Globally Normalized Transition-Based Neural Networks

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn,
Alessandro Presta, Kuzman Ganchev, Slav Petrov, Michael Collins

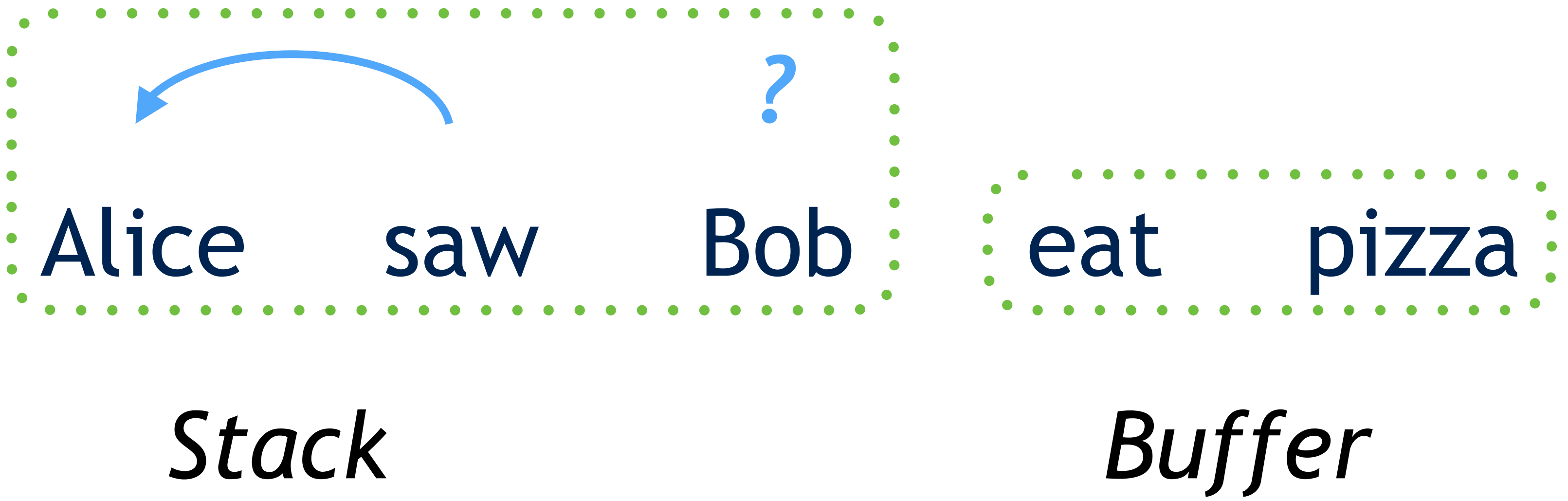
Research
at Google

Parsey McParseface Now Has 40 Multi-lingual Cousins!

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn,
Alessandro Presta, Kuzman Ganchev, Slav Petrov, Michael Collins

