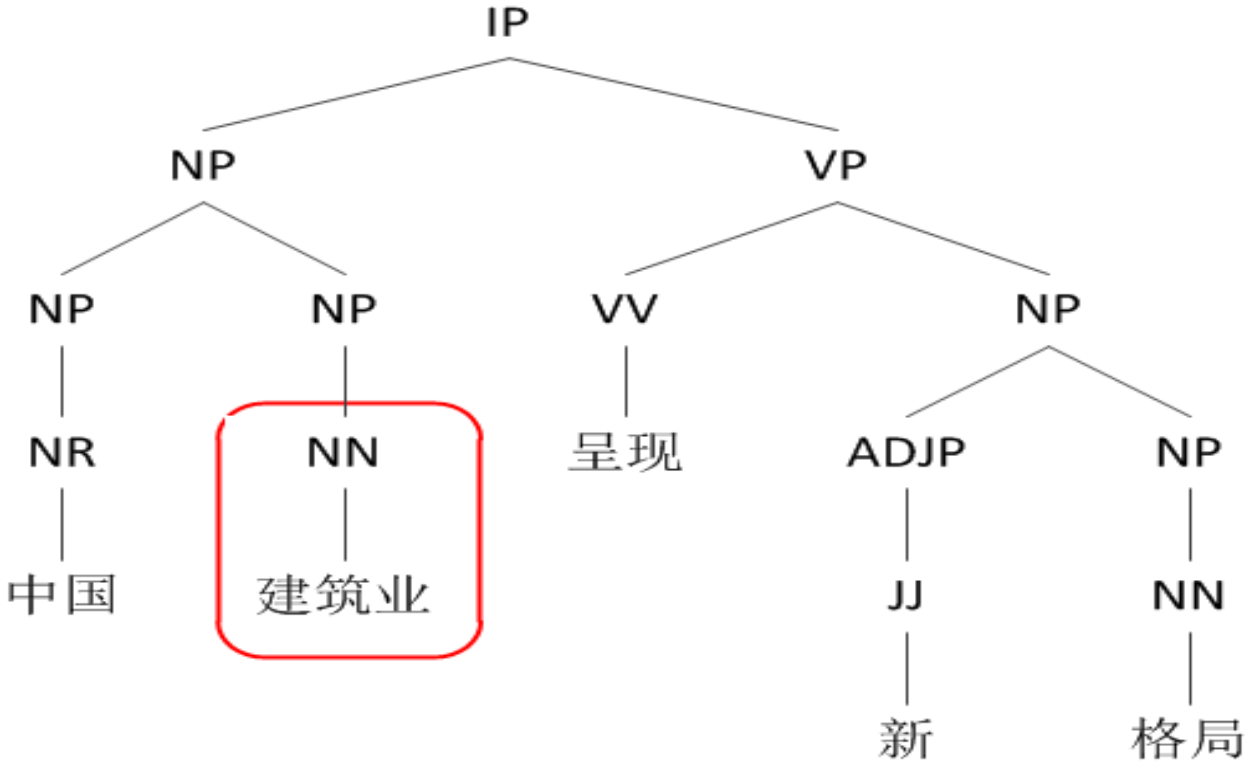
Chinese Parsing Exploiting Characters

Meishan Zhang[1], Yue Zhang[2],
Wanxiang Che[1], Ting Liu[1]

Research Center for Social Computing and Information Retrieval[1]
Harbin Institute of Technology, China
{mszhang, car, tliu}@ir.hit.edu.cn
Singapore University of Technology and Design[2]
yue_zhang@sutd.edu.sg

Traditional: Word-based Chinese Parsing

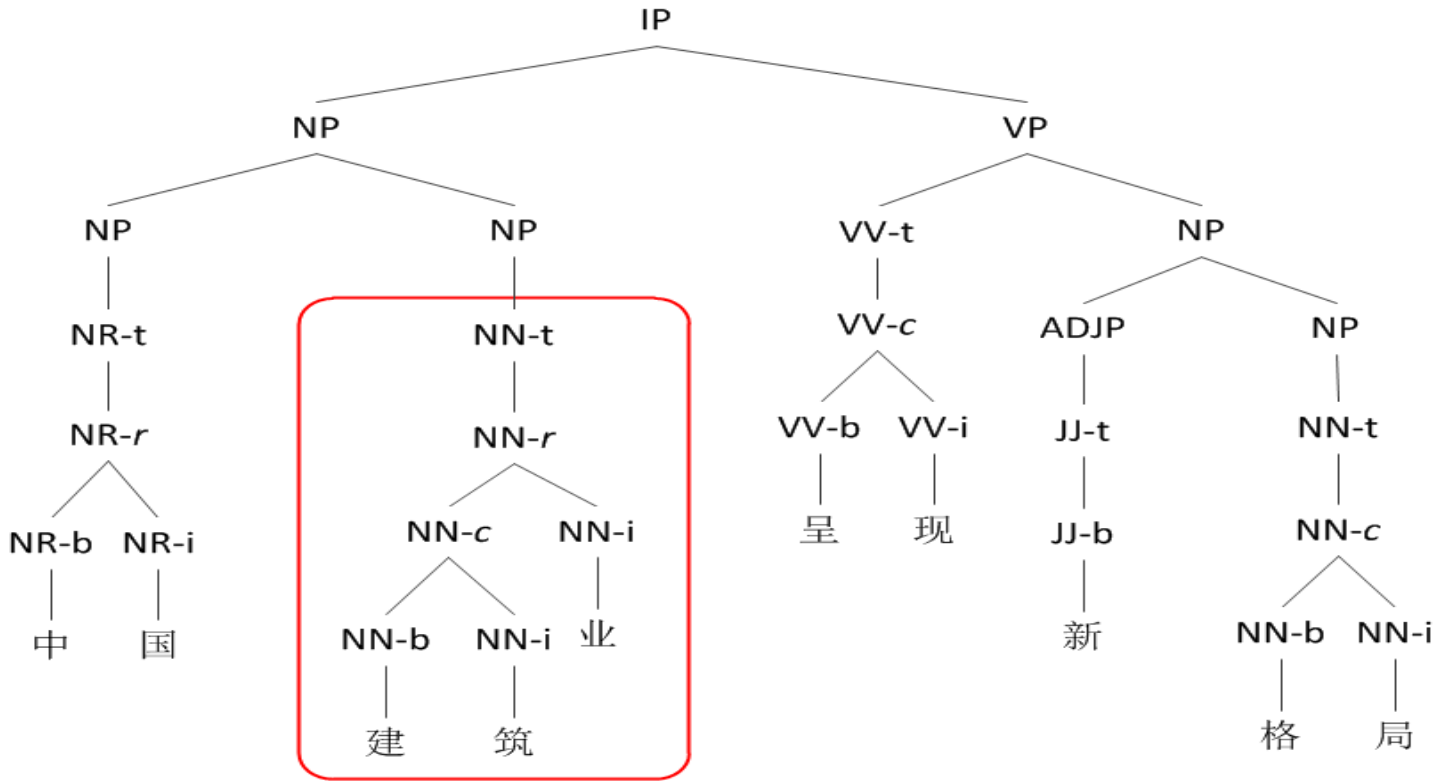
Traditional: Word-based Chinese Parsing



CTB-style word-based syntax tree for “中国 (China) 建筑业 (architecture industry) 呈现 (show) 新 (new) 格局 (pattern)”.

This work: Character-based Chinese Parsing

This work: Character-based Chinese Parsing



Character-level syntax tree with hierarchal word structures for “中 (middle) 国 (nation) 建 (construction) 筑 (building) 业 (industry) 呈 (present) 现 (show) 新 (new) 格 (style) 局 (situation)”.

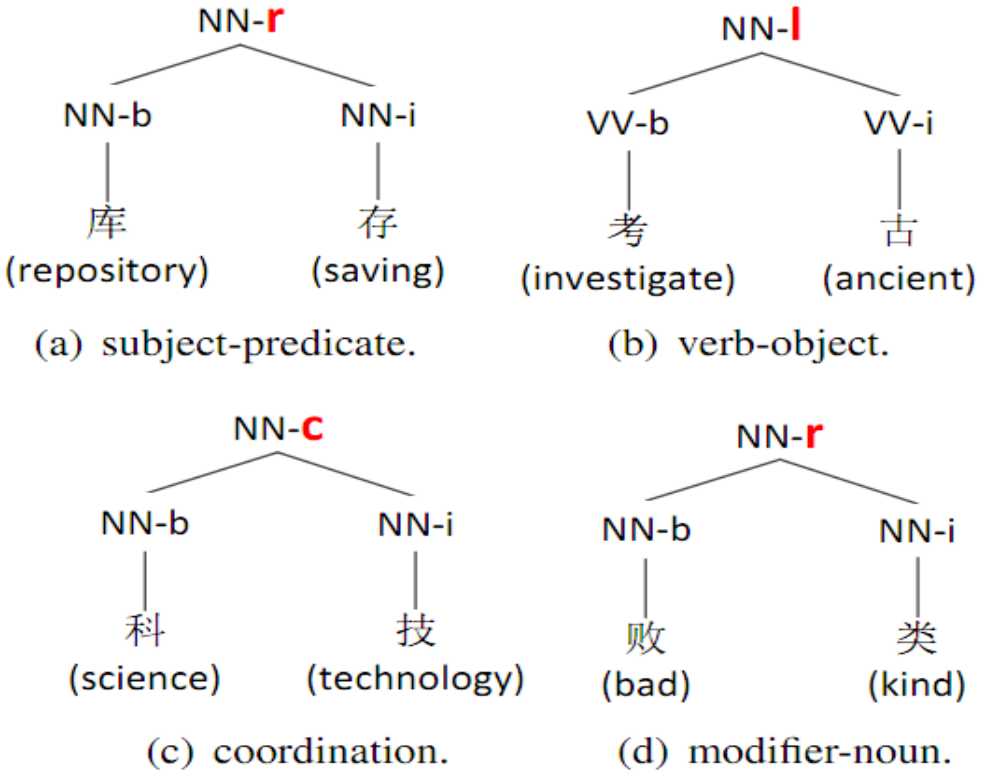
Why Character-based ?

Why Character-based ?

- Chinese words have syntactic structures.

