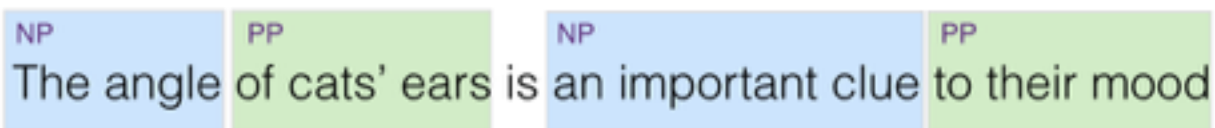
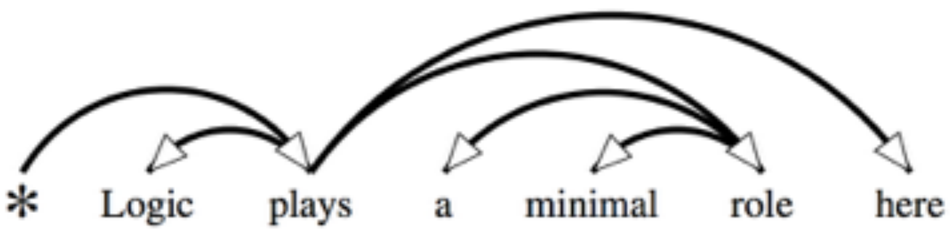
Neural Representation Learning in Linguistic Structured Prediction

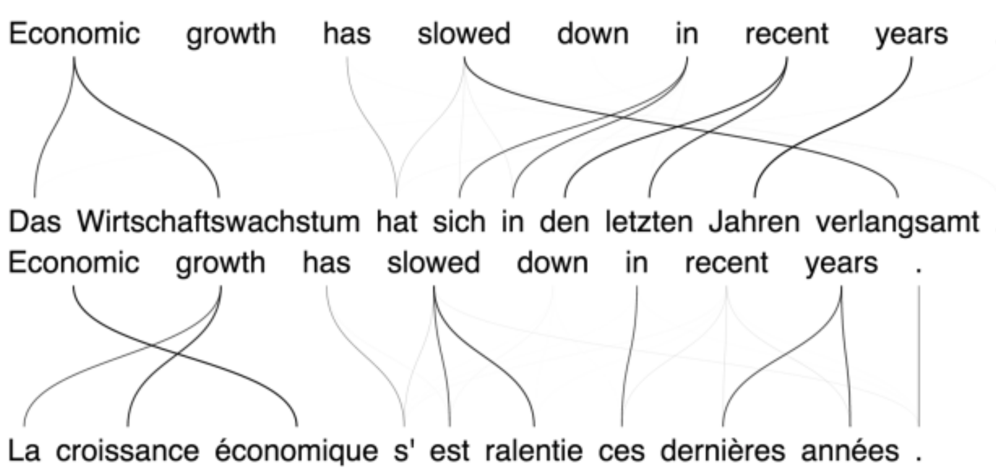
Lingpeng Kong
Carnegie Mellon University

Thesis Defense
9/18/2017

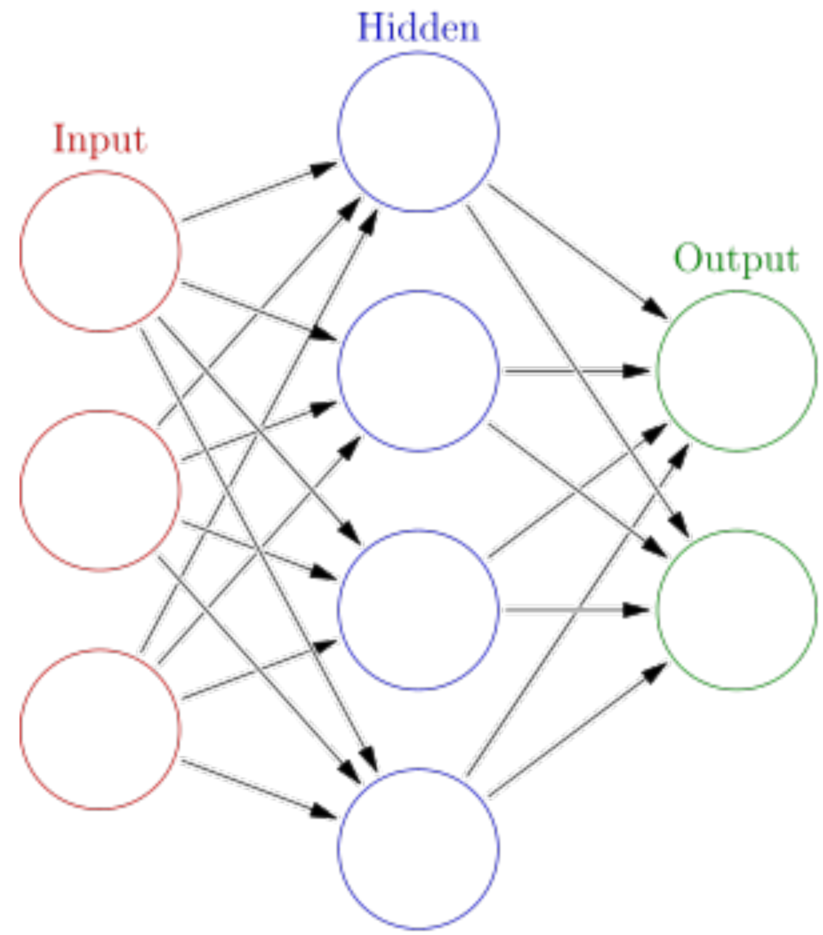
Linguistic Structure

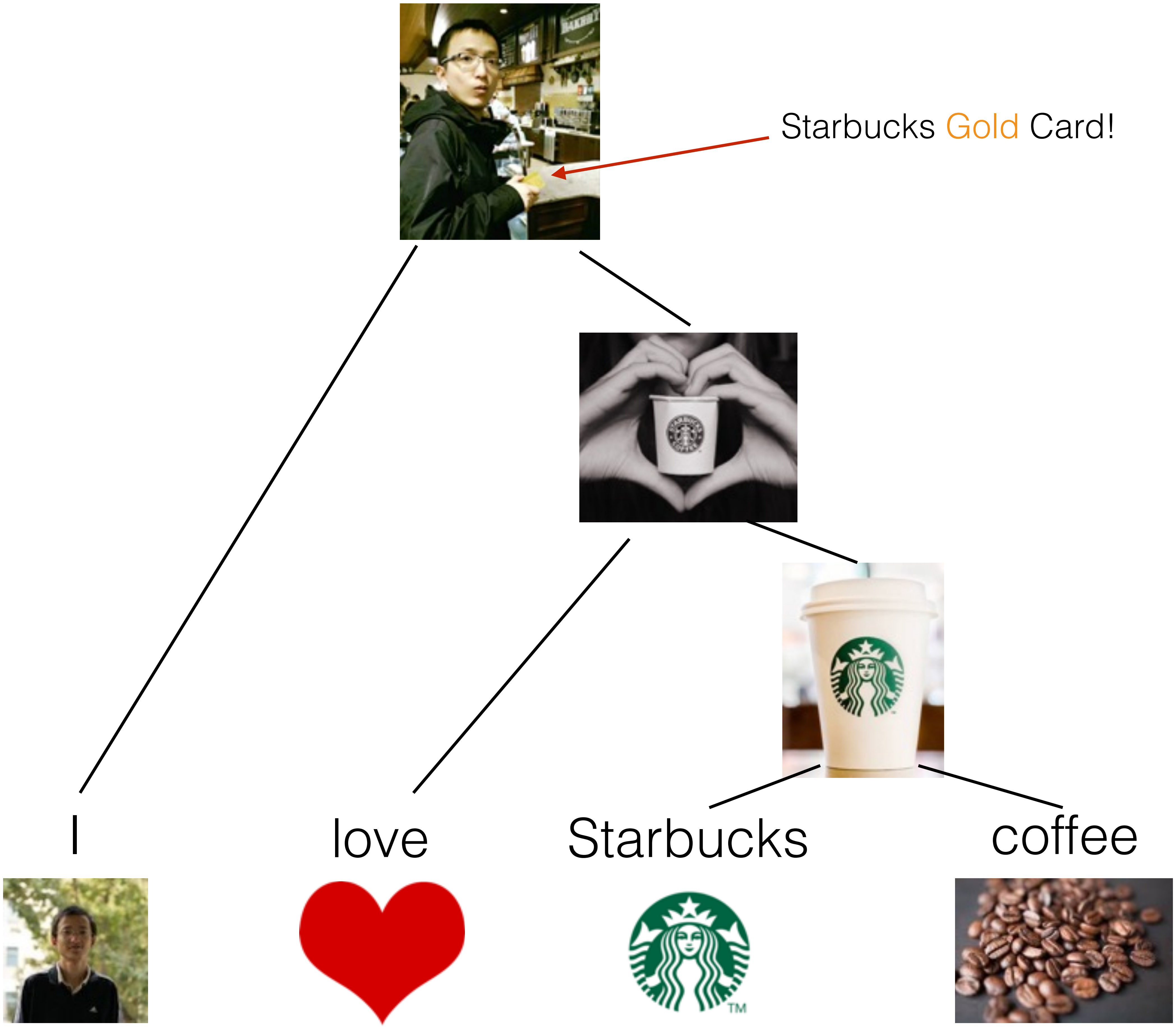


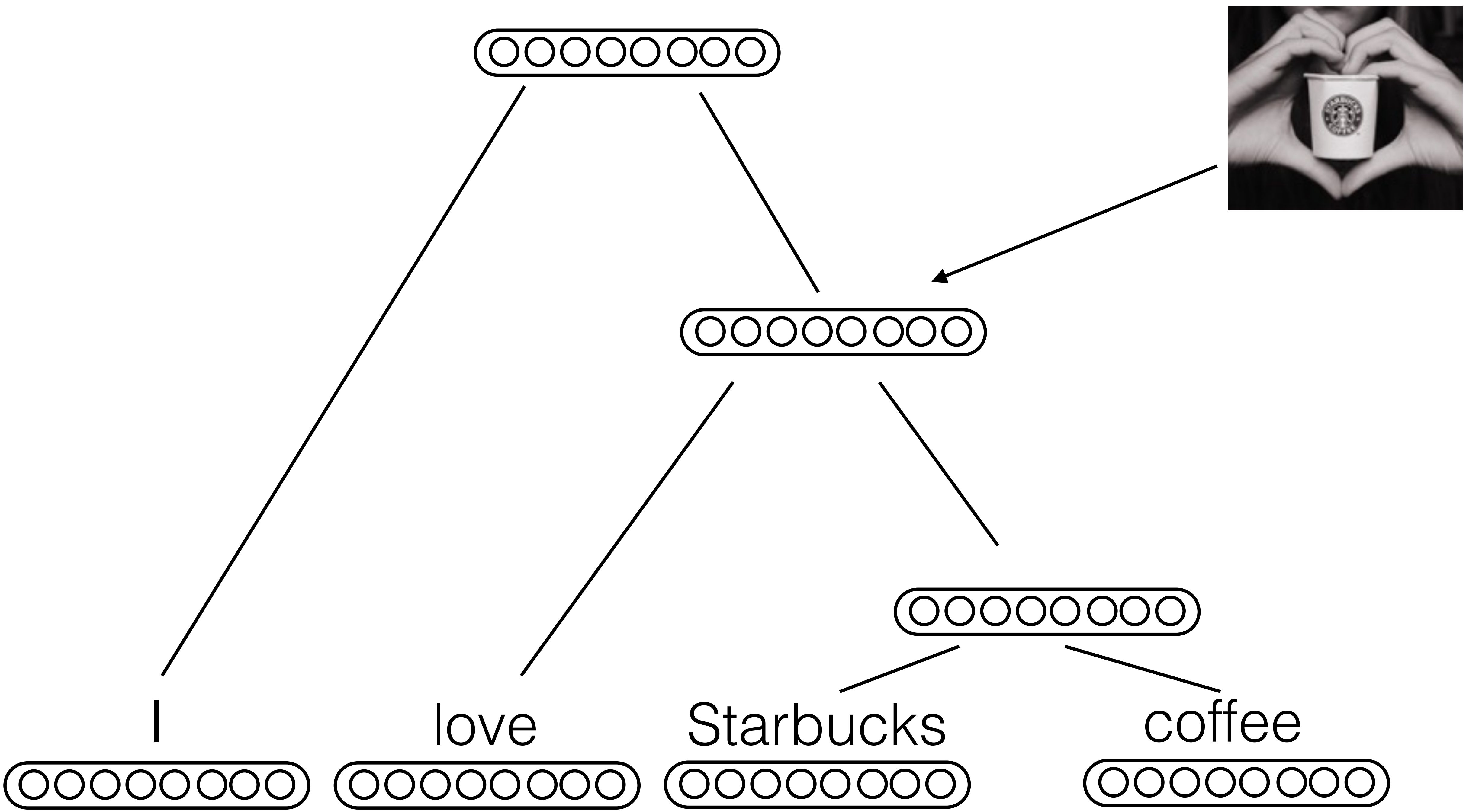




Neural Representation Learning




Starbucks Gold Card!
I
love
Starbucks
coffee

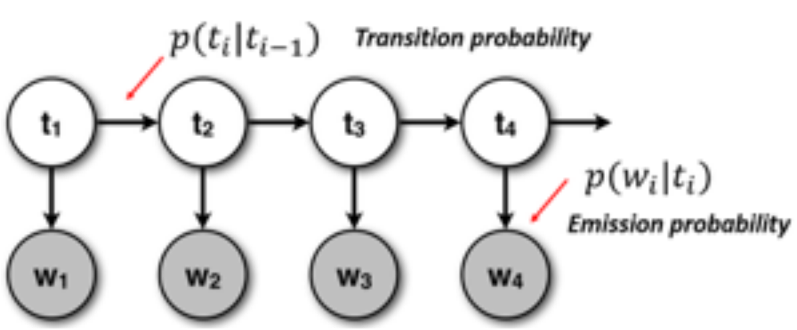

I
love
Starbucks
coffee

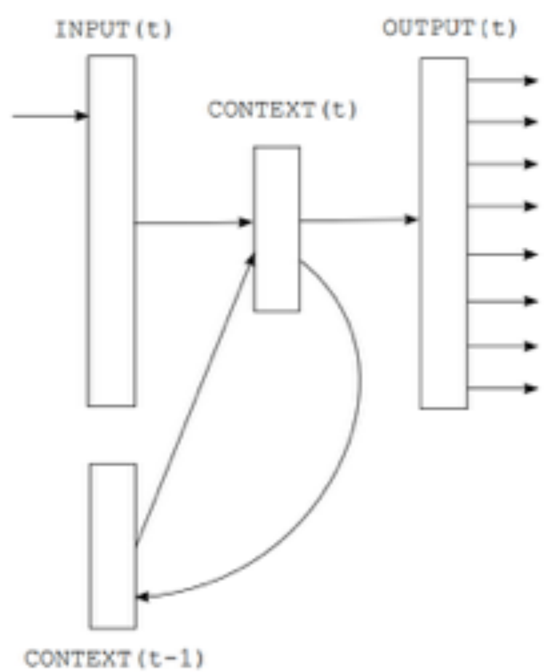
Inductive bias (Mitchell, 1980)

