Chinese Parsing Exploiting Characters

Meishan Zhang\textsuperscript{1}, Yue Zhang\textsuperscript{2}, Wanxiang Che\textsuperscript{1}, Ting Liu\textsuperscript{1}

Research Center for Social Computing and Information Retrieval\textsuperscript{1}
Harbin Institute of Technology, China
\{mszhang, car, tliu\}@ir.hit.edu.cn

Singapore University of Technology and Design\textsuperscript{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)”.

Harbin Institute of Technology, China
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)”.

Harbin Institute of Technology, China
Why Character-based?
Why Character-based?

- Chinese words have syntactic structures.
Why Character-based?

- Chinese words have syntactic structures.

(a) subject-predicate.
(b) verb-object.
(c) coordination.
(d) modifier-noun.
Why Character-based?

- Chinese words have syntactic structures.

```
  NN-C
    / \     / \
   NN-r  NN-r
   / \   / \
  NN-b NN-i  NN-i  NN-i
   \   \   \   \   \   \   \   
  卧 (crouching)  虎 (tiger)  藏 (hidden)  龙 (dragon)
```
Why Character-based?

- Deep character information of word structures.
Why Character-based?

- Deep character information of word structures.

---

Harbin Institute of Technology, China
Why Character-based?

- Deep character information of word structures.

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

Harbin Institute of Technology, China
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.

![Diagram showing binarized tree structure for words in Chinese and English]
**Word Structure Annotation**

- Binarized tree structure for each word.

```
NN-l
  NN-c
  NN-b
  朋
  (friend)

NN-i
  友
  (friend)

NN-i
  们
  (plural)

NN-c
  教
  (teach)

NN-i
  育
  (education)

NN-r
  NN-c
  NN-b
  界
  (field)
```

- 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.

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

- 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

- 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.

- **NN-r**
  - **NN-b**
    - 制
  - **NN-i**
    - 服

- **VV-I**
  - **VV-b**
    - 制
  - **VV-i**
    - 服

uniform dress

*Harbin Institute of Technology, China*
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

- \textsc{shift-separate}(t)
Transition Actions

- **SHIFT-SEPARATE**(t)

```
stack                                queue
...  NP                              ...  建筑
   |                                (construction) (building)
  NR-t                             
   |                                
  NR-r                             
   |                                
  NR-b    NR-i                      
   |                                
  中  国  (middle) (nation)
```

Harbin Institute of Technology, China
Transition Actions

- **SHIFT-SEPARATE**($t$)
Transition Actions

- SHIFT-APPEND
Transition Actions

- SHIFT-APPEND

(stack)  queue

NP  NN-b
NR-t  建 (construction)
NR-r  筑 (building)
NR-b  中 (middle)
NR-i  国 (nation)

...  ...
Transition Actions

- **SHIFT-APPEND**

```
<table>
<thead>
<tr>
<th></th>
<th>NP</th>
<th>NN-b</th>
<th>NN-i</th>
<th>NR-r</th>
<th>NR-t</th>
</tr>
</thead>
<tbody>
<tr>
<td>stack</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>queue</td>
<td>筑 (building)</td>
<td>业 (industry)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(construction)</td>
<td>(nation)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```

```
<table>
<thead>
<tr>
<th></th>
<th>NP</th>
<th>NN-b</th>
<th>NN-i</th>
<th>NR-r</th>
</tr>
</thead>
<tbody>
<tr>
<td>stack</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>queue</td>
<td>业 (industry)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```

Harbin Institute of Technology, China
Transition Actions

- REDUCE-SUBWORD($d$)
Transition Actions

- REDUCE-SUBWORD($d$)
Transition Actions

- REDUCE-SUBWORD($d$)
Transition Actions

- REDUCE-WORD
Transition Actions

- REDUCE-WORD

```
... NP
   /   
NR-t  
   / 
NR-r  
   /   
NR-b  NR-i
    /   
  建筑
    /   
  (construction) (building)
```

```
 呈
(present)
```

Chinese: 呈
English: present
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**($l$)

```
stack          queue
...
NP            呈...
  |   
  NR-t        
  |       
  NR-r      
  |     
  NR-b       
  |       
  NR-i       
  |       
  中 (middle)  国 (nation)
  |   
  建 (construction)  筑 (building)
  |       
  NN-b       
  |       
  NN-t       
  |       
  NN-r       
  |       
  NN-c       
  |       
  NN-i       
```