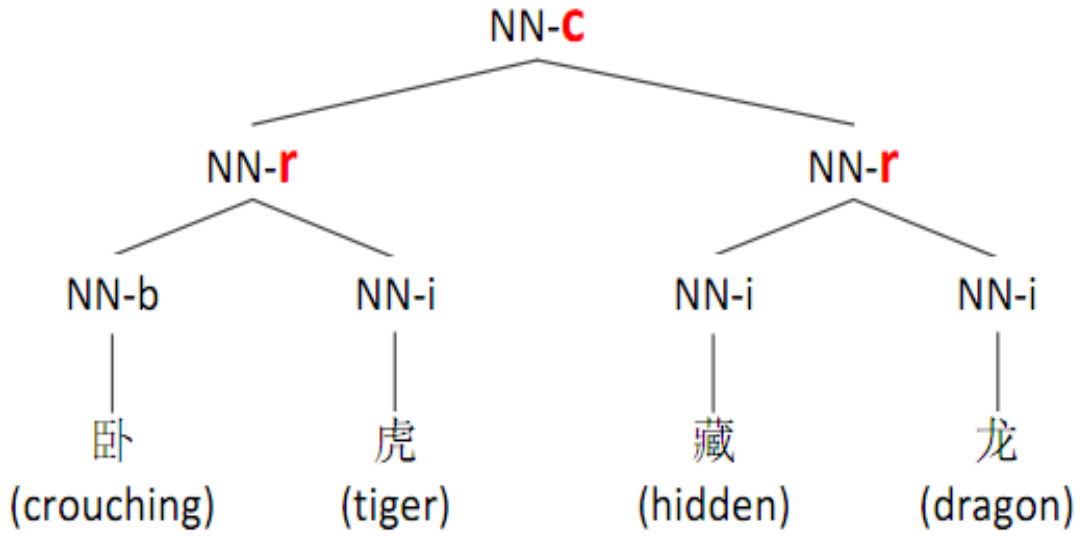
Why Character-based ?

- Chinese words have syntactic structures.



Why Character-based ?

- Chinese words have syntactic structures.

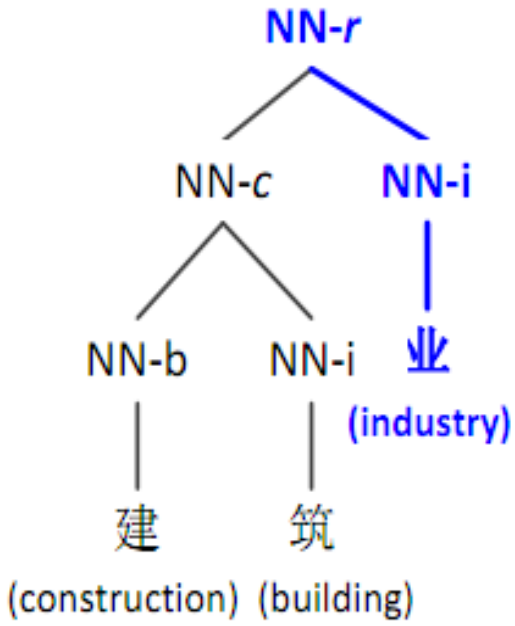


Why Character-based ?

- Deep character information of word structures.

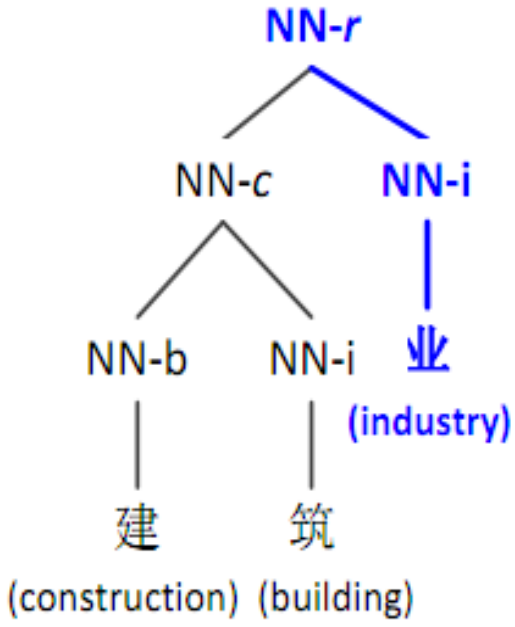
Why Character-based ?

- Deep character information of word structures.



Why Character-based ?

- Deep character information of word structures.



Representing the whole word by a character, which is less sparse.

Why Character-based ?

- Build syntax tree from character sequences.
  - Not require segmentation or POS-tagging as input.
  - Benefit from joint framework, avoid error propagation.

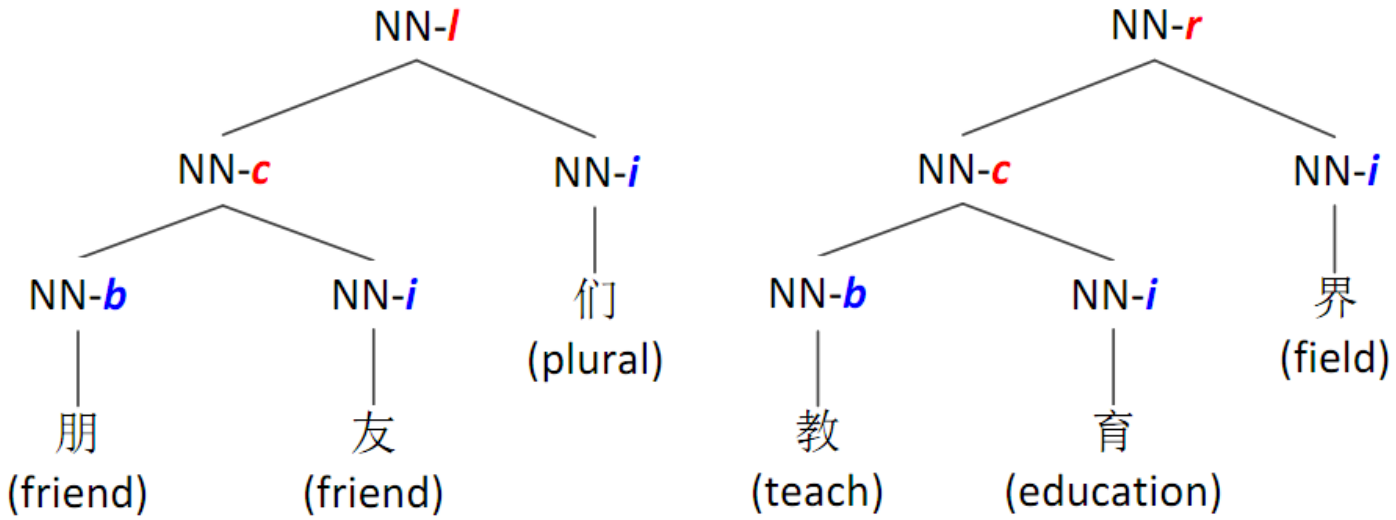
Word Structure Annotation

Word Structure Annotation

- Binarized tree structure for each word.

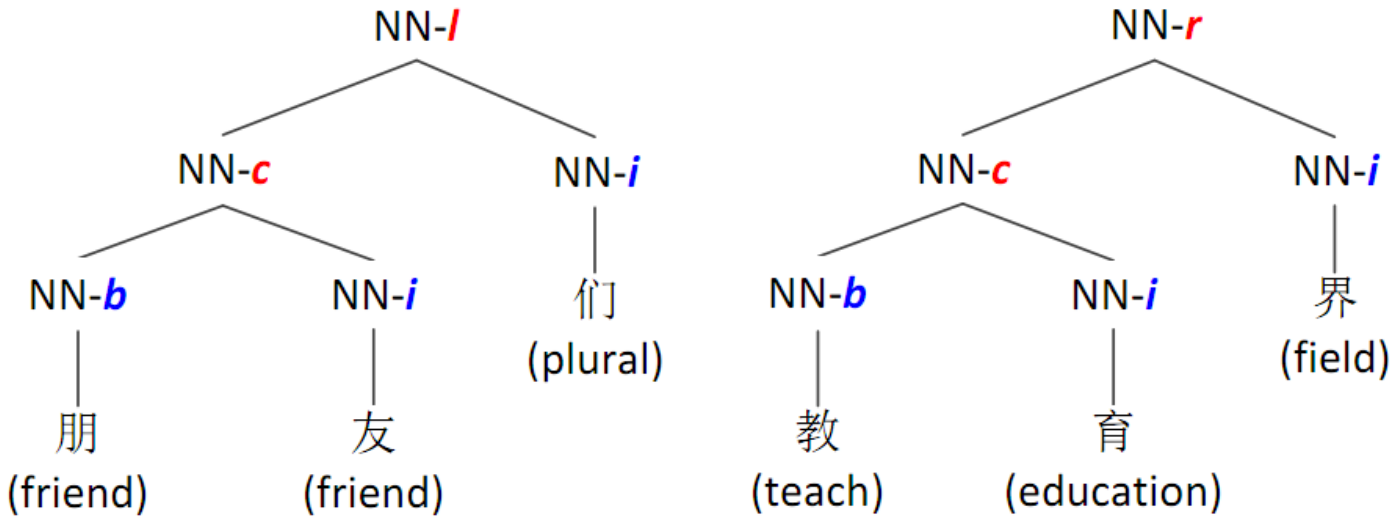
Word Structure Annotation

- Binarized tree structure for each word.



Word Structure Annotation

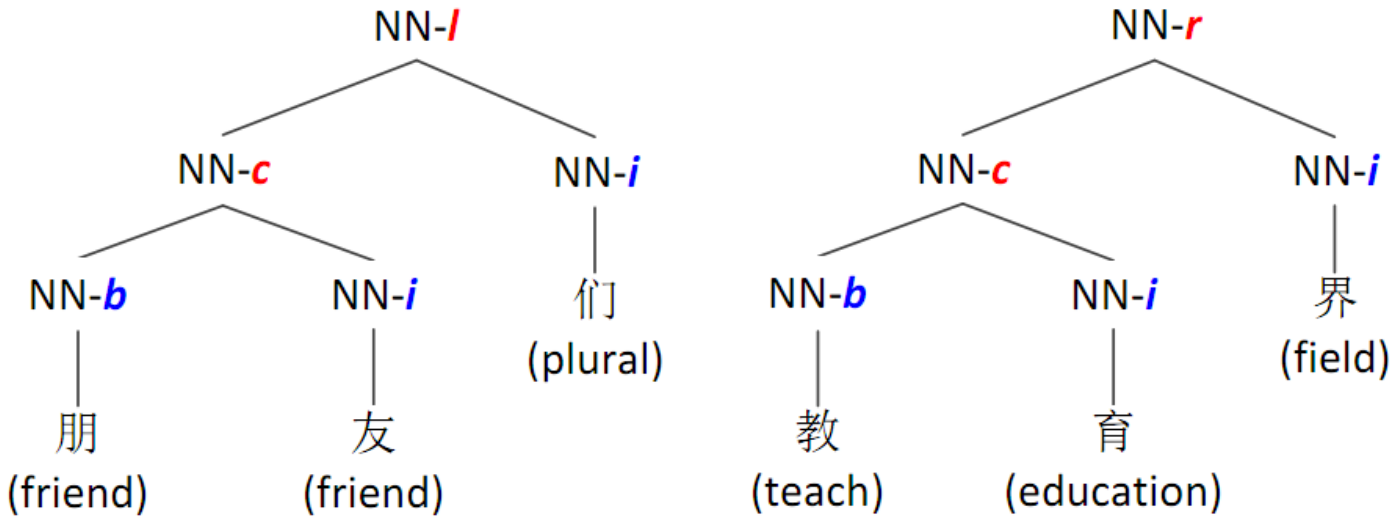
- Binarized tree structure for each word.



