Deep Learning and Lexical, Syntactic and Semantic Analysis

Wanxiang Che (HIT)
Yue Zhang (SUTD)
Part 5: Beam-search Decoding
A transition system

- Automata
  - State
    - Start state — an empty structure
    - End state — the output structure
    - Intermediate states — partially constructed structures
  - Actions
    - Change one state to another
A transition system

- Automata
A transition system

• Automata
A transition system

• Automata
A transition system

• Automata
A transition system

- Automata
A transition system

- Automata
A transition system

• Automata
A transition system

• State
  • Corresponds to partial results during decoding
    • start state, end state, $S_i$

• Actions
  • The operations that can be applied for state transition
  • Construct output incrementally
    • $a_i$
A transition-based POS-tagging example

• POS tagging
  I like reading books → I/PRON like/VERB reading/VERB books/NOUN

• Transition system
  • State
    • Partially labeled word-POS pairs
    • Unprocessed words

• Actions
  • TAG(t) \( w_1/t_1 \cdots w_i/t_i \rightarrow w_1/t_1 \cdots w_i/t_i w_{i+1}/t \)
A transition-based POS-tagging example

• Start State

I like reading books
A transition-based POS-tagging example

• TAG(PRON)

I/PRON like reading books
A transition-based POS-tagging example

• TAG(VERB)

I/PRON like/VERB reading books
A transition-based POS-tagging example

• TAG(VERB)

I/PRON like/VERB reading/VERB

books
A transition-based POS-tagging example

- TAG (NOUN)

I/PRON like/VERB reading/VERB books/NOUN
A transition-based POS-tagging example

• End State

I/PRON like/VERB reading/VERB books/NOUN
Word segmentation

- State
  - Partially segmented results
  - Unprocessed characters

- Two candidate actions
  - Separate   ##  ##  →  ##  ##  #
  - Append     ##  ##  →  ##  ##  #
Word segmentation

• Initial State

Word segmentation

• Separate

我 喜欢读书

Word segmentation

• Separate

Word segmentation

- Append

Word segmentation

• Separate

Word segmentation

• Separate

Word segmentation

• End State

我 喜欢 读书

The arc-eager transition system

• State
  • A stack to hold partial structures
  • A queue of next incoming words

• Actions
  • SHIFT, REDUCE, ARC-LEFT, ARC-RIGHT
The arc-eager transition system

- State
The arc-eager transition system

• Actions
  • Shift
The arc-eager transition system

- **Actions**
  - **Shift**
    - Pushes stack

![Diagram of the arc-eager transition system]

The stack

N0 N1 N2 N3 ...
The arc-eager transition system

- Actions
  - Reduce
The arc-eager transition system

- **Actions**
  - **Reduce**
    - Pops stack
The arc-eager transition system

- Actions
- Arc-Left
The arc-eager transition system

- **Actions**
  - **Arc-Left**
    - Pops stack
    - Adds link
The arc-eager transition system

- Actions
  - Arc-right

![Diagram of an arc-eager transition system]

The stack

The input
The arc-eager transition system

- **Actions**
  - **Arc-right**
    - Pushes stack
    - Adds link

```
  N0LC
  N0  ST  STP
  N1 N2 N3 ...
  The input

  The stack
  N0LC  STRC  STLC

  ...```
The arc-eager transition system

- An example
  - S – Shift
  - R – Reduce
  - AL – ArcLeft
  - AR – ArcRight

  He does it here
The arc-eager transition system

- An example
  - S – Shift
  - R – Reduce
  - AL – ArcLeft
  - AR – ArcRight

He does it here $\rightarrow$ S

He does it here
The arc-eager transition system

- An example
  - S – Shift
  - R – Reduce
  - AL – ArcLeft
  - AR – ArcRight

The arc-eager transition system
The arc-eager transition system

- An example
  - S – Shift
  - R – Reduce
  - AL – ArcLeft
  - AR – ArcRight

He does it here → S → He does it here → AL → does it here → S → does it here
The arc-eager transition system

- An example
  - S – Shift
  - R – Reduce
  - AL – ArcLeft
  - AR – ArcRight

He does it here \( \xrightarrow{S} \) He \( \xrightarrow{AL} \) does it here \( \xrightarrow{S} \) does it here

He \( \xrightarrow{AR} \) it here

He \( \xrightarrow{AR} \) here

He
The arc-eager transition system

- An example
  - S – Shift
  - R – Reduce
  - AL – ArcLeft
  - AR – ArcRight

```
He does it here   S   He does it here   AL   He does it here   S   does here
```

