# Deep Learning and Lexical, Syntactic and Semantic Analysis

Wanxiang Che (HIT)
Yue Zhang (SUTD)

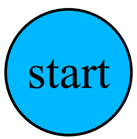
# Part 5: Beam-search Decoding

# 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

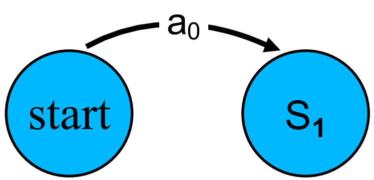
# A transition system

- Automata



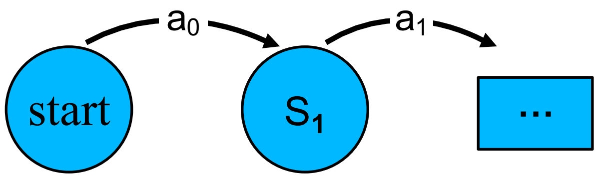
# A transition system

- Automata



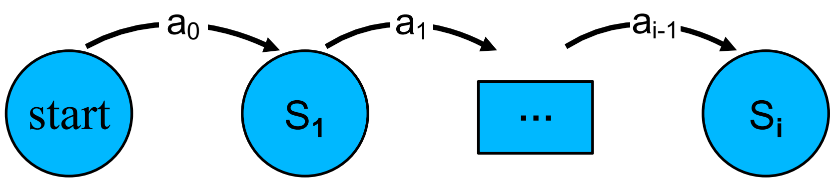
# A transition system

- Automata



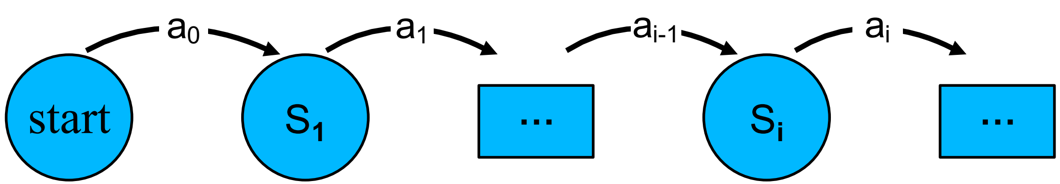
# A transition system

- Automata



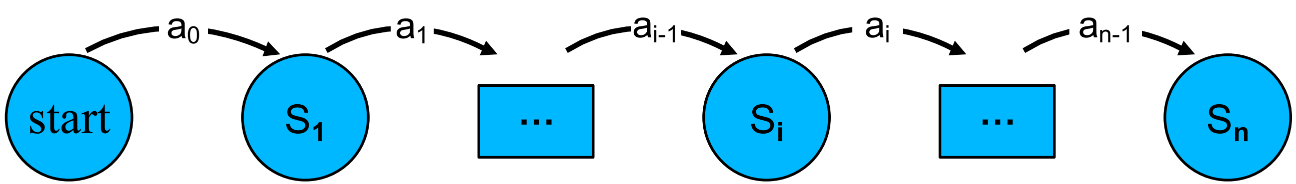
# A transition system

- Automata



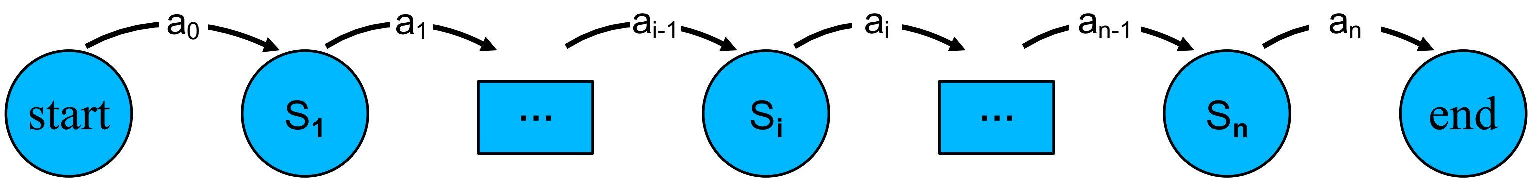
# A transition system

- Automata



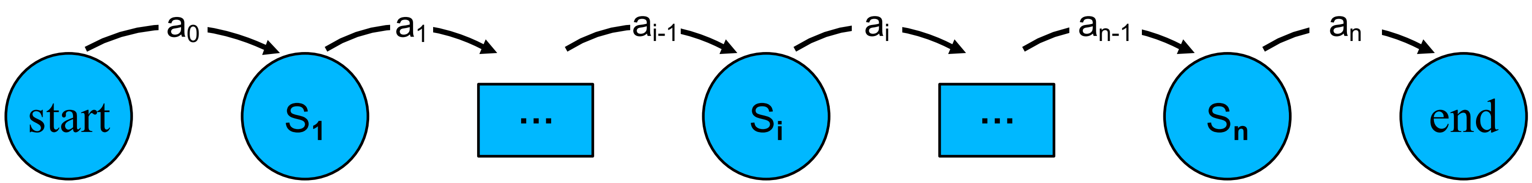
# A transition system

- 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 $\rightarrow$ I/PRON like/VERB reading/VERB books/NOUN

- Transition system
  - State
    - Partially labeled word-POS pairs
    - Unprocessed words
  - Actions
    - TAG(t) $w_1/t_1 \cdots wi/t_i \rightarrow w_1/t_1 \cdots w_i/t_i \, w_{i+1}/t$

# 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

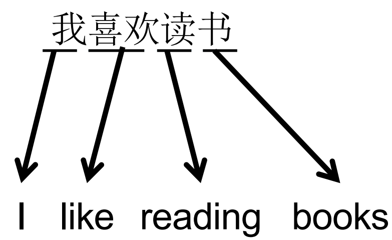
- State
  - Partially segmented results
  - Unprocessed characters

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

Zhang, Y., & Clark, S. (2007). Chinese Segmentation Using a Word-Based Perceptron Algorithm. ACL.

# Word segmentation

- Initial State




Zhang, Y., & Clark, S. (2007). Chinese Segmentation Using a Word-Based Perceptron Algorithm. ACL.

# Word segmentation

- Separate

我

喜欢读书

Zhang, Y., & Clark, S. (2007). Chinese Segmentation Using a Word-Based Perceptron Algorithm. ACL.

# Word segmentation

- Separate

我　喜

欢读书

Zhang, Y., & Clark, S. (2007). Chinese Segmentation Using a Word-Based Perceptron Algorithm. ACL.

# Word segmentation

- Append

| 我　喜欢 | 读书 |
|---|---|

Zhang, Y., & Clark, S. (2007). Chinese Segmentation Using a Word-Based Perceptron Algorithm. ACL.

# Word segmentation

- Separate

我　喜欢　读　　　　　　　　书

Zhang, Y., & Clark, S. (2007). Chinese Segmentation Using a Word-Based Perceptron Algorithm. ACL.

# Word segmentation

- Separate

我　喜欢　读　书

Zhang, Y., & Clark, S. (2007). Chinese Segmentation Using a Word-Based Perceptron Algorithm. ACL.

# Word segmentation

- End State

我 喜欢 读 书

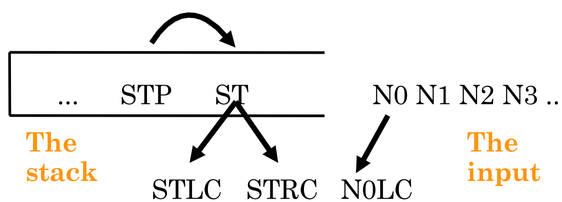
Zhang, Y., & Clark, S. (2007). Chinese Segmentation Using a Word-Based Perceptron Algorithm. ACL.