- **b, i denote whether the below character is at a word's beginning position.**
- **l, r, c denote the head direction of current node, respectively left, right and coordination.**

Word Structure Annotation

- Binarized tree structure for each word.



- **b, i denote whether the below character is at a word's beginning position.**
- **l, r, c denote the head direction of current node, respectively left, right and coordination.**

We extend word-based phrase-structures into character-based syntax trees using the word structures demonstrated above.

Word Structure Annotation

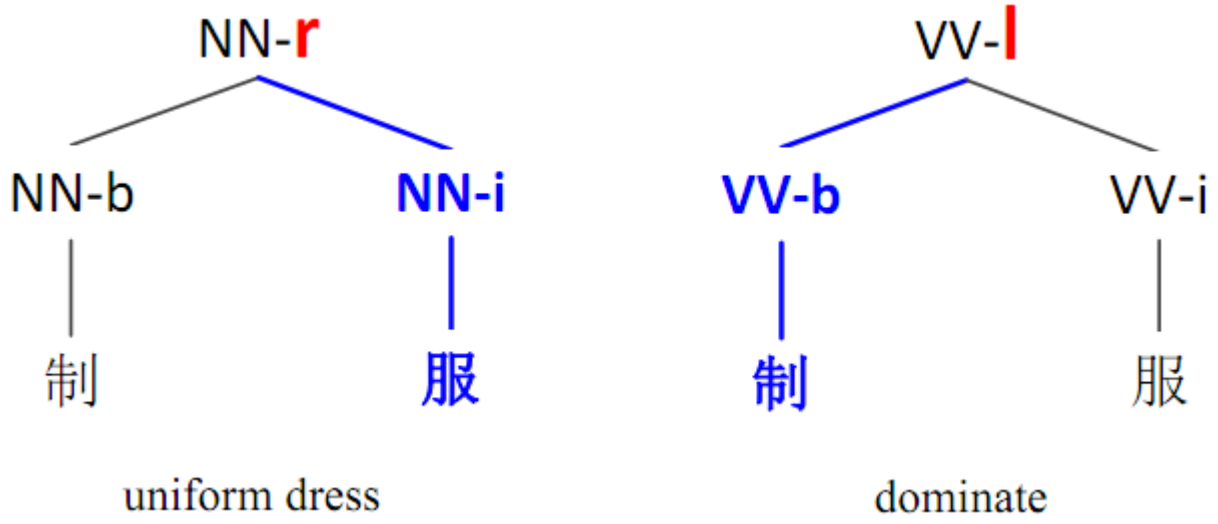
- Annotation input: a word and its POS.

  A word may have different structures according to different POS.

Word Structure Annotation

- Annotation input: a word and its POS.

  A word may have different structures according to different POS.



Outline

- Our Chinese Parsing Model
- Experiments
- Conclusion

Outline

- Our Chinese Parsing Model
- Experiments
- Conclusion

The Character-based Parser

The Character-based Parser

- A Transition-based Parser using Beam-search Decoding Algorithm.
  - Extended from Zhang and Clark (2009), a word-based transition parser.

The Character-based Parser

- A Transition-based Parser using Beam-search Decoding Algorithm.
  - Extended from Zhang and Clark (2009), a word-based transition parser.
- Incorporating features of a word-based parser as well as a joint SEG&POS system.

The Character-based Parser

- A Transition-based Parser using Beam-search Decoding Algorithm.
  - Extended from Zhang and Clark (2009), a word-based transition parser.
- Incorporating features of a word-based parser as well as a joint SEG&POS system.
- Adding the deep character information from word structures.

The Transition System

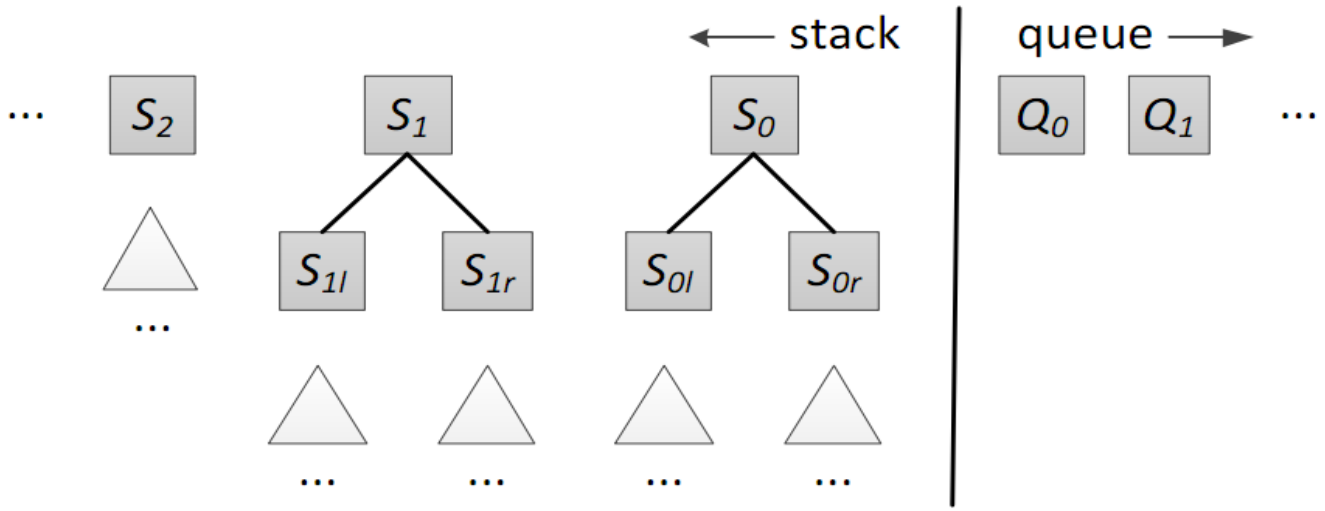
The Transition System

- State

- Actions:

The Transition System

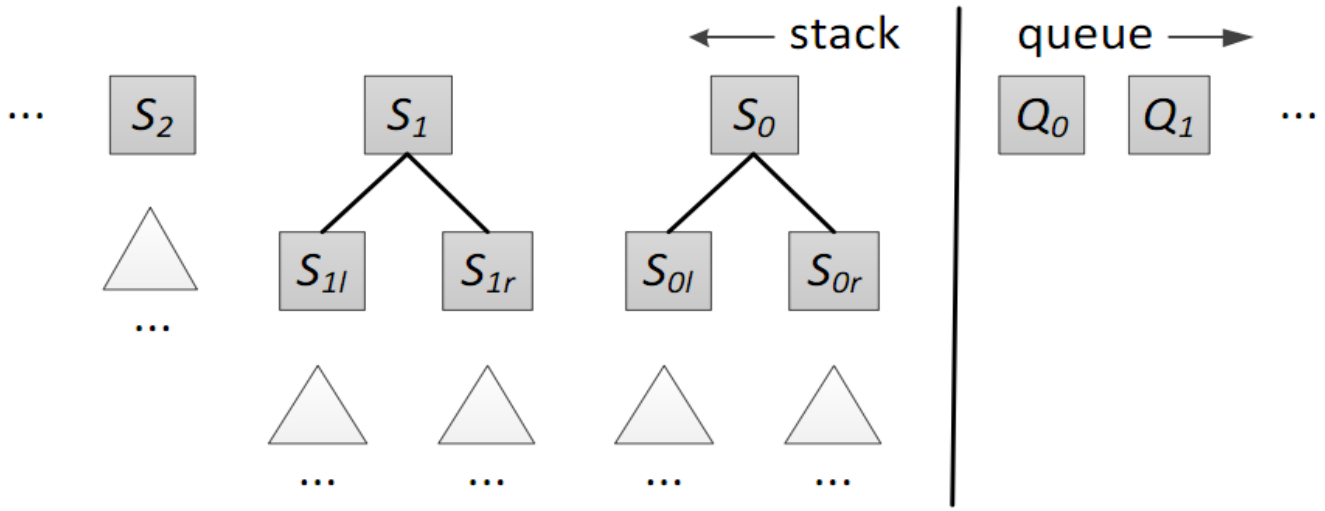
- State



- Actions:

The Transition System

- State



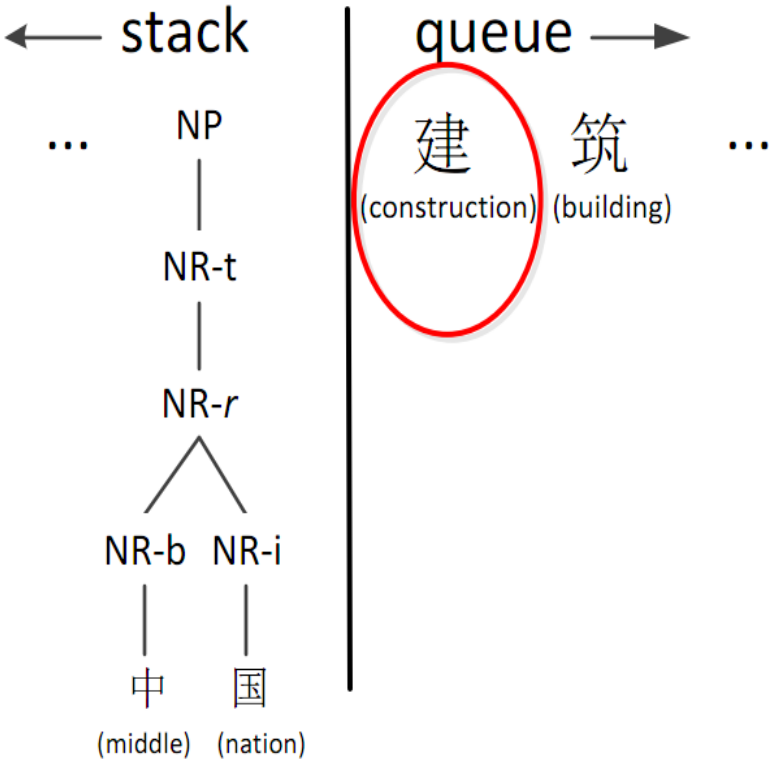
- Actions:
  - SHIFT-SEPARATE(*t*), SHIFT-APPEND, REDUCE-SUBWORD(*d*), REDUCE-WORD, REDUCE-BINARY(*d;l*), REDUCE-UNARY(*l*), TERMINATE

Transition Actions

- SHIFT-SEPARATE(*t*)