```
He
```

```
He
```

```
He
```

```
He
```

```
He
```

```
He
```
The arc-eager transition system

- An example
  - S – Shift
  - R – Reduce
  - AL – ArcLeft
  - AR – ArcRight

\[
\text{He does it here} \xrightarrow{S} \text{He does it here} \xrightarrow{AL} \text{does it here} \xrightarrow{S} \text{does it here}
\]
The arc-eager transition system

An example

- S – Shift
- R – Reduce
- AL – ArcLeft
- AR – ArcRight

He does it here → S → He does it here → AL → He does it here → S → He does it here

He does here ← R → He does here ← AR → He does here ← R → He does it here
Other examples

• Language generation

• Translation
  • Word by word
  • Phrase by phrase
  • Syntax tree synthesis
Part 5.1: Beam-search Decoding — learning to search
(Zhang and Clark, 2011)
Find the best sequence of actions
Beam-search decoding

Beam-search decoding

Beam-search decoding

Beam-search decoding

Beam-search decoding

Beam-search decoding

Beam-search decoding

function BEAM-SEARCH(problem, agenda, candidates, B)

candidates ← \{START\ITEM(problem)\}
agenda ← CLEAR(agenda)

loop do
    for each candidate in candidates
        agenda ← INSERT(EXPAND(candidate, problem), agenda)
        best ← TOP(agenda)
    if GOAL\TEST(problem, best)
        then return best
    candidates ← TOP-B(agenda, B)
    agenda ← CLEAR(agenda)

Beam-search decoding

- Our parser
- Decoding

Beam-search decoding

• Our parser

• Decoding

Beam-search decoding

• Our parser

• Decoding

Beam-search decoding

- Our parser
- Decoding

Beam-search decoding

• Our parser

• Decoding

Beam-search decoding

• Our parser

• Decoding

Beam-search decoding

• Our parser

• Decoding
Beam-search decoding

• Our parser

• Decoding
Online learning

Online learning

Online learning

Online learning

Online learning

Online learning

Inputs: training examples \((x_i, y_i = \{S_0^i S_1^i \cdots S_m^i\})\) is a state sequence \(1^N\)
Initialization: set \( \vec{w} = 0 \)
Algorithm:
for \( r = 1 \cdots P, i = 1 \cdots N \) do
  candidates \( \leftarrow \{S_0^i\} \)
  agenda \( \leftarrow \) CLEAR(agenda)
  for \( k = 1 \cdots m \), \( m \) corresponds to a specific training example. \( \textbf{do} \)
    for each candidate in candidates \( \textbf{do} \)
      agenda \( \leftarrow \) INSERT(EXPAND(candidate), agenda)
      candidates \( \leftarrow \) TOP \(-\) B(agenda, B)
      best \( \leftarrow \) TOP(agenda)
      if \( S_k^i \) is not in candidates or \((\text{best} \neq S_m^i \text{ and } k \text{ equals } m)\) then
        \( \vec{w} = \vec{w} + \Phi(S_k^i) - \Phi(\text{best}) \)
      end if
    end for
  end for
end for
Output: \( \vec{w} \)

The main strengths

• Fast
• Arbitrary nonlocal features
• Learning fixes search

State-of-the-art results

• Chinese
  • Word segmentation

State-of-the-art results

• Chinese
  • Joint segmentation and POS-tagging


State-of-the-art results

• Chinese
  • Joint segmentation, POS-tagging and chunking

State-of-the-art results

• Chinese
  • Joint segmentation, POS-tagging and dependency parsing

State-of-the-art results

• Chinese
  • Joint segmentation, POS-tagging and constituent parsing

State-of-the-art results

• Chinese
  • Joint segmentation, POS-tagging and normalization

  • Tao Qian, Yue Zhang, Meishan Zhang and Donghong Ji. A Transition-based Model for Joint Segmentation, POS-tagging and Normalization. In proceedings of EMNLP 2015, Lisboa, Portugal, September.
State-of-the-art results

• All Languages
  • Constituent parsing


State-of-the-art results

• All Languages
  • Dependency parsing

State-of-the-art results

• All Languages
  • CCG parsing


State-of-the-art results

• All Languages
  • Natural language synthesis

• Yijia Liu, Yue Zhang, Wanxiang Che and Bing Qin. Transition-Based Syntactic Linearization. In Proceedings of NAACL 2015, Denver, Colorado, USA, May.
State-of-the-art results

• All Languages
  • Joint morphological generation and text linearization

State-of-the-art results

• All Languages
  • Joint entity and relation extraction

Part 5.2: A Neural Network Version
Neural Network Model

• Use NN to substitute perceptron

• Why?
  - Better non-linear power
  - Unsupervised word embeddings
  - Automatic feature combination
  - Shown useful in greedy models
Word segmentation

<table>
<thead>
<tr>
<th>step</th>
<th>action</th>
<th>buffer($\cdots w_{-1}w_0$)</th>