# 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

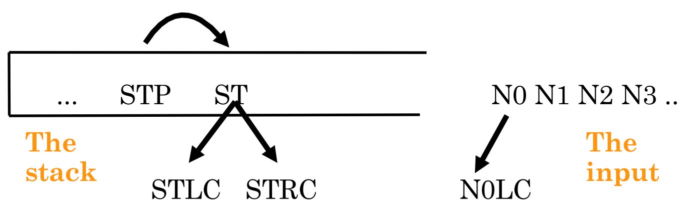
# The arc-eager transition system

- State



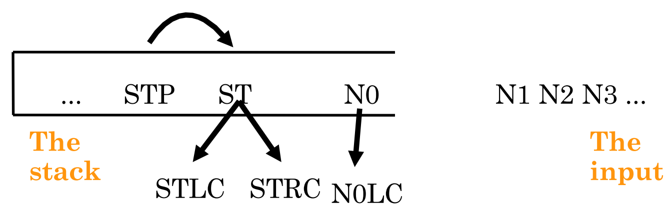
# The arc-eager transition system

- Actions
  - Shift



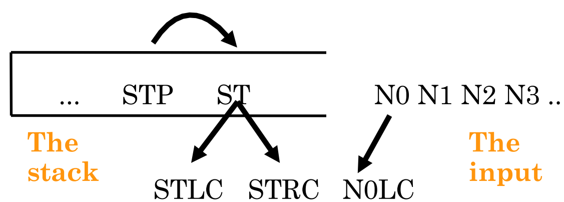
# The arc-eager transition system

- Actions
  - Shift
    - ➢ Pushes stack



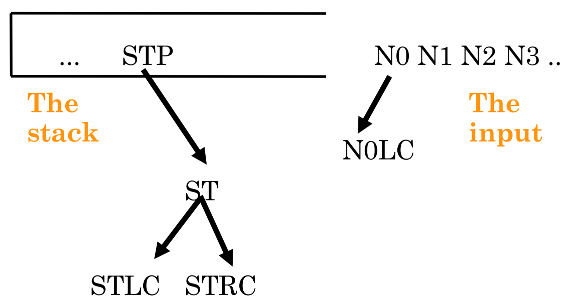
# The arc-eager transition system

- Actions
  - Reduce



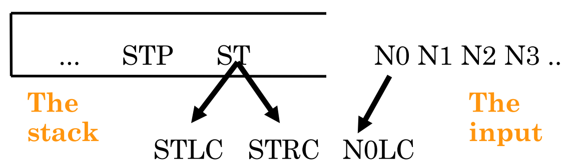
# The arc-eager transition system

- Actions
  - Reduce
    - ➢ Pops stack



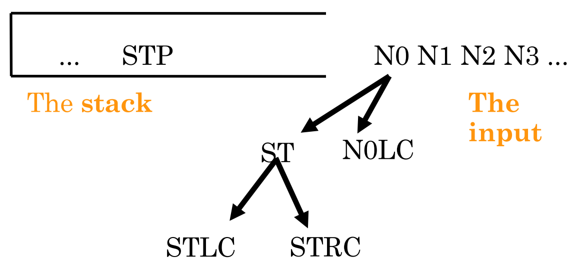
# The arc-eager transition system

- Actions
  - Arc-Left



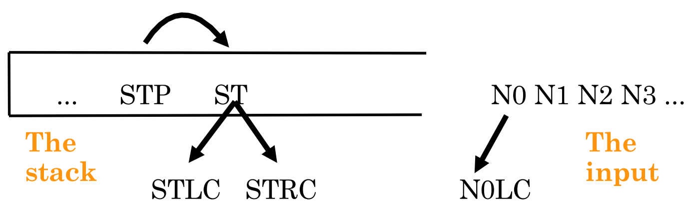
# The arc-eager transition system

- Actions
  - Arc-Left
    - ➢ Pops stack
    - ➢ Adds link



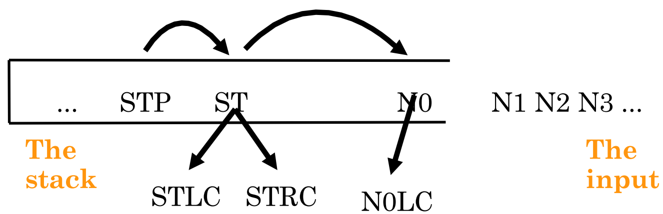
# The arc-eager transition system

- Actions
  - Arc-right



# The arc-eager transition system

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



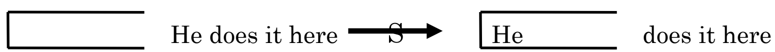
# The arc-eager transition system

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

He does it here

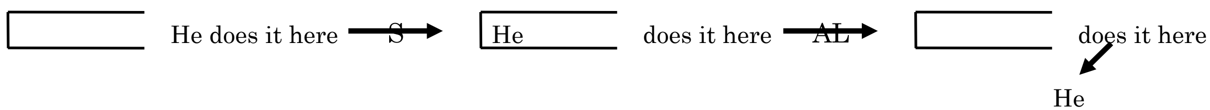
# The arc-eager transition system

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



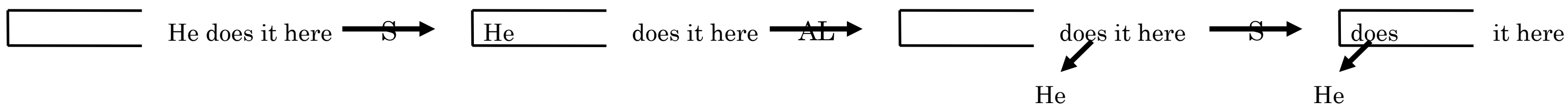
# The arc-eager transition system

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



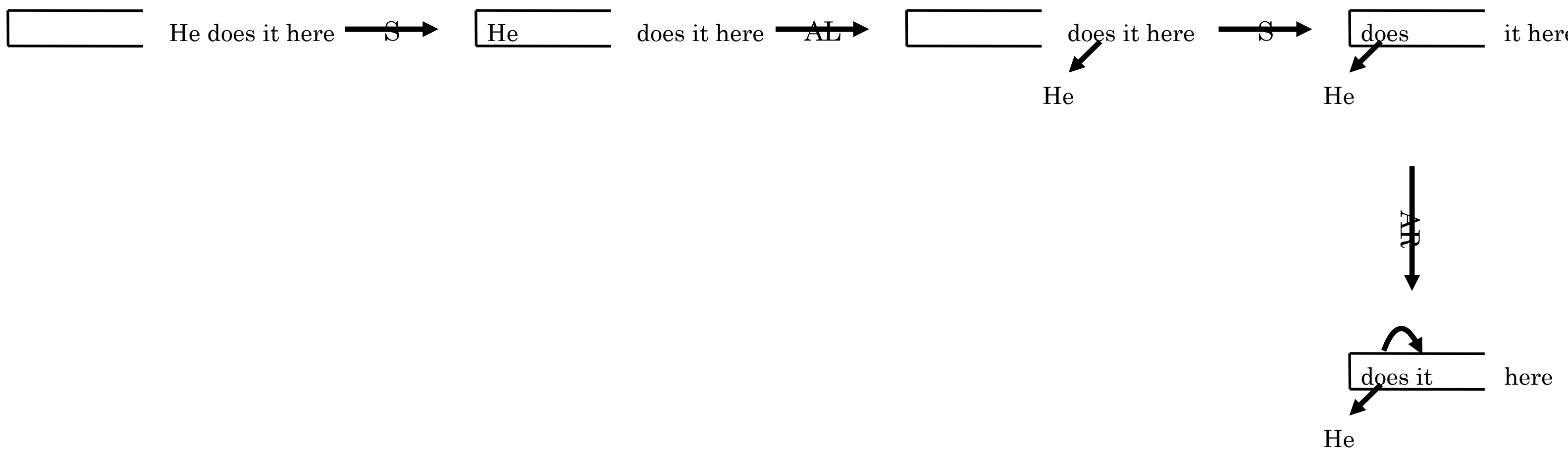
# The arc-eager transition system

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



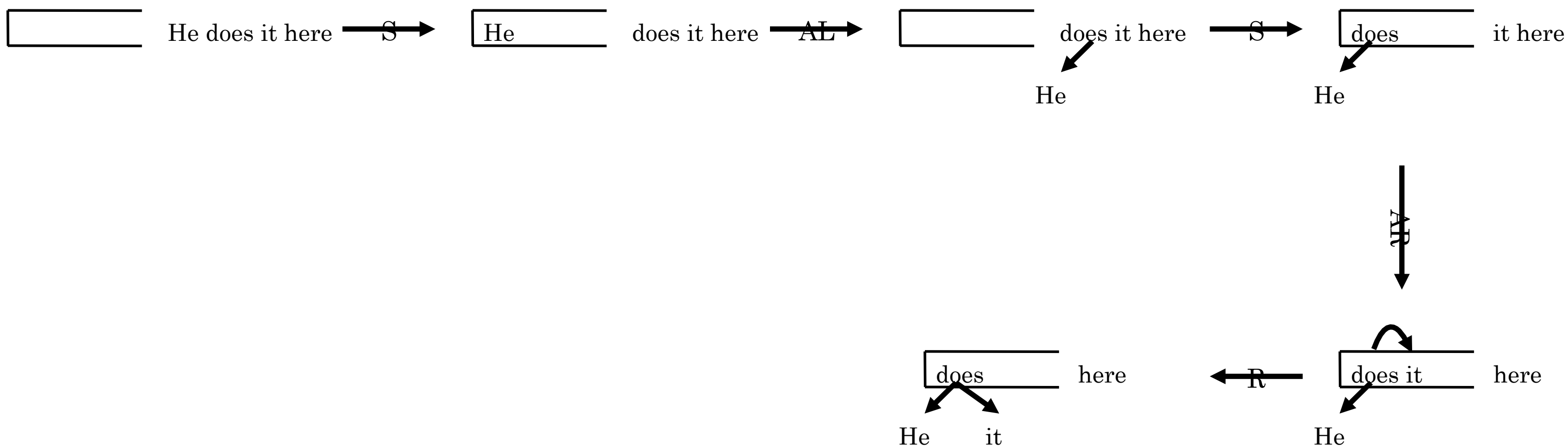
# The arc-eager transition system

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



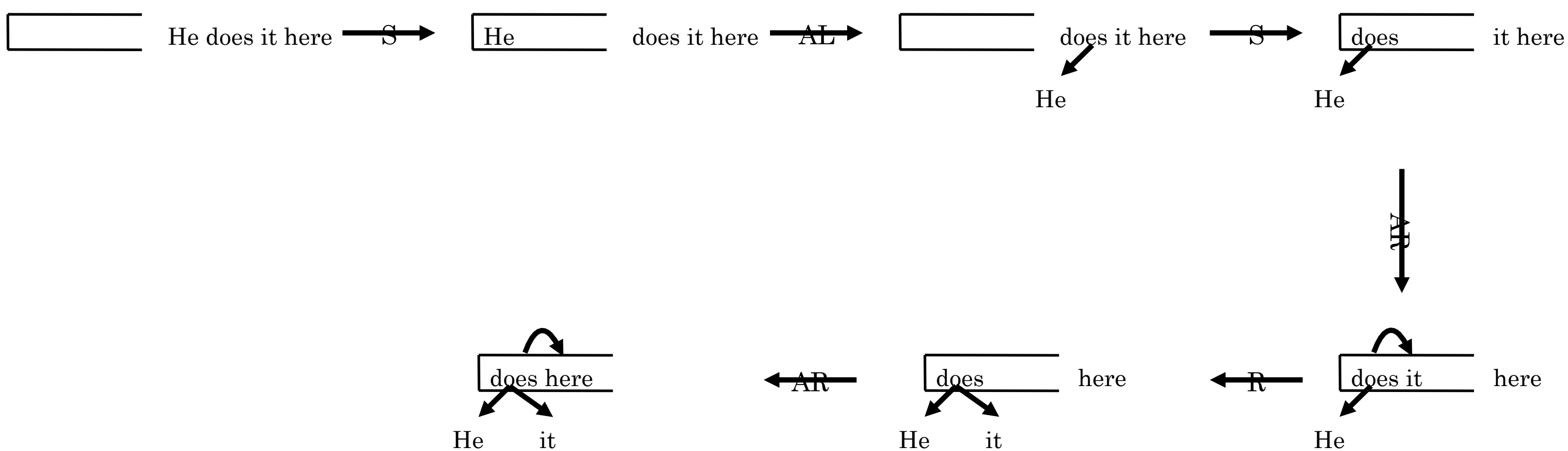
# The arc-eager transition system

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



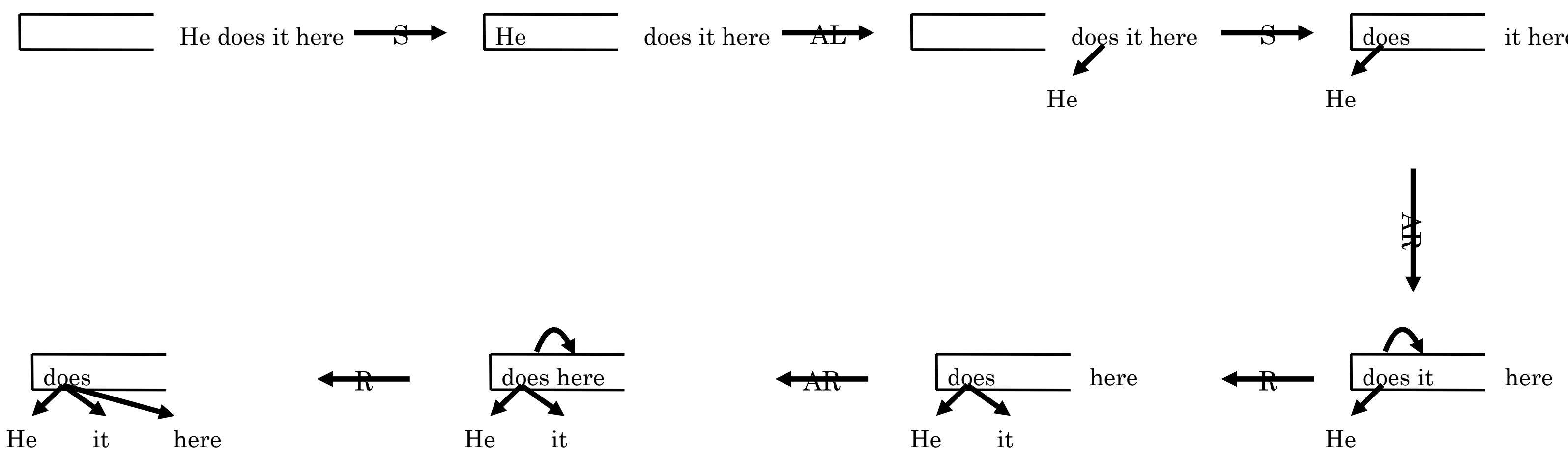
# The arc-eager transition system

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



# The arc-eager transition system

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



# Other examples

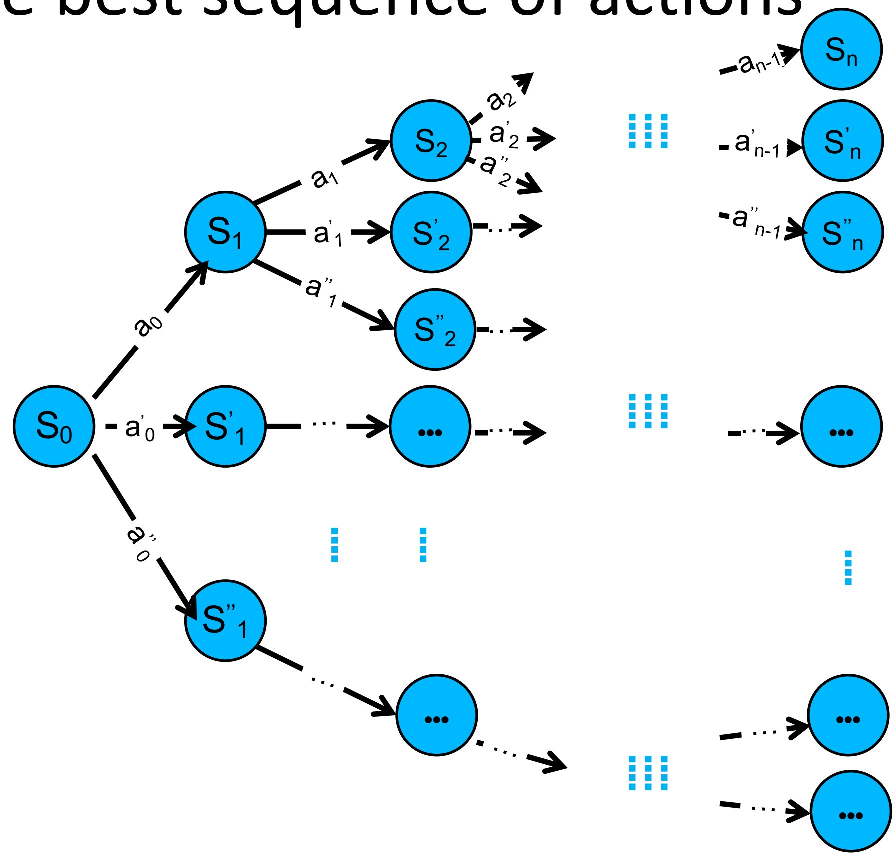
- Language generation

- Translation
  - Word by word
  - Phrase by phrase
  - Syntax tree synthesis

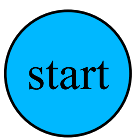
# Part 5.1: Beam-search Decoding
## ——learning to search
(Zhang and Clark,2011)