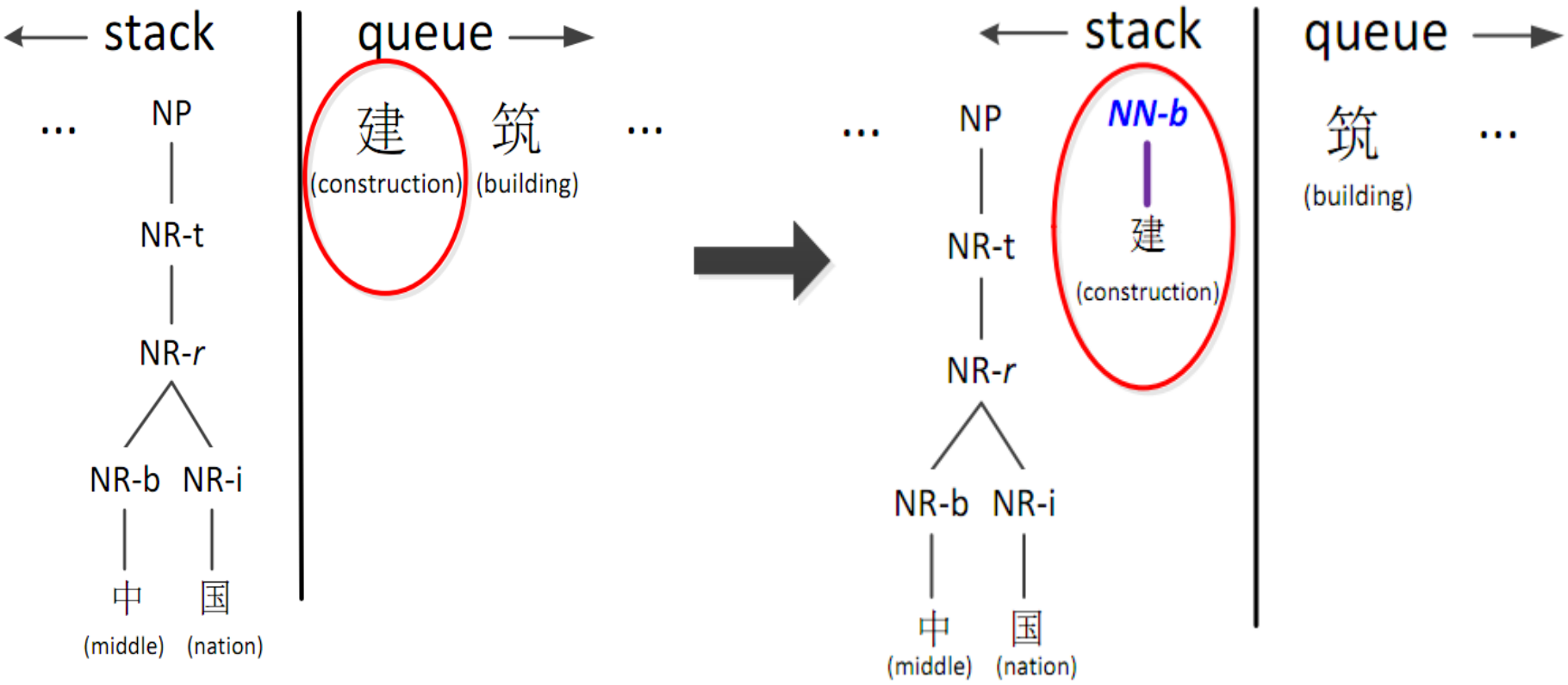
Transition Actions

- SHIFT-SEPARATE(*t*)



Transition Actions

- SHIFT-SEPARATE(*t*)

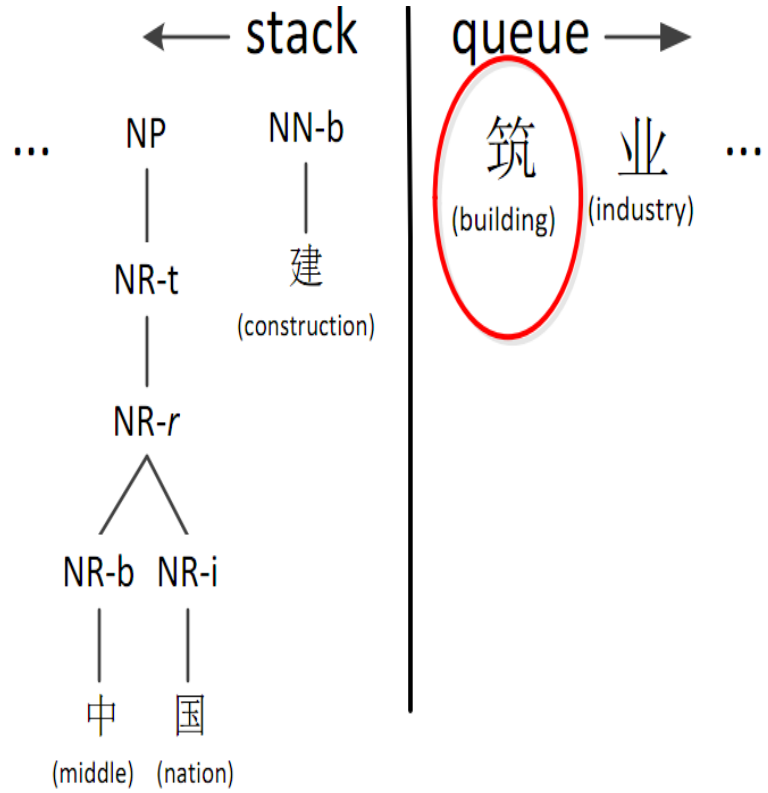


Transition Actions

- SHIFT-APPEND

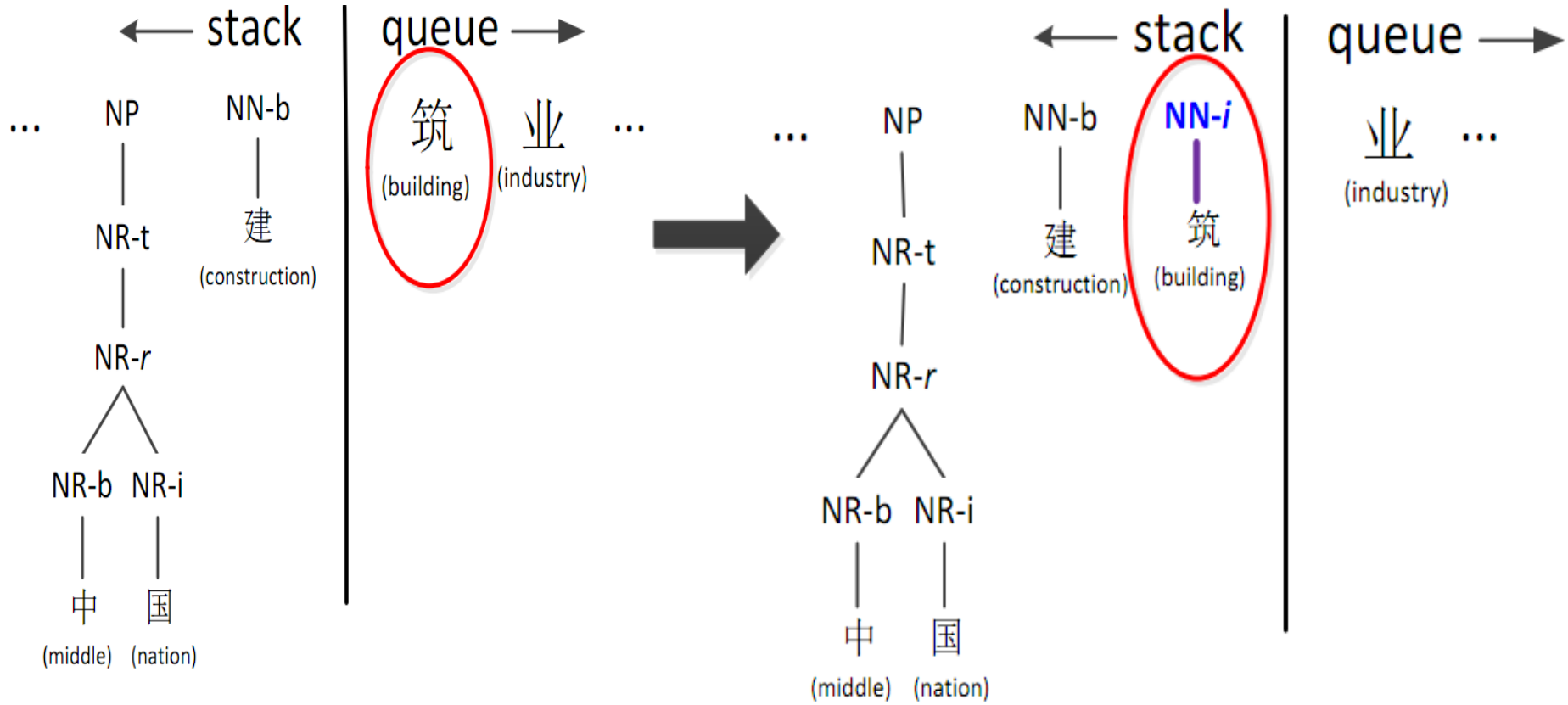
Transition Actions

- SHIFT-APPEND



Transition Actions

- SHIFT-APPEND

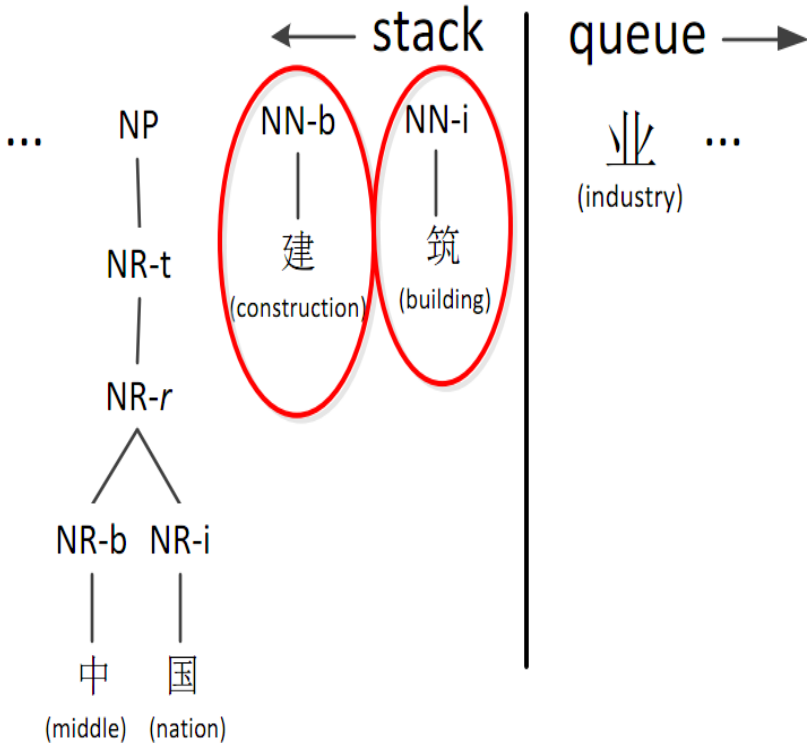


Transition Actions

- REDUCE-SUBWORD(***d***)