<th>queue($c_0c_1\cdots$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>$\phi$</td>
<td>中国 ...</td>
</tr>
<tr>
<td>1</td>
<td>SEP</td>
<td></td>
<td>中国 外 ...</td>
</tr>
<tr>
<td>2</td>
<td>APP</td>
<td>中国</td>
<td>外企 ...</td>
</tr>
<tr>
<td>3</td>
<td>SEP</td>
<td>中国 外企</td>
<td>业务 ...</td>
</tr>
<tr>
<td>4</td>
<td>APP</td>
<td>中国 外企 业务</td>
<td>发展 ...</td>
</tr>
<tr>
<td>5</td>
<td>SEP</td>
<td>业务 发展</td>
<td>迅速</td>
</tr>
<tr>
<td>6</td>
<td>APP</td>
<td>业务 发展</td>
<td>迅速</td>
</tr>
<tr>
<td>7</td>
<td>SEP</td>
<td>业务 发展</td>
<td>迅速</td>
</tr>
<tr>
<td>8</td>
<td>APP</td>
<td>业务 发展</td>
<td>迅速</td>
</tr>
<tr>
<td>9</td>
<td>SEP</td>
<td>业务 发展</td>
<td>迅速</td>
</tr>
<tr>
<td>10</td>
<td>APP</td>
<td>业务 发展</td>
<td>迅速</td>
</tr>
</tbody>
</table>

## Word segmentation

<table>
<thead>
<tr>
<th>Feature templates</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_{-1}c_0$</td>
<td>$APP, SEP$</td>
</tr>
<tr>
<td>$w_{-1}, w_{-1}w_{-2}, w_{-1}c_0, w_{-2}len(w_{-1})$</td>
<td></td>
</tr>
<tr>
<td>$start(w_{-1})c_0, end(w_{-1})c_0$</td>
<td></td>
</tr>
<tr>
<td>$start(w_{-1})end(w_{-1}), end(w_{-2})end(w_{-1})$</td>
<td>$SEP$</td>
</tr>
<tr>
<td>$w_{-2}len(w_{-1}), len(w_{-2})w_{-1}$</td>
<td></td>
</tr>
<tr>
<td>$w_{-1}, \text{ where } len(w_{-1}) = 1$</td>
<td></td>
</tr>
</tbody>
</table>

Word segmentation

## Word segmentation

<table>
<thead>
<tr>
<th>Models</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>word-based models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>discrete</td>
<td>95.29</td>
<td>95.26</td>
<td>95.28</td>
</tr>
<tr>
<td>neural</td>
<td>95.34</td>
<td>94.69</td>
<td>95.01</td>
</tr>
<tr>
<td>combined</td>
<td><strong>96.11</strong></td>
<td><strong>95.79</strong></td>
<td><strong>95.95</strong></td>
</tr>
<tr>
<td><strong>character-based models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>discrete</td>
<td>95.38</td>
<td>95.12</td>
<td>95.25</td>
</tr>
<tr>
<td>neural</td>
<td>94.59</td>
<td>94.92</td>
<td>94.76</td>
</tr>
<tr>
<td>combined</td>
<td>95.63</td>
<td>95.60</td>
<td>95.61</td>
</tr>
<tr>
<td><strong>other models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang et al. (2014)</td>
<td>N/A</td>
<td>N/A</td>
<td>95.71</td>
</tr>
<tr>
<td>Wang et al. (2011)</td>
<td>95.83</td>
<td>95.75</td>
<td>95.79</td>
</tr>
<tr>
<td>Zhang and Clark (2011)</td>
<td>95.46</td>
<td>94.78</td>
<td>95.13</td>
</tr>
</tbody>
</table>

Main results on CTB60 test dataset

## Word segmentation

<table>
<thead>
<tr>
<th>Models</th>
<th>PKU</th>
<th>MSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>our word-based models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>discrete</td>
<td>95.1</td>
<td>97.3</td>
</tr>
<tr>
<td>neural</td>
<td>95.1</td>
<td>97.0</td>
</tr>
<tr>
<td>combined</td>
<td>95.7</td>
<td><strong>97.7</strong></td>
</tr>
<tr>
<td>character-based models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>discrete</td>
<td>94.9</td>
<td>96.8</td>
</tr>
<tr>
<td>neural</td>
<td>94.4</td>
<td>97.2</td>
</tr>
<tr>
<td>combined</td>
<td>95.4</td>
<td>97.2</td>
</tr>
<tr>
<td>other models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cai and Zhao (2016)</td>
<td>95.5</td>
<td>96.5</td>
</tr>
<tr>
<td>Ma and Hinrichs (2015)</td>
<td>95.1</td>
<td>96.6</td>
</tr>
<tr>
<td>Pei et al. (2014)</td>
<td>95.2</td>
<td>97.2</td>
</tr>
<tr>
<td>Zhang et al. (2013a)</td>
<td><strong>96.1</strong></td>
<td>97.5</td>
</tr>
<tr>
<td>Sun et al. (2012)</td>
<td>95.4</td>
<td>97.4</td>
</tr>
<tr>
<td>Zhang and Clark (2011)</td>
<td>95.1</td>
<td>97.1</td>
</tr>
<tr>
<td>Sun (2010)</td>
<td>95.2</td>
<td>96.9</td>
</tr>
<tr>
<td>Sun et al. (2009)</td>
<td>95.2</td>
<td>97.3</td>
</tr>
</tbody>
</table>

Main results on PKU and MSR test dataset

Word segmentation

• Cai and Zhao (2016) presents a similar idea

Dependency Parsing

• Zhang & Nivre (2011)

\[ y = \arg \max_{y' \in \text{GEN}(x)} \text{score}(y') \]

\[ \text{score}(y) = \sum_{a \in y} \theta \cdot \Phi(a) \]

Dependency Parsing

• Chen and Manning (2014)

\[ h = (W_1 x + b_1)^3 \]

\[ p = \text{softmax}(o) \]

\[ o = W_2 h \]

Dependency Parsing

• What does not work

\[ s(y) = \sum_{a \in y} \log p_a \]

\[ L(\theta) = \max(0, \delta - s(y_g) + s(y_p)) + \frac{\lambda}{2} \| \theta \|^2 \]

Dependency Parsing

• Sentence-level log likelihood

\[
p(y_i \mid x, \theta) = \frac{e^{f(x, \theta)_i}}{\sum_{y_j \in \text{GEN}(x)} e^{f(x, \theta)_j}}