Harbin Institute of Technology, China
Transition Actions

- REDUCE-UNARY(l)

```
  stack          queue
...
NP              呈 ...
  \|--\          ...}
   |  \--\         \--\          ...
   |   |            |   |
  NN-t | NN-r         NR-t | NN-r         NP
   \--\                  \--\      
   |                      |    |
  NN-c     NN-i           NR-b    NR-i
   \--\    \--\            \--\    \--\  
   |     |    |             |     |    |
  NN-b   NN-i  (industry)  NR-b  NR-i  (industry)
   \--\   \--\               \--\   \--\  
   |    |    (construction) (building) (middle) (nation)
       |    
       建筑 (construction) (building)
```

Harbin Institute of Technology, China
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)
Features

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

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

Harbin Institute of Technology, China
Features
Outline

- Our Chinese Parsing Model
- Experiments
- Conclusion
Experiments
Experiments

- Penn Chinese Treebank 5 (CTB-5)
Experiments

- Penn Chinese Treebank 5 (CTB-5)

<table>
<thead>
<tr>
<th></th>
<th>CTB files</th>
<th># sent.</th>
<th># words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1-270</td>
<td>18089</td>
<td>493,939</td>
</tr>
<tr>
<td></td>
<td>400-1151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop</td>
<td>301-325</td>
<td>350</td>
<td>6,821</td>
</tr>
<tr>
<td>Test</td>
<td>271-300</td>
<td>348</td>
<td>8,008</td>
</tr>
</tbody>
</table>
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

```
  NN-C
   /\                  /\                /\                /\                /\                /\
  NN-r                NN-r              NN-r              NN-r              \\
  /\                  /\                /\                /\                /\
 NN-b                NN-i              NN-i              NN-i              NN-i
 /\                  /\                /\                /\                /\
 卧 (crouching)      虎 (tiger)        藏 (hidden)        龙 (dragon)      
```

Harbin Institute of Technology, China
Results
## Results

<table>
<thead>
<tr>
<th>Task</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipeline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seg</td>
<td>97.35</td>
<td>98.02</td>
<td>97.69</td>
</tr>
<tr>
<td>Tag</td>
<td>93.51</td>
<td>94.15</td>
<td>93.83</td>
</tr>
<tr>
<td>Parse</td>
<td>81.58</td>
<td>82.95</td>
<td>82.26</td>
</tr>
</tbody>
</table>
# Results

<table>
<thead>
<tr>
<th></th>
<th>Task</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pipeline</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seg</td>
<td></td>
<td>97.35</td>
<td>98.02</td>
<td>97.69</td>
</tr>
<tr>
<td>Tag</td>
<td></td>
<td>93.51</td>
<td>94.15</td>
<td>93.83</td>
</tr>
<tr>
<td>Parse</td>
<td></td>
<td>81.58</td>
<td>82.95</td>
<td>82.26</td>
</tr>
<tr>
<td><strong>Flat word structures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seg</td>
<td></td>
<td>97.32</td>
<td>98.13</td>
<td>97.73</td>
</tr>
<tr>
<td>Tag</td>
<td></td>
<td>94.09</td>
<td>94.88</td>
<td>94.48</td>
</tr>
<tr>
<td>Parse</td>
<td></td>
<td>83.39</td>
<td>83.84</td>
<td>83.61</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Task</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pipeline</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seg</td>
<td>97.35</td>
<td>98.02</td>
<td>97.69</td>
</tr>
<tr>
<td>Tag</td>
<td>93.51</td>
<td>94.15</td>
<td>93.83</td>
</tr>
<tr>
<td>Parse</td>
<td>81.58</td>
<td>82.95</td>
<td>82.26</td>
</tr>
<tr>
<td><strong>Flat word structures</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seg</td>
<td>97.32</td>
<td>98.13</td>
<td>97.73</td>
</tr>
<tr>
<td>Tag</td>
<td>94.09</td>
<td>94.88</td>
<td>94.48</td>
</tr>
<tr>
<td>Parse</td>
<td>83.39</td>
<td>83.84</td>
<td>83.61</td>
</tr>
<tr>
<td><strong>Annotated word structures</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seg</td>
<td>97.49</td>
<td>98.18</td>
<td>97.84</td>
</tr>
<tr>
<td>Tag</td>
<td>94.46</td>
<td>95.14</td>
<td>94.80</td>
</tr>
<tr>
<td>Parse</td>
<td>84.42</td>
<td>84.43</td>
<td>84.43</td>
</tr>
<tr>
<td><strong>WS</strong></td>
<td>94.02</td>
<td>94.69</td>
<td>94.35</td>
</tr>
</tbody>
</table>
Influence of Deep Character Features
## Influence of Deep Character Features