# Search

- Find the best sequence of actions

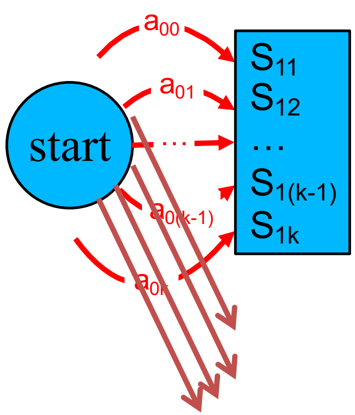


# Beam-search decoding



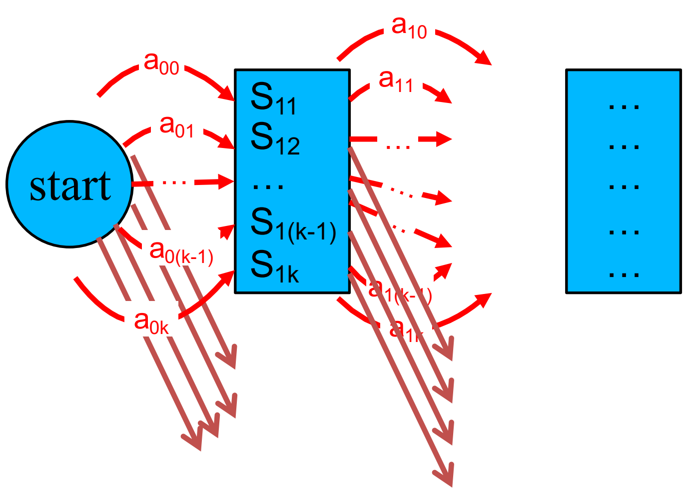

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Beam-search decoding



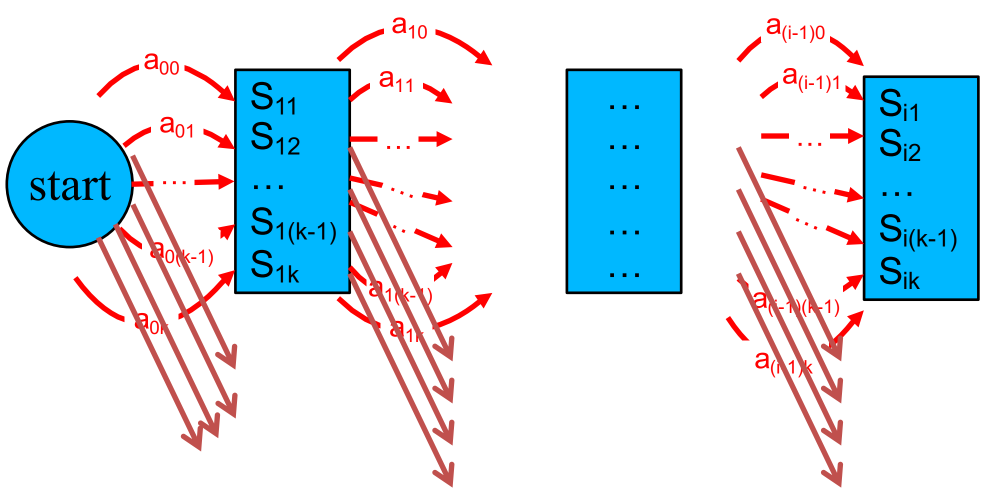

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Beam-search decoding



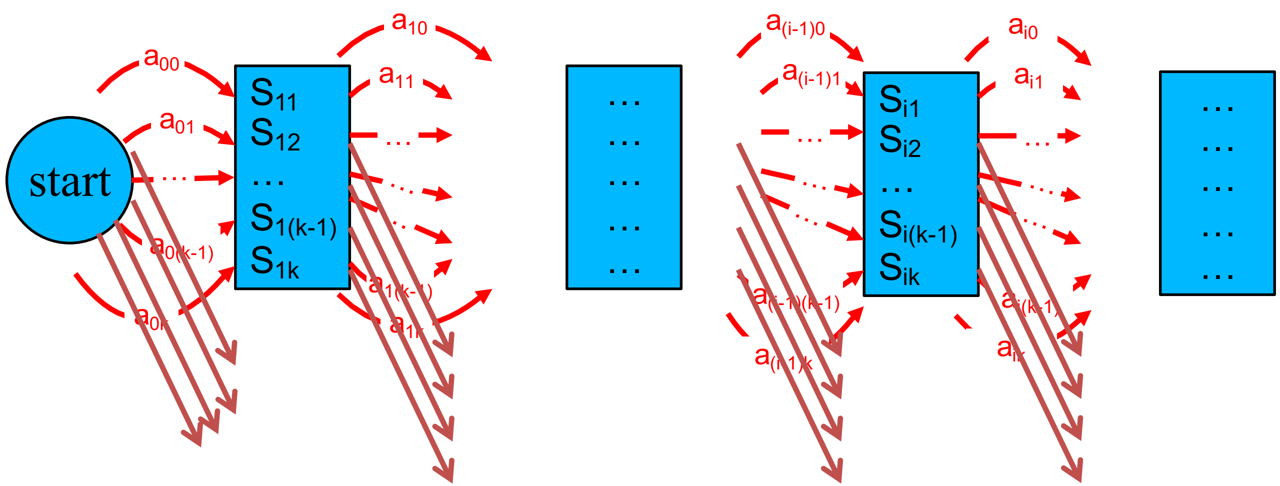

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Beam-search decoding



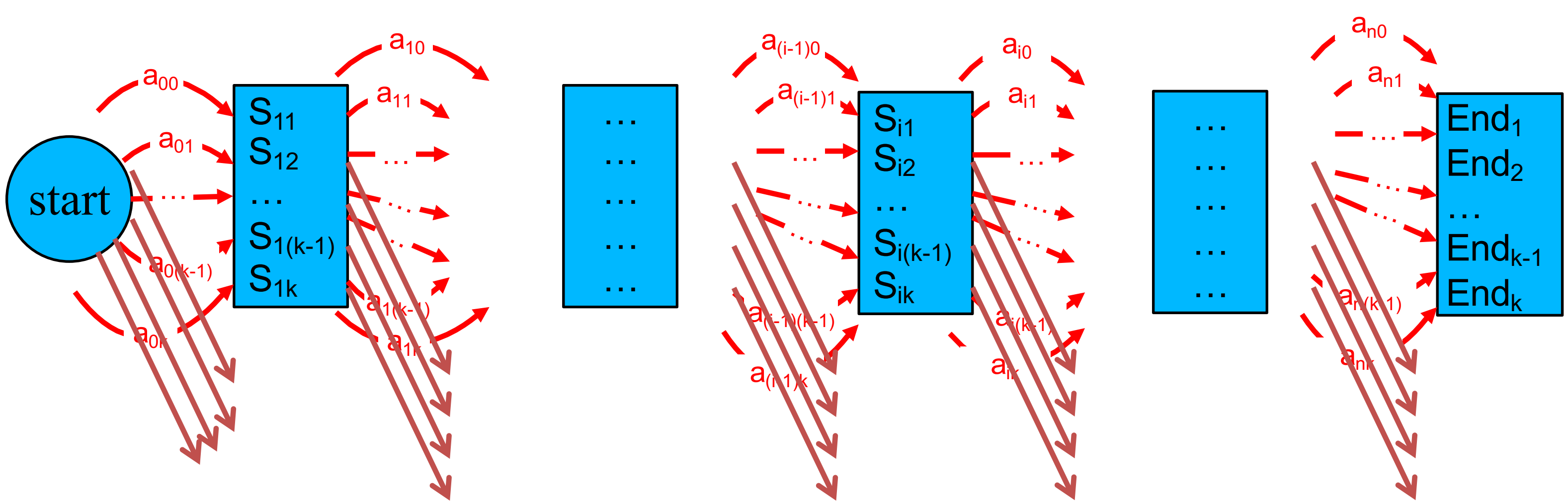

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Beam-search decoding




Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Beam-search decoding




Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Beam-search decoding

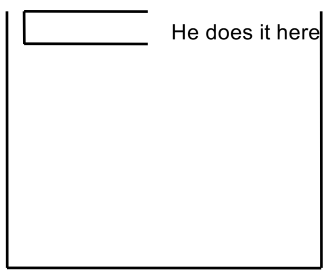
```
function BEAM-SEARCH(problem, agenda, candidates, B)

  candidates ← {STARTITEM(problem)}
  agenda ← CLEAR(agenda)
  loop do
    for each candidate in candidates
      agenda ← INSERT(EXPAND(candidate, problem), agenda)
    best ← TOP(agenda)
    if GOALTEST(problem, best)
      then return best
    candidates ← TOP-B(agenda, B)
    agenda ← CLEAR(agenda)
```

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Beam-search decoding

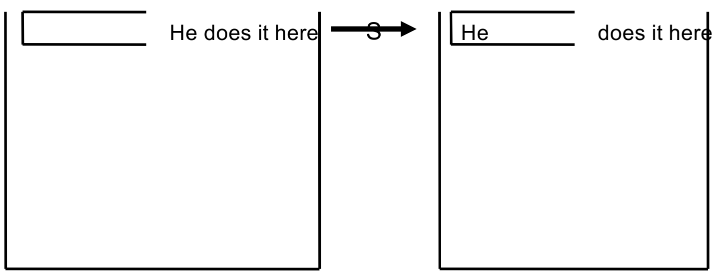
- Our parser
  - Decoding




Zhang, Y., & Clark, S. (2008). A Tale of Two Parsers: Investigating and Combining Graph-Based And transition-Based Dependency Parsing Using Beam-search. EMNLP.

# Beam-search decoding

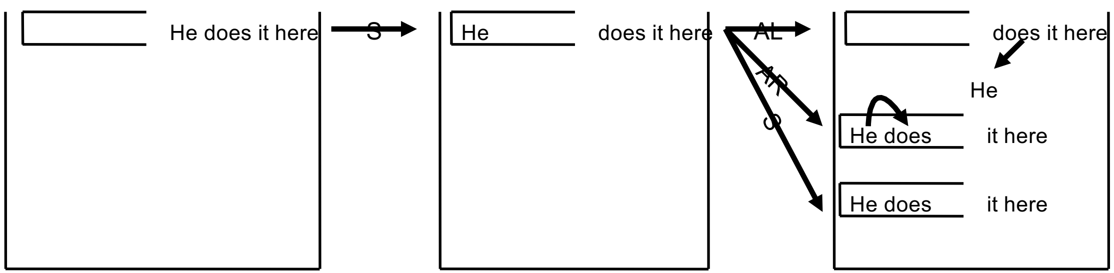
- Our parser
  - Decoding




Zhang, Y., & Clark, S. (2008). A Tale of Two Parsers: Investigating and Combining Graph-Based And transition-Based Dependency Parsing Using Beam-search. EMNLP.

# Beam-search decoding

- Our parser
  - Decoding

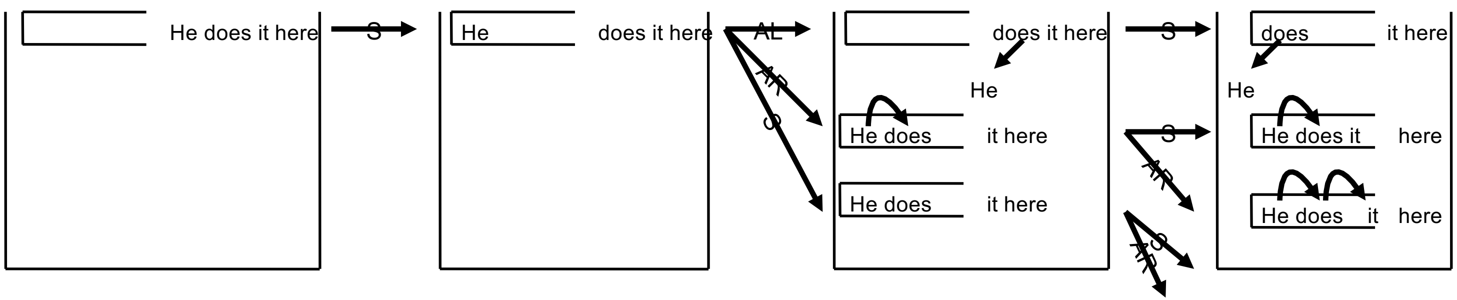




Zhang, Y., & Clark, S. (2008). A Tale of Two Parsers: Investigating and Combining Graph-Based And transition-Based Dependency Parsing Using Beam-search. EMNLP.

# Beam-search decoding

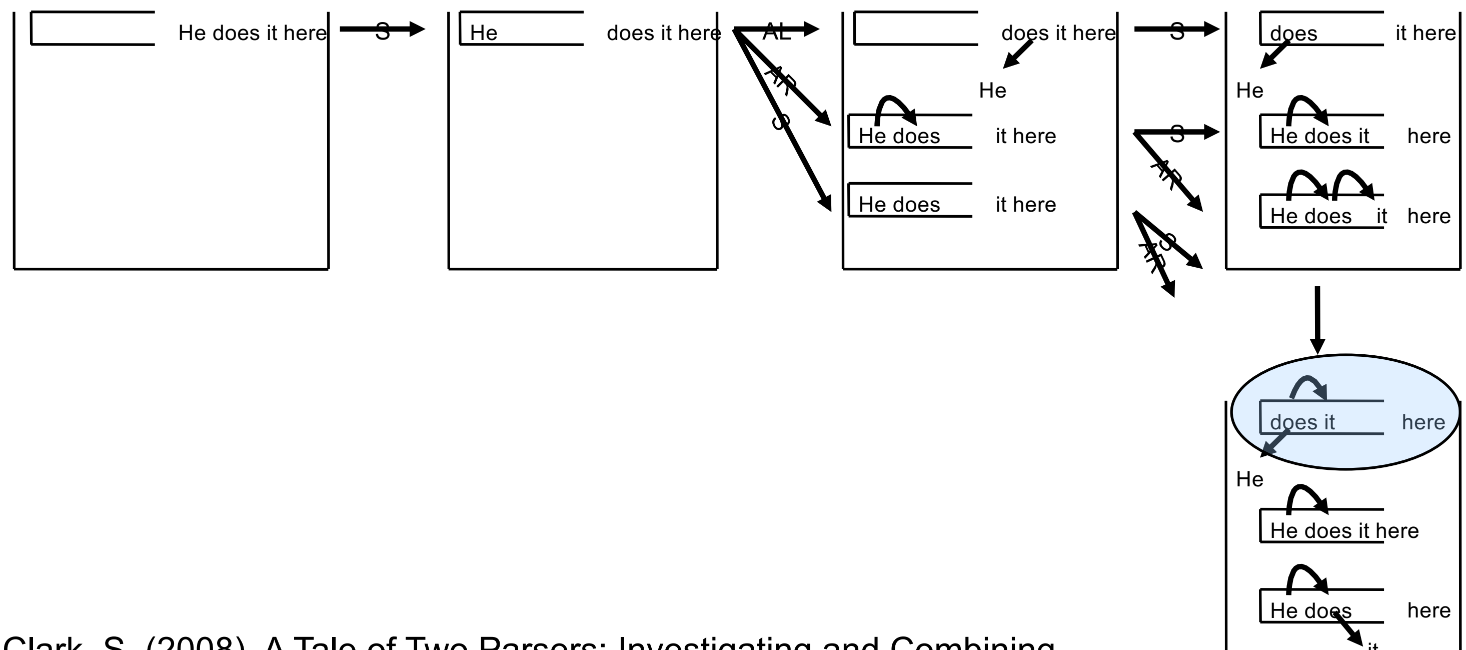
- Our parser
  - Decoding




Zhang, Y., & Clark, S. (2008). A Tale of Two Parsers: Investigating and Combining Graph-Based And transition-Based Dependency Parsing Using Beam-search. EMNLP.