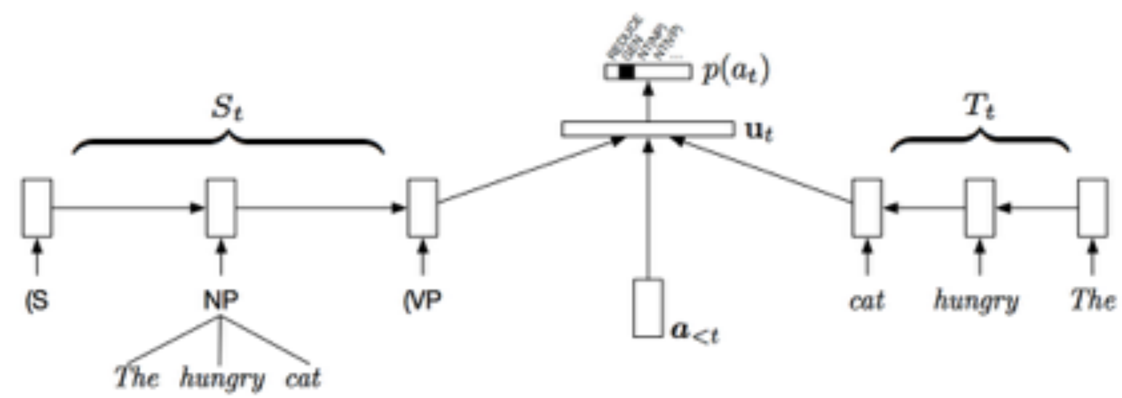
(Dyer et al, 2016)

(Mikolov et al, 2010)

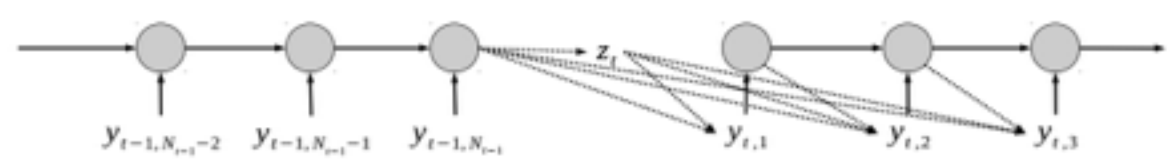
(Chen and Goodman, 1980)



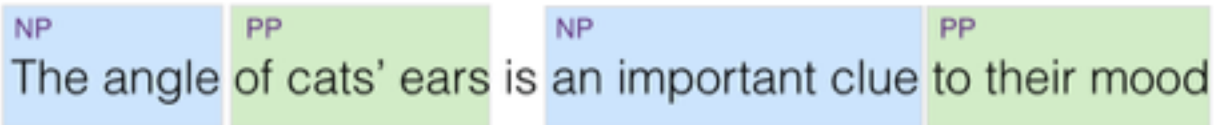




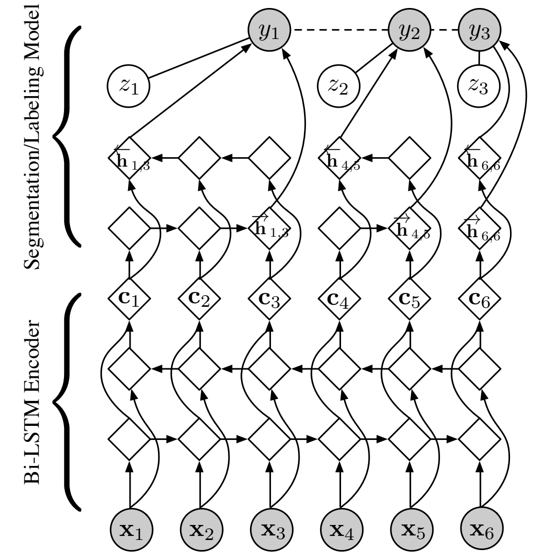
(Ji et al, 2016)



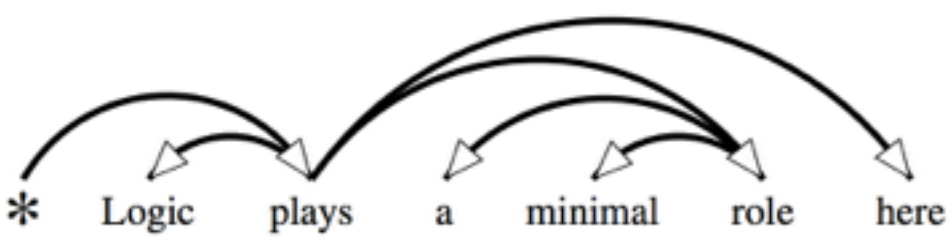
z — linguistic structures



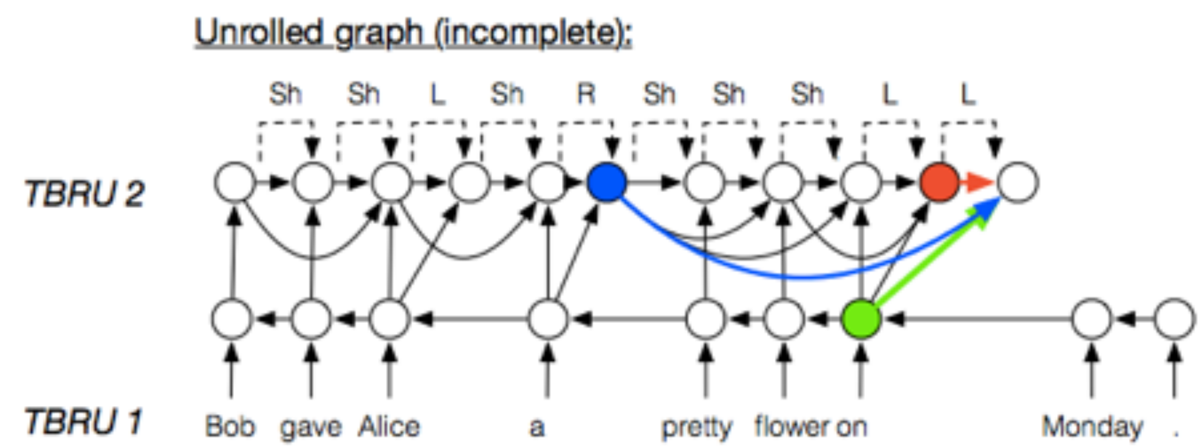
z = segment structures



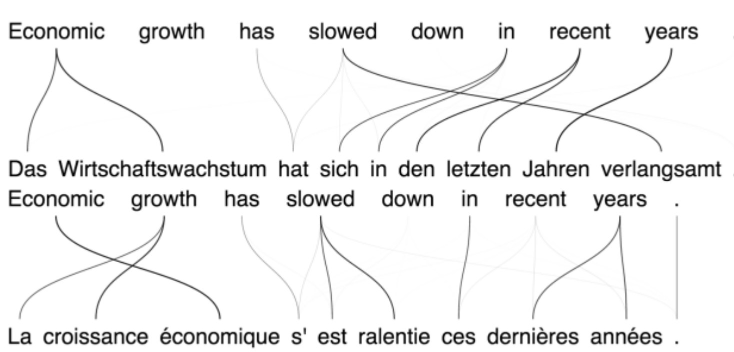
Part I: Segmental RNNs



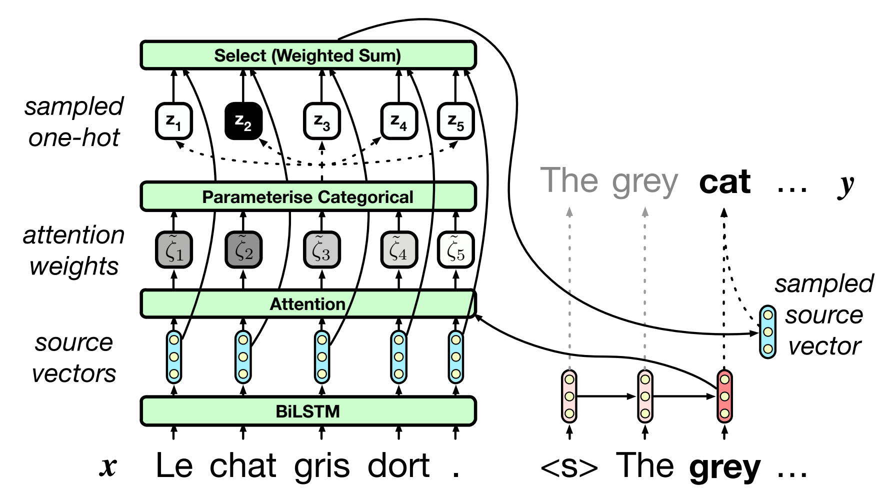
z = parse tree structures



Part II: DRAGNN



z = alignment structures



Part III: Stochastic Attention

Outline

- Introduction
- Part I — Segmental Recurrent Neural Networks
- Part II — A Transition-based Framework for Dynamically Connected Neural Networks
- Part III — Inference and Regularization in Sequence to Sequence Models with Stochastic Attention
- Conclusion and Future Work

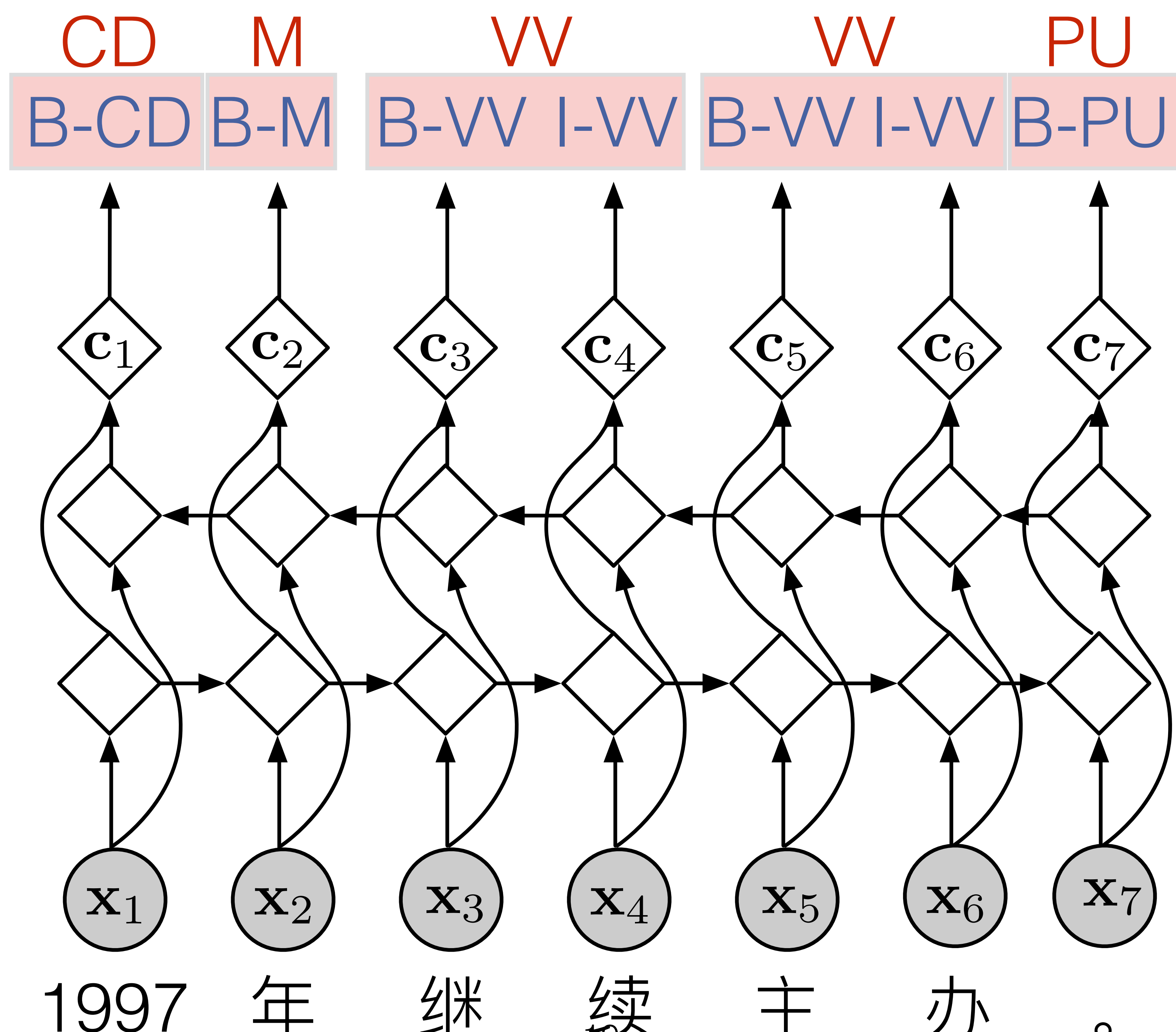
Outline

- Introduction
- Part I — Segmental Recurrent Neural Networks
- Part II — A Transition-based Framework for Dynamically Connected Neural Networks
- Part III — Inference and Regularization in Sequence to Sequence Models with Stochastic Attention
- Conclusion and Future Work