\]

\[
f(x, \theta)_i = \sum_{a_k \in y_i} o(x, y_i, k, a_k)
\]

Dependency Parsing

• Contrastive Estimation

\[
L(\theta) = -\sum_{(x_i, y_i) \in (X, Y)} \log p(y_i \mid x_i, \theta)
\]

\[
= -\sum_{(x_i, y_i) \in (X, Y)} \log \frac{e^{f(x_i, \theta)_i}}{Z(x_i, \theta)}
\]

\[
= \sum_{(x_i, y_i) \in (X, Y)} \log Z(x_i, \theta) - f(x_i, \theta)_i
\]

\[
Z(x, \theta) = \sum_{y_j \in \text{GEN}(x)} e^{f(x, \theta)_j}
\]

Dependency Parsing

• Contrastive Estimation

\[ L'(\theta) = - \sum_{(x_i, y_i) \in (X,Y)} \log p'(y_i \mid x_i, \theta) \]

\[ = - \sum_{(x_i, y_i) \in (X,Y)} \log \frac{e^{f(x_i, \theta)_i}}{Z'(x_i, \theta)} \]

\[ = \sum_{(x_i, y_i) \in (X,Y)} \log Z'(x_i, \theta) - f(x_i, \theta)_i \]

\[ Z'(x, \theta) = \sum_{y_j \in \text{BEAM}(x)} e^{f(x, \theta)_j} \]

Dependency Parsing

• Results

<table>
<thead>
<tr>
<th>Description</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>91.63</td>
</tr>
<tr>
<td>beam = 1</td>
<td>74.90</td>
</tr>
<tr>
<td>beam = 4</td>
<td>84.64</td>
</tr>
<tr>
<td>beam = 16</td>
<td>91.53</td>
</tr>
<tr>
<td>beam = 64</td>
<td>93.12</td>
</tr>
<tr>
<td>beam = 100</td>
<td>93.23</td>
</tr>
</tbody>
</table>

Dependency Parsing

• Results

<table>
<thead>
<tr>
<th>Description</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>greedy neural parser</td>
<td>91.47</td>
</tr>
<tr>
<td>ranking model</td>
<td>89.08</td>
</tr>
<tr>
<td>beam contrastive learning</td>
<td>93.28</td>
</tr>
</tbody>
</table>

Dependency Parsing

• Results

<table>
<thead>
<tr>
<th>System</th>
<th>UAS</th>
<th>LAS</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline greedy parser</td>
<td>91.47</td>
<td>90.43</td>
<td>0.001</td>
</tr>
<tr>
<td>Huang and Sagae (2010)</td>
<td>92.10</td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>Zhang and Nivre (2011)</td>
<td>92.90</td>
<td>91.80</td>
<td>0.03</td>
</tr>
<tr>
<td>Choi and McCallum (2013)</td>
<td>92.96</td>
<td>91.93</td>
<td>0.009</td>
</tr>
<tr>
<td>Ma et al. (2014)</td>
<td>93.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bohnet and Nivre (2012)††</td>
<td>93.67</td>
<td>92.68</td>
<td>0.4</td>
</tr>
<tr>
<td>Suzuki et al. (2009)†</td>
<td>93.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Koo et al. (2008)†</td>
<td>93.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chen et al. (2014)†</td>
<td>93.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>beam size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>training</td>
<td>100</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>decoding</td>
<td>100</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>93.28</td>
<td>92.35</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>93.20</td>
<td>92.27</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>92.40</td>
<td>91.95</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Google

• Andor et al. follows this method
  • Offers theorem
  • Tries more tasks
  • Get better results

## Dependency parsing

<table>
<thead>
<tr>
<th>Method</th>
<th>WSJ UAS</th>
<th>WSJ LAS</th>
<th>Union-News UAS</th>
<th>Union-News LAS</th>
<th>Union-Web UAS</th>
<th>Union-Web LAS</th>
<th>Union-QTB UAS</th>
<th>Union-QTB LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martins et al. (2013)*</td>
<td>92.89</td>
<td>90.55</td>
<td>93.10</td>
<td>91.13</td>
<td>88.23</td>
<td>85.04</td>
<td>94.21</td>
<td>91.54</td>
</tr>
<tr>
<td>Zhang and McDonald (2014)*</td>
<td>93.22</td>
<td>91.02</td>
<td>93.32</td>
<td>91.48</td>
<td>88.65</td>
<td>85.59</td>
<td>93.37</td>
<td>90.69</td>
</tr>
<tr>
<td>Weiss et al. (2015)</td>
<td>93.99</td>
<td>92.05</td>
<td>93.91</td>
<td>92.25</td>
<td>89.29</td>
<td>86.44</td>
<td>94.17</td>
<td>92.06</td>
</tr>
<tr>
<td>Alberti et al. (2015)</td>
<td>94.23</td>
<td>92.36</td>
<td>94.10</td>
<td>92.55</td>
<td>89.55</td>
<td>86.85</td>
<td>94.74</td>
<td>93.04</td>
</tr>
<tr>
<td>Our Local (B=1)</td>
<td>92.95</td>
<td>91.02</td>
<td>93.11</td>
<td>91.46</td>
<td>88.42</td>
<td>85.58</td>
<td>92.49</td>
<td>90.38</td>
</tr>
<tr>
<td>Our Local (B=32)</td>
<td>93.59</td>
<td>91.70</td>
<td>93.65</td>
<td>92.03</td>
<td>88.96</td>
<td>86.17</td>
<td>93.22</td>
<td>91.17</td>
</tr>
<tr>
<td>Our Global (B=32)</td>
<td><strong>94.61</strong></td>
<td><strong>92.79</strong></td>
<td><strong>94.44</strong></td>
<td><strong>92.93</strong></td>