# Beam-search decoding

- Our parser
  - Decoding

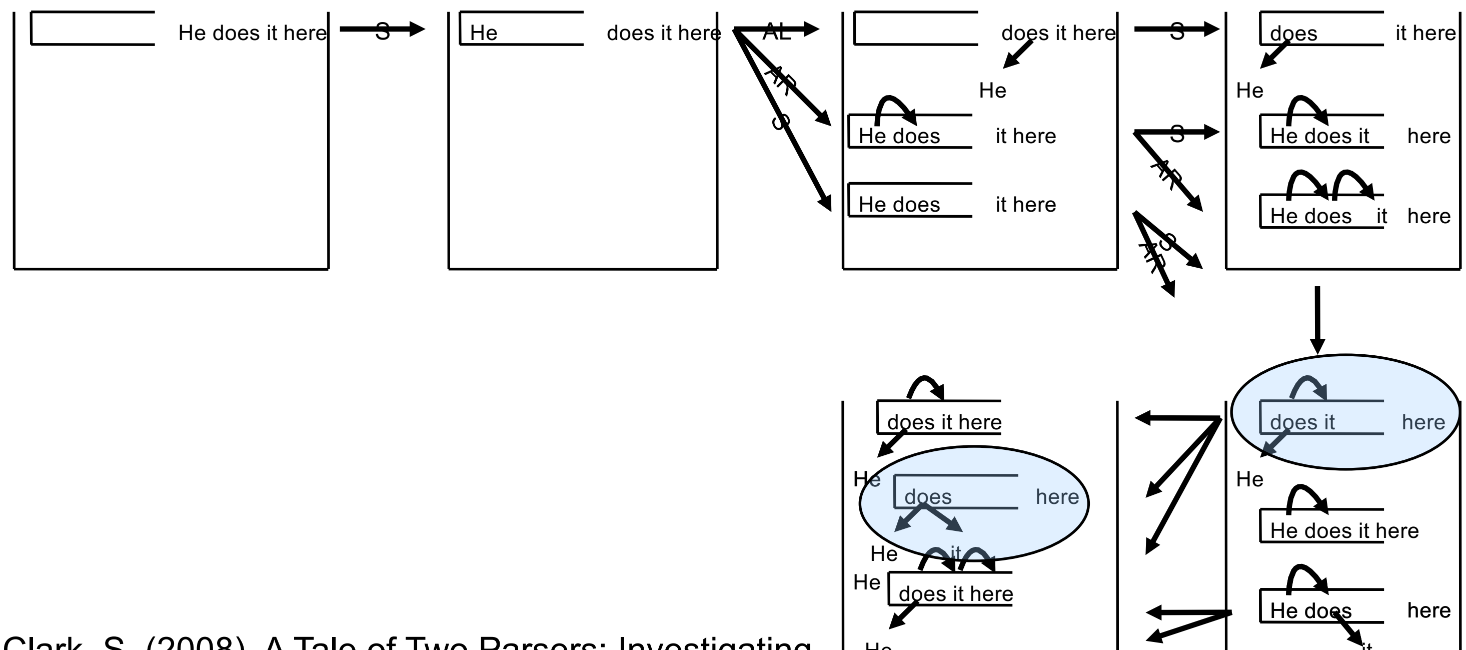




Zhang, Y., & Clark, S. (2008). A Tale of Two Parsers: Investigating and Combining Graph-Based And transition-Based Dependency Parsing Using Beam-search. EMNLP.

# Beam-search decoding

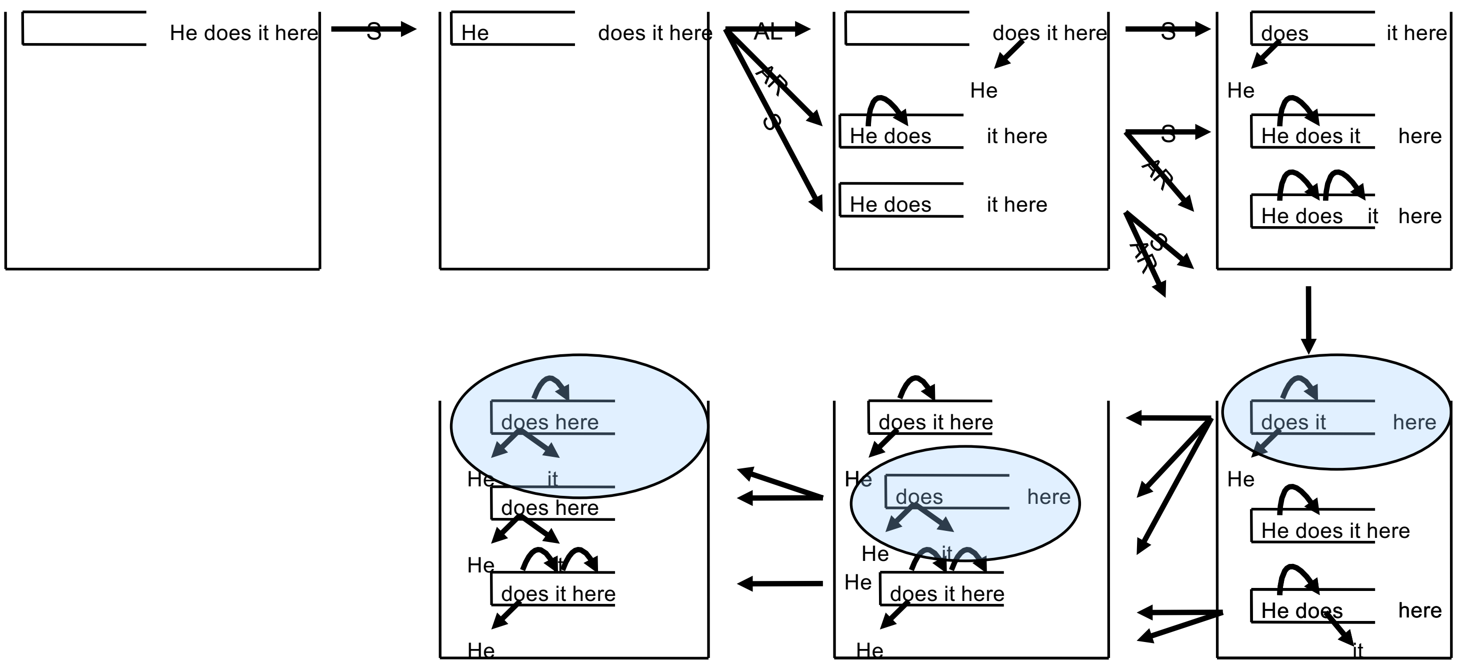
- Our parser
  - Decoding




Zhang, Y., & Clark, S. (2008). A Tale of Two Parsers: Investigating and Combining Graph-Based And transition-Based Dependency Parsing Using Beam-search. EMNLP.