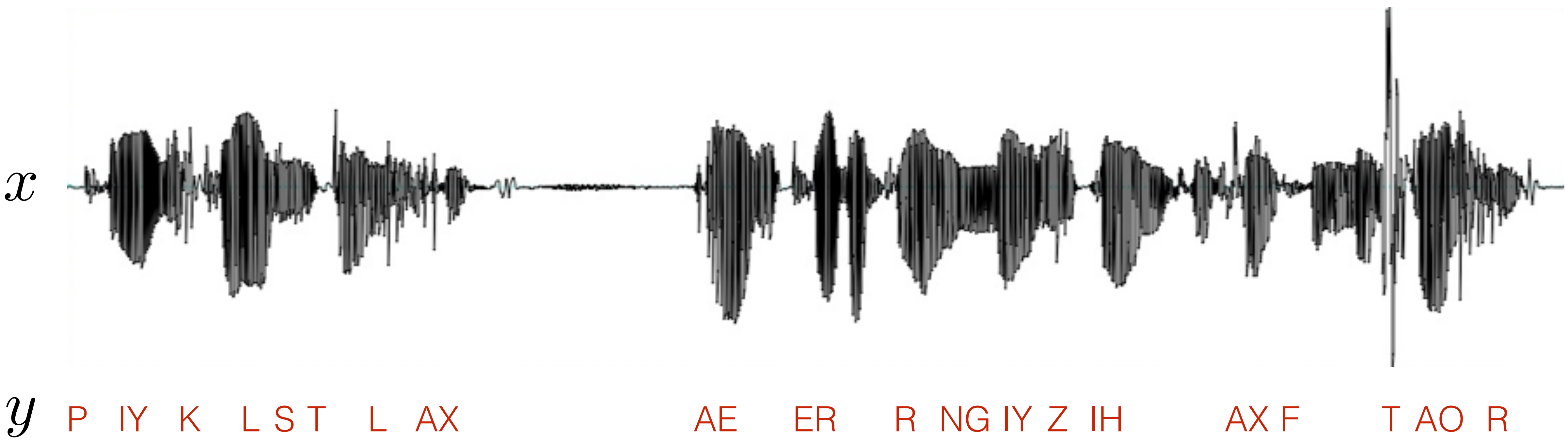
Joint Word Segmentation and POS Tagging

x	1997	年	继 续	主 办	。
y	CD	M	VV	VV	PU

Bi-LSTMs Tagger



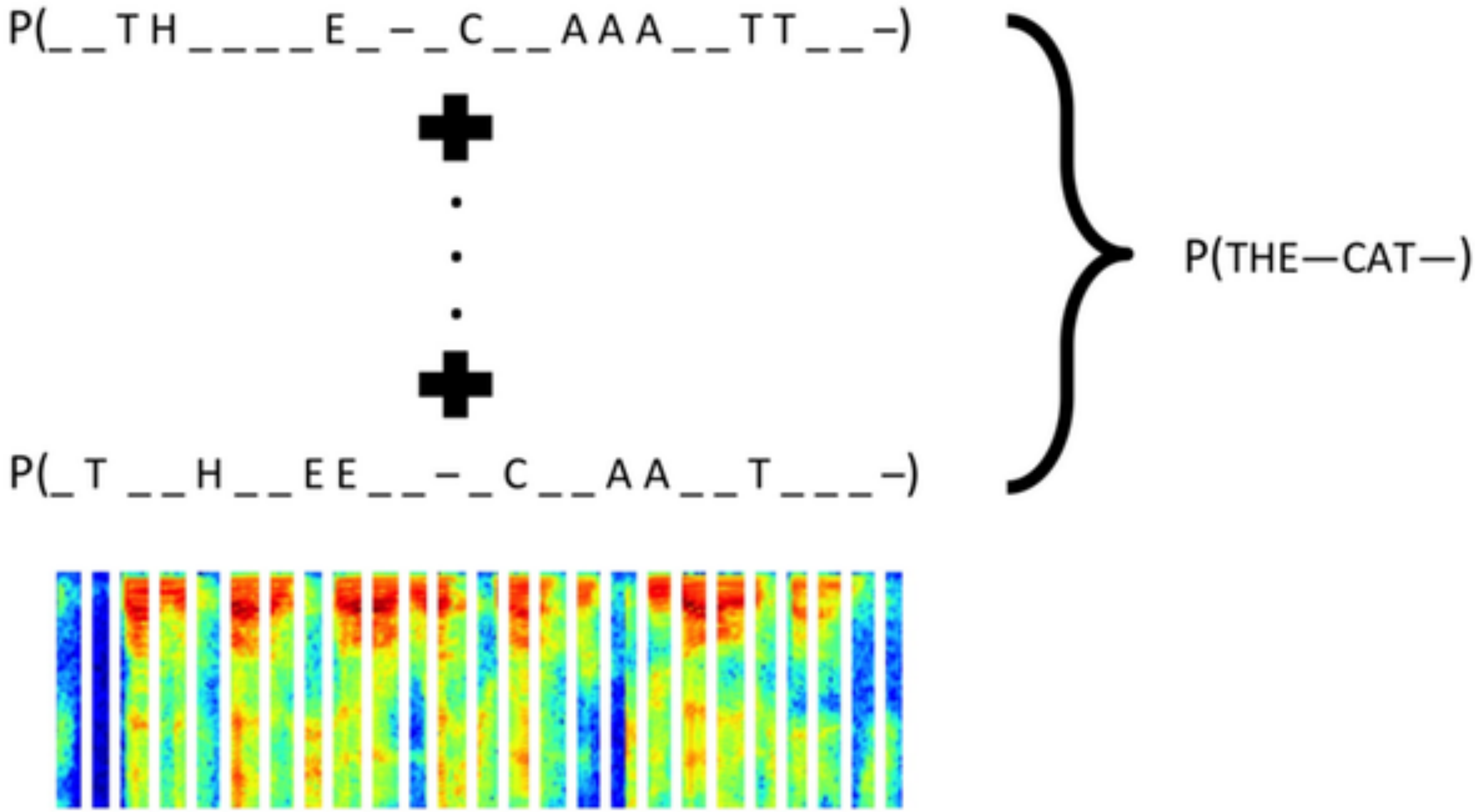
Speech Recognition



Text: *Please call Stella. Ask her to bring these things with her from the store.*

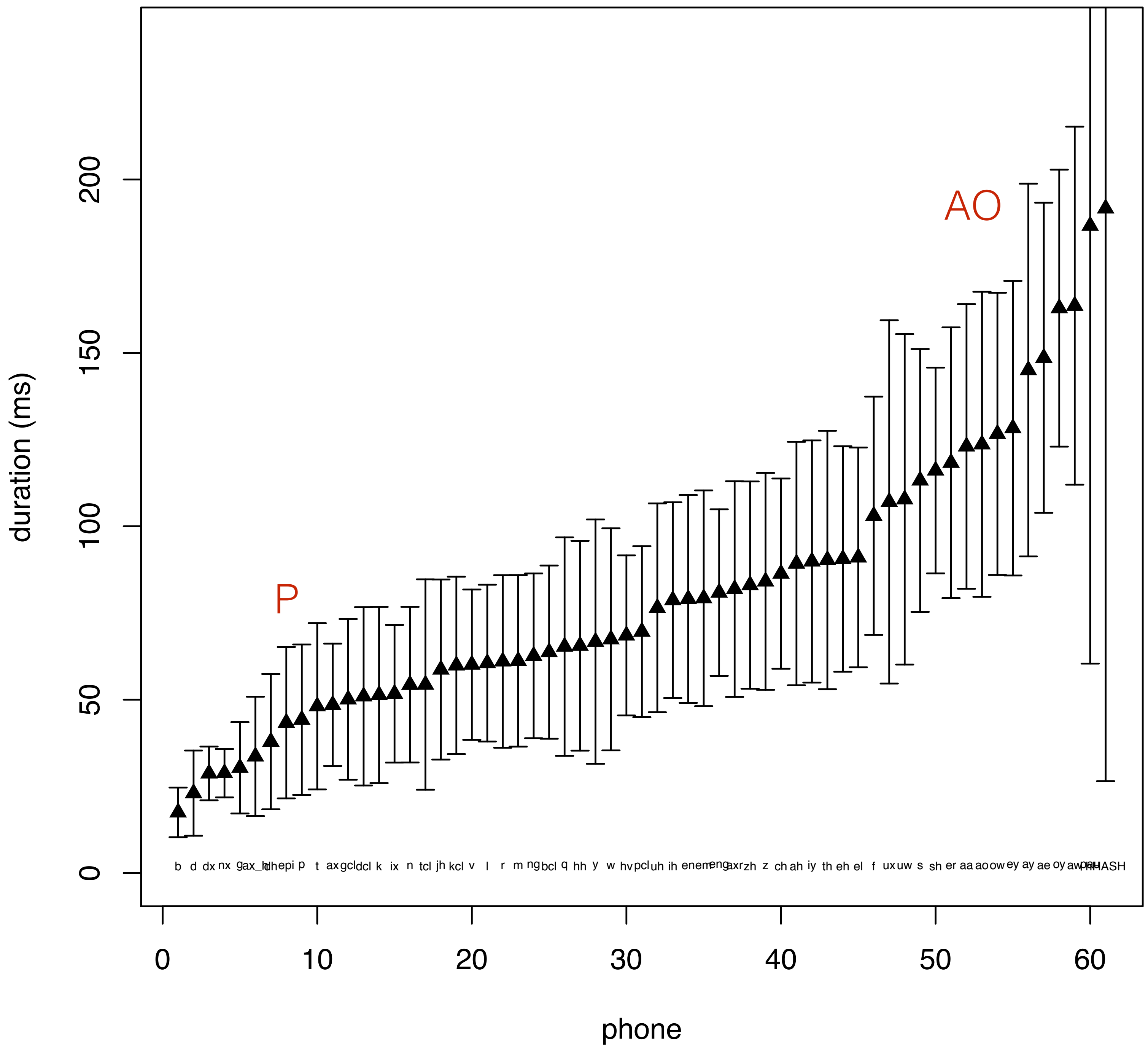
http://groups.linguistics.northwestern.edu/documentation/images/praat_aligned.jpg

Connectionist Temporal Classification (CTC)



(Graves et al, 2013)

Duration Features



Segmental Recurrent Neural Networks (SRNNs)

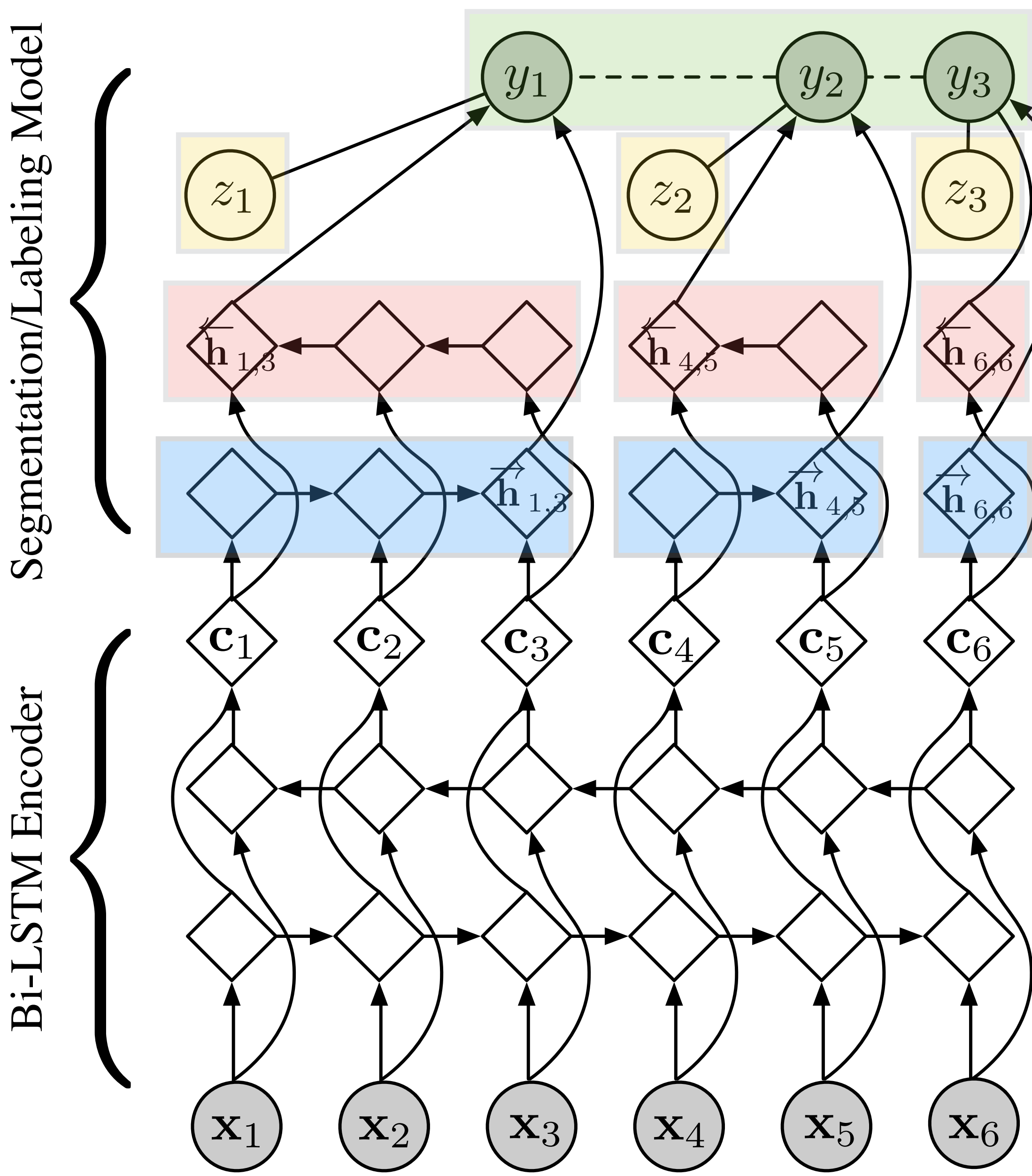
x	1997	年	继	续	主	办	。
z	1	1	2		2		1
y	CD	M	VV		VV		PU

SRNNs — $p(y, z|x)$

$$y^* = \arg\max_y \sum_z p(y, z \mid x)$$

$$\approx \arg\max_y \max_z p(y, z \mid x)$$

Segmental Recurrent Neural Networks (SRNNs)



Forward Segment Embedding

Backward Segment Embedding

Duration Embedding

Label Embedding

$$p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{i=1}^{|\boldsymbol{y}|} \exp f(y_{i-k:i}, z_i, \mathbf{x})$$

$$f(y_{i-k:i}, z_i, \mathbf{x}_{s_i:s_i+z_i-1}) = \mathbf{w}^\top \phi(\mathbf{V}[\mathbf{g}_y(y_{i-k}); \ldots; \mathbf{g}_y(y_i); \mathbf{g}_z(z_i); \overrightarrow{\mathrm{RNN}}(\mathbf{c}_{s_i:s_i+z_i-1}); \overleftarrow{\mathrm{RNN}}(\mathbf{c}_{s_i:s_i+z_i-1})] + \mathbf{a}) + b$$

(Sarawagi and Cohen, 2005)

Parameter Learning

Fully Supervised

$$\mathcal{L} = \sum_{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\in\mathcal{D}} -\log p(\boldsymbol{y},\boldsymbol{z} \mid \boldsymbol{x})$$

$$= \sum_{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\in\mathcal{D}} \underline{\log Z(\boldsymbol{x})} - \log Z(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$$

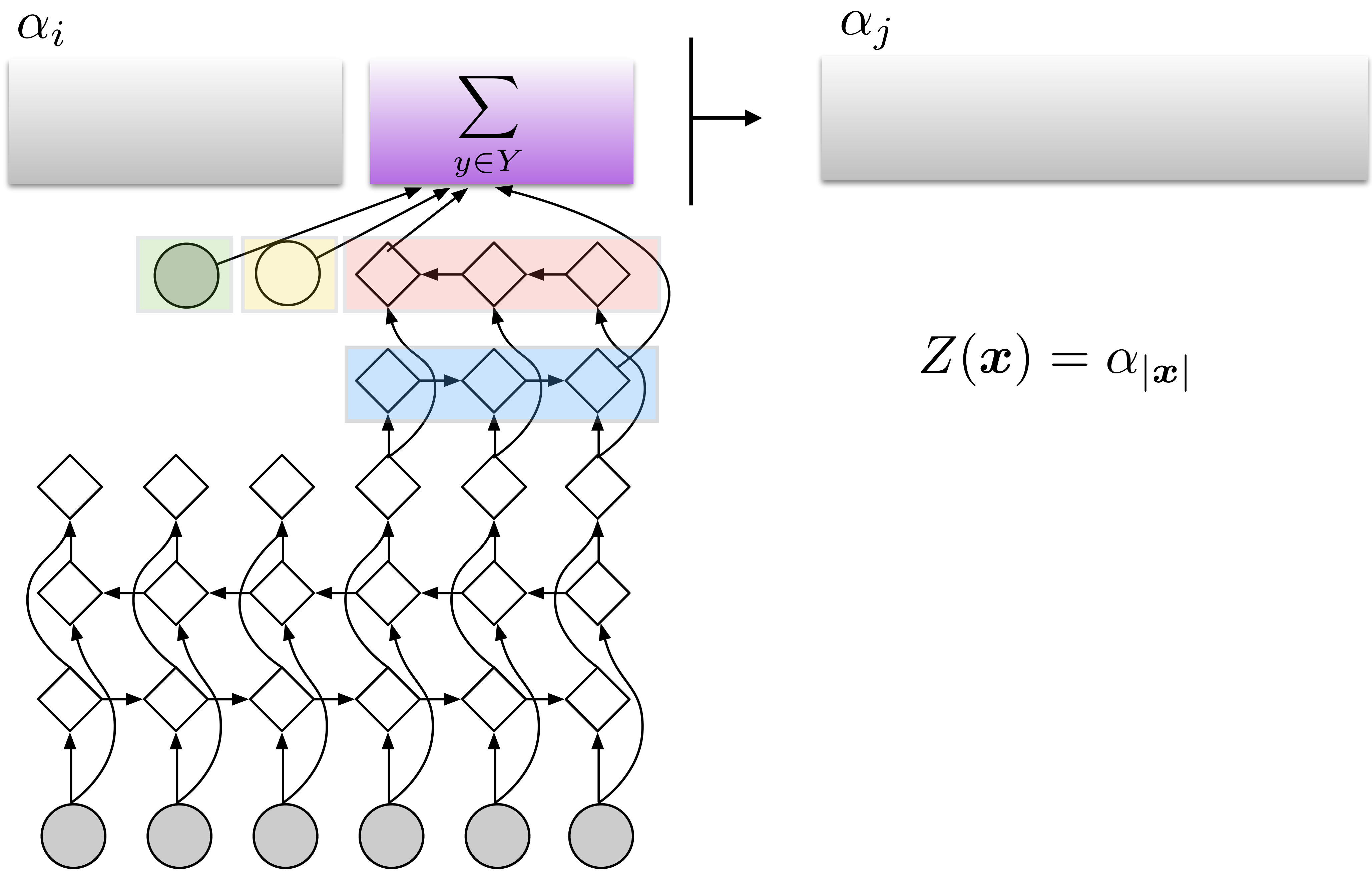
Partially Supervised

$$\mathcal{L} = \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} -\log p(\boldsymbol{y} \mid \boldsymbol{x})$$

$$= \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \sum_{\boldsymbol{z}\in\mathcal{Z}(\boldsymbol{x},\boldsymbol{y})} -\log p(\boldsymbol{y},\boldsymbol{z} \mid \boldsymbol{x})$$