<td><strong>90.17</strong></td>
<td><strong>87.54</strong></td>
<td><strong>95.40</strong></td>
<td><strong>93.64</strong></td>
</tr>
<tr>
<td>Parsey McParseface (B=8)</td>
<td>-</td>
<td>-</td>
<td>94.15</td>
<td>92.51</td>
<td>89.08</td>
<td>86.29</td>
<td>94.77</td>
<td>93.17</td>
</tr>
</tbody>
</table>

### Dependency parsing

<table>
<thead>
<tr>
<th>Method</th>
<th>Catalan UAS</th>
<th>Catalan LAS</th>
<th>Chinese UAS</th>
<th>Chinese LAS</th>
<th>Czech UAS</th>
<th>Czech LAS</th>
<th>English UAS</th>
<th>English LAS</th>
<th>German UAS</th>
<th>German LAS</th>
<th>Japanese UAS</th>
<th>Japanese LAS</th>
<th>Spanish UAS</th>
<th>Spanish LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Shared Task Result</td>
<td>-</td>
<td>87.86</td>
<td>-</td>
<td>79.17</td>
<td>-</td>
<td>80.38</td>
<td>-</td>
<td>89.88</td>
<td>-</td>
<td>87.48</td>
<td>-</td>
<td>92.57</td>
<td>-</td>
<td>87.64</td>
</tr>
<tr>
<td>Ballesteros et al. (2015)</td>
<td>90.22</td>
<td>86.42</td>
<td>80.64</td>
<td>76.52</td>
<td>79.87</td>
<td>73.62</td>
<td>90.56</td>
<td>88.01</td>
<td>88.83</td>
<td>86.10</td>
<td>93.47</td>
<td>92.55</td>
<td>90.38</td>
<td>86.59</td>
</tr>
<tr>
<td>Zhang and McDonald (2014)</td>
<td>91.41</td>
<td>87.91</td>
<td>82.87</td>
<td>78.57</td>
<td>86.62</td>
<td>80.59</td>
<td>92.69</td>
<td>90.01</td>
<td>89.88</td>
<td>87.38</td>
<td>92.82</td>
<td>91.87</td>
<td>90.82</td>
<td>87.34</td>
</tr>
<tr>
<td>Lei et al. (2014)</td>
<td>91.33</td>
<td>87.22</td>
<td>81.67</td>
<td>76.71</td>
<td>88.76</td>
<td>81.77</td>
<td>92.75</td>
<td>90.00</td>
<td>90.81</td>
<td>87.81</td>
<td>94.04</td>
<td>91.84</td>
<td>91.16</td>
<td>87.38</td>
</tr>
<tr>
<td>Bohnet and Nivre (2012)</td>
<td>92.44</td>
<td>89.60</td>
<td>82.52</td>
<td>78.51</td>
<td>88.82</td>
<td>83.73</td>
<td>92.87</td>
<td>90.60</td>
<td>91.37</td>
<td>89.38</td>
<td>93.67</td>
<td>92.63</td>
<td>92.24</td>
<td>89.60</td>
</tr>
<tr>
<td>Alberti et al. (2015)</td>
<td>92.31</td>
<td>89.17</td>
<td>83.57</td>
<td>79.90</td>
<td>88.45</td>
<td>83.57</td>
<td>92.70</td>
<td>90.56</td>
<td>90.58</td>
<td>88.20</td>
<td>93.99</td>
<td>93.10</td>
<td>92.26</td>
<td>89.33</td>
</tr>
<tr>
<td>Our Local (B=1)</td>
<td>91.24</td>
<td>88.21</td>
<td>81.29</td>
<td>77.29</td>
<td>85.78</td>
<td>80.63</td>
<td>91.44</td>
<td>89.29</td>
<td>89.12</td>
<td>86.95</td>
<td>93.71</td>
<td>92.85</td>
<td>91.01</td>
<td>88.14</td>
</tr>
<tr>
<td>Our Local (B=16)</td>
<td>91.91</td>
<td>88.93</td>
<td>82.22</td>
<td>78.26</td>
<td>86.25</td>
<td>81.28</td>
<td>92.16</td>
<td>90.05</td>
<td>89.53</td>
<td>87.4</td>
<td>93.61</td>
<td>92.74</td>
<td>91.64</td>
<td>88.88</td>
</tr>
<tr>
<td>Our Global (B=16)</td>
<td><strong>92.67</strong></td>
<td><strong>89.83</strong></td>
<td><strong>84.72</strong></td>
<td><strong>80.85</strong></td>
<td><strong>88.94</strong></td>
<td><strong>84.56</strong></td>
<td><strong>93.22</strong></td>
<td><strong>91.23</strong></td>
<td><strong>90.91</strong></td>
<td><strong>89.15</strong></td>
<td><strong>93.65</strong></td>
<td><strong>92.84</strong></td>
<td><strong>92.62</strong></td>
<td><strong>89.95</strong></td>
</tr>
</tbody>
</table>

### POS-tagging

<table>
<thead>
<tr>
<th>Method</th>
<th>En WSJ</th>
<th>En-Union News</th>
<th>En-Union Web</th>
<th>En-Union QTB</th>
<th>CoNLL ’09 Ca</th>
<th>CoNLL ’09 Ch</th>
<th>CoNLL ’09 Cz</th>
<th>CoNLL ’09 En</th>
<th>CoNLL ’09 Ge</th>
<th>CoNLL ’09 Ja</th>
<th>CoNLL ’09 Sp</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear CRF</td>
<td>97.17</td>
<td>97.60</td>
<td>94.58</td>
<td>96.04</td>
<td>98.81</td>
<td>94.45</td>
<td>98.90</td>
<td>97.50</td>
<td>97.14</td>
<td>97.90</td>
<td>98.79</td>
<td>97.17</td>
</tr>
<tr>
<td>Ling et al. (2015)</td>
<td><strong>97.78</strong></td>
<td>97.44</td>
<td>94.03</td>
<td>96.18</td>
<td>98.77</td>
<td>94.38</td>
<td>99.00</td>
<td>97.60</td>
<td><strong>97.84</strong></td>
<td>97.06</td>
<td>98.71</td>
<td>97.16</td>
</tr>
<tr>
<td>Our Local (B=1)</td>
<td>97.44</td>
<td>97.66</td>
<td>94.46</td>
<td>96.59</td>
<td>98.91</td>
<td>94.56</td>
<td>98.96</td>
<td>97.36</td>
<td>97.35</td>
<td>98.02</td>
<td>98.88</td>
<td>97.29</td>
</tr>
<tr>
<td>Our Local (B=8)</td>
<td>97.45</td>
<td>97.69</td>
<td>94.46</td>
<td>96.64</td>
<td>98.88</td>
<td>94.56</td>
<td>98.96</td>
<td>97.40</td>
<td>97.35</td>
<td>98.02</td>
<td>98.89</td>