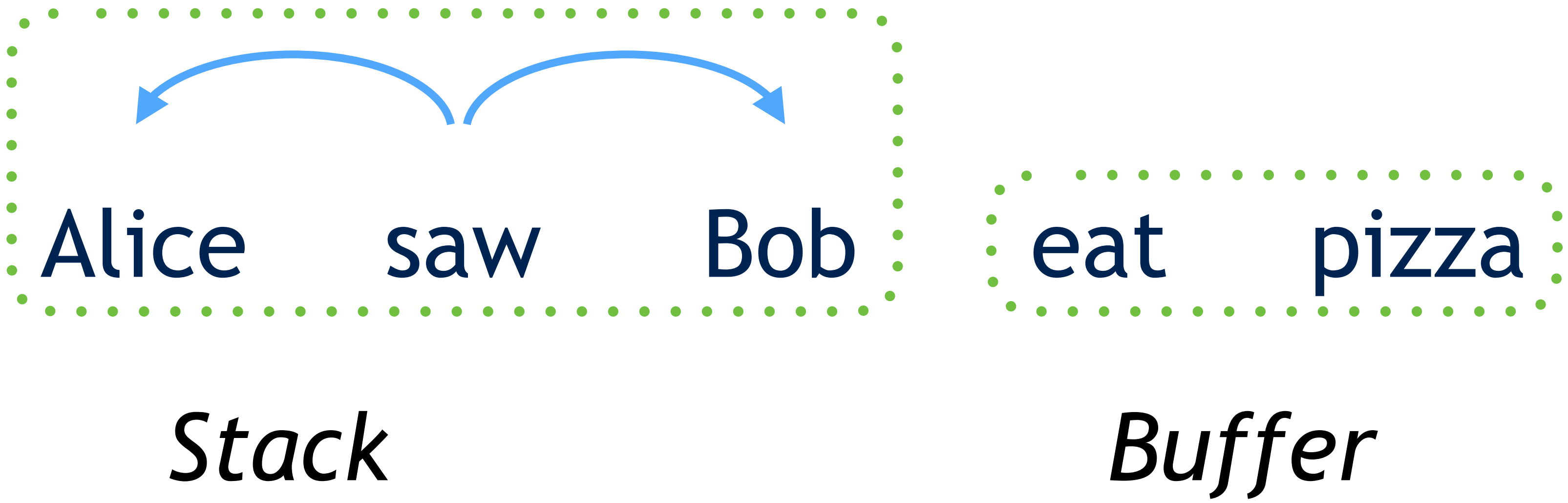
Research at Google

Transition-Based Parsing



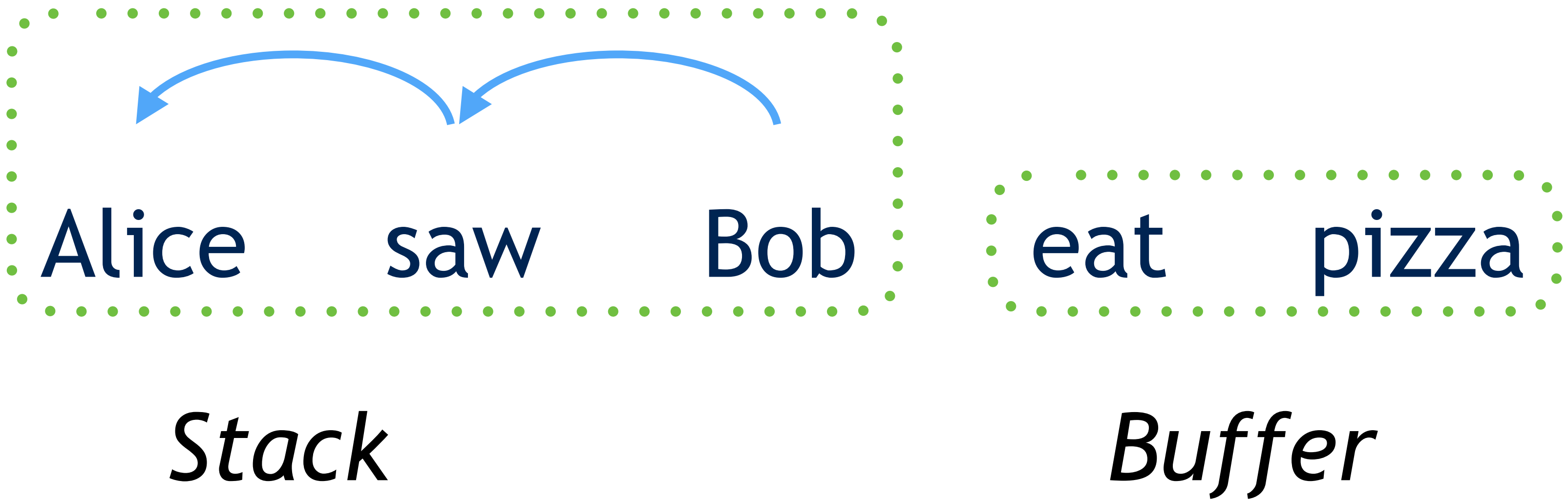
Transition-Based Parsing

RIGHT-ARC



Transition-Based Parsing