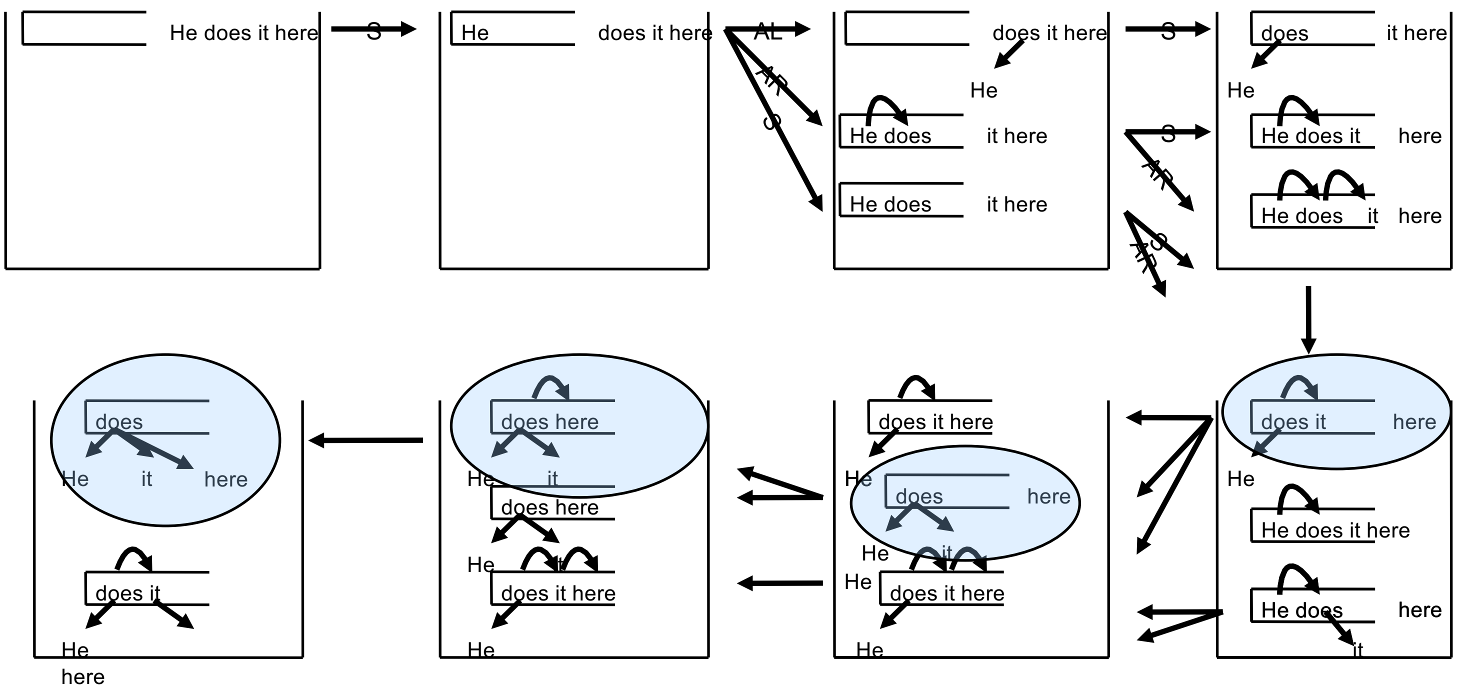
# Beam-search decoding

- Our parser
  - Decoding

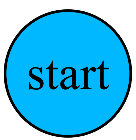


# Beam-search decoding

- Our parser
  - Decoding

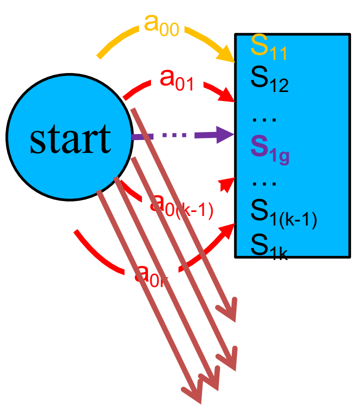


# Online learning



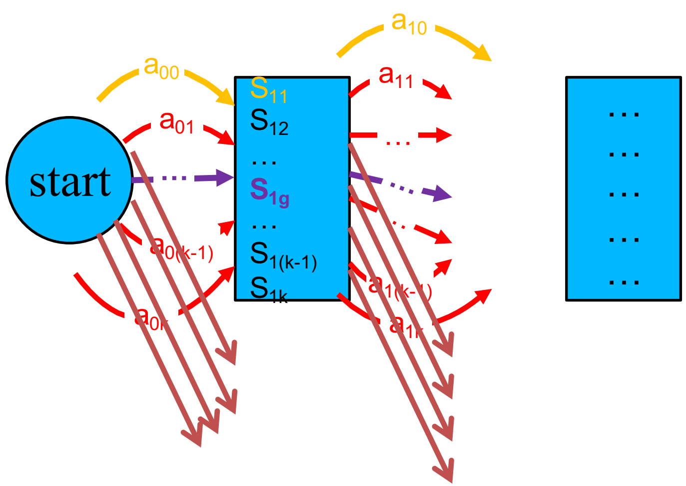

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Online learning



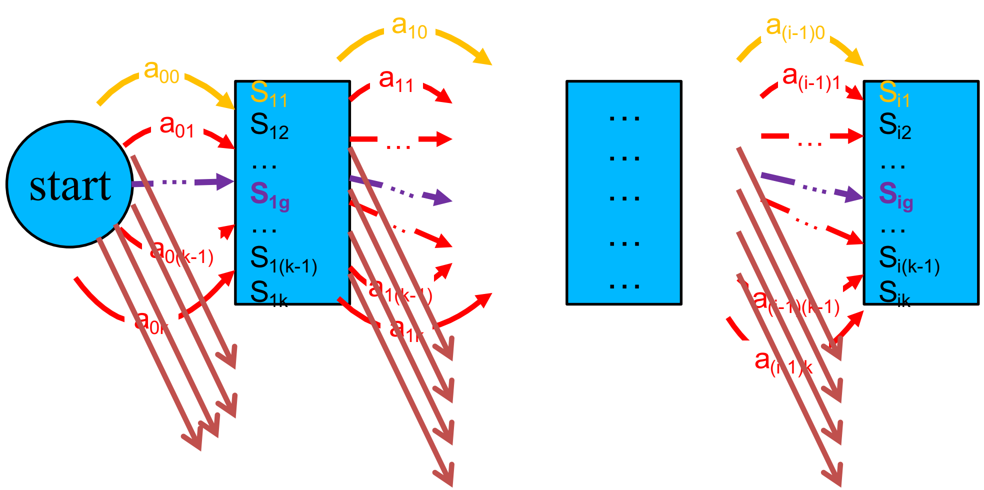

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Online learning



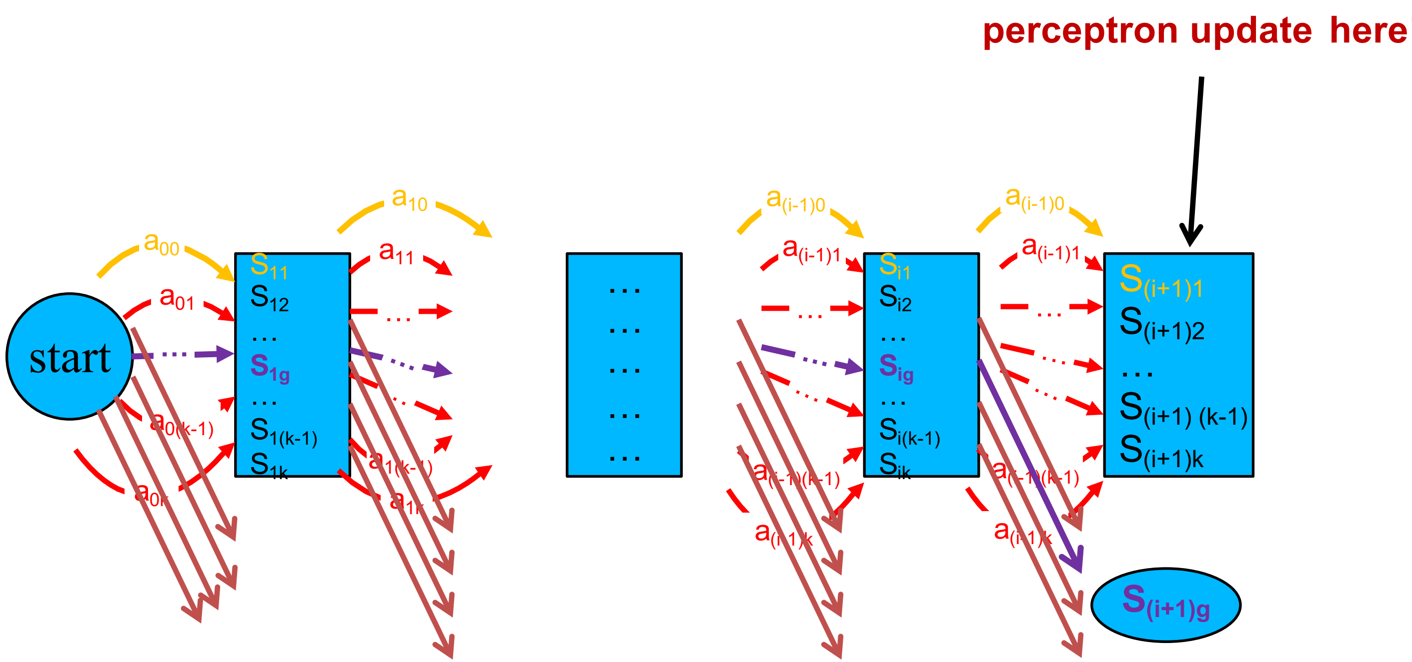

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Online learning




Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# Online learning




Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# 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 $\overrightarrow{w} = 0$
**Algorithm:**
**for** $r = 1 \cdots P, i = 1 \cdots N$ **do**
  $candidates \leftharpoonup \{S_0^i\}$
  $agenda \leftharpoonup \mathrm{CLEAR}(agenda)$
  **for** $k = 1 \cdots m$, $m$ corresponds to a specific training example. **do**
    **for** each $candidate$ in $candidates$ **do**
      $agenda \leftharpoonup \mathrm{INSERT}(\mathrm{EXPAND}(candidate), agenda)$
      $candidates \leftharpoonup \mathrm{TOP-B}(agenda, B)$
      $best \leftharpoonup \mathrm{TOP}(agenda)$
      **if** $S_k^i$ is not in $candidates$ or ($best \neq S_m^i$ and $k$ equals $m$) **then**
        $\overrightarrow{w} = \overrightarrow{w} + \Phi(S_k^i) - \Phi(best)$
      **end if**
    **end for**
  **end for**
**end for**
**Output:** $\overrightarrow{w}$

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# The main strengths

- Fast
- Arbitrary nonlocal features
- Learning fixes search

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. CL.

# State-of-the-art results

- Chinese
  - Word segmentation
    - Yue Zhang and Stephen Clark. Chinese Segmentation Using a Word-Based Perceptron Algorithm. In proceedings of ACL 2007. Prague, Czech Republic. June.

# State-of-the-art results

- Chinese
  - Joint segmentation and POS-tagging
    - Yue Zhang and Stephen Clark. Joint Word Segmentation and POS Tagging Using a Single Perceptron. In proceedings of ACL 2008. Ohio, USA. June.
    - Yue Zhang and Stephen Clark. A Fast Decoder for Joint Word Segmentation and POS-tagging Using a Single Discriminative Model. In proceedings of EMNLP 2010. Massachusetts, USA. October.

# State-of-the-art results

- Chinese
  - Joint segmentation, POS-tagging and chunking

    - Chen Lyu, Yue Zhang and Donghong Ji. Joint Word Segmentation, POS-Tagging and Syntactic Chunking. In Proceedings of the AAAI 2016, Phoenix, Arizona, USA, February

# State-of-the-art results

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

    - Meishan Zhang, Yue Zhang, Wanxiang Che and Ting Liu. Character-Level Chinese Dependency Parsing. In Proceedings of ACL 2014. Baltimore, USA, June.

# State-of-the-art results

- Chinese
  - Joint segmentation, POS-tagging and constituent parsing
    - Meishan Zhang, Yue Zhang, Wanxiang Che and Ting Liu. Chinese Parsing Exploiting Characters. In proceedings of ACL 2013. Sophia, Bulgaria. August.

# 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
    - Yue Zhang and Stephen Clark. Transition-Based Parsing of the Chinese Treebank Using a Global Discriminative Model. In proceedings of IWPT 2009. Paris, France. October.
    - Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang and Jingbo Zhu. Fast and Accurate Shift-Reduce Constituent Parsing. In proceedings of ACL 2013. Sophia, Bulgaria. August.

# State-of-the-art results

- All Languages
  - Dependency parsing
    - Yue Zhang and Stephen Clark. Joint Word Segmentation and POS Tagging Using a Single Perceptron. In proceedings of ACL 2008. Ohio, USA. June.
    - Yue Zhang and Joakim Nivre. Transition-Based Dependency Parsing with Rich Non-Local Features. In proceedings of ACL 2011, short papers. Portland, USA. June.
    - Yue Zhang and Joakim Nivre. Analyzing the Effect of Global Learning and Beam-Search for Transition-Based Dependency Parsing. In proceedings of COLING 2012, posters. Mumbai, India. December.
    - Ji Ma, Yue Zhang and Jingbo Zhu. Punctuation Processing for Projective Dependency Parsing. In Proceedings of ACL 2014. Baltimore, USA, June.

# State-of-the-art results

- All Languages
  - CCG parsing
    - Yue Zhang and Stephen Clark. Shift-Reduce CCG Parsing. In proceedings of ACL 2011. Portland, USA. June.
    - Wenduan Xu, Stephen Clark and Yue Zhang. Shift-Reduce CCG Parsing with a Dependency Model. In Proceedings of ACL 2014. Baltimore, USA, June.

# 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.
    - Jiangming Liu and Yue Zhang. An Empirical Comparison Between N-gram and Syntactic Language Models for Word Ordering. In proceedings of EMNLP 2015, Lisboa, Portugal, September.
    - Ratish Puduppully, Yue Zhang and Manish Shrivastava. Transition-Based Syntactic Linearization with Lookahead Features. In Proceedings of the NAACL 2016, San Diego, USA, June.

# State-of-the-art results

- All Languages
  - Joint morphological generation and text linearization
    - Linfeng Song, Yue Zhang, Kai Song and Qun Liu. Joint Morphological Generation and Syntactic Linearization. In Proceedings of AAAI 2014. Quebec City, Canada, July.

# State-of-the-art results

- All Languages
  - Joint entity and relation extraction
    - Fei Li, Yue Zhang, Meishan Zhang and Donghong Ji. Joint Models for Extracting Adverse Drug Events from Biomedical Text. In Proceedings of IJCAI 2016. New York City, USA, July.

# 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

| step | action | buffer($\cdots w_{-1} w_0$) | queue($c_0 c_1 \cdots$) |
|---|---|---|---|
| 0 | - | $\phi$ | 中 国 ··· |
| 1 | *SEP* | 中 | 国 外 ··· |
| 2 | *APP* | 中国 | 外 企 ··· |
| 3 | *SEP* | 中国 外 | 企 业 ··· |
| 4 | *APP* | 中国 外企 | 业 务 ··· |
| 5 | *SEP* | 中国 外企 业 | 务 发 ··· |
| 6 | *APP* | 中国 外企 业务 | 发 展 ··· |
| 7 | *SEP* | ··· 业务 发 | 展 迅 速 |
| 8 | *APP* | ··· 业务 发展 | 迅 速 |
| 9 | *SEP* | ··· 发展 迅 | 速 |
| 10 | *APP* | ··· 发展 迅速 | $\phi$ |

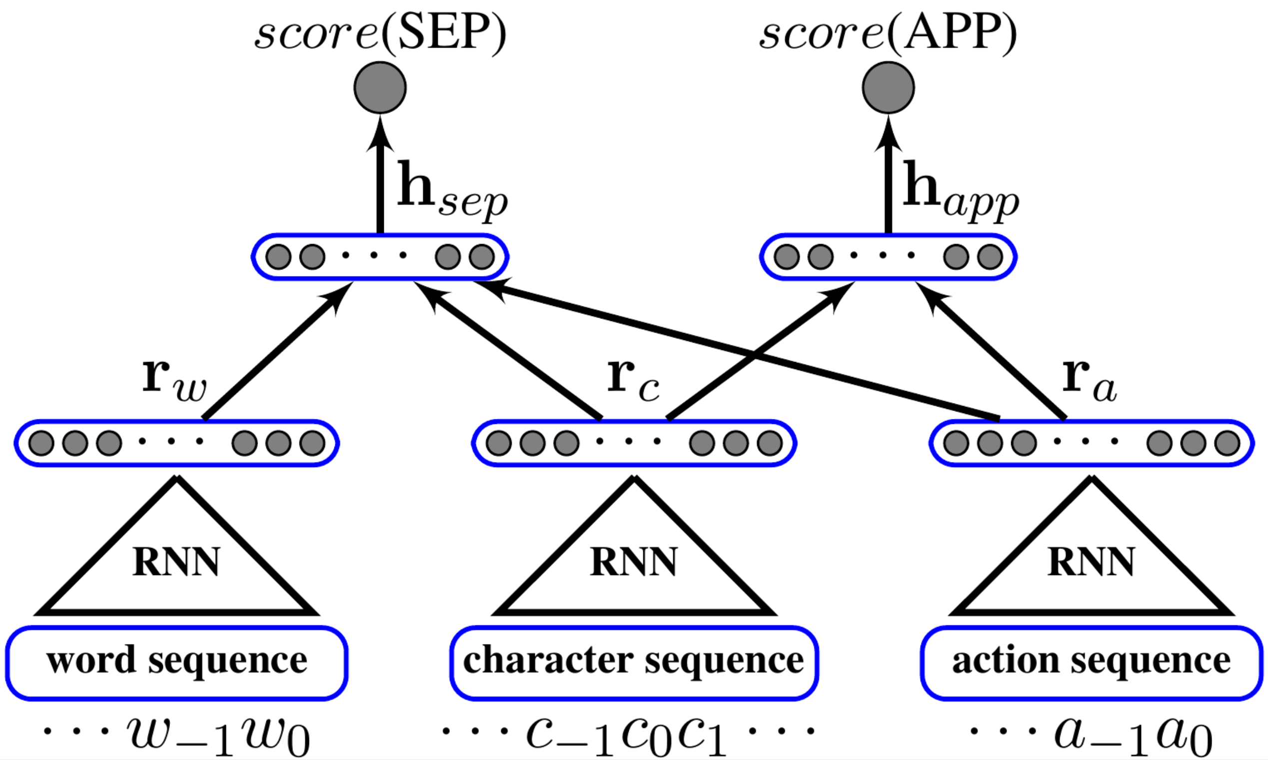
Zhang, M., Zhang, Y., & Fu, G. (2016). Transition-Based Neural Word Segmentation. ACL.