<td>97.30</td>
</tr>
<tr>
<td>Our Global (B=8)</td>
<td>97.44</td>
<td><strong>97.77</strong></td>
<td><strong>94.80</strong></td>
<td><strong>96.86</strong></td>
<td><strong>99.03</strong></td>
<td><strong>94.72</strong></td>
<td><strong>99.02</strong></td>
<td><strong>97.65</strong></td>
<td>97.52</td>
<td><strong>98.37</strong></td>
<td><strong>98.97</strong></td>
<td><strong>97.47</strong></td>
</tr>
<tr>
<td>Parsley McParseface</td>
<td>-</td>
<td>97.52</td>
<td>94.24</td>
<td>96.45</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Google

Compression

<table>
<thead>
<tr>
<th>Method</th>
<th>Generated corpus</th>
<th>Human eval</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>F1</td>
<td>read</td>
</tr>
<tr>
<td>Filippova et al. (2015)</td>
<td>35.36</td>
<td>82.83</td>
<td>4.66</td>
</tr>
<tr>
<td>Automatic</td>
<td>-</td>
<td>-</td>
<td>4.31</td>
</tr>
<tr>
<td>Our Local (B=1)</td>
<td>30.51</td>
<td>78.72</td>
<td>4.58</td>
</tr>
<tr>
<td>Our Local (B=8)</td>
<td>31.19</td>
<td>75.69</td>
<td>-</td>
</tr>
<tr>
<td>Our Global (B=8)</td>
<td>35.16</td>
<td>81.41</td>
<td><strong>4.67</strong></td>
</tr>
</tbody>
</table>

Part 5.3: Similar methods by others
Other methods (I)

• Constituent parsing

Watanabe, T., & Sumita, E. (2016). Transition-based Neural Constituent Parsing. ACL.
Other methods (I)

- Update at max-violation

\[ j^* = \arg \min_j \left\{ \rho_\theta(y_0^j) - \max_{d \in B_j} \rho_\theta(d) \right\} \]

- Using expected loss from all violations

\[ L(w, y; B, \theta) = \max \left\{ 0, 1 - \rho_\theta(y_0^{j^*}) + \mathbb{E}_{\tilde{B}_{j^*}}[\rho_\theta] \right\} \]

\[ \tilde{B}_{j^*} = \left\{ d \in B_{j^*} | \rho_\theta(d) > \rho_\theta(y_0^{j^*}) \right\} \]

\[ p_\theta(d) = \frac{\exp(\rho_\theta(d))}{\sum_{d' \in \tilde{B}_{j^*}} \exp(\rho_\theta(d'))} \]

\[ \mathbb{E}_{\tilde{B}_{j^*}}[\rho_\theta] = \sum_{d \in \tilde{B}_{j^*}} p_\theta(d) \rho_\theta(d). \]

Watanabe, T., & Sumita, E. (2016). Transition-based Neural Constituent Parsing. ACL.
Other methods (I)

<table>
<thead>
<tr>
<th>parser</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collins (Collins, 1997)</td>
<td>87.8</td>
</tr>
<tr>
<td>Berkeley (Petrov and Klein, 2007)</td>
<td>90.1</td>
</tr>
<tr>
<td>SSN (Henderson, 2004)</td>
<td>90.1</td>
</tr>
<tr>
<td>ZPar (Zhu et al., 2013)</td>
<td>90.4</td>
</tr>
<tr>
<td>CVG (Socher et al., 2013)</td>
<td>90.4</td>
</tr>
<tr>
<td>Charniak-R (Charniak and Johnson, 2005)</td>
<td><strong>91.0</strong></td>
</tr>
<tr>
<td>This work: TNCP</td>
<td>90.7</td>
</tr>
</tbody>
</table>

Watanabe, T., & Sumita, E. (2016). Transition-based Neural Constituent Parsing. ACL.
Other methods (I)

<table>
<thead>
<tr>
<th>parser</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZPar (Zhu et al., 2013)</td>
<td>83.2</td>
</tr>
<tr>
<td>Berkeley (Petrov and Klein, 2007)</td>
<td>83.3</td>
</tr>
<tr>
<td>Joint (Wang and Xue, 2014)</td>
<td><strong>84.9</strong></td>
</tr>
<tr>
<td>This work: TNCP</td>
<td>84.3</td>
</tr>
</tbody>
</table>

Watanabe, T., & Sumita, E. (2016). Transition-based Neural Constituent Parsing. ACL.
Other methods (II)

• CCG Parsing

• expected F1 training

\[
J(\theta) = -xF1(\theta) \\
= - \sum_{y_i \in \Lambda(x_n)} p(y_i | \theta) F1(\Delta_{y_i}, \Delta_{x_n}^G)
\]

\[
p(y_i | \theta) = \frac{\exp\{\rho(y_i)\}}{\sum_{y \in \Lambda(x_n)} \exp\{\rho(y)\}}
\]

Other methods (II)

\[
\frac{\partial J(\theta)}{\partial \theta} = - \sum_{y_i \in \Lambda(x_n)} \sum_{y_{ij} \in y_i} \frac{\partial J(\theta)}{\partial s_\theta(y_{ij})} \frac{\partial s_\theta(y_{ij})}{\partial \theta}
\]

\[
= - \sum_{y_i \in \Lambda(x_n)} \sum_{y_{ij} \in y_i} \delta_{y_{ij}} \frac{\partial s_\theta(y_{ij})}{\partial \theta},
\]

where

\[
\delta_{y_{ij}} = \frac{- \partial xF1(\theta)}{\partial s_\theta(y_{ij})}
\]

\[
= - \frac{\partial (G(\theta)/Z(\theta))}{\partial s_\theta(y_{ij})}
\]

\[
= \frac{G(\theta)Z'(\theta) - G'(\theta)Z(\theta)}{Z^2(\theta)}
\]

\[
= \frac{\exp\{\rho(y_i)\}}{Z(\theta)}(xF1(\theta) - F1(\Delta_{y_i, \Delta^G_{x_n}})) \frac{1}{s_\theta(y_{ij})}
\]

\[
= p(y_i | \theta)(xF1(\theta) - F1(\Delta_{y_i, \Delta^G_{x_n}})) \frac{1}{s_\theta(y_{ij})},
\]

Other methods (\textit{II})

<table>
<thead>
<tr>
<th>Model</th>
<th>Section 00</th>
<th></th>
<th></th>
<th>Section 23</th>
<th></th>
<th></th>
<th></th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LP</td>