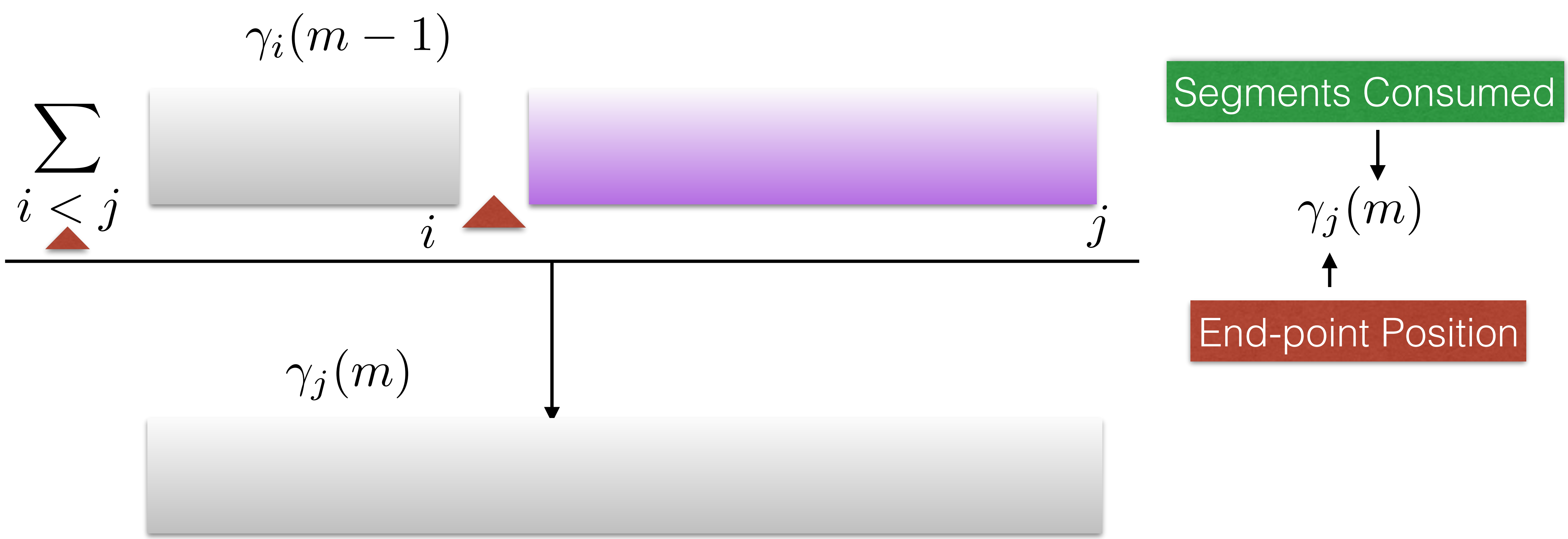
$$= \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \underline{\log Z(\boldsymbol{x})} - \underline{\log Z(\boldsymbol{x},\boldsymbol{y})}$$

Dynamic Programming



$$Z(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}$$

Dynamic Programming



$$Z(\boldsymbol{x}, \boldsymbol{y}) = \gamma_{|\boldsymbol{x}|}(|\boldsymbol{y}|)$$

Experiments

Online Hand Writing Recognition

	P (seg)	R (seg)	F (seg)	Error
CTC	-	-	-	13.8%
SRNNs(Full)	98.9%	98.6%	98.6%	5.4%
SRNNs (Partial)	99.2%	99.1%	99.2%	2.7%

(Kassel, 1995)
(Taskar et al. 2004)

Experiments

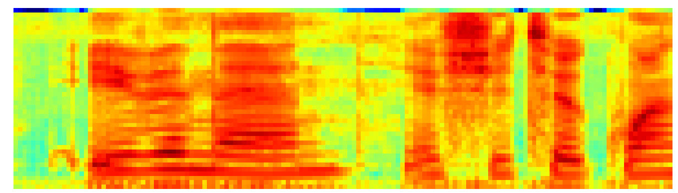
Joint Chinese word segmentation and POS tagging

1997	年	继 续	主 办	。
CD	M	VV	VV	PU

	P (seg)	R (seg)	F (seg)
BiRNNs	94.7%	95.2%	95.0%
SRNNs	95.3%	95.8%	95.5%

	P (tag)	R (tag)	F (tag)
BiRNNs	88.1%	88.5%	88.3%
SRNNs	89.8%	90.3%	90.3%

Experiments

Speech Recognition 

Multi-task Learning with CTC and Segmental CRF for Speech Recognition [INTERSPEECH 2016]

Segmental Recurrent Neural Networks for End-to-end Speech Recognition [INTERSPEECH 2017]

Outline

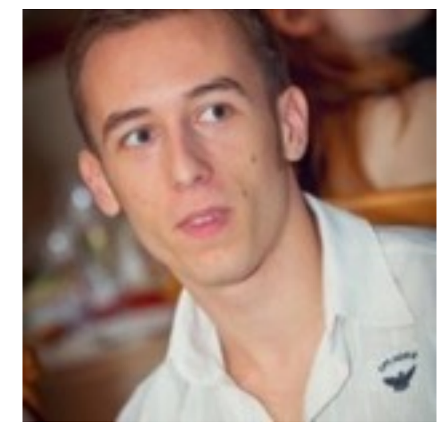
- Introduction
- Part I — Segmental Recurrent Neural Networks
- Part II — A Transition-based Framework for Dynamically Connected Neural Networks
- Part III — Inference and Regularization in Sequence to Sequence Models with Stochastic Attention
- Conclusion and Future Work

Dynamic **R**ecurrent **A**cyclic **G**raphical **N**eural **N**etworks (DRAGNN)

David Weiss

Chris Alberti

Daniel Andor

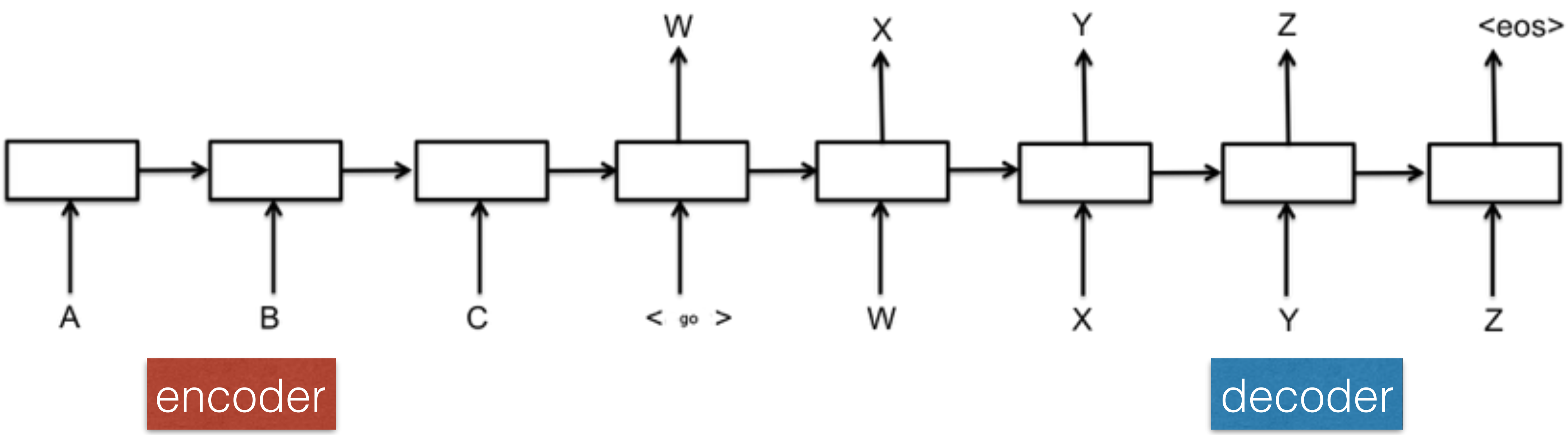


Ivan Bogatyy

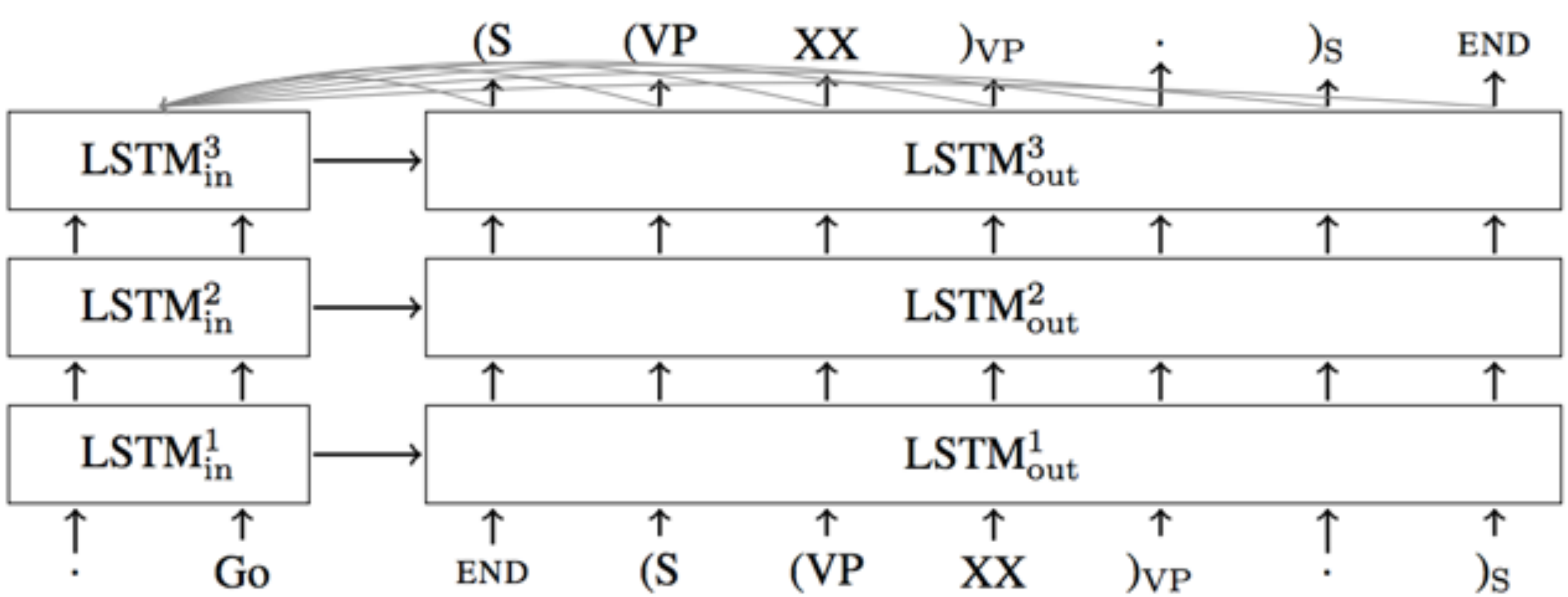
https://github.com/tensorflow/models/tree/master/syntaxnet/dragnn