# Word segmentation

| **Feature templates** | **Action** |
|---|---|
| $c_{-1}c_0$ | *APP, SEP* |
| $w_{-1}, w_{-1}w_{-2}, w_{-1}c_0, w_{-2}len(w_{-1})$<br>$start(w_{-1})c_0, end(w_{-1})c_0$<br>$start(w_{-1})end(w_{-1}), end(w_{-2})end(w_{-1})$<br>$w_{-2}len(w_{-1}), len(w_{-2})w_{-1}$<br>$w_{-1}$, where $len(w_{-1}) = 1$ | *SEP* |

Zhang, M., Zhang, Y., & Fu, G. (2016). Transition-Based Neural Word Segmentation. ACL.

# Word segmentation




Zhang, M., Zhang, Y., & Fu, G. (2016). Transition-Based Neural Word Segmentation. ACL.

# Word segmentation

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

Main results on CTB60 test dataset

Zhang, M., Zhang, Y., & Fu, G. (2016). Transition-Based Neural Word Segmentation. ACL.

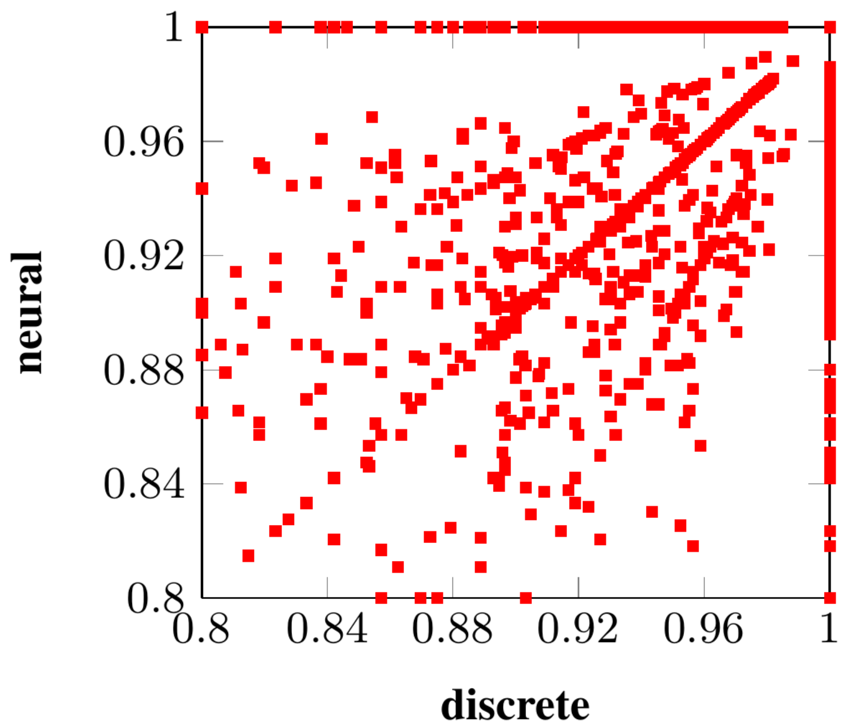
# Word segmentation

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

Main results on PKU and MSR test dataset

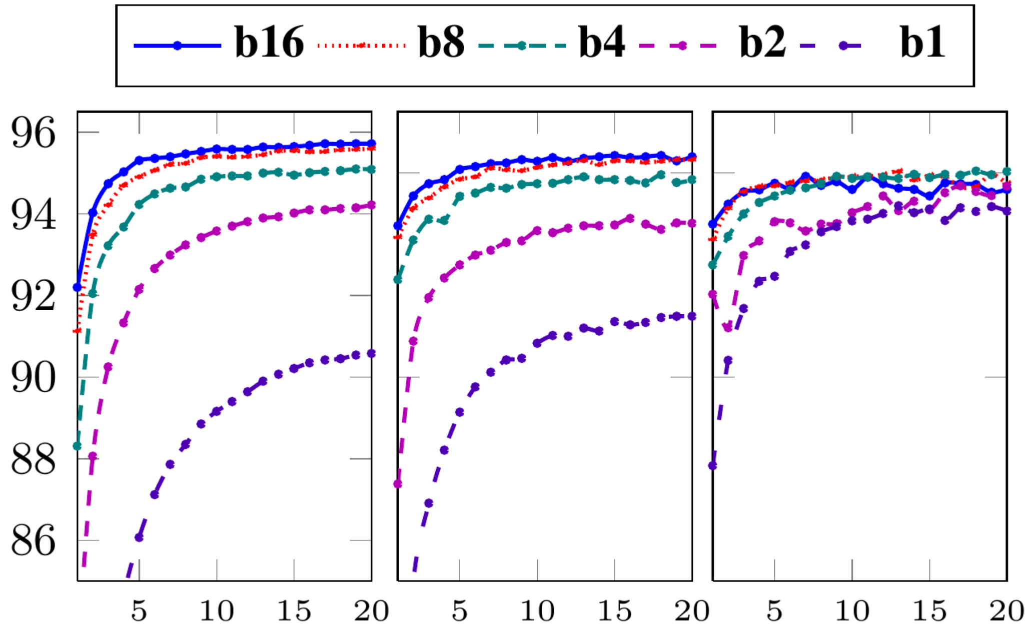
Zhang, M., Zhang, Y., & Fu, G. (2016). Transition-Based Neural Word Segmentation. ACL.

# Word segmentation




Zhang, M., Zhang, Y., & Fu, G. (2016). Transition-Based Neural Word Segmentation. ACL.

# Word segmentation



(a) discrete (b) neural(-tune) (c) neural(+tune)

Zhang, M., Zhang, Y., & Fu, G. (2016). Transition-Based Neural Word Segmentation. ACL.

# Word segmentation

- Cai and Zhao (2016) presents a similar idea

Cai, D., & Zhao, H. (2016). Neural Word Segmentation Learning for Chinese. ACL.

# Dependency Parsing

- Zhang & Nivre (2011)

$$y = \arg\max_{y' \in \mathrm{GEN}(x)} score(y')$$

$$score(y) = \sum_{a \in y} \theta \cdot \Phi(a)$$

Zhou, H., Zhang, Y., Huang, S., & Chen, J. (2015). A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. ACL.

# Dependency Parsing

- Chen and Manning (2014)

$$h = (W_1 x + b_1)^3$$

$$p = softmax(o)$$

$$o = W_2 h$$

Zhou, H., Zhang, Y., Huang, S., & Chen, J. (2015). A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. ACL.

# Dependency Parsing

- What does not work

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

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

Zhou, H., Zhang, Y., Huang, S., & Chen, J. (2015). A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. ACL.

# Dependency Parsing

- Sentence-level log likelihood

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

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

Zhou, H., Zhang, Y., Huang, S., & Chen, J. (2015). A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. ACL.

# Dependency Parsing

- Contrastive Estimation

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

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

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

$$Z(x, \theta) = \sum_{y_j \in \mathrm{GEN}(x)} e^{f(x,\ \theta)_j}$$

Zhou, H., Zhang, Y., Huang, S., & Chen, J. (2015). A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. ACL.

# Dependency Parsing

- Contrastive Estimation

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

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

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

$$Z'(x, \theta) = \sum_{y_j \in \mathrm{BEAM}(x)} e^{f(x,\, \theta)_j}$$

Zhou, H., Zhang, Y., Huang, S., & Chen, J. (2015). A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. ACL.

# Dependency Parsing

- Results

| Description | UAS | |
|---|---|---|
| Baseline | 91.63 | |
| | structured | greedy |
| beam = 1 | 74.90 | 91.63 |
| beam = 4 | 84.64 | 91.92 |
| beam = 16 | 91.53 | 91.90 |
| beam = 64 | 93.12 | 91.84 |
| beam = 100 | 93.23 | 91.81 |

Zhou, H., Zhang, Y., Huang, S., & Chen, J. (2015). A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. ACL.

# Dependency Parsing

- Results

| Description | UAS |
|---|---|
| greedy neural parser | 91.47 |
| ranking model | 89.08 |
| beam contrastive learning | 93.28 |

Zhou, H., Zhang, Y., Huang, S., & Chen, J. (2015). A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. ACL.

# Dependency Parsing

- Results

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

Zhou, H., Zhang, Y., Huang, S., & Chen, J. (2015). A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. ACL.

# Google

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

Andor, D., Alberti, Chris., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., & Collins, M. (2016). Globally Normalized Transition-Based Neural Networks. ACL.

# Google

- Dependency parsing

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

Andor, D., Alberti, Chris., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., & Collins, M. (2016). Globally Normalized Transition-Based Neural Networks. ACL.

# Google

- Dependency parsing

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

Andor, D., Alberti, Chris., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., & Collins, M. (2016). Globally Normalized Transition-Based Neural Networks. ACL.

# Google

- POS-tagging

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

Andor, D., Alberti, Chris., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., & Collins, M. (2016). Globally Normalized Transition-Based Neural Networks. ACL.

# Google

- Compression

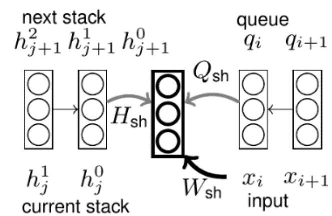
| Method | Generated corpus A | F1 | Human eval read | info |
|---|---|---|---|---|
| Filippova et al. (2015) | **35.36** | **82.83** | 4.66 | 4.03 |
| Automatic | - | - | 4.31 | 3.77 |
| Our Local (B=1) | 30.51 | 78.72 | 4.58 | 4.03 |
| Our Local (B=8) | 31.19 | 75.69 | - | - |
| Our Global (B=8) | 35.16 | 81.41 | **4.67** | **4.07** |

Andor, D., Alberti, Chris., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., & Collins, M. (2016). Globally Normalized Transition-Based Neural Networks. ACL.

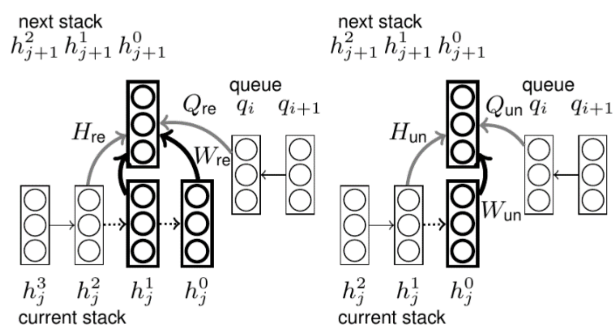
# Part 5.3: Similar methods by others

# Other methods ( $I$ )

- Constituent parsing




(a) shift-$X$ action




(b) reduce-$X$ action (c) unary-$X$ action

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

# Other methods ( $I$ )

- Update at max-violation

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

- Using expected loss from all violations

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

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

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

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

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

# Other methods ( $I$ )

| parser | test |
|---|---|
| Collins (Collins, 1997) | 87.8 |
| Berkeley (Petrov and Klein, 2007) | 90.1 |
| SSN (Henderson, 2004) | 90.1 |
| ZPar (Zhu et al., 2013) | 90.4 |
| CVG (Socher et al., 2013) | 90.4 |
| Charniak-R (Charniak and Johnson, 2005) | **91.0** |
| This work: TNCP | 90.7 |

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

# Other methods ( $I$ )

| parser | test |
|---|---|
| ZPar (Zhu et al., 2013) | 83.2 |
| Berkeley (Petrov and Klein, 2007) | 83.3 |
| Joint (Wang and Xue, 2014) | **84.9** |
| This work: TNCP | 84.3 |

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

# Other methods ( II )

- CCG Parsing
- expected F1 training

$$J(\theta) = -\mathrm{xF1}(\theta)$$
$$= -\sum_{y_i \in \Lambda(x_n)} p(y_i|\theta)\mathrm{F1}(\Delta_{y_i}, \Delta^G_{x_n})$$

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

Xu, W., Auli, M., & Clark, S. (2016). Expected F-Measure Training for Shift-Reduce Parsing with Recurrent Neural Networks. NAACL.

# Other methods ( $II$ )

$$
\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta} &= - \sum_{y_i \in \Lambda(x_n)} \sum_{y_{ij} \in y_i} \frac{\partial J(\theta)}{\partial s_\theta(y_{ij})} \frac{\partial s_\theta(y_{ij})}{\partial \theta} \\
&= - \sum_{y_i \in \Lambda(x_n)} \sum_{y_{ij} \in y_i} \delta_{y_{ij}} \frac{\partial s_\theta(y_{ij})}{\partial \theta},
\end{aligned}
$$

$$
\begin{aligned}
\delta_{y_{ij}} &= -\frac{\partial \mathrm{xF1}(\theta)}{\partial s_\theta(y_{ij})} \\
&= -\frac{\partial (G(\theta)/Z(\theta))}{\partial s_\theta(y_{ij})} \\
&= \frac{G(\theta)Z'(\theta) - G'(\theta)Z(\theta)}{Z^2(\theta)} \\
&= \frac{\exp\{\rho(y_i)\}}{Z(\theta)} (\mathrm{xF1}(\theta) - \mathrm{F1}(\Delta_{y_i}, \Delta^G_{x_n})) \frac{1}{s_\theta(y_{ij})} \\
&= p(y_i|\theta)(\mathrm{xF1}(\theta) - \mathrm{F1}(\Delta_{y_i}, \Delta^G_{x_n})) \frac{1}{s_\theta(y_{ij})},
\end{aligned}
$$

Xu, W., Auli, M., & Clark, S. (2016). Expected F-Measure Training for Shift-Reduce Parsing with Recurrent Neural Networks. NAACL.