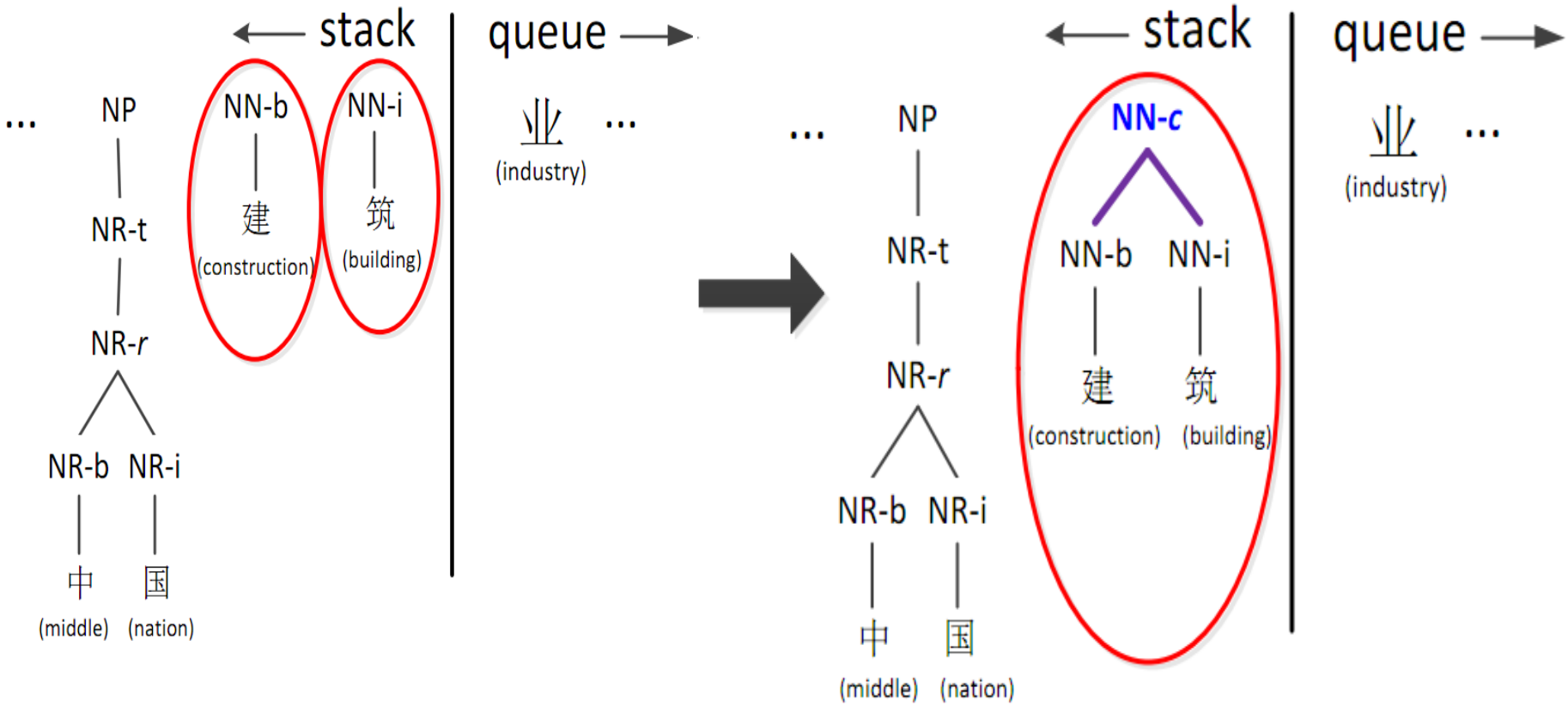
Transition Actions

- REDUCE-SUBWORD(*d*)



Transition Actions

- REDUCE-SUBWORD(*d*)

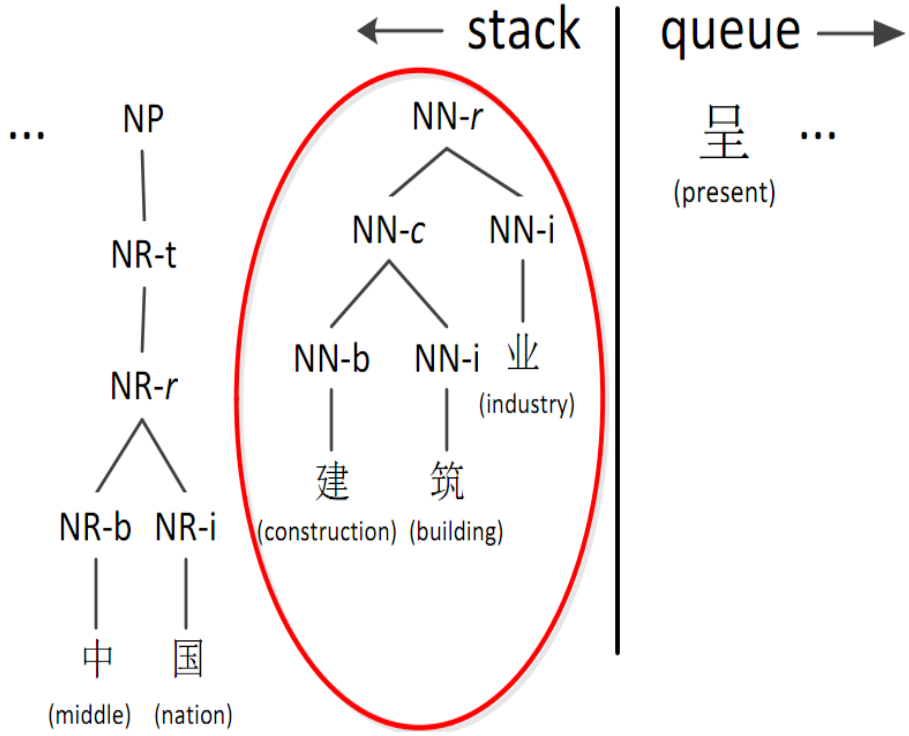


Transition Actions

- REDUCE-WORD

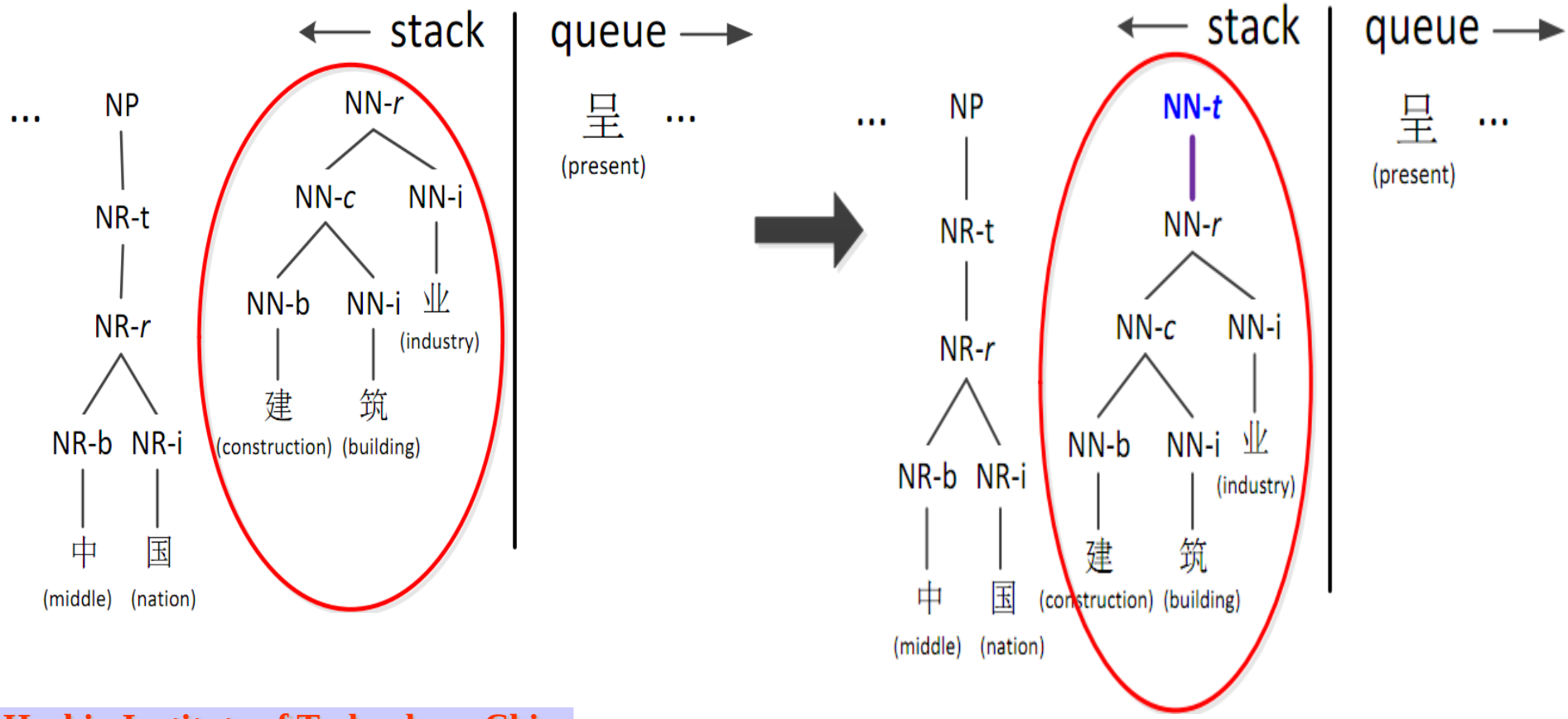
Transition Actions

- REDUCE-WORD



Transition Actions

- REDUCE-WORD

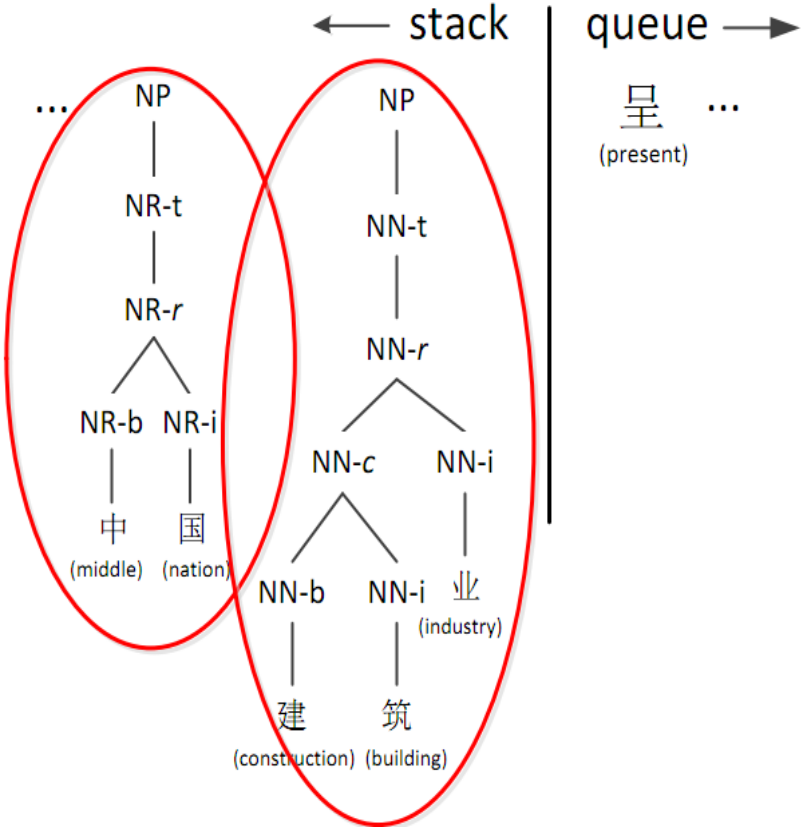


Transition Actions

- REDUCE-BINARY(*d; l*)