Research at Google

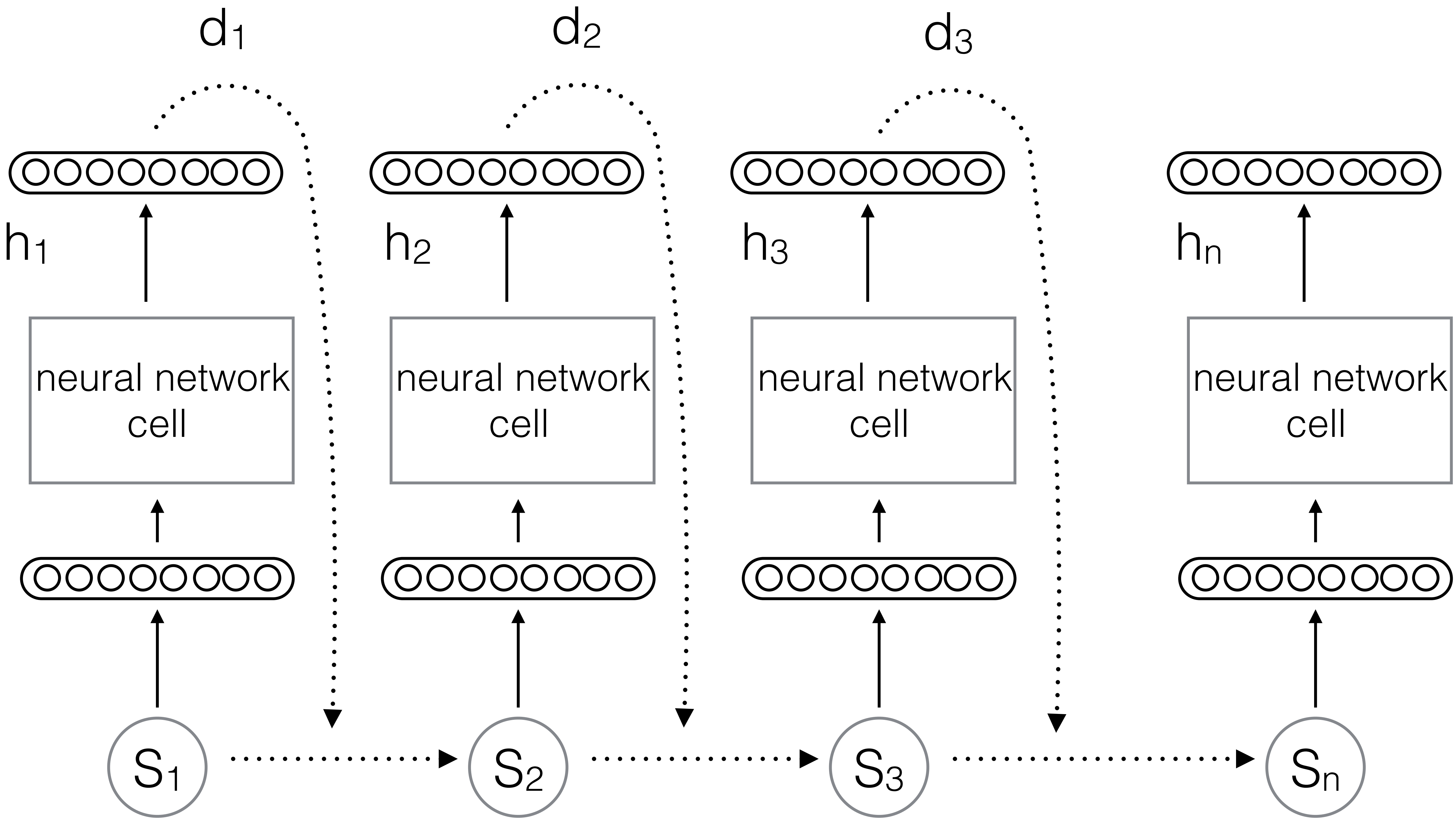
Sequence-to-sequence model

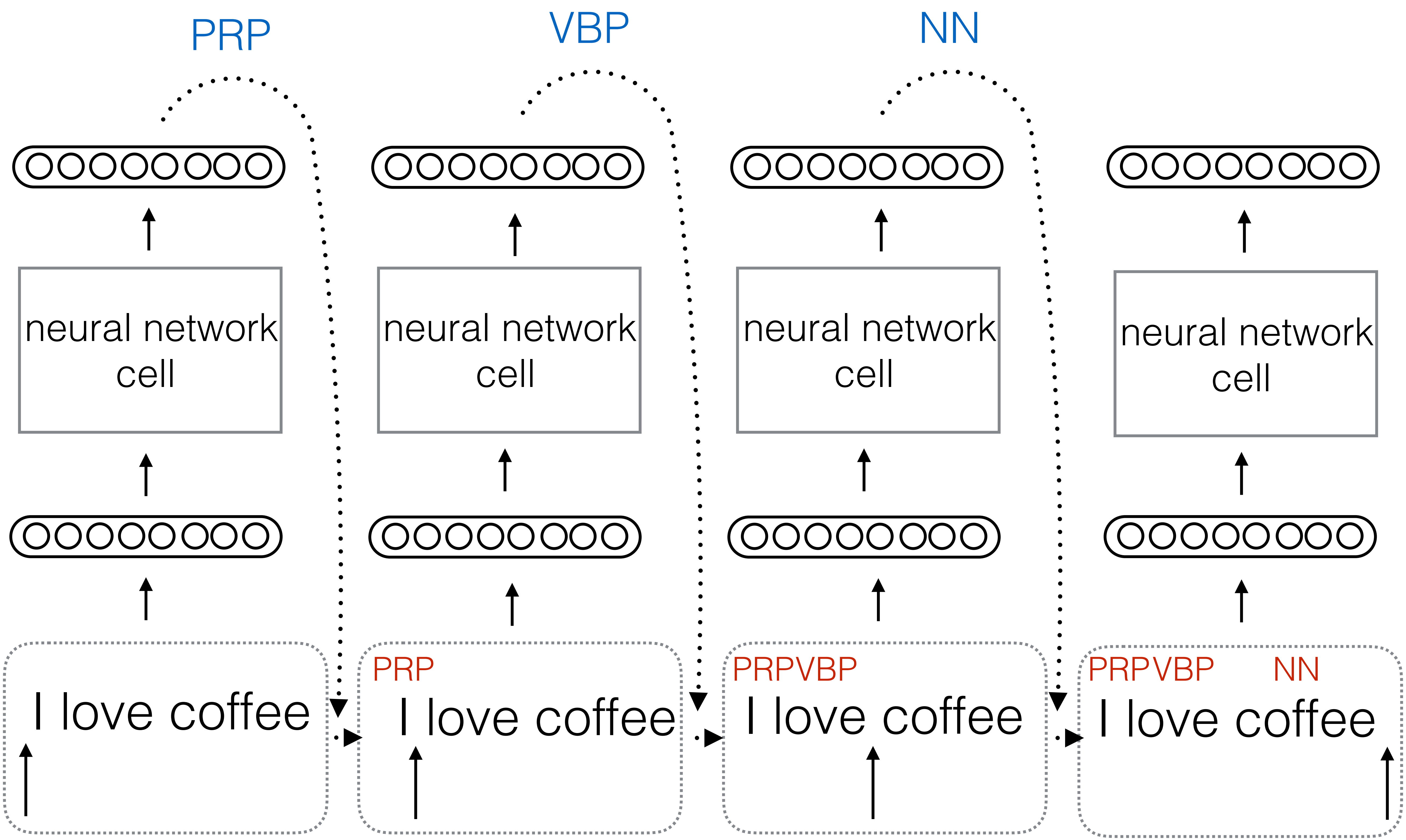


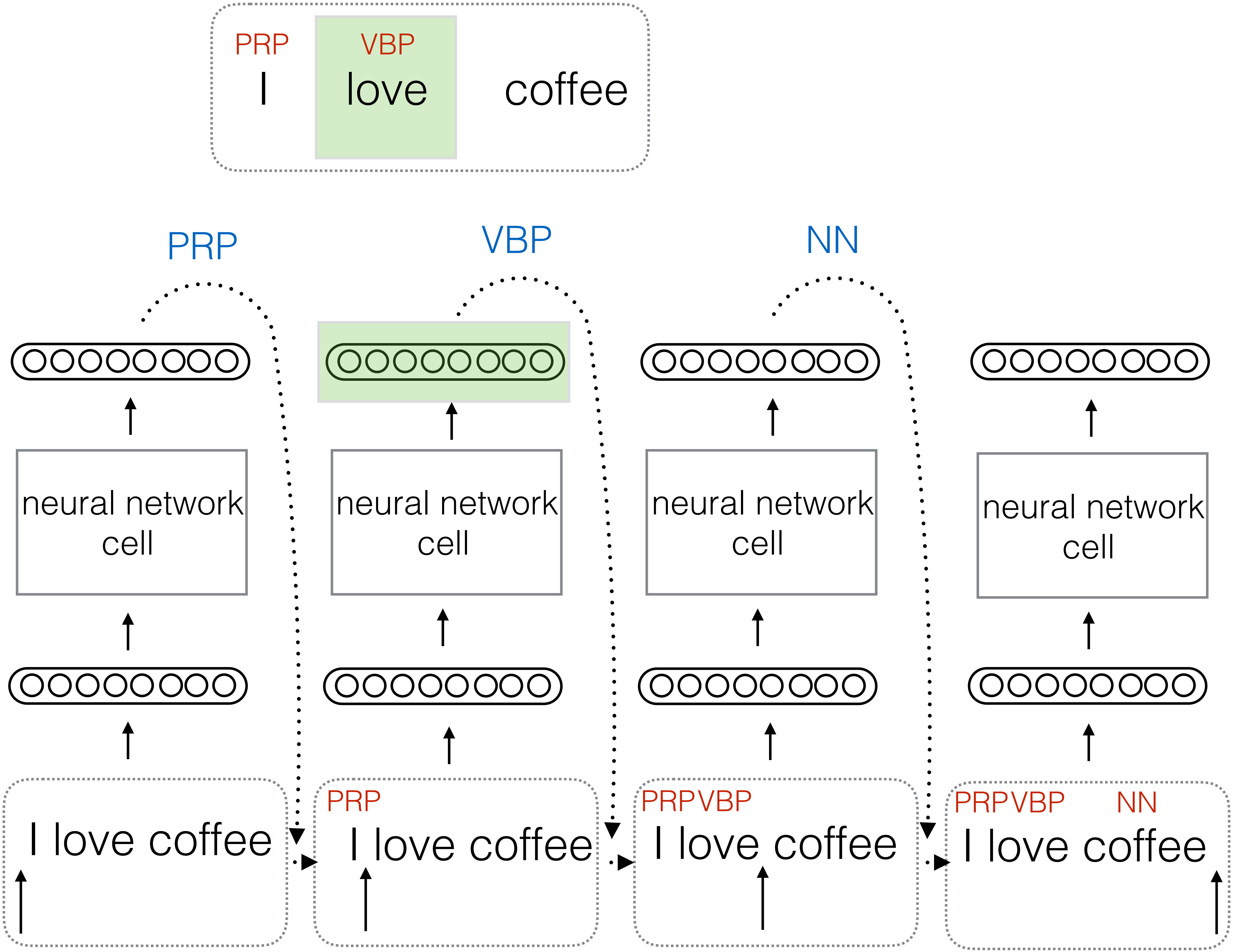
(Cho et al, 2014)



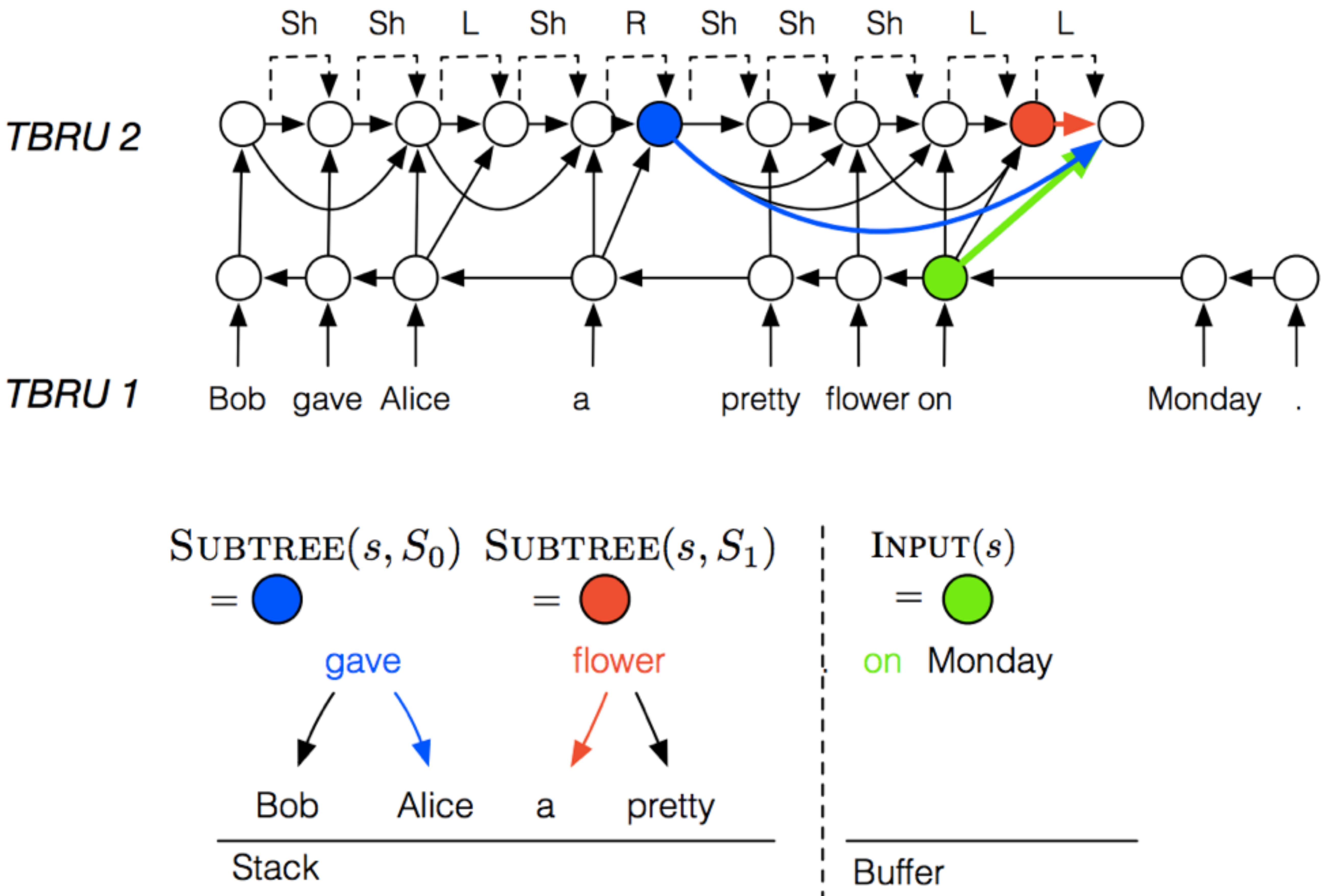
(Vinyals et al, 2015)


d1
d2
d3
h1
h2
h3
hn
neural network cell
neural network cell
neural network cell
neural network cell
S1
S2
S3
Sn


PRP
VBP
NN
neural network cell
neural network cell
neural network cell
neural network cell
I love coffee
PRP
I love coffee
PRPVBP
I love coffee
PRPVBP NN
I love coffee

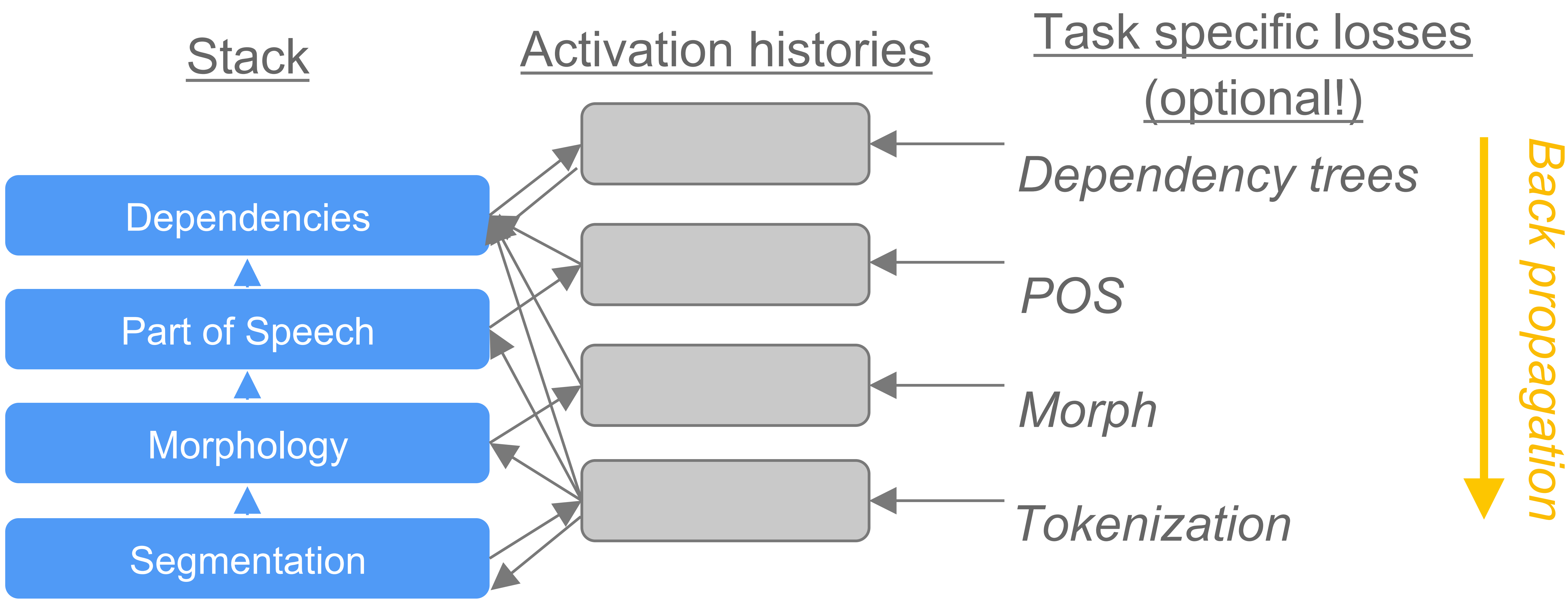


PRP
I
VBP
love
coffee
PRP
VBP
NN
neural network cell
neural network cell
neural network cell
neural network cell
I love coffee
PRP
I love coffee
PRPVBP
I love coffee
PRPVBP NN
I love coffee

