
ADP	63.48%	69.75%	66.47%	ADP	60.11%	87.90%	71.40%
ADV	80.08%	76.40%	78.20%	ADV	75.00%	70.80%	72.84%
AUX	59.84%	81.56%	69.03%	AUX	64.71%	86.03%	73.86%
CCONJ	92.68%	58.46%	71.70%	CCONJ	92.68%	58.46%	71.70%
DET	96.81%	68.94%	80.53%	DET	91.22%	86.45%	88.77%
INTJ	100.00%	0.00%	0.00%	INTJ	100.00%	20.00%	33.33%
NOUN	88.17%	82.27%	85.12%	NOUN	77.87%	85.56%	81.54%
NUM	63.92%	98.41%	77.50%	NUM	65.14%	93.65%	76.84%
PART	84.03%	91.74%	87.72%	PART	91.56%	94.50%	93.00%
PRON	94.06%	93.14%	93.60%	PRON	92.47%	88.24%	90.30%
PROPN	38.17%	89.29%	53.48%	PROPN	54.05%	71.43%	61.54%
PUNCT	99.84%	99.84%	99.84%	PUNCT	99.84%	100.00%	99.92%
SCONJ	100.00%	0.00%	0.00%	SCONJ	20.00%	4.35%	7.14%
SYM	100.00%	0.00%	0.00%	SYM	100.00%	100.00%	100.00%
VERB	76.29%	77.56%	76.92%	VERB	83.31%	76.41%	79.71%
TOTAL	81.85%	81.62%	81.74%	TOTAL	88.85%	88.70%	88.78%

Table 3 F-score of word level POS (UPOS) for our word-based tagger

Table 4 F-score of word level POS (XPOS) for our character-based tagger after the recombination

Explications

— — —

N’y avait-il aucun INTJ, SCONJ et SYM dans le corpus de mots ?

Moins bon score :

ADJ

ADV

NOUN

Explication:

Inconsistance dans la création du corpus
CB. (ex 活动 : VV, le mot est sensé être
un N, mais dans le corpus constitué, ce
mot est parfois N, parfois un V)

Entraînement du parser : Expériences

Category	Precision	Recall	F-score
case	89.66%	96.30%	92.86%
cc	70.31%	95.74%	81.08%
clf	89.71%	90.39%	90.04%
comp	80.82%	84.96%	82.83%
compound	66.67%	77.42%	71.64%
conj	56.04%	44.74%	49.76%
det	96.21%	93.38%	94.78%
discourse	93.62%	84.62%	88.89%
mark	76.71%	78.87%	77.78%
mod	90.71%	78.86%	84.37%
obl	45.10%	62.16%	52.27%
parataxis	5.13%	11.11%	7.02%
punct	99.53%	100.00%	99.76%
root	85.34%	85.34%	85.34%
subj	79.27%	84.12%	81.62%
vocative	100.00%	0.00%	0.00%
TOTAL	81.41%	75.49%	78.33%

Table 7 F-score of the most frequent dependency relations of the word-based parser

Category	Precision	Recall	F-score
case	85.19%	85.19%	85.19%
cc	73.77%	95.74%	83.33%
clf	91.94%	91.94%	91.94%
comp	78.74%	85.19%	81.84%
compound	62.93%	78.49%	69.86%
conj	62.32%	37.72%	46.99%
det	96.27%	94.85%	95.56%
discourse	97.78%	84.62%	90.72%
mark	71.43%	84.51%	77.42%
mod	90.94%	78.93%	84.51%
obl	62.00%	70.27%	65.88%
parataxis	47.02%	44.44%	45.69%
punct	99.68%	100.00%	99.84%
root	86.64%	86.64%	86.64%
subj	79.08%	86.35%	82.56%
vocative	81.82%	47.37%	60.00%
TOTAL	83.67%	78.81%	81.17%

Table 8 F-score of the most frequent dependency relations of the character-based parser after the recombination of characters

	cc	clf	comp	compound	conj	dep	det	discourse	dislocated	flat	mark	mod	morphe*	obj	obl	parataxis
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
0	23	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	791	0	4	7	3	0	2	0	0	0	23	0	0	1	1
0	0	5	0	136	1	20	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	35	0	0	0	0	0	0	1	0	0	0	1
0	0	1	0	6	1	265	1	0	0	0	1	6	0	0	1	0
0	0	0	0	0	0	3	106	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	38	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	5	0	0	0	1	0	0	55	1	0	0	0	0
0	0	1	0	0	4	1	0	2	0	0	0	449	0	0	0	0
0	0	0	0	1	0	1	0	0	0	0	0	0	2099	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	2	0	0	0	0	0	4	0	1	23	0
0	0	11	0	0	2	1	0	0	0	0	0	4	0	0	0	8
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
11	39	215	0	33	63	138	29	8	2	35	15	143	4	6	12	7

Matrice de confusion - dépendances
entre les mots

Evaluation d'annotation des dépendances

Mesures:

LAS: Labeled attachment score (pourcentage de mots attachés à la **bonne tête** et avec le **bon label**)

UAS: Unlabeled attachment score (pourcentage de mots attachés à la **bonne tête**)

OLS: Orthogonal Label Unattached Score (pourcentage de mots attachés au gouverneur avec le **bon label**, peu importe le gouverneur)

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

Table 5 Comparison between the results of WB and CB parser

Erreur sur les étiquettes Morph (m:)

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

=> Faible impact, peu de confusion entre Morphe et Deprel.

Nos remarques

— — —

Reproductibilité: Corpus trouvable en ligne; mais pas les algorithmes de constitution du corpus d'entraînement. Il faut réécrire les algorithmes.

Données d'entraînement: Inconsistance dans les données après les pré-traitements. Méthodes à améliorer.

Entraînement: utilisation d'un système pré-existant (Dozat Parser) pour les expériences. Peut-être pas 100% adapté au problème, mais point (+), ce modèle est conçu pour une analyse caractère par caractère.

Utilisation: pas de modèle publié.

Conclusion

Il s'agit plutôt de tester une méthode innovante d'analyse du chinois. De nos jours, les modèles ont tendance à se baser sur les caractères et plus sur les mots issus de la tokenisation.