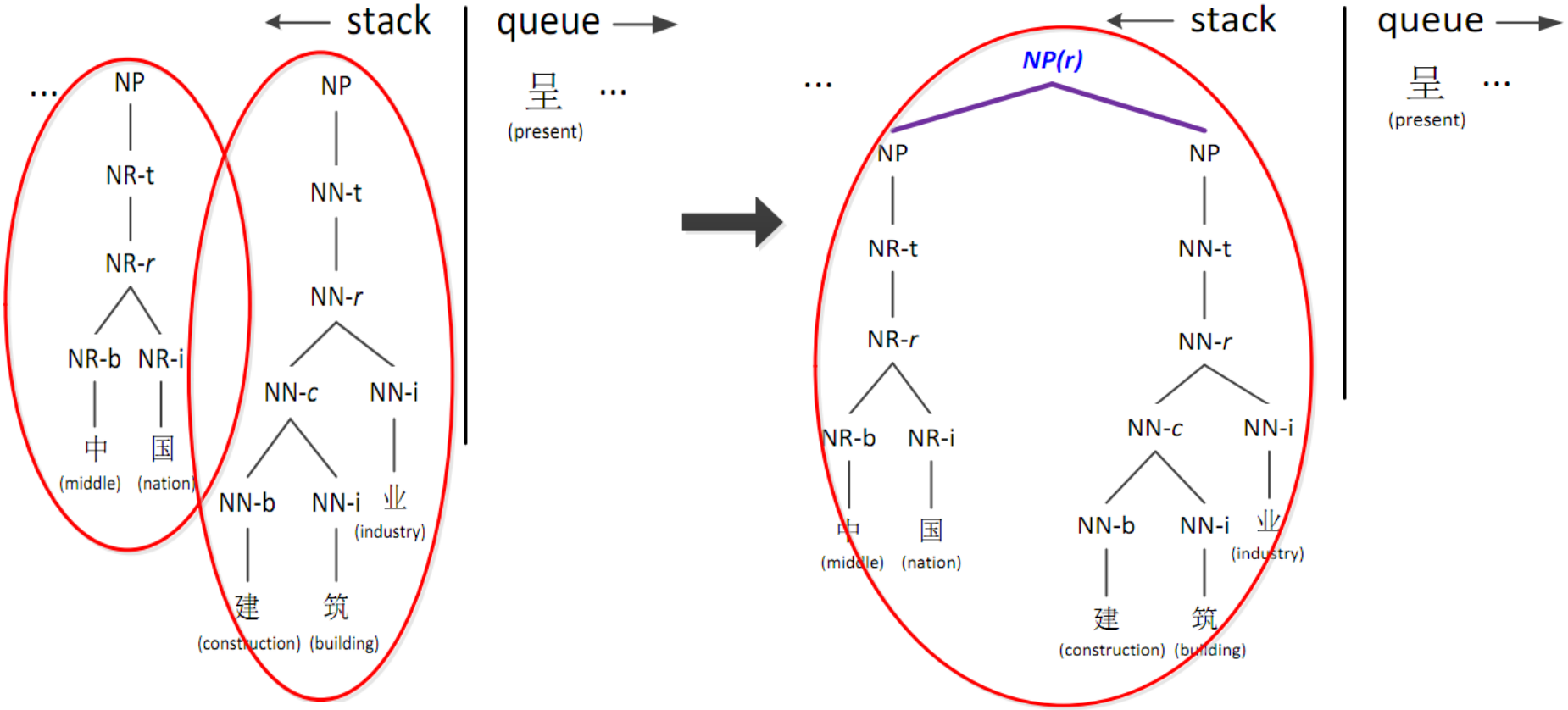
Transition Actions

- REDUCE-BINARY(*d; l*)



Transition Actions

- REDUCE-BINARY(*d; l*)

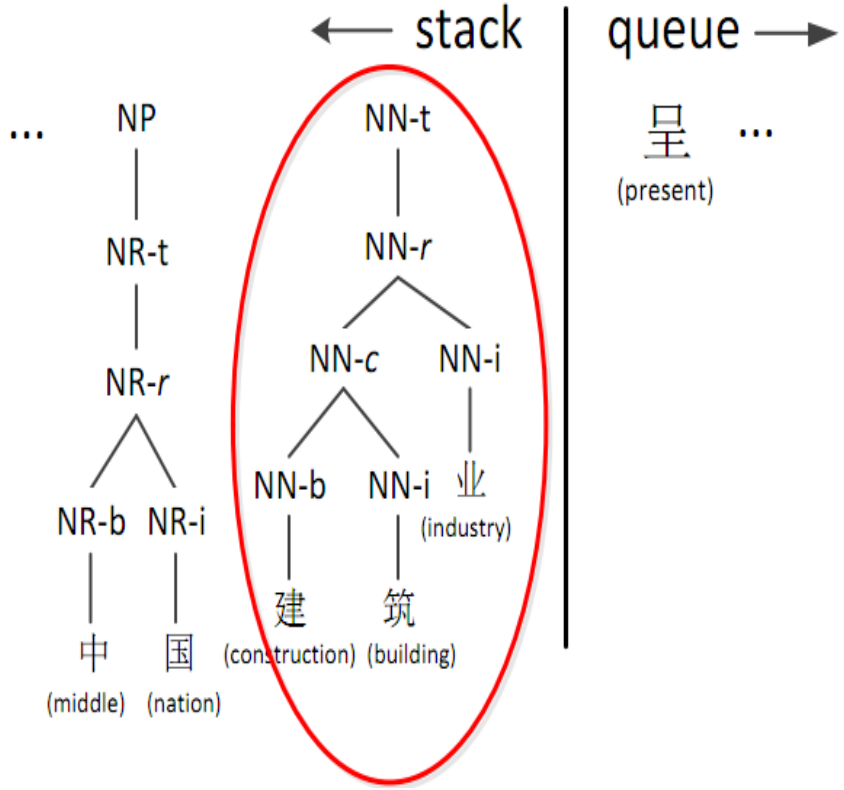


Transition Actions

- REDUCE-UNARY(*l*)

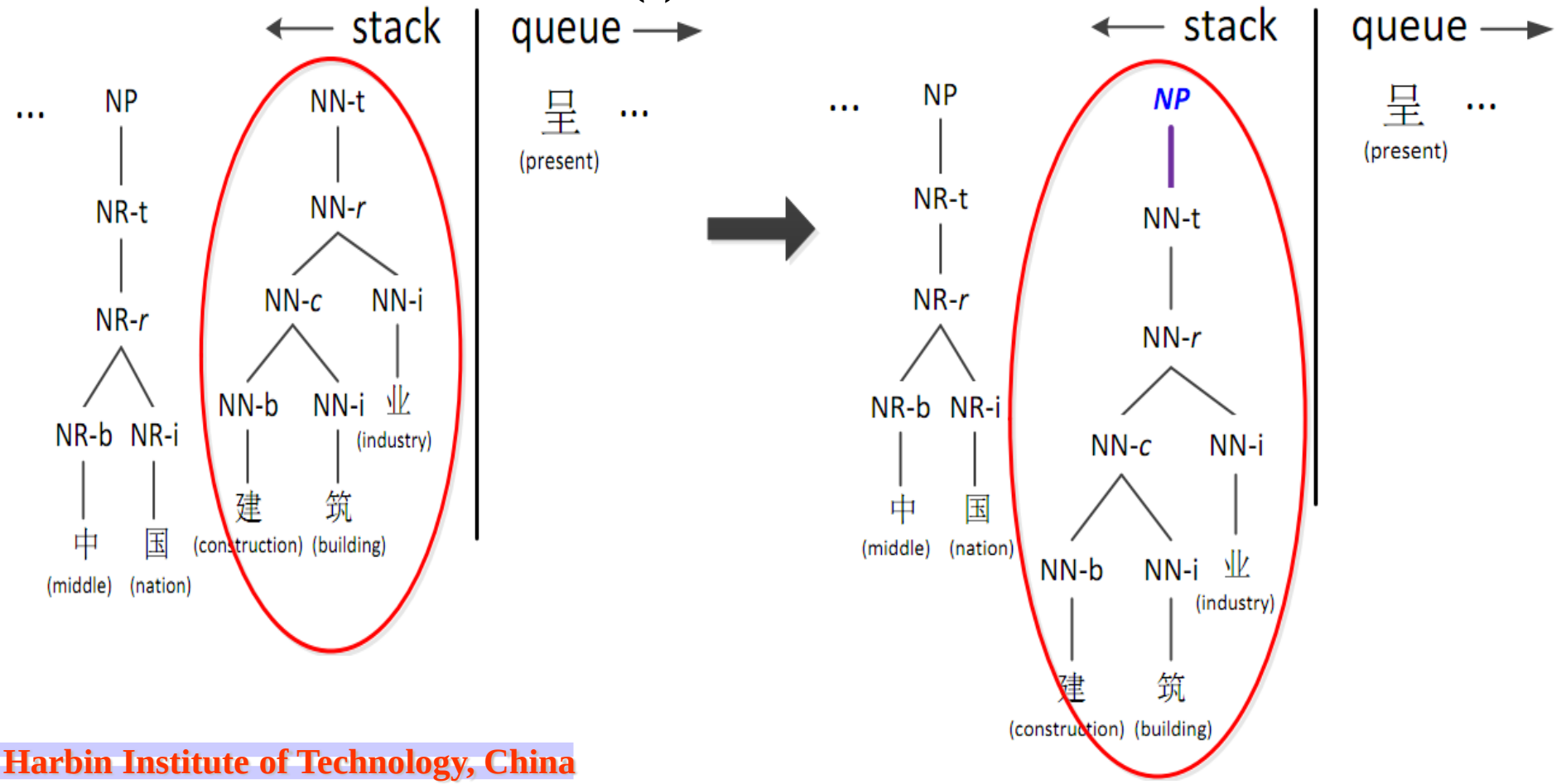
Transition Actions

- REDUCE-UNARY(*l*)



Transition Actions

- REDUCE-UNARY(*l*)