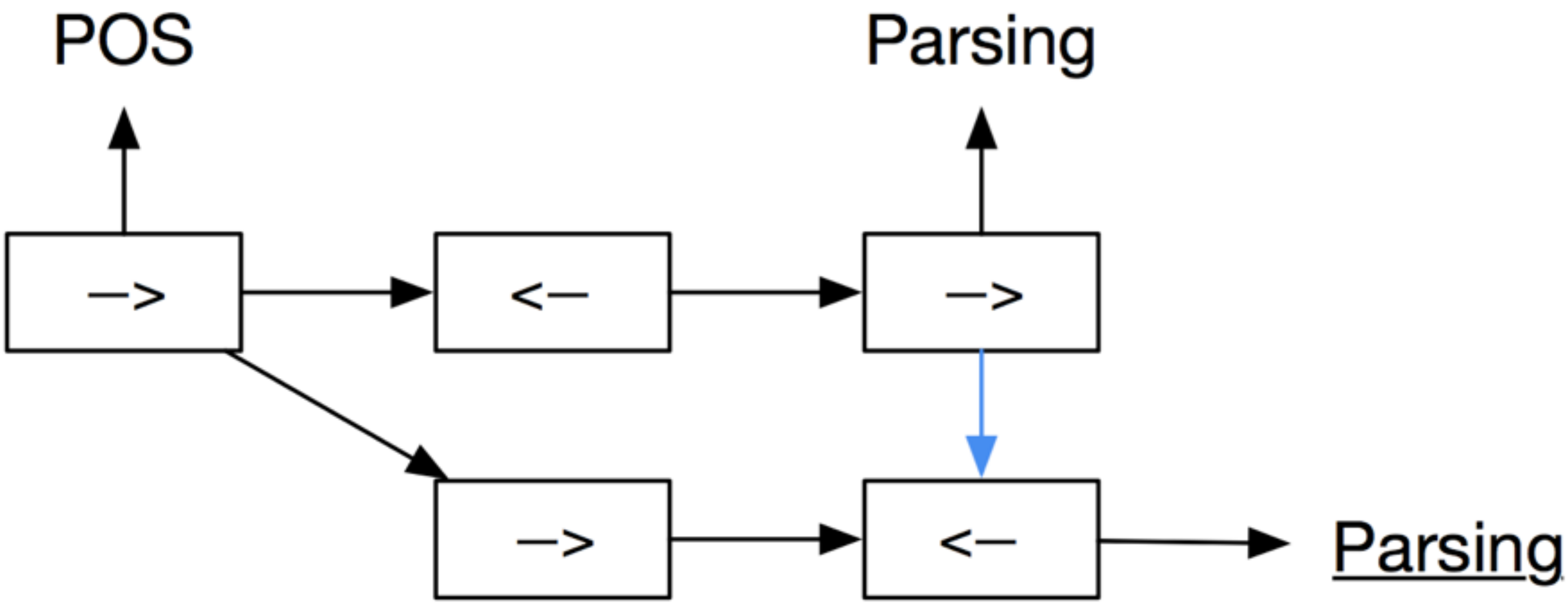


Sh Sh L Sh R Sh Sh Sh L L
TBRU 2
TBRU 1
Bob gave Alice a pretty flower on Monday .
SUBTREE(s, S0) =
SUBTREE(s, S1) =
INPUT(s) =
gave
flower
on Monday
Bob Alice a pretty
Stack
Buffer

Deep multi-task learning with DRAGNN

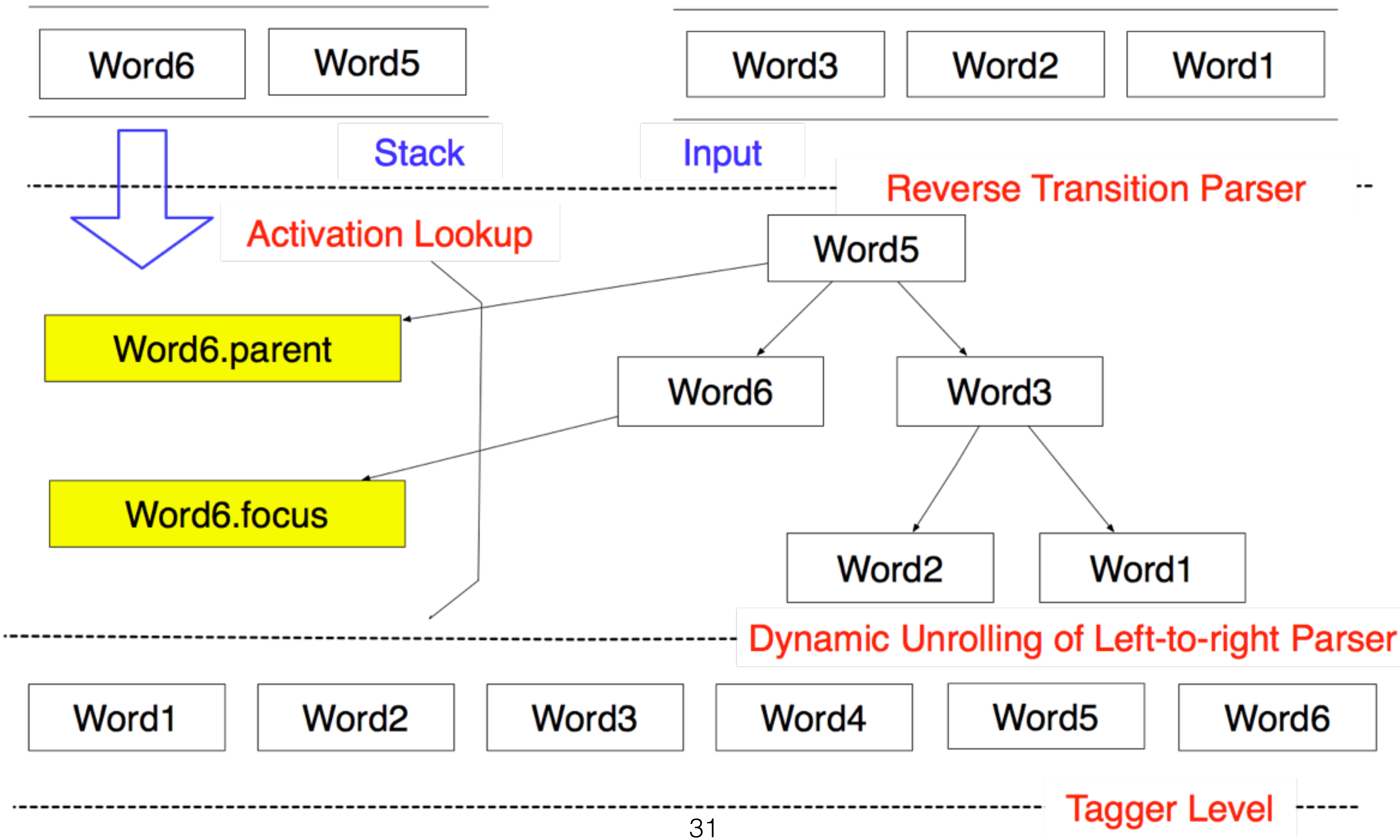


Bi-directional Parsing



(Attardi and Dell'Orletta, 2009)

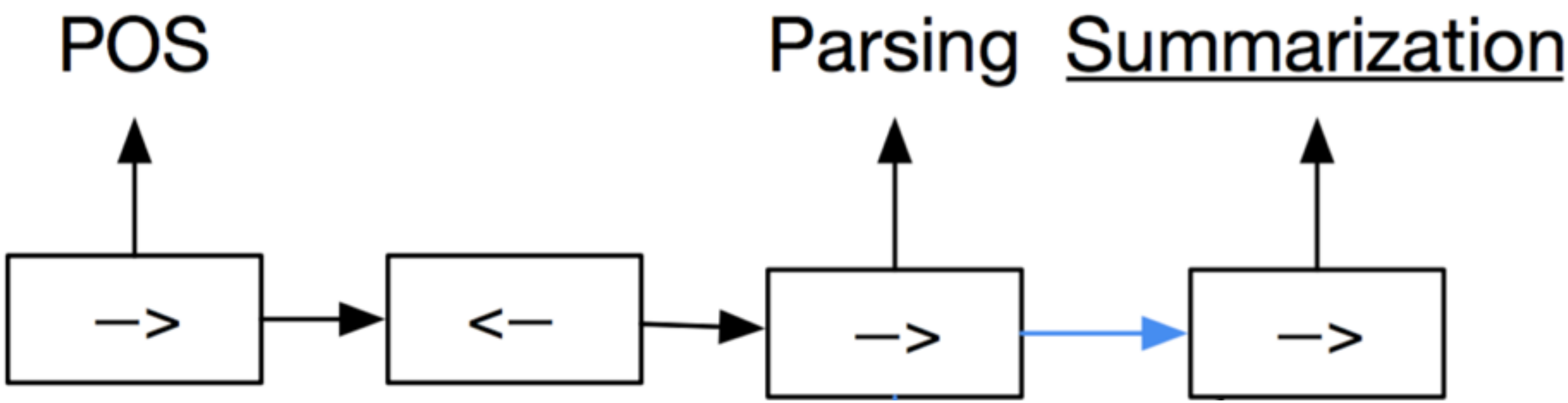
Bi-directional Parsing



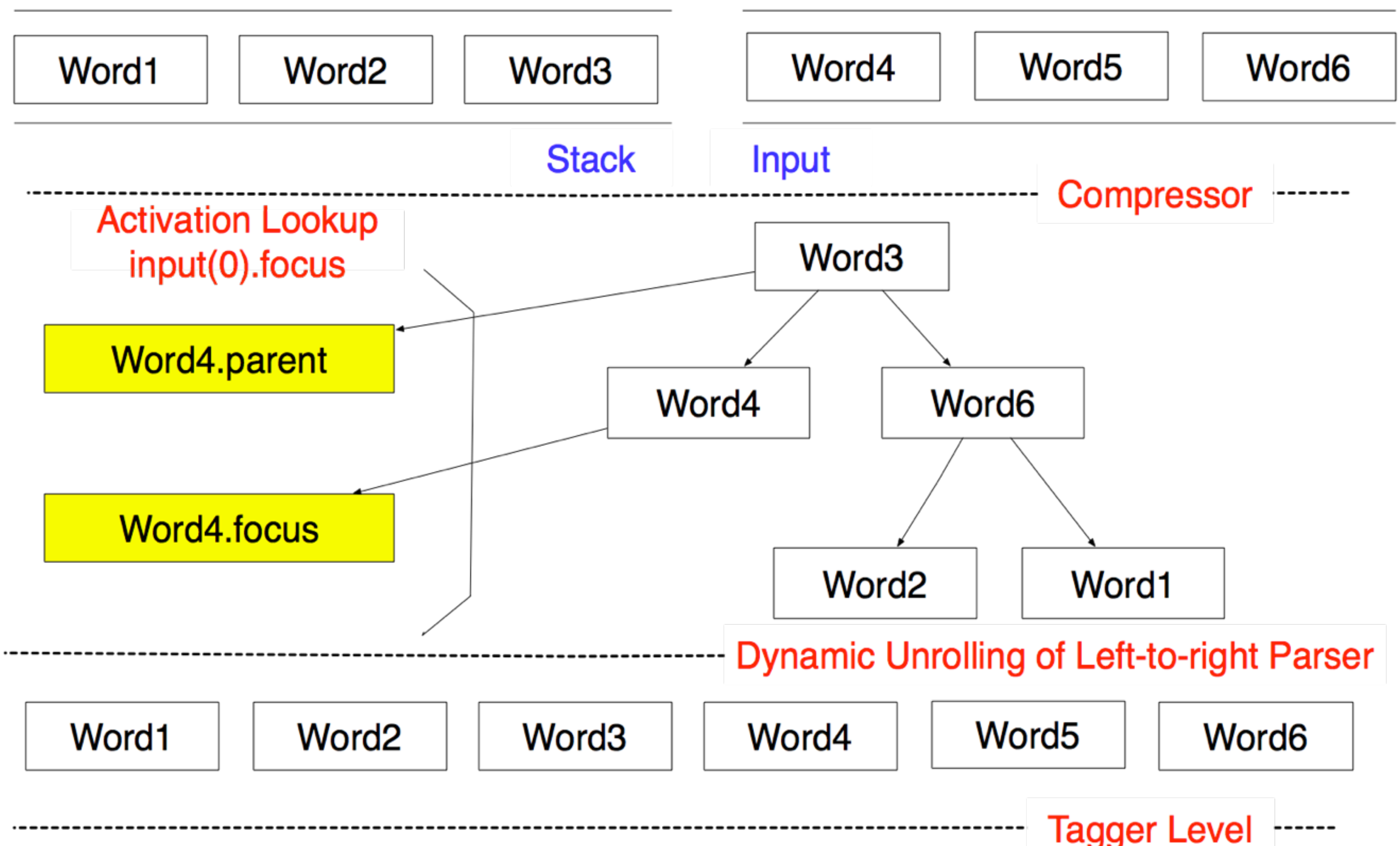
Bi-directional Parsing

	Union-News			Union-Web			Union-QTB		
Model	**UAS**	**LAS**	**POS**	**UAS**	**LAS**	**POS**	**UAS**	**LAS**	**POS**
Andor et al. (2016)	94.44	92.93	97.77	90.17	87.54	94.80	95.40	93.64	96.86
Left-to-right Parsing	94.60	93.17	97.88	90.09	87.50	94.75	95.62	94.06	96.76
Bi-directional Parsing	**94.66**	**93.23**	**98.09**	**90.22**	**87.67**	**95.06**	**96.05**	**94.51**	**97.25**

Compressor Pipeline



Compressor Pipeline



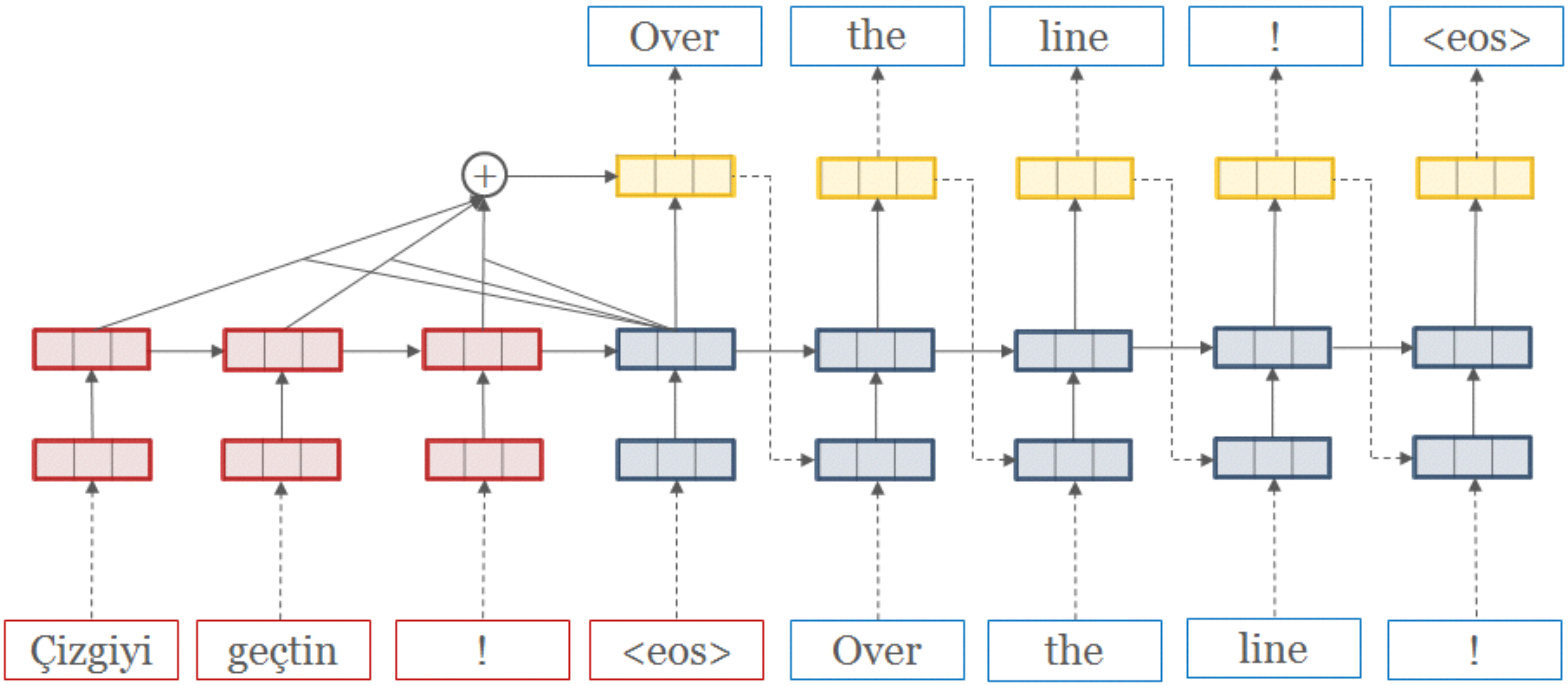
Compressor Pipeline

Model		A(%)	F1(%)	LAS(%)
Single LSTM	Right-to-left → Summarize	28.93	79.75	–
Bi-LSTM	Right-to-left, Left-to-right → Summarize	29.51	80.03	–
Multi-task LSTM (Luong et al., 2015)	Right-to-left → Parse, Summarize	30.07	80.31	**89.42**
Parse sub-trees	Right-to-left → Parse → Summarize	**30.56**	**80.74**	89.13