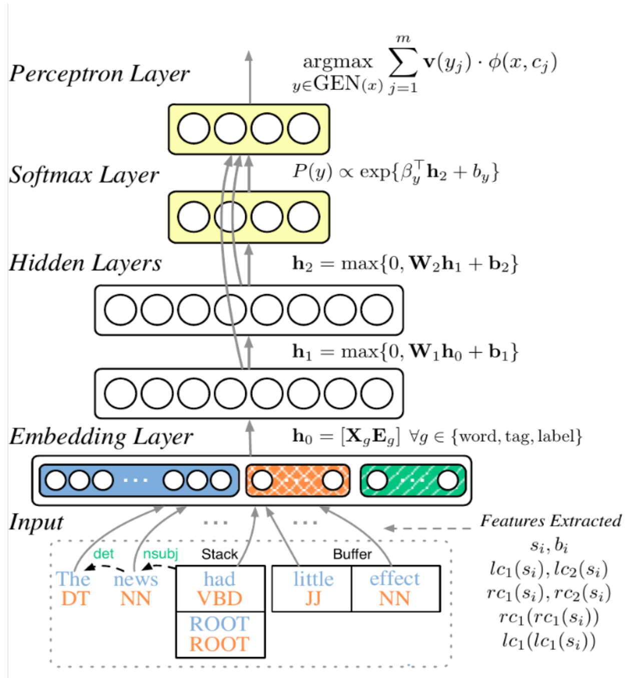
# Other methods ( $II$ )

| | Section 00 | | | | Section 23 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | LP | LR | LF | CAT | LP | LR | LF | CAT | Speed |
| C&C (normal) | 85.18 | 82.53 | 83.83 | 92.39 | 85.45 | 83.97 | 84.70 | 92.83 | 97.90 |
| C&C (hybrid) | 86.07 | 82.77 | 84.39 | 92.57 | 86.24 | 84.17 | 85.19 | 93.00 | 95.25 |
| Zhang and Clark (2011) ($b = 16$) | 87.15 | 82.95 | 85.00 | 92.77 | 87.43 | 83.61 | 85.48 | 93.12 | - |
| Zhang and Clark (2011)* ($b = 16$) | 86.76 | 83.15 | 84.92 | 92.64 | 87.04 | 84.14 | 85.56 | 92.95 | 49.54 |
| Xu et al. (2014) ($b = 128$) | 86.29 | **84.09** | 85.18 | 92.75 | 87.03 | **85.08** | 86.04 | 93.10 | 12.85 |
| RNN-greedy ($b = 1$) | 88.12 | 81.38 | 84.61 | 93.42 | 88.53 | 81.65 | 84.95 | 93.57 | 337.45 |
| RNN-greedy ($b = 6$) | 87.96 | 82.27 | 85.02 | 93.47 | 88.54 | 82.77 | 85.56 | 93.68 | 96.04 |
| RNN-xF1 ($b = 8$) | **88.20** | 83.40 | **85.73** | **93.56** | **88.74** | 84.22 | **86.42** | **93.87** | 67.65 |

Xu, W., Auli, M., & Clark, S. (2016). Expected F-Measure Training for Shift-Reduce Parsing with Recurrent Neural Networks. NAACL.

# Other methods (Ⅲ)

- Dependency parsing




Weiss, D., Alberti, C., Collins, M., & Petrov, S. (2015). Structured Training for Neural Network Transition-Based Parsing. ACL.

# Other methods (Ⅲ)

- Using Chen and Manning features for perceptron training
- Back-propagation pre-training

$$L(\Theta) = -\sum_j \log P(y_j \mid c_j, \Theta) + \lambda \sum_i \|\mathbf{W}_i\|_2^2$$

- Structured perceptron training

$$(h_1, h_2,\ P(y))$$

Weiss, D., Alberti, C., Collins, M., & Petrov, S. (2015). Structured Training for Neural Network Transition-Based Parsing. ACL.

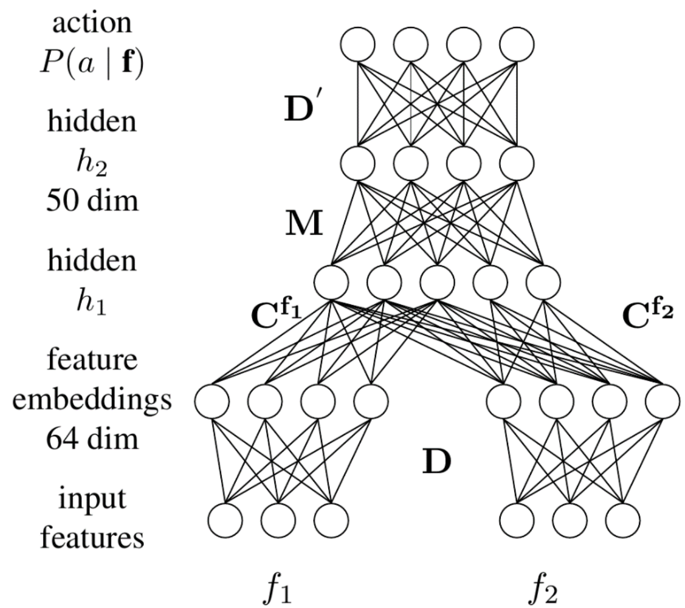
# Other methods (Ⅲ)

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

Weiss, D., Alberti, C., Collins, M., & Petrov, S. (2015). Structured Training for Neural Network Transition-Based Parsing. ACL.

# Other methods (*IV*)

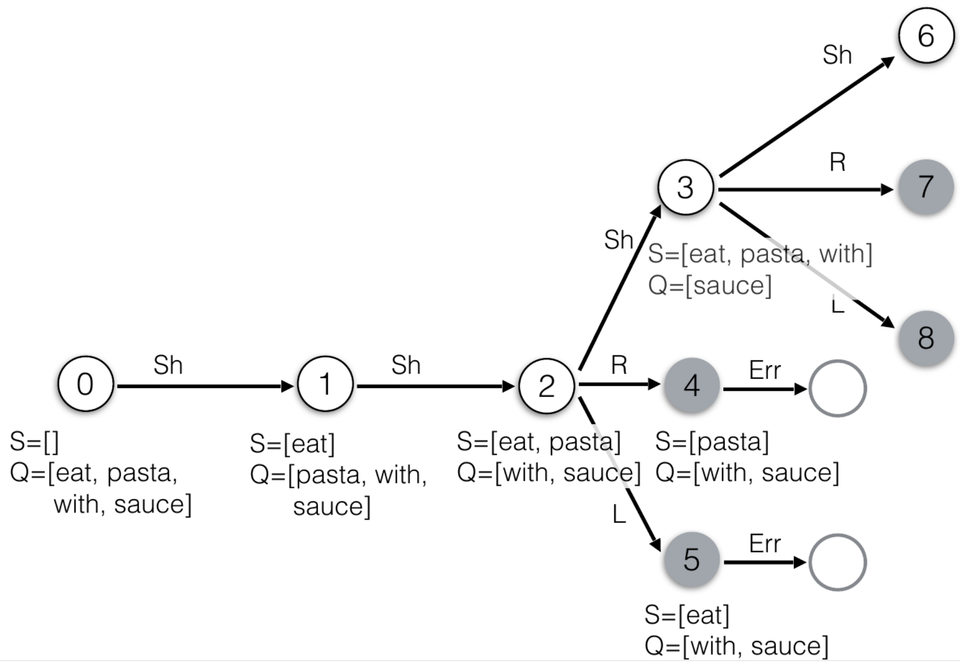
- Dependency parsing





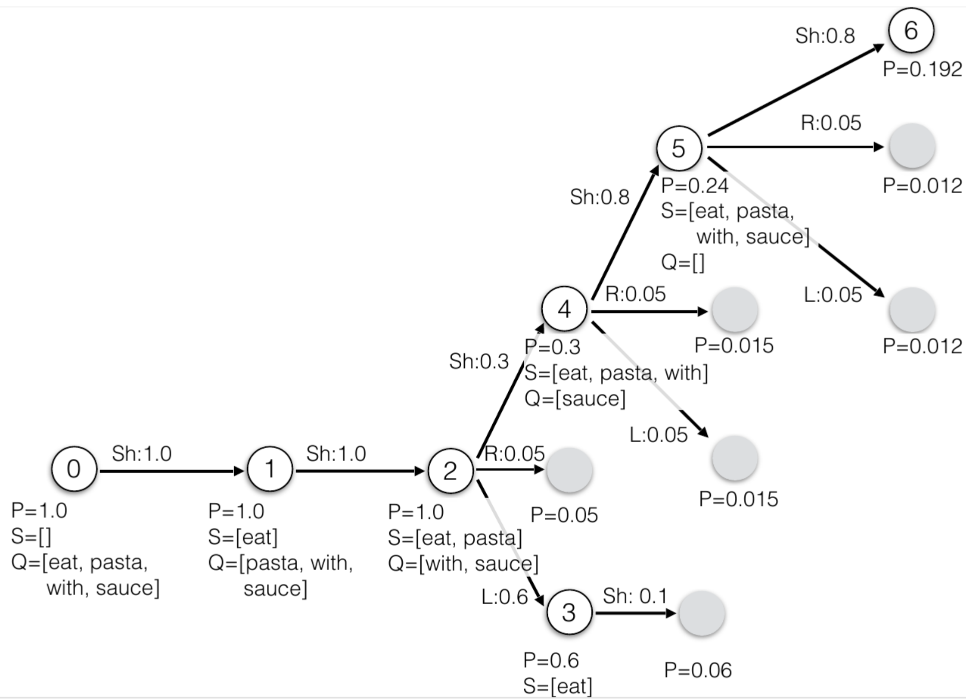
Vaswani, A., & Sagae, K. (2016). Efficient Structured Inference for Transition-Based Parsing with Neural Networks and Error States. TACL.

# 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 ($IV$)

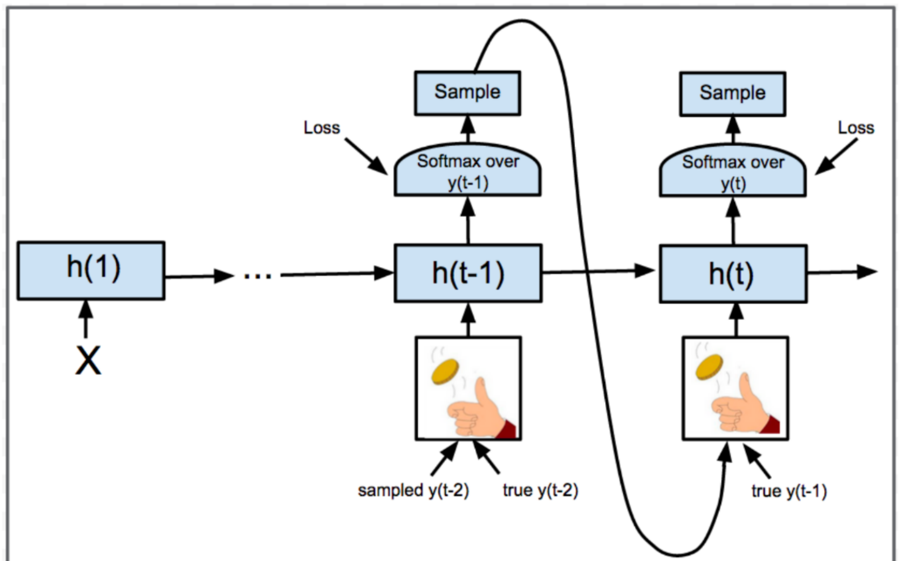
| System | wsj23-S | wsj23-YM |
|---|---|---|
| **ErrSt–25–rand** | **92.17** | **92.16** |
| **ErrSt–25–pre*** | **93.61** | **93.21** |
| Chen & Manning* | 91.8 | – |
| Huang & Sagae | – | 92.1 |
| Zhang & Nivre | 93.5 | 92.9 |
| Weiss et al.* | 93.99 | – |
| Zhang & McDonald | 93.71 | 93.57 |
| Martins et al. | 92.82 | 93.07 |
| Koo et al. (dep2c)* | – | 93.16 |

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

Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled Sampling for Sequence Prediction with Recurrent Nueral Networks. NIPS.