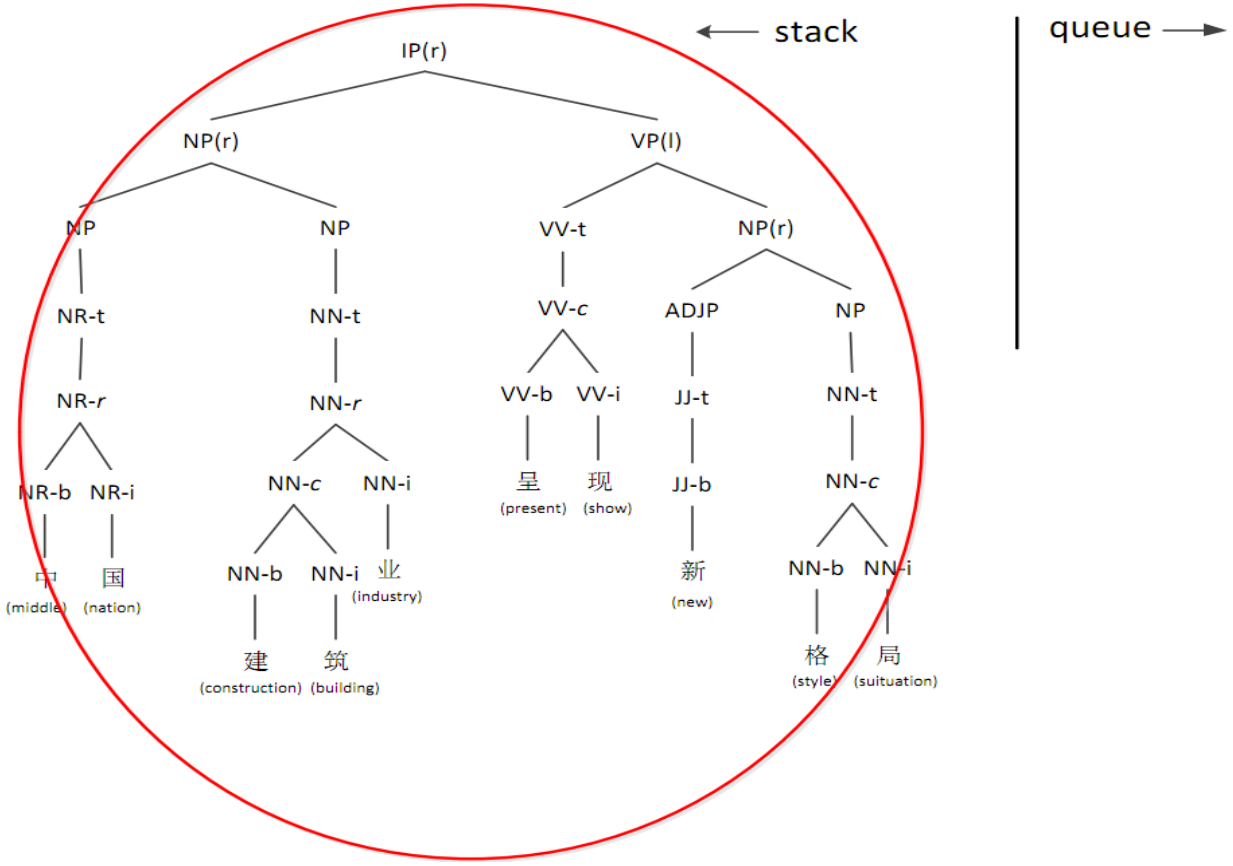


Transition Actions

- TERMINATE

Transition Actions

- TERMINATE



Features

Features

- From word-based parser (Zhang and Clark, 2009)

Features

- From word-based parser (Zhang and Clark, 2009)
- From joint SEG&POS-Tagging (Zhang and Clark, 2010)

Features

- From word-based parser (Zhang and Clark, 2009)
- From joint SEG&POS-Tagging (Zhang and Clark, 2010)

Word-based Features

Features

- From word-based parser (Zhang and Clark, 2009)
- From joint SEG&POS-Tagging (Zhang and Clark, 2010)

Word-based Features

- Deep character features (new)

Features

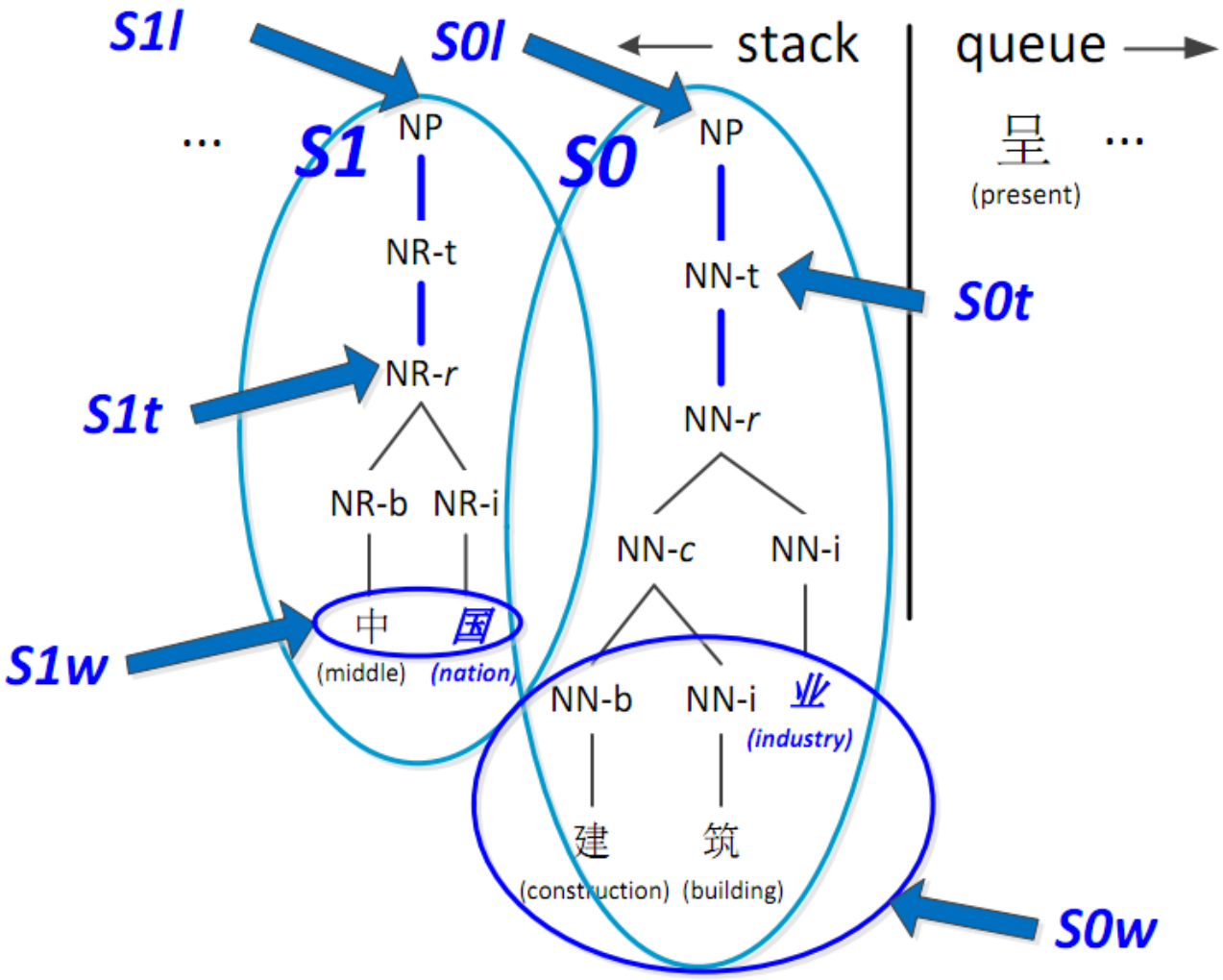
- From word-based parser (Zhang and Clark, 2009)
- From joint SEG&POS-Tagging (Zhang and Clark, 2010)

Word-based Features

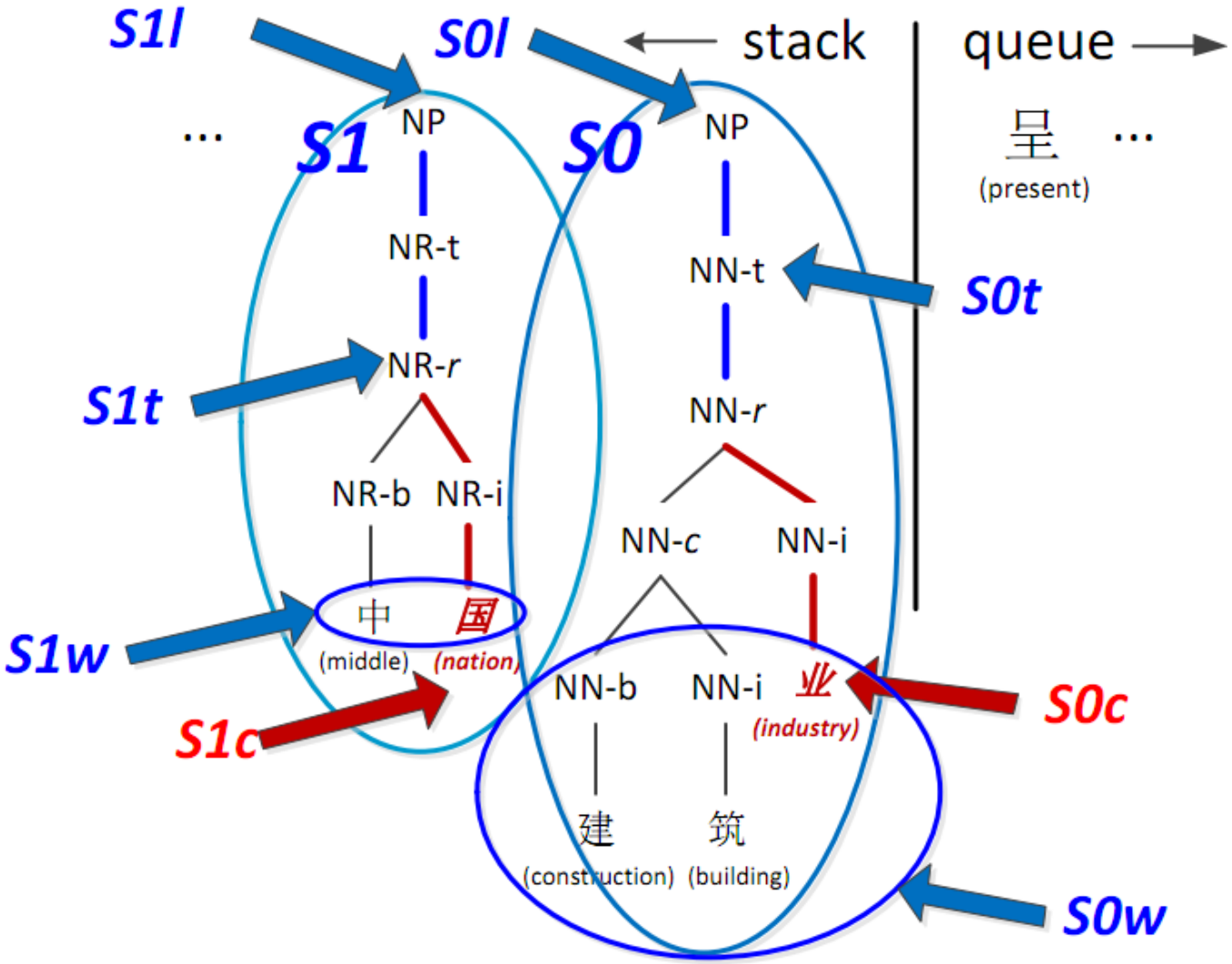
- Deep character features (new)