<table>
<thead>
<tr>
<th>With deep character features</th>
<th>Task</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seg</td>
<td>96.71</td>
<td>96.81</td>
<td>96.76</td>
<td></td>
</tr>
<tr>
<td>Tag</td>
<td>94.12</td>
<td>94.22</td>
<td>94.17</td>
<td></td>
</tr>
<tr>
<td>Parse</td>
<td>85.08</td>
<td>85.60</td>
<td>85.34</td>
<td></td>
</tr>
<tr>
<td>WS</td>
<td>93.13</td>
<td>93.22</td>
<td>93.17</td>
<td></td>
</tr>
</tbody>
</table>
## Influence of Deep Character Features

<table>
<thead>
<tr>
<th></th>
<th>Task</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>With deep character features</strong></td>
<td>Seg</td>
<td>96.71</td>
<td>96.81</td>
<td>96.76</td>
</tr>
<tr>
<td></td>
<td>Tag</td>
<td>94.12</td>
<td>94.22</td>
<td>94.17</td>
</tr>
<tr>
<td></td>
<td>Parse</td>
<td>85.08</td>
<td>85.60</td>
<td>85.34</td>
</tr>
<tr>
<td></td>
<td>WS</td>
<td>93.13</td>
<td>93.22</td>
<td>93.17</td>
</tr>
<tr>
<td><strong>Without deep character features</strong></td>
<td>Seg</td>
<td>96.59</td>
<td>96.46</td>
<td>96.53</td>
</tr>
<tr>
<td></td>
<td>Tag</td>
<td>93.80</td>
<td>93.68</td>
<td>93.74</td>
</tr>
<tr>
<td></td>
<td>Parse</td>
<td>84.60</td>
<td>84.90</td>
<td>84.75</td>
</tr>
<tr>
<td></td>
<td>WS</td>
<td>92.76</td>
<td>92.64</td>
<td>92.70</td>
</tr>
</tbody>
</table>
Compare with Other Systems
## Compare with Other Systems

<table>
<thead>
<tr>
<th>Task</th>
<th>Seg</th>
<th>Tag</th>
<th>Parse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kruengkrai+ ’09</td>
<td>97.87</td>
<td>93.67</td>
<td>–</td>
</tr>
<tr>
<td>Sun ’11</td>
<td>98.17</td>
<td>94.02</td>
<td>–</td>
</tr>
<tr>
<td>Wang+ ’11</td>
<td>98.11</td>
<td>94.18</td>
<td>–</td>
</tr>
<tr>
<td>Li ’11</td>
<td>97.3</td>
<td>93.5</td>
<td>79.7</td>
</tr>
<tr>
<td>Li+ ’12</td>
<td>97.50</td>
<td>93.31</td>
<td>–</td>
</tr>
<tr>
<td>Hatori+ ’12</td>
<td>98.26</td>
<td>94.64</td>
<td>–</td>
</tr>
<tr>
<td>Qian+ ’12</td>
<td>97.96</td>
<td>93.81</td>
<td>82.85</td>
</tr>
</tbody>
</table>
# Compare with Other Systems

<table>
<thead>
<tr>
<th>Task</th>
<th>Seg</th>
<th>Tag</th>
<th>Parse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kruengkrai+ ’09</td>
<td>97.87</td>
<td>93.67</td>
<td>–</td>
</tr>
<tr>
<td>Sun ’11</td>
<td>98.17</td>
<td>94.02</td>
<td>–</td>
</tr>
<tr>
<td>Wang+ ’11</td>
<td>98.11</td>
<td>94.18</td>
<td>–</td>
</tr>
<tr>
<td>Li ’11</td>
<td>97.3</td>
<td>93.5</td>
<td>79.7</td>
</tr>
<tr>
<td>Li+ ’12</td>
<td>97.50</td>
<td>93.31</td>
<td>–</td>
</tr>
<tr>
<td>Hatori+ ’12</td>
<td>98.26</td>
<td>94.64</td>
<td>–</td>
</tr>
<tr>
<td>Qian+ ’12</td>
<td>97.96</td>
<td>93.81</td>
<td>82.85</td>
</tr>
<tr>
<td>Ours pipeline</td>
<td>97.69</td>
<td>93.83</td>
<td>82.26</td>
</tr>
<tr>
<td>Ours joint flat</td>
<td>97.73</td>
<td>94.48</td>
<td>83.61</td>
</tr>
<tr>
<td>Ours joint annotated</td>
<td>97.84</td>
<td>94.80</td>
<td>84.43</td>
</tr>
</tbody>
</table>
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 can 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](https://github.com/zhangmeishan/wordstructures).

- **Code**