<td>LR</td>
<td>LF</td>
<td>CAT</td>
<td>LP</td>
<td>LR</td>
<td>LF</td>
<td>CAT</td>
</tr>
<tr>
<td>C&amp;C (normal)</td>
<td>85.18</td>
<td>82.53</td>
<td>83.83</td>
<td>92.39</td>
<td>85.45</td>
<td>83.97</td>
<td>84.70</td>
<td>92.83</td>
</tr>
<tr>
<td>C&amp;C (hybrid)</td>
<td>86.07</td>
<td>82.77</td>
<td>84.39</td>
<td>92.57</td>
<td>86.24</td>
<td>84.17</td>
<td>85.19</td>
<td>93.00</td>
</tr>
<tr>
<td>Zhang and Clark (2011) ($b = 16$)</td>
<td>87.15</td>
<td>82.95</td>
<td>85.00</td>
<td>92.77</td>
<td>87.43</td>
<td>83.61</td>
<td>85.48</td>
<td>93.12</td>
</tr>
<tr>
<td>Zhang and Clark (2011)* ($b = 16$)</td>
<td>86.76</td>
<td>83.15</td>
<td>84.92</td>
<td>92.64</td>
<td>87.04</td>
<td>84.14</td>
<td>85.56</td>
<td>92.95</td>
</tr>
<tr>
<td>Xu et al. (2014) ($b = 128$)</td>
<td>86.29</td>
<td>84.09</td>
<td>85.18</td>
<td>92.75</td>
<td>87.03</td>
<td>85.08</td>
<td>86.04</td>
<td>93.10</td>
</tr>
<tr>
<td>RNN-greedy ($b = 1$)</td>
<td>88.12</td>
<td>81.38</td>
<td>84.61</td>
<td>93.42</td>
<td>88.53</td>
<td>81.65</td>
<td>84.95</td>
<td>93.57</td>
</tr>
<tr>
<td>RNN-greedy ($b = 6$)</td>
<td>87.96</td>
<td>82.27</td>
<td>85.02</td>
<td>93.47</td>
<td>88.54</td>
<td>82.77</td>
<td>85.56</td>
<td>93.68</td>
</tr>
<tr>
<td>RNN-xF1 ($b = 8$)</td>
<td><strong>88.20</strong></td>
<td>83.40</td>
<td><strong>85.73</strong></td>
<td><strong>93.56</strong></td>
<td><strong>88.74</strong></td>
<td>84.22</td>
<td><strong>86.42</strong></td>
<td><strong>93.87</strong></td>
</tr>
</tbody>
</table>

Other methods (III)

• Dependency parsing

Other methods (Ⅲ)

• Using Chen and Manning features for perceptron training

• Back-propagation pre-training

\[
L(\Theta) = -\sum_j \log P(y_j \mid c_j, \Theta) + \lambda \sum_i ||W_i||_2^2
\]

• Structured perceptron training

\[(h_1, h_2, P(y))\]

Other methods (III)

<table>
<thead>
<tr>
<th>Method</th>
<th>UAS</th>
<th>LAS</th>
<th>Beam</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Graph-based</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bohnet (2010)</td>
<td>92.88</td>
<td>90.71</td>
<td>n/a</td>
</tr>
<tr>
<td>Martins et al. (2013)</td>
<td>92.89</td>
<td>90.55</td>
<td>n/a</td>
</tr>
<tr>
<td>Zhang and McDonald (2014)</td>
<td>93.22</td>
<td>91.02</td>
<td>n/a</td>
</tr>
<tr>
<td><strong>Transition-based</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*Zhang and Nivre (2011)</td>
<td>93.00</td>
<td>90.95</td>
<td>32</td>
</tr>
<tr>
<td>Bohnet and Kuhn (2012)</td>
<td>93.27</td>
<td>91.19</td>
<td>40</td>
</tr>
<tr>
<td>Chen and Manning (2014)</td>
<td>91.80</td>
<td>89.60</td>
<td>1</td>
</tr>
<tr>
<td>S-LSTM (Dyer et al., 2015)</td>
<td>93.20</td>
<td>90.90</td>
<td>1</td>
</tr>
<tr>
<td>Our Greedy</td>
<td>93.19</td>
<td>91.18</td>
<td>1</td>
</tr>
<tr>
<td>Our Perceptron</td>
<td>93.99</td>
<td>92.05</td>
<td>8</td>
</tr>
<tr>
<td><strong>Tri-training</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*Zhang and Nivre (2011)</td>
<td>92.92</td>
<td>90.88</td>
<td>32</td>
</tr>
<tr>
<td>Our Greedy</td>
<td>93.46</td>
<td>91.49</td>
<td>1</td>
</tr>
<tr>
<td>Our Perceptron</td>
<td>94.26</td>
<td>92.41</td>
<td>8</td>
</tr>
</tbody>
</table>
Other methods (IV)

• Dependency parsing

Vaswani, A., & Sagae, K. (2016). Efficient Structured Inference for Transition-Based Parsing with Neural Networks and Error States. TACL.
Other methods \( (IV) \)

Vaswani, A., & Sagae, K. (2016). Efficient Structured Inference for Transition-Based Parsing with Neural Networks and Error States. TACL.
Other methods (IV)

Vaswani, A., & Sagae, K. (2016). Efficient Structured Inference for Transition-Based Parsing with Neural Networks and Error States. TACL.
### Other methods (IV)

<table>
<thead>
<tr>
<th>System</th>
<th>wsj23-S</th>
<th>wsj23-YM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ErrSt–25–rand</td>
<td>92.17</td>
<td>92.16</td>
</tr>
<tr>
<td>ErrSt–25–pre*</td>
<td>93.61</td>
<td>93.21</td>
</tr>
<tr>
<td>Chen &amp; Manning*</td>
<td>91.8</td>
<td>–</td>
</tr>
<tr>
<td>Huang &amp; Sagae</td>
<td>–</td>
<td>92.1</td>
</tr>
<tr>
<td>Zhang &amp; Nivre</td>
<td>93.5</td>
<td>92.9</td>
</tr>
<tr>
<td>Weiss et al.*</td>
<td>93.99</td>
<td>–</td>
</tr>
<tr>
<td>Zhang &amp; McDonald</td>
<td>93.71</td>
<td>93.57</td>
</tr>
<tr>
<td>Martins et al.</td>
<td>92.82</td>
<td>93.07</td>
</tr>
<tr>
<td>Koo et al. (dep2c)*</td>