Deep Character Features

Features


S1l
S0l
stack
queue
...
S1
S0
NP
NR-t
NR-r
NR-b
NR-i
中
(middle)
国
(nation)
NP
NN-t
NN-r
NN-c
NN-i
呈
(present)
...
S0t
S1t
S1w
NN-b
NN-i
业
(industry)
建
(construction)
筑
(building)
S0w

Features


S1l
S0l
stack
queue
...
S1
S0
NP
NR-t
NR-r
NR-b
NR-i
中
(middle)
国
(nation)
NP
NN-t
NN-r
NN-c
NN-i
NN-b
NN-i
业
(industry)
建
(construction)
筑
(building)
呈
(present)
...
S1t
S1w
S1c
S0t
S0c
S0w

Outline

- Our Chinese Parsing Model
- Experiments
- Conclusion

Experiments

Experiments

- Penn Chinese Treebank 5 (CTB-5)

Experiments

- Penn Chinese Treebank 5 (CTB-5)

	CTB files	# sent.	# words
Training	1-270 400-1151	18089	493,939
Develop	301-325	350	6,821
Test	271-300	348	8,008

Experiments

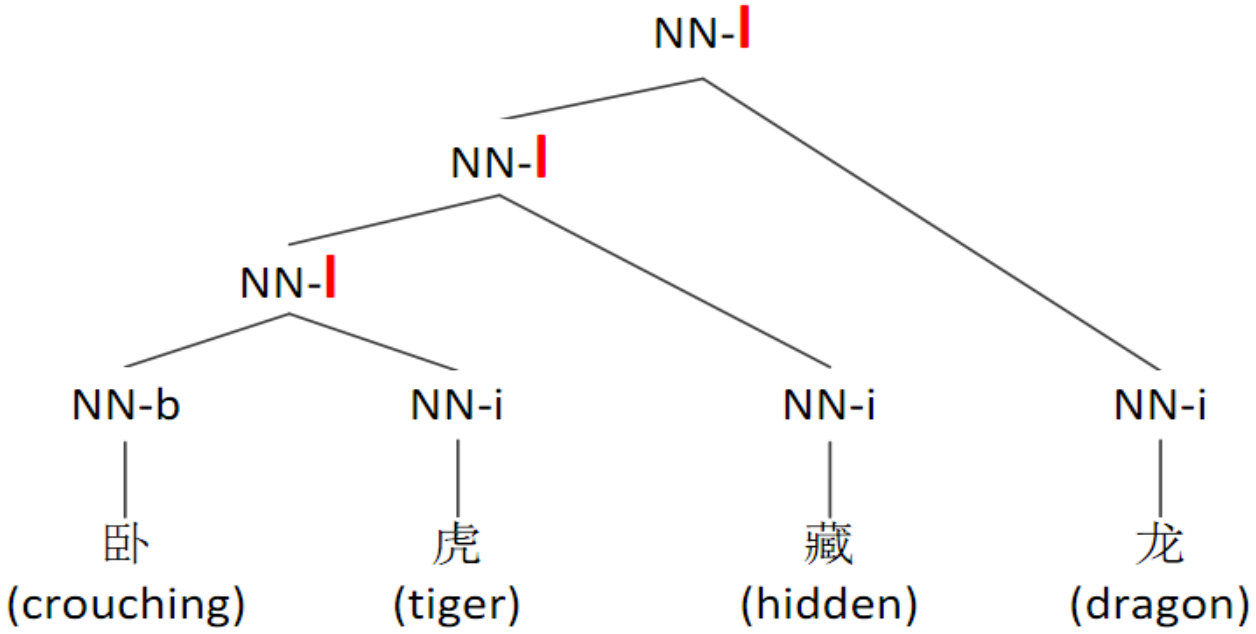
- Baseline models
  - Pipeline model including:
    - Joint SEG&POS-Tagging model (Zhang and Clark, 2010).
    - Word-based constituent parser (Zhang and Clark, 2009).

Experiments

- Our proposed models
  - Joint model with flat word structures.

Experiments

- Our proposed models
  - Joint model with flat word structures.

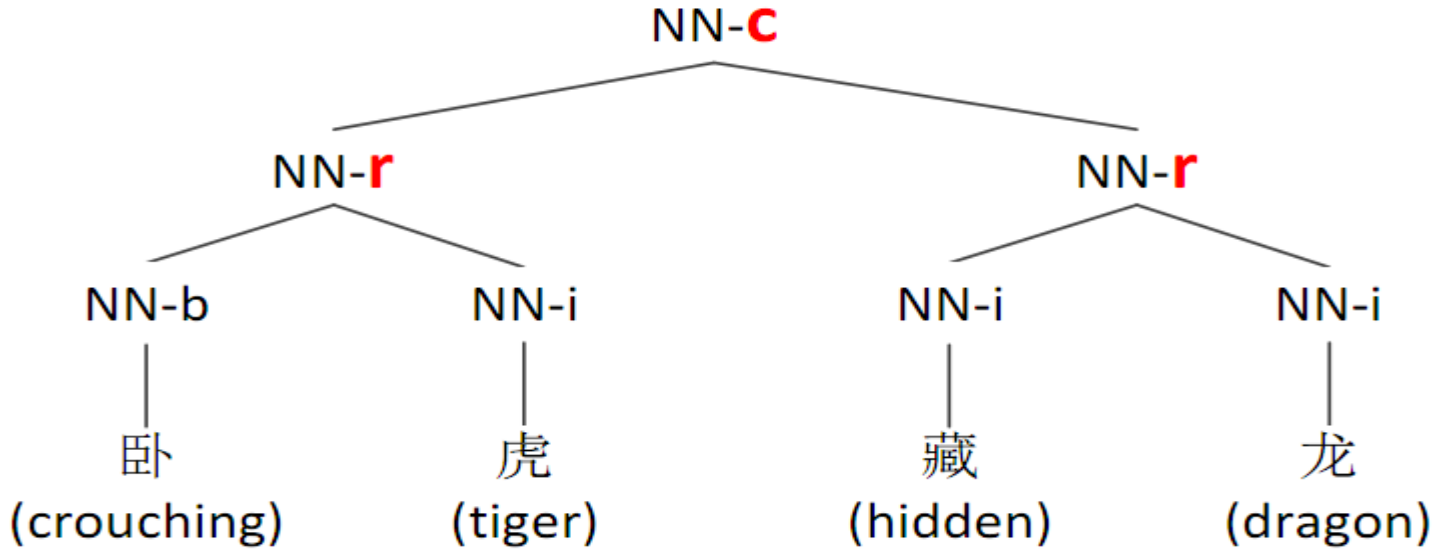


Experiments

- Our proposed models
  - Joint model with flat word structures
  - Joint model with annotated word structures

Experiments

- Our proposed models
  - Joint model with flat word structures
  - Joint model with annotated word structures



Results

Results

	Task	P	R	F
Pipeline	Seg	97.35	98.02	97.69
	Tag	93.51	94.15	93.83
	Parse	81.58	82.95	82.26

Results

	Task	P	R	F
Pipeline	Seg	97.35	98.02	97.69
	Tag	93.51	94.15	93.83
	Parse	81.58	82.95	82.26
Flat word structures	Seg	97.32	98.13	97.73
	Tag	94.09	94.88	94.48
	Parse	83.39	83.84	83.61

Results

	Task	P	R	F
Pipeline	Seg	97.35	98.02	97.69
	Tag	93.51	94.15	93.83
	Parse	81.58	82.95	82.26
Flat word structures	Seg	97.32	98.13	97.73
	Tag	94.09	94.88	94.48
	Parse	83.39	83.84	83.61
Annotated word structures	Seg	97.49	98.18	97.84
	Tag	94.46	95.14	94.80
	Parse	84.42	84.43	84.43
	WS	94.02	94.69	94.35

Influence of Deep Character Features

Influence of Deep Character Features

	Task	P	R	F
With deep character features	Seg	96.71	96.81	96.76
	Tag	94.12	94.22	94.17
	Parse	85.08	85.60	85.34
	WS	93.13	93.22	93.17

Influence of Deep Character Features

	Task	P	R	F
With deep character features	Seg	96.71	96.81	96.76
	Tag	94.12	94.22	94.17
	Parse	85.08	85.60	85.34
	WS	93.13	93.22	93.17
Without deep character features	Seg	96.59	96.46	96.53
	Tag	93.80	93.68	93.74
	Parse	84.60	84.90	84.75
	WS	92.76	92.64	92.70

Compare with Other Systems

Compare with Other Systems

Task	Seg	Tag	Parse
Kruengkrai+ ’09	97.87	93.67	–
Sun ’11	98.17	94.02	–
Wang+ ’11	98.11	94.18	–
Li ’11	97.3	93.5	79.7
Li+ ’12	97.50	93.31	–
Hatori+ ’12	98.26	94.64	–
Qian+ ’12	97.96	93.81	82.85

Compare with Other Systems

Task	Seg	Tag	Parse
Kruengkrai+ ’09	97.87	93.67	–
Sun ’11	98.17	94.02	–
Wang+ ’11	98.11	94.18	–
Li ’11	97.3	93.5	79.7
Li+ ’12	97.50	93.31	–
Hatori+ ’12	98.26	94.64	–
Qian+ ’12	97.96	93.81	82.85
Ours pipeline	97.69	93.83	82.26
Ours joint flat	97.73	94.48	83.61
Ours joint annotated	97.84	94.80	84.43

Outline

- Our Chinese Parsing Model
- Experiments
- Conclusion

Conclusion

Conclusion

- We annotated a number of word structures which are useful for syntax parsing.

Conclusion

- We annotated a number of word structures which are useful for syntax parsing.
- We developed a high-performance character-level transition-based parser that can jointly parse the word structures and the phrase structures.

Conclusion

- We annotated a number of word structures which are useful for syntax parsing.
- We developed an high-performance character-level transition-based parser that cab jointly parse the word structures and the phrase structures.
- We proposed a set of deep character features for our parser that are effective for POS-tagging and syntax parsing.

Thank you

- Data
  - https://github.com/zhangmeishan/wordstructures .
- Code
  - http://sourceforge.net/projects/zpar/ , version 0.6.