Outline

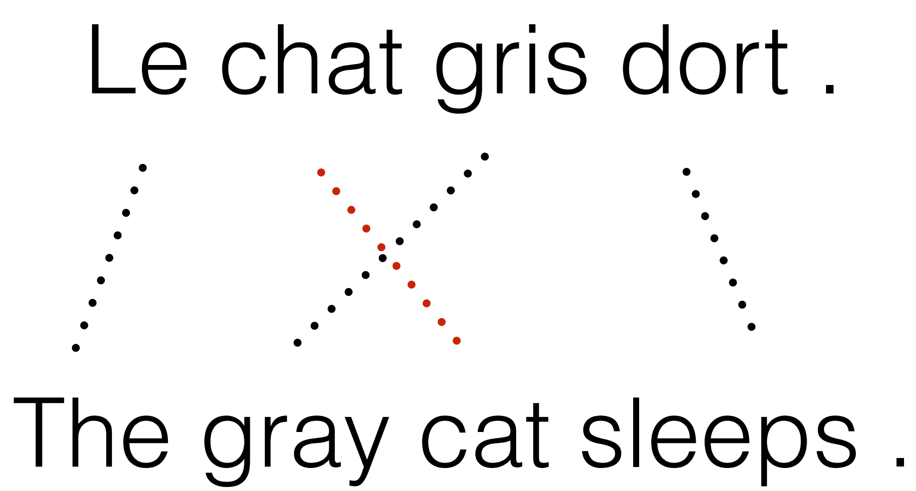
- Introduction
- Part I — Segmental Recurrent Neural Networks
- Part II — A Transition-based Framework for Dynamically Connected Neural Networks
- Part III — Inference and Regularization in Sequence to Sequence Models with Stochastic Attention
- Conclusion and Future Work

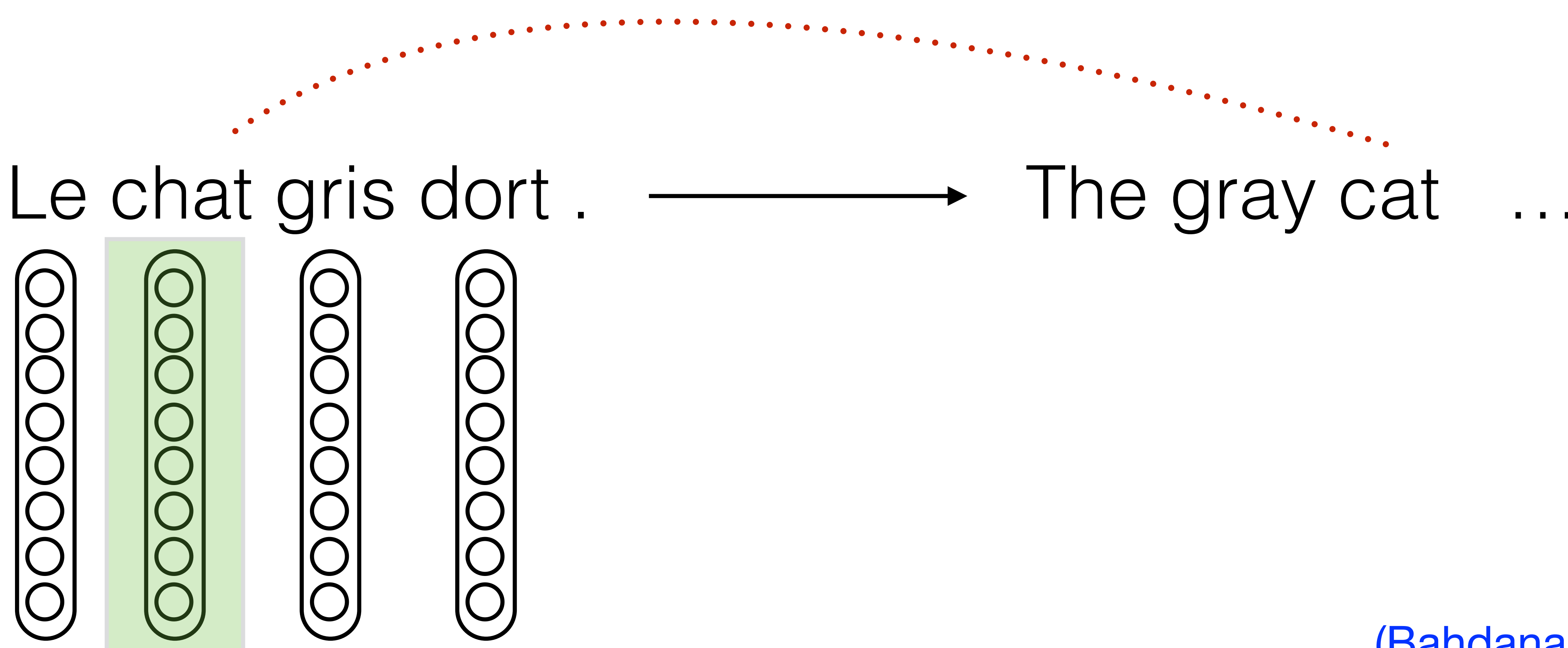
Neural Machine Translation



http://opennmt.net/

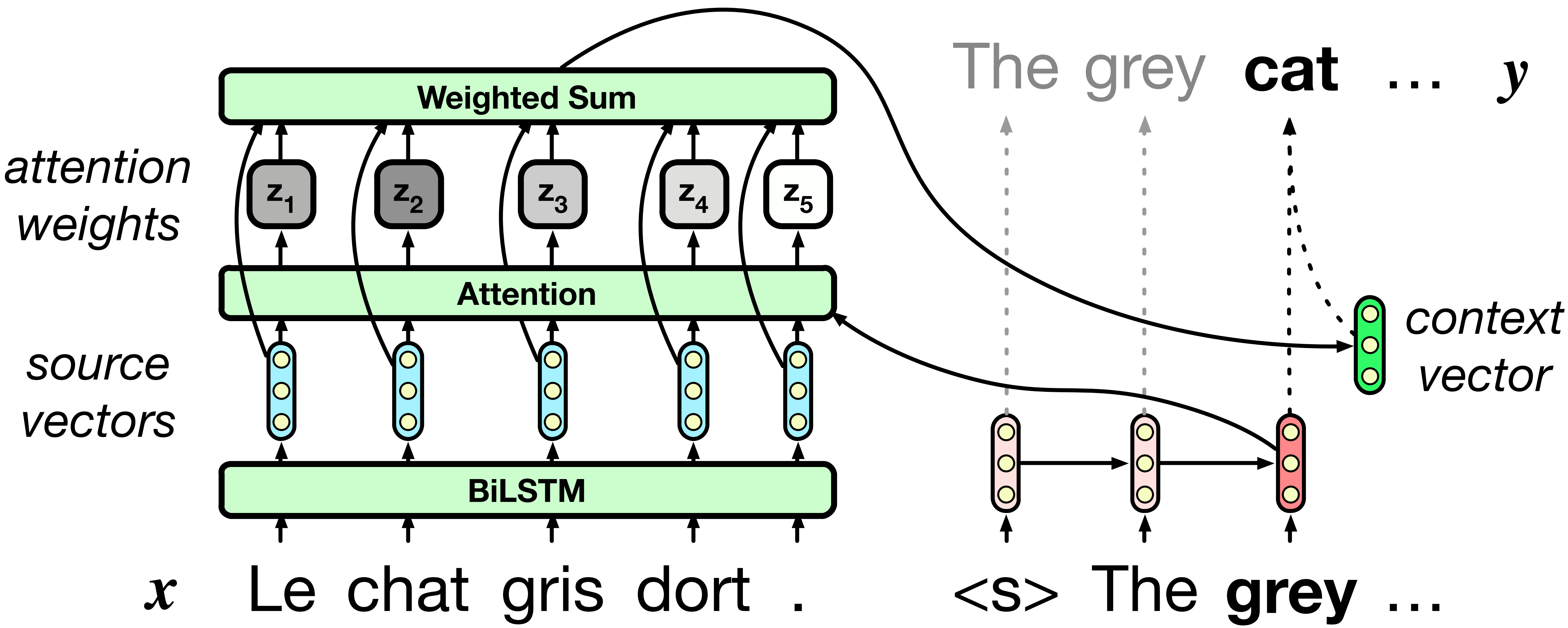
Attention Mechanism



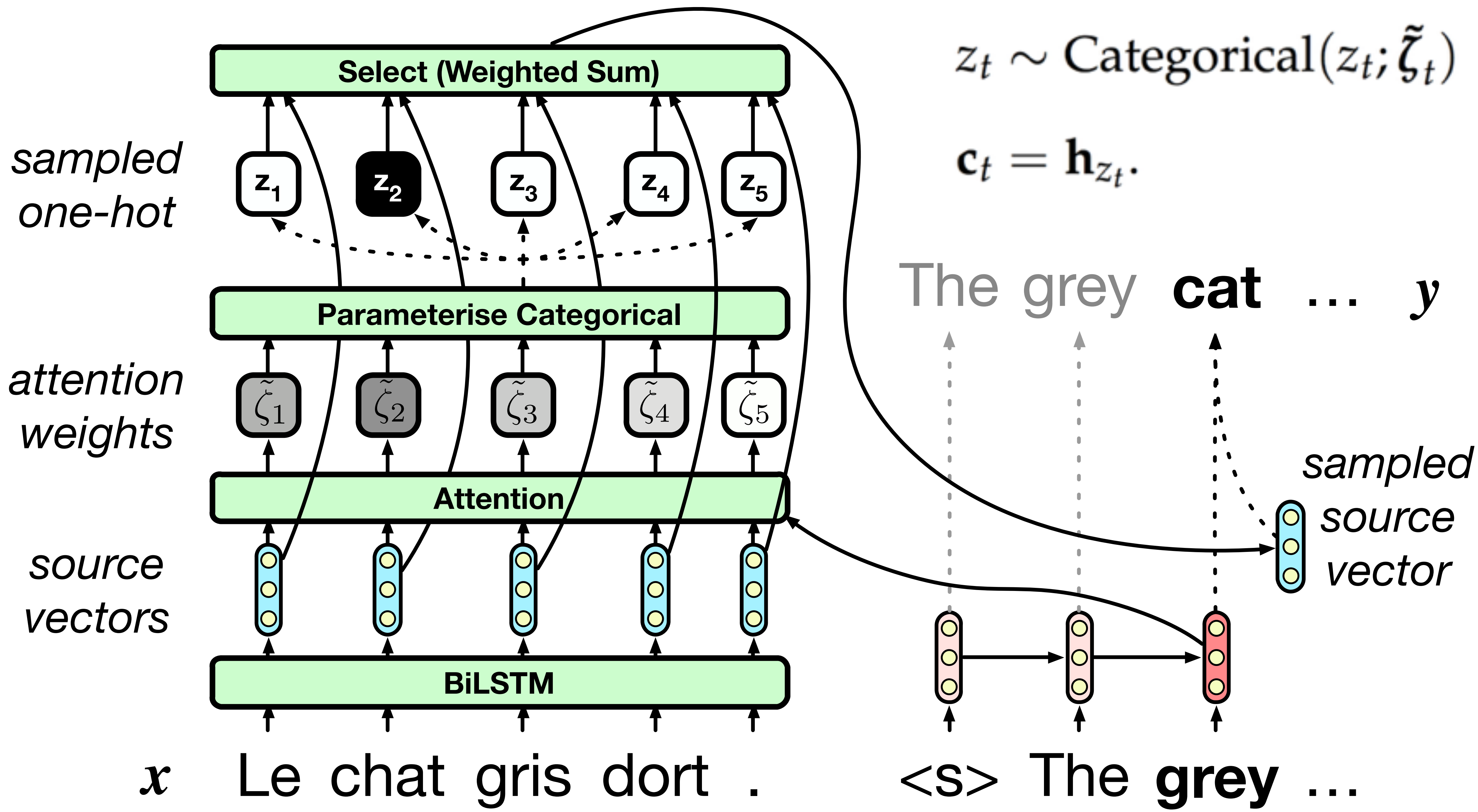


(Bahdanau et al., 2014)

Deterministic Attention



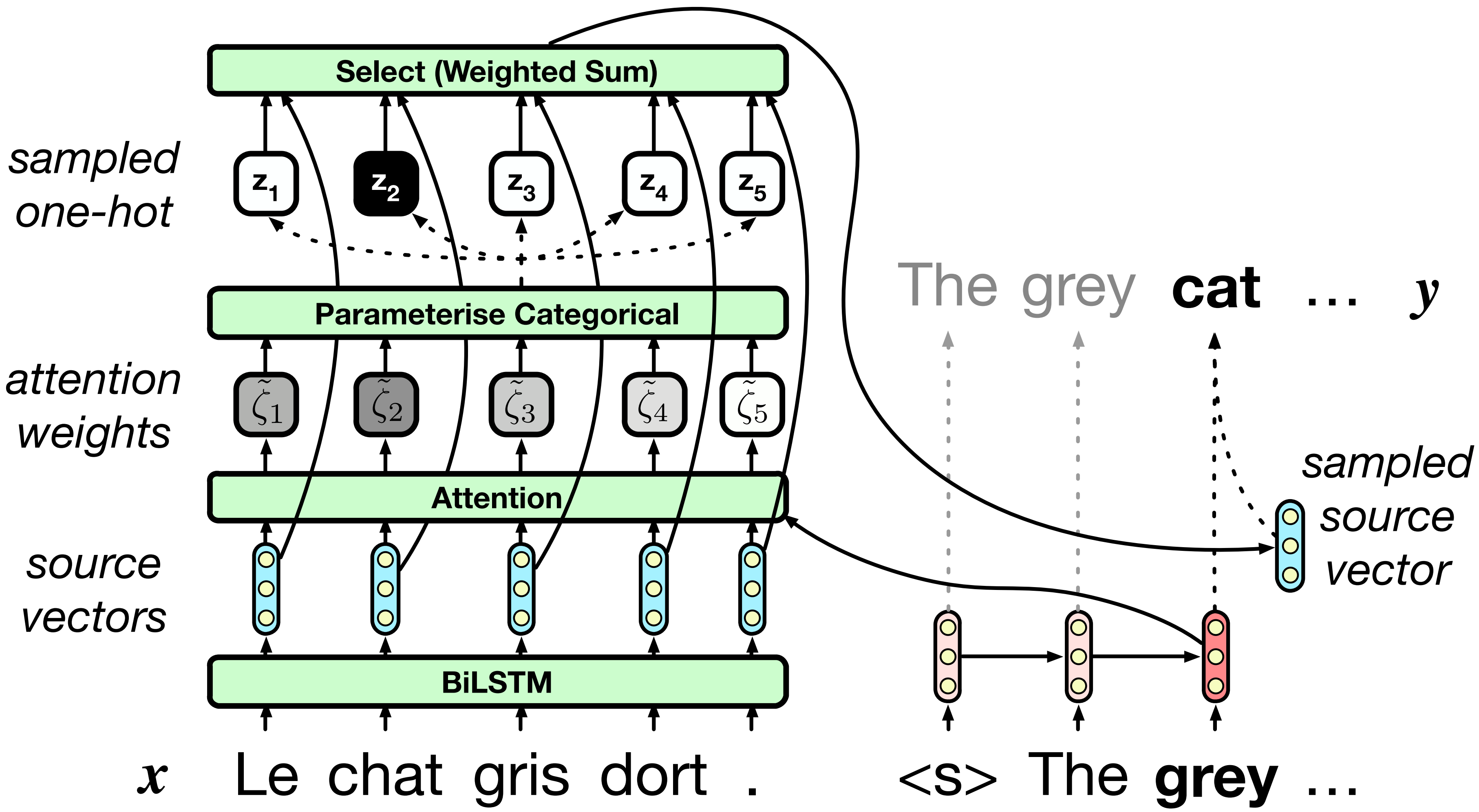
Stochastic Attention



$$z_t \sim \text{Categorical}(z_t; \tilde{\zeta}_t)$$

$$\mathbf{c}_t = \mathbf{h}_{z_t}.$$

Marginal Likelihood and Training Objective



$$
\begin{aligned}
p(\boldsymbol{y} \mid \boldsymbol{x}) &= \sum_{z} p(\boldsymbol{y}, z \mid \boldsymbol{x}) \\
&= \prod_{t=1}^{M} \sum_{z_t=1}^{N} p(z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) p(y_t \mid z_t, \boldsymbol{x}, \boldsymbol{y}_{<t})
\end{aligned}
$$