# Sequence to sequence ( $I$ )

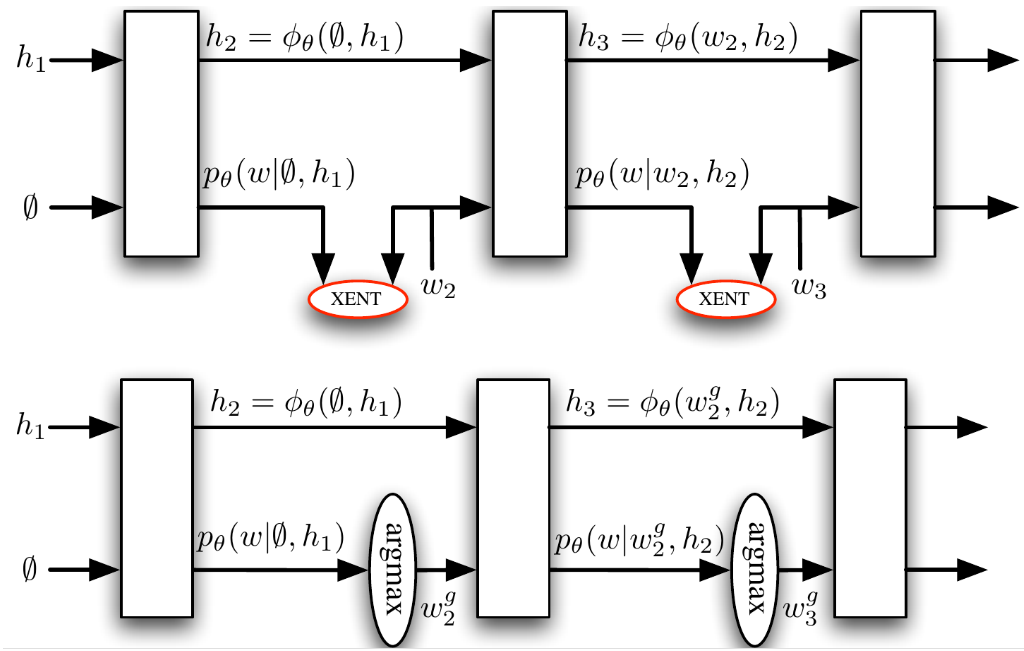
- Scheduled Sampling

| Approach | F1 |
|---|---|
| Baseline LSTM | 86.54 |
| Baseline LSTM with Dropout | 87.0 |
| Always Sampling | - |
| Scheduled Sampling | **88.08** |
| Scheduled Sampling with Dropout | **88.68** |

Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled Sampling for Sequence Prediction with Recurrent Nueral Networks. NIPS.

# Sequence to sequence ( II )

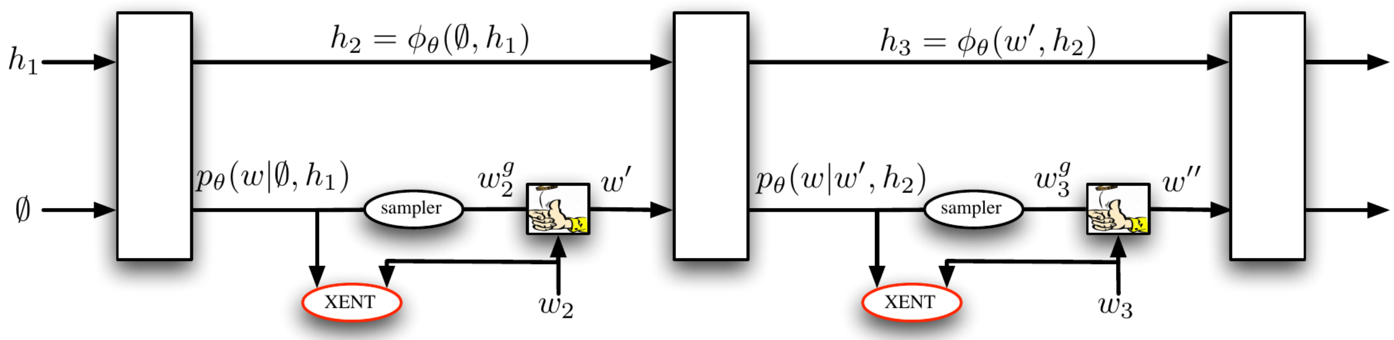
- Sequence-level training




Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence Level Training with Recurrent Neural Networks. ICLR.

# Sequence to sequence ( $II$ )

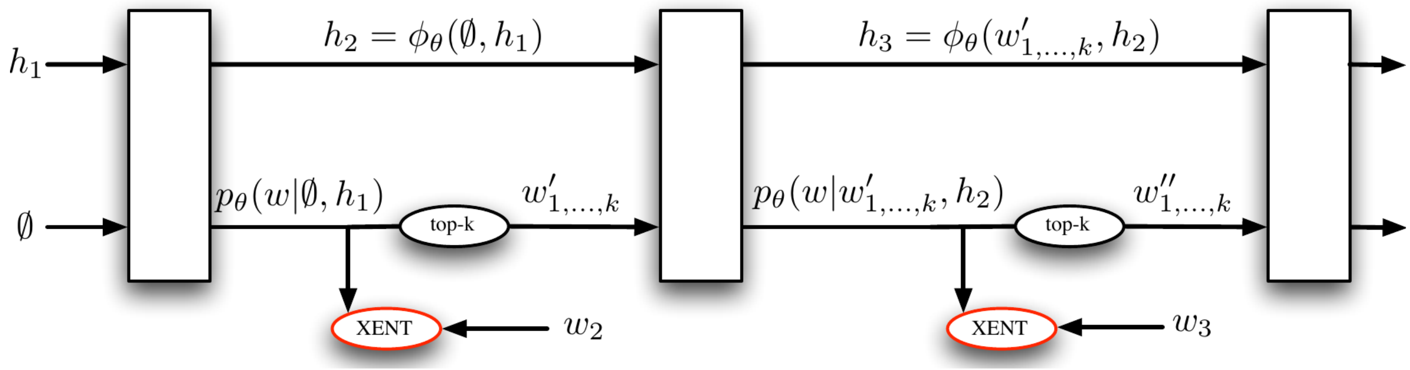
- Sequence-level training




Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence Level Training with Recurrent Neural Networks. ICLR.

# Sequence to sequence ( *II* )

- Sequence-level training




Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence Level Training with Recurrent Neural Networks. ICLR.

# Sequence to sequence ( Ⅱ )

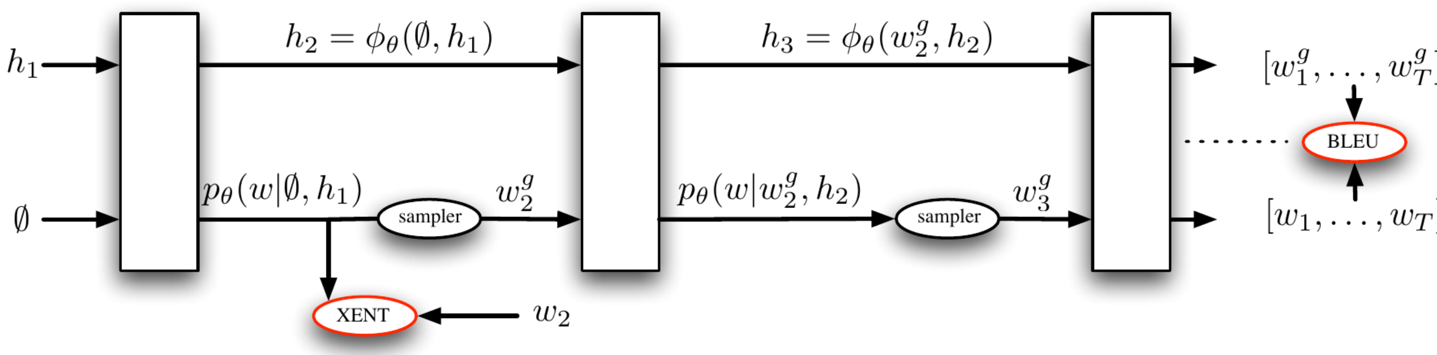
- Reinforce

$$L_\theta = -\sum_{w_1^g,\ldots,w_T^g} p_\theta(w_1^g,\ldots,w_T^g)\, r(w_1^g,\ldots,w_T^g) = -\mathbb{E}_{[w_1^g,\ldots w_T^g]\sim p_\theta} r(w_1^g,\ldots,w_T^g)$$

Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence Level Training with Recurrent Neural Networks. ICLR.

# Sequence to sequence ( $II$ )

- Mixer





Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence Level Training with Recurrent Neural Networks. ICLR.

# 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**

Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence Level Training with Recurrent Neural Networks. ICLR.

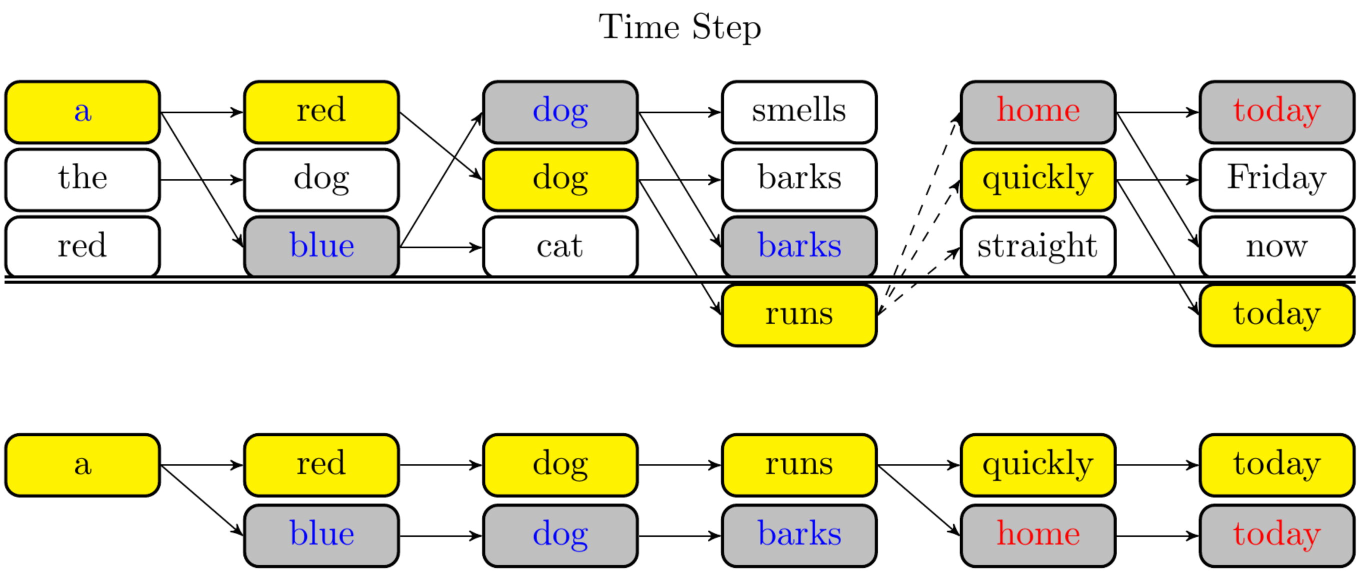
# Sequence to sequence ( *II* )

| *TASK* | XENT | DAD | E2E | MIXER |
|---|---|---|---|---|
| *summarization* | 13.01 | 12.18 | 12.78 | **16.22** |
| *translation* | 17.74 | 20.12 | 17.77 | **20.73** |
| *image captioning* | 27.8 | 28.16 | 26.42 | **29.16** |

Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence Level Training with Recurrent Neural Networks. ICLR.

# Sequence to sequence (Ⅲ)

- Learning for Search




Wiseman, S., & Rush, A. (2016). Sequence-to-Sequence Learning as Beam-Search Optimization. arxiv.

# Sequence to sequence (Ⅲ)

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

Wiseman, S., & Rush, A. (2016). Sequence-to-Sequence Learning as Beam-Search Optimization. arxiv.

# Sequence to sequence (Ⅲ)

- Need greedy pre-training

Wiseman, S., & Rush, A. (2016). Sequence-to-Sequence Learning as Beam-Search Optimization. arxiv.

# Sequence to sequence (Ⅲ)

- Curriculum beam increase

Wiseman, S., & Rush, A. (2016). Sequence-to-Sequence Learning as Beam-Search Optimization. arxiv.

# Sequence to sequence (Ⅲ)

| | Word Ordering (BLEU) | | |
|---|---|---|---|
| | $K_{te} = 1$ | $K_{te} = 5$ | $K_{te} = 10$ |
| seq2seq | 25.2 | 29.8 | 31.0 |
| BSO | 28.0 | 33.2 | 34.3 |
| ConBSO | **28.6** | **34.3** | **34.5** |
| LSTM-LM | 15.4 | - | 26.8 |

Wiseman, S., & Rush, A. (2016). Sequence-to-Sequence Learning as Beam-Search Optimization. arxiv.

# Sequence to sequence (Ⅲ)

| | Dependency Parsing (UAS/LAS) | | |
|---|---|---|---|
| | $K_{te} = 1$ | $K_{te} = 5$ | $K_{te} = 10$ |
| seq2seq | **87.33/82.26** | 88.53/84.16 | 88.66/84.33 |
| BSO | 86.91/82.11 | 91.00/**87.18** | 91.17/**87.41** |
| ConBSO | 85.11/79.32 | **91.25**/86.92 | **91.57**/87.26 |
| Andor | 93.17/91.18 | - | - |

Wiseman, S., & Rush, A. (2016). Sequence-to-Sequence Learning as Beam-Search Optimization. arxiv.

# Sequence to sequence (Ⅲ)

| | Machine Translation (BLEU) | | |
|---|---|---|---|
| | $K_{te} = 1$ | $K_{te} = 5$ | $K_{te} = 10$ |
| seq2seq | 22.53 | 24.03 | 23.87 |
| BSO, SB-$\Delta$ | **23.83** | **26.36** | **25.48** |
| XENT | 17.74 | $\leq 20.5$ | $\leq 20.5$ |
| DAD | 20.12 | $\leq 22.5$ | $\leq 23.0$ |
| MIXER | 20.73 | - | $\leq 22.0$ |

Wiseman, S., & Rush, A. (2016). Sequence-to-Sequence Learning as Beam-Search Optimization. arxiv.