<td>–</td>
<td>93.16</td>
</tr>
</tbody>
</table>

Vaswani, A., & Sagae, K. (2016). Efficient Structured Inference for Transition-Based Parsing with Neural Networks and Error States. TACL.
Part 5.4: Beam-search Decoding for Sequence to Sequence Models
Sequence to sequence (Ⅰ)

• Scheduled Sampling

Beam Search Inference

Sequence to sequence (\( I \))

- Scheduled Sampling

<table>
<thead>
<tr>
<th>Approach</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline LSTM</td>
<td>86.54</td>
</tr>
<tr>
<td>Baseline LSTM with Dropout</td>
<td>87.0</td>
</tr>
<tr>
<td>Always Sampling</td>
<td>-</td>
</tr>
<tr>
<td>Scheduled Sampling</td>
<td>88.08</td>
</tr>
<tr>
<td>Scheduled Sampling with Dropout</td>
<td>88.68</td>
</tr>
</tbody>
</table>

Sequence to sequence ($II$)

- **Sequence-level training**

Sequence to sequence (II)

• Sequence-level training

Sequence to sequence (\( II \))

• Sequence-level training

Sequence to sequence (II)

• Reinforce

\[ L_\theta = - \sum_{w^g_1, \ldots, w^g_T} p_\theta(w^g_1, \ldots, w^g_T) \rho(w^g_1, \ldots, w^g_T) = -\mathbb{E}_{w^g_1, \ldots, w^g_T \sim p_\theta} \rho(w^g_1, \ldots, w^g_T) \]

Sequence to sequence (Ⅱ)

- Mixer

Sequence to sequence (II)

Data: a set of sequences with their corresponding context.
Result: RNN optimized for generation.

Initialize RNN at random and set $N^{\text{XENT}}$, $N^{\text{XE+R}}$ and $\Delta$;
for $s = T, 1, -\Delta$ do
    if $s == T$ then
        train RNN for $N^{\text{XENT}}$ epochs using XENT only;
    else
        train RNN for $N^{\text{XE+R}}$ epochs. Use XENT loss in the first $s$ steps, and REINFORCE (sampling from the model) in the remaining $T - s$ steps;
    end
end

Sequence to sequence (II)

<table>
<thead>
<tr>
<th>TASK</th>
<th>XENT</th>
<th>DAD</th>
<th>E2E</th>
<th>MIXER</th>
</tr>
</thead>
<tbody>
<tr>
<td>summarization</td>
<td>13.01</td>
<td>12.18</td>
<td>12.78</td>
<td><strong>16.22</strong></td>
</tr>
<tr>
<td>translation</td>
<td>17.74</td>
<td>20.12</td>
<td>17.77</td>
<td><strong>20.73</strong></td>
</tr>
<tr>
<td>image captioning</td>
<td>27.8</td>
<td>28.16</td>
<td>26.42</td>
<td><strong>29.16</strong></td>
</tr>
</tbody>
</table>

Sequence to sequence (Ⅲ)

• Learning for Search

Sequence to sequence (III)

\[ \mathcal{L}(f) = \sum_{t=1}^{T} \Delta(\hat{y}^{(K)}_{1:t}) \left[ 1 - f(y_t, h_{t-1}) + f(\hat{y}^{(K)}_t, \hat{h}^{(K)}_{t-1}) \right] \]

Sequence to sequence (III)

• Need greedy pre-training

Sequence to sequence (Ⅲ)

- Curriculum beam increase

## Sequence to sequence (III)

<table>
<thead>
<tr>
<th>Model</th>
<th>$K_{te} = 1$</th>
<th>$K_{te} = 5$</th>
<th>$K_{te} = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq2seq</td>
<td>25.2</td>
<td>29.8</td>
<td>31.0</td>
</tr>
<tr>
<td>BSO</td>
<td>28.0</td>
<td>33.2</td>
<td>34.3</td>
</tr>
<tr>
<td>ConBSO</td>
<td><strong>28.6</strong></td>
<td><strong>34.3</strong></td>
<td><strong>34.5</strong></td>
</tr>
<tr>
<td>LSTM-LM</td>
<td>15.4</td>
<td>-</td>
<td>26.8</td>
</tr>
</tbody>
</table>

## Sequence to sequence (III)

<table>
<thead>
<tr>
<th>Model</th>
<th>UAS</th>
<th>LAS</th>
<th>UAS</th>
<th>LAS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_{te} = 1$</td>
<td>$K_{te} = 5$</td>
<td>$K_{te} = 10$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq2seq</td>
<td>87.33/82.26</td>
<td>88.53/84.16</td>
<td>88.66/84.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSO</td>
<td>86.91/82.11</td>
<td>91.00/87.18</td>
<td>91.17/87.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ConBSO</td>
<td>85.11/79.32</td>
<td>91.25/86.92</td>
<td>91.57/87.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andor</td>
<td>93.17/91.18</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Sequence to sequence (Ⅲ)

<table>
<thead>
<tr>
<th>Model</th>
<th>Machine Translation ($K_{te} = 1$)</th>
<th>Machine Translation ($K_{te} = 5$)</th>
<th>Machine Translation ($K_{te} = 10$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq2seq</td>
<td>22.53</td>
<td>24.03</td>
<td>23.87</td>
</tr>
<tr>
<td>BSO, SB-Δ</td>
<td><strong>23.83</strong></td>
<td><strong>26.36</strong></td>
<td><strong>25.48</strong></td>
</tr>
<tr>
<td>XENT</td>
<td>17.74</td>
<td>≤ 20.5</td>
<td>≤ 20.5</td>
</tr>
<tr>
<td>DAD</td>
<td>20.12</td>
<td>≤ 22.5</td>
<td>≤ 23.0</td>
</tr>
<tr>
<td>MIXER</td>
<td>20.73</td>
<td>-</td>
<td>≤ 22.0</td>
</tr>
</tbody>
</table>