Approximating the Marginal Likelihood

Variational lower bound:

$$
\begin{aligned}
\log p(\boldsymbol{y} \mid \boldsymbol{x}) &= \log \sum_{z} p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x}) \\
&= \sum_{t=1}^{M} \log \sum_{z_t=1}^{N} p(z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) p(y_t \mid z_t, \boldsymbol{x}, \boldsymbol{y}_{<t}) \\
&= \sum_{t=1}^{M} \log \sum_{z_t=1}^{N} p(y_t, z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) \\
&\geq \sum_{t=1}^{M} \mathbb{E}_{q} \log \frac{p(y_t, z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t})}{q(z_t)} \\
&= \sum_{t=1}^{M} \mathbb{E}_{q} \log p(y_t, z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) + H(q)
\end{aligned}
$$

Approximating the Marginal Likelihood

REINFORCE:

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{t=1}^{M} \sum_{z_t=1}^{N} p(z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) p(y_t \mid z_t, \boldsymbol{x}, \boldsymbol{y}_{<t})$$

$$= \prod_{t=1}^{M} p(\tilde{z}_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) p(y_t \mid \tilde{z}_t, \boldsymbol{x}, \boldsymbol{y}_{<t})$$

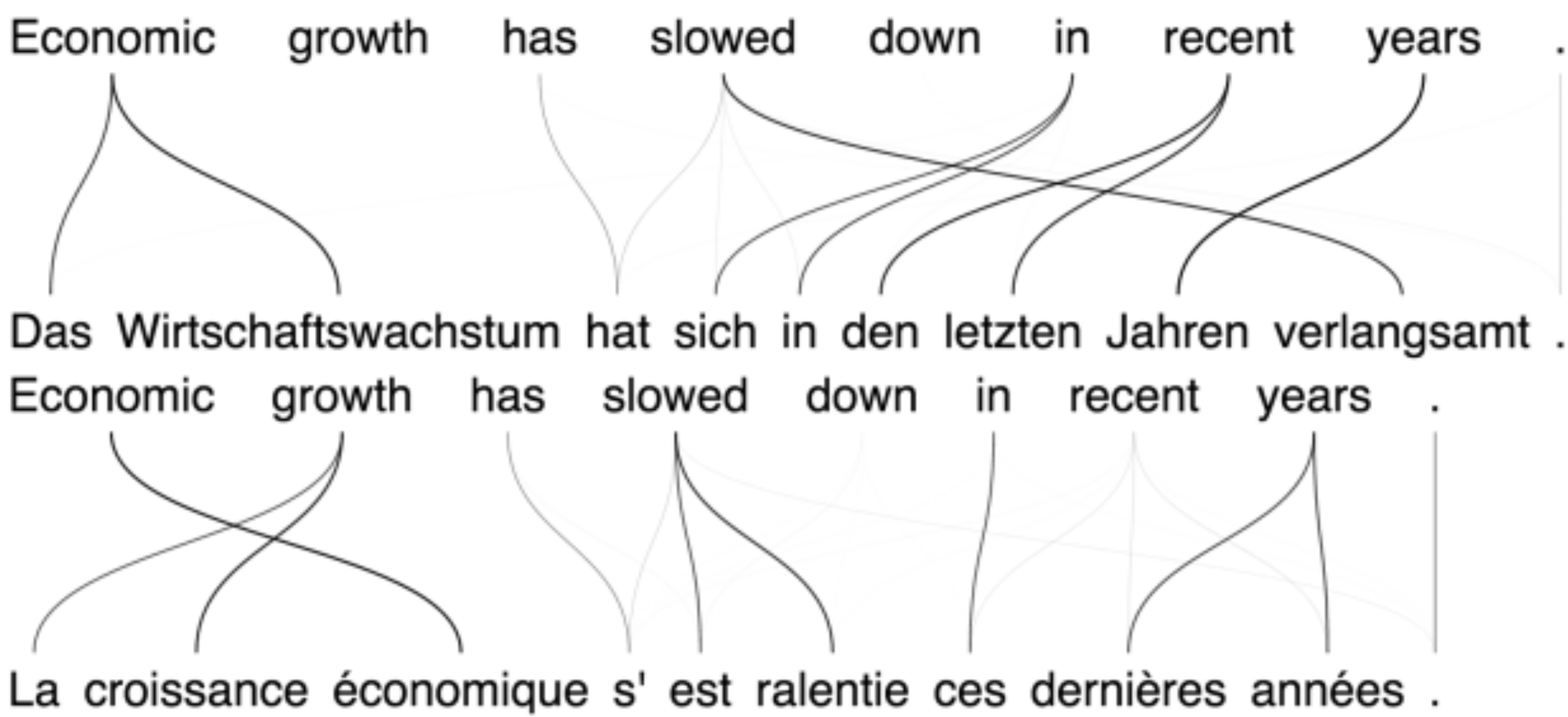
One-sample approximation

Experiments: Deterministic vs. Stochastic Attention

Model	Inference	BLEU	PPL
Deterministic	-	31.87	5.25
Stochastic	exact	**31.91**	**4.65**
Stochastic	variational	30.10	5.40
Stochastic	REINFORCE	29.85	5.31

- **Let's not give up yet!**
  - Neural nets can fit noisy data (Zhang, Bengio, Hardt, Recht, Vinyals, ICLR 2017).
  - (Stochastic) attention should be sensible, not just a random fit
  - Let's **regularize** the posterior distributions so they look more like what we expect posteriors to be (Ganchev, Graça, Gillenwater, Taskar, JMLR 2010)
- **Strategy:**
  - Apply KL penalty (true PR penalty)
  - Use variants of IS (biased estimator) using the expected posterior as the instrumental distribution

IBM Models



Great q distribution!

Posterior Regularization

Exact:

$$\mathcal{L} = -\log \sum_{\boldsymbol{z}} p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x}) + \gamma \times D_{KL}(p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \mid\mid \tilde{q}(\boldsymbol{z}))$$

$$p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x})}{\sum_{\boldsymbol{z}} p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x})}$$

Posterior Regularization

Importance Sampling:

$$\begin{aligned}
\log p(\boldsymbol{y} \mid \boldsymbol{x}) &= \log \sum_{z} p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x}) \\
&= \sum_{t=1}^{M} \log \sum_{z_t=1}^{N} p(z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) p(y_t \mid z_t, \boldsymbol{x}, \boldsymbol{y}_{<t}) \\
&= \sum_{t=1}^{M} \log \sum_{z_t=1}^{N} p(y_t, z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) \\
&= \sum_{t=1}^{M} \log \sum_{z_t=1}^{N} \tilde{q}(z_t) w(z_t, \boldsymbol{x}, \boldsymbol{y}) \\
&= \sum_{t=1}^{M} \mathbb{E}_{\tilde{q}} w(z_t, \boldsymbol{x}, \boldsymbol{y}),
\end{aligned}$$

Monte Carlo approximation

$$w(z_t, \boldsymbol{x}, \boldsymbol{y}) = \frac{p(y_t, z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t})}{\tilde{q}(z_t)}$$

Posterior Regularization

Jensen IS:

$$\begin{aligned}
\log p(\boldsymbol{y} \mid \boldsymbol{x}) &= \log \sum_{z} p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x}) \\
&= \sum_{t=1}^{M} \log \sum_{z_t=1}^{N} p(z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) p(y_t \mid z_t, \boldsymbol{x}, \boldsymbol{y}_{<t}) \\
&= \sum_{t=1}^{M} \log \sum_{z_t=1}^{N} p(y_t, z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) \\
&\geq \sum_{t=1}^{M} \mathop{\mathbb{E}}_{\tilde{q}} \log \frac{p(y_t, z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t})}{\tilde{q}(z_t)} \\
&= \sum_{t=1}^{M} \mathop{\mathbb{E}}_{\tilde{q}} \log p(y_t, z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}) - H(\tilde{q})
\end{aligned}$$

Importance Sampling

$$w(z_t, \boldsymbol{x}, \boldsymbol{y}) = \frac{p(y_t, z_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t})}{\tilde{q}(z_t)}$$

sample from a fixed distribution

Mnih and Rezende (2016)

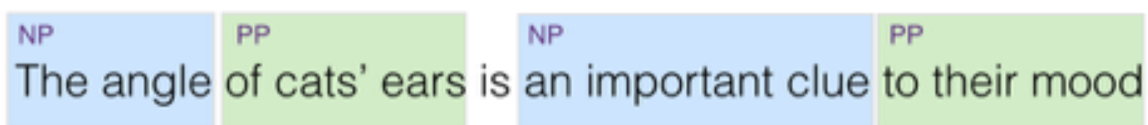
Experiments: Posterior regularization

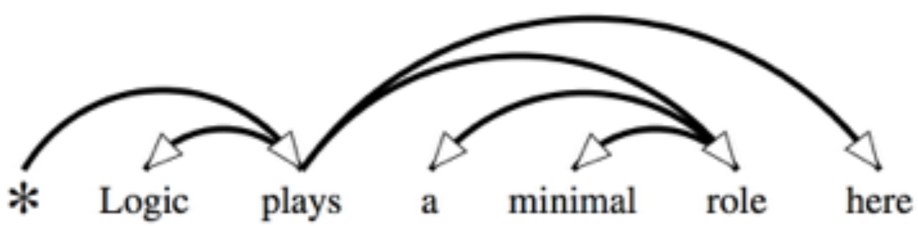
Model	Inference	PR	BLEU	PPL
Deterministic	exact	none	31.87	5.25
Stochastic	exact	none	31.91	4.65
Deterministic Chen et al. (2016)	exact	full	32.48	5.20
Stochastic	exact	full	35.17	4.03
Stochastic	IS with q	approximate	34.68	4.04
Stochastic	Jensen bound IS with q	approximate	**35.40**	**3.94**

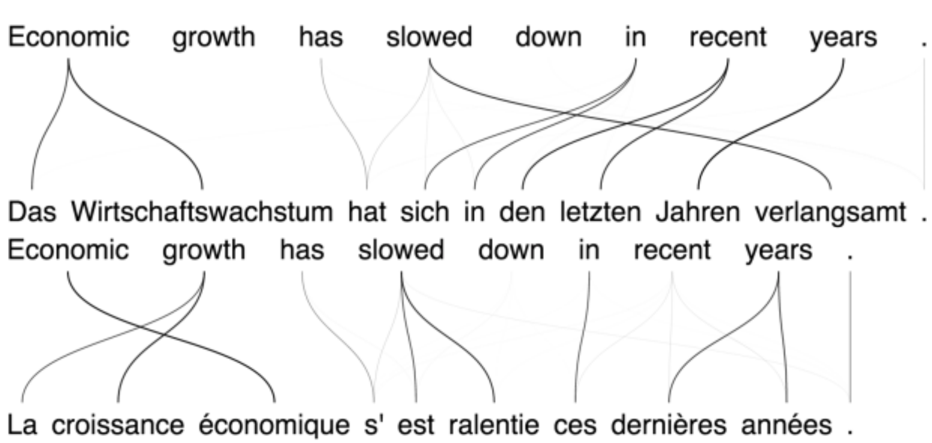
Outline

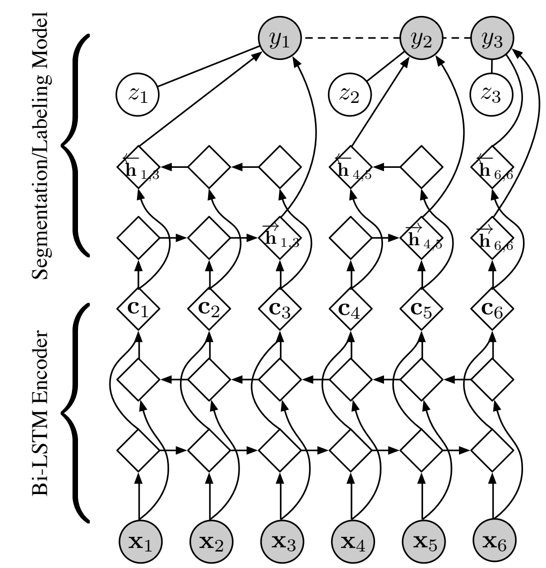
- Introduction
- Part I — Segmental Recurrent Neural Networks
- Part II — A Transition-based Framework for Dynamically Connected Neural Networks
- Part III — Inference and Regularization in Sequence to Sequence Models with Stochastic Attention
- Conclusion and Future Work

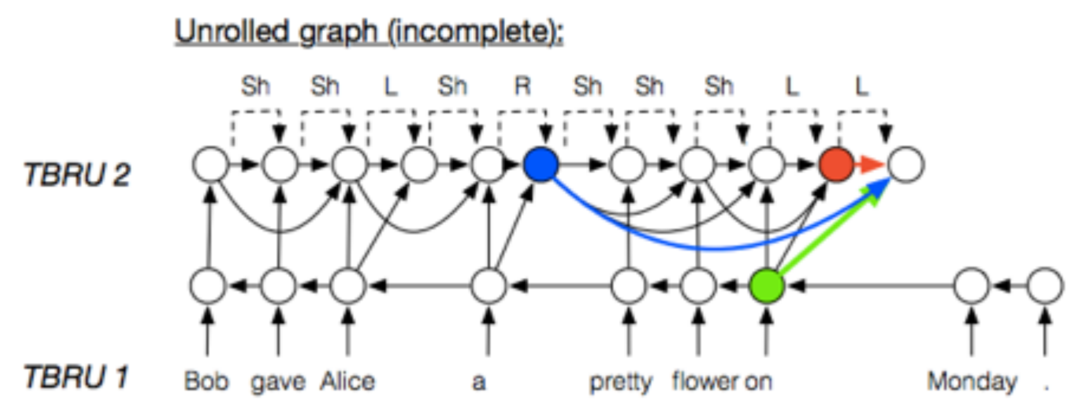
Conclusion and Future Work

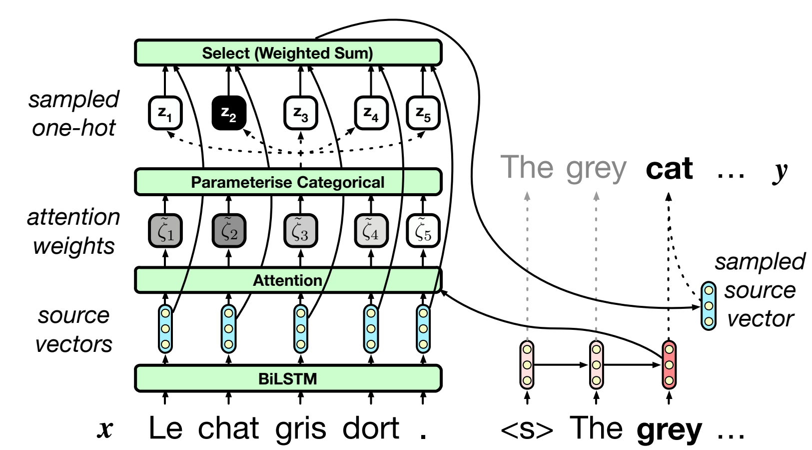












Conclusion and Future Work

- Segment Structures in Neural Machine Translation
- Automatic Linguistic Structure Discovery
- Hard Constraints in Attention Mechanism

Thank you!