LEFT-ARC



Alice saw Bob

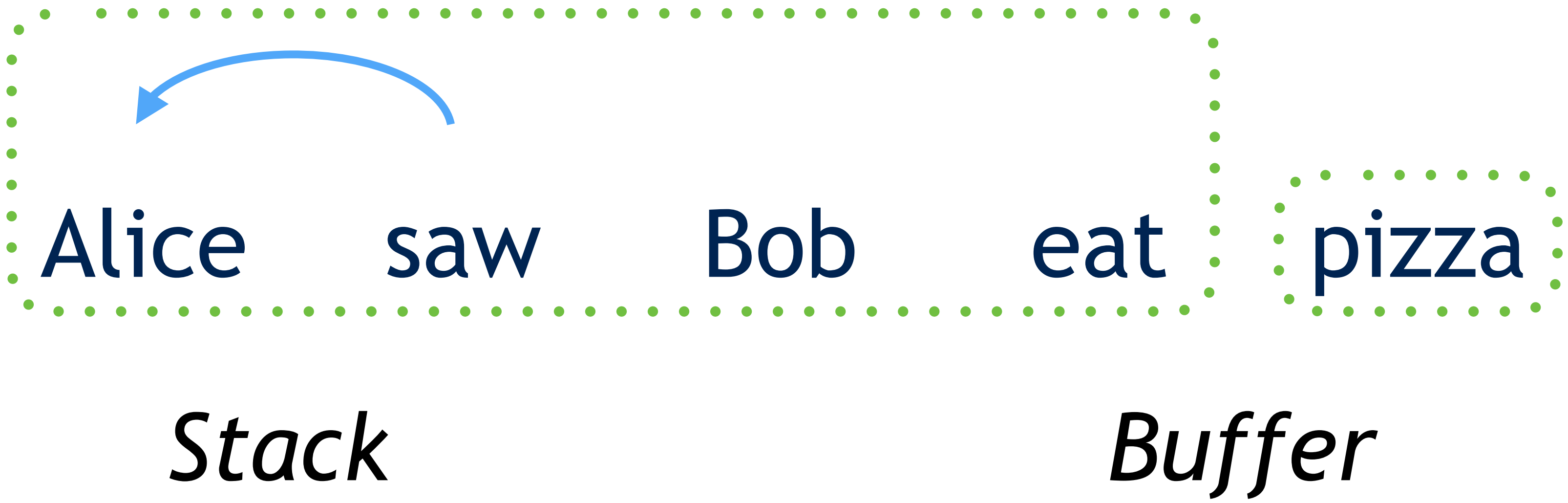
Stack

eat pizza

Buffer

Transition-Based Parsing

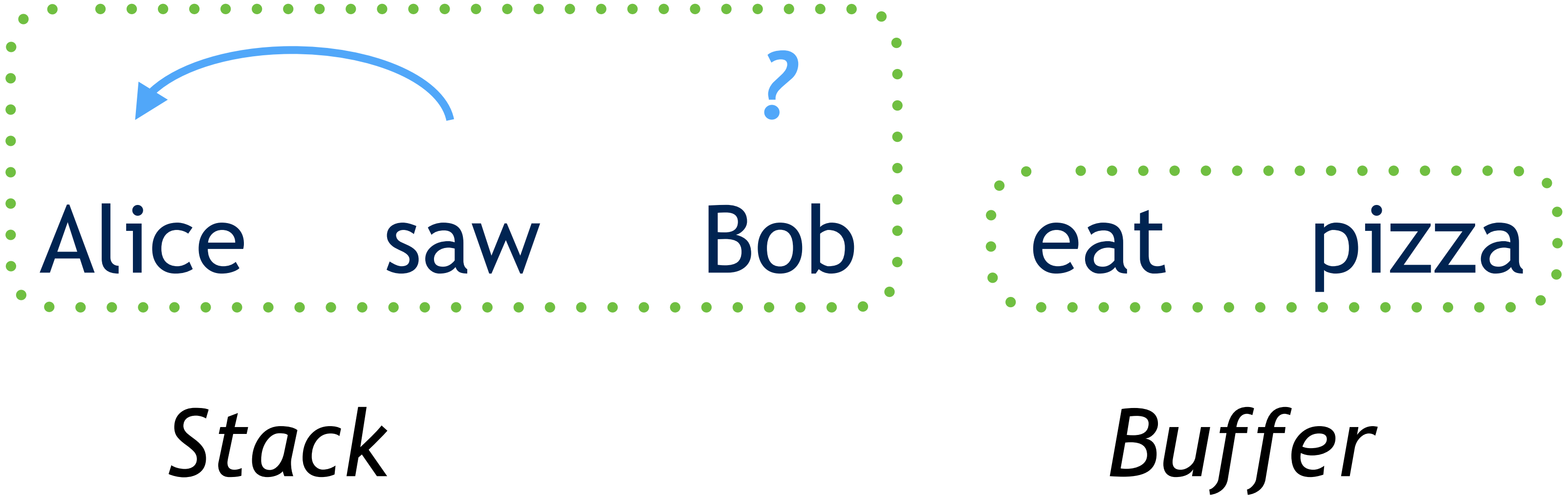
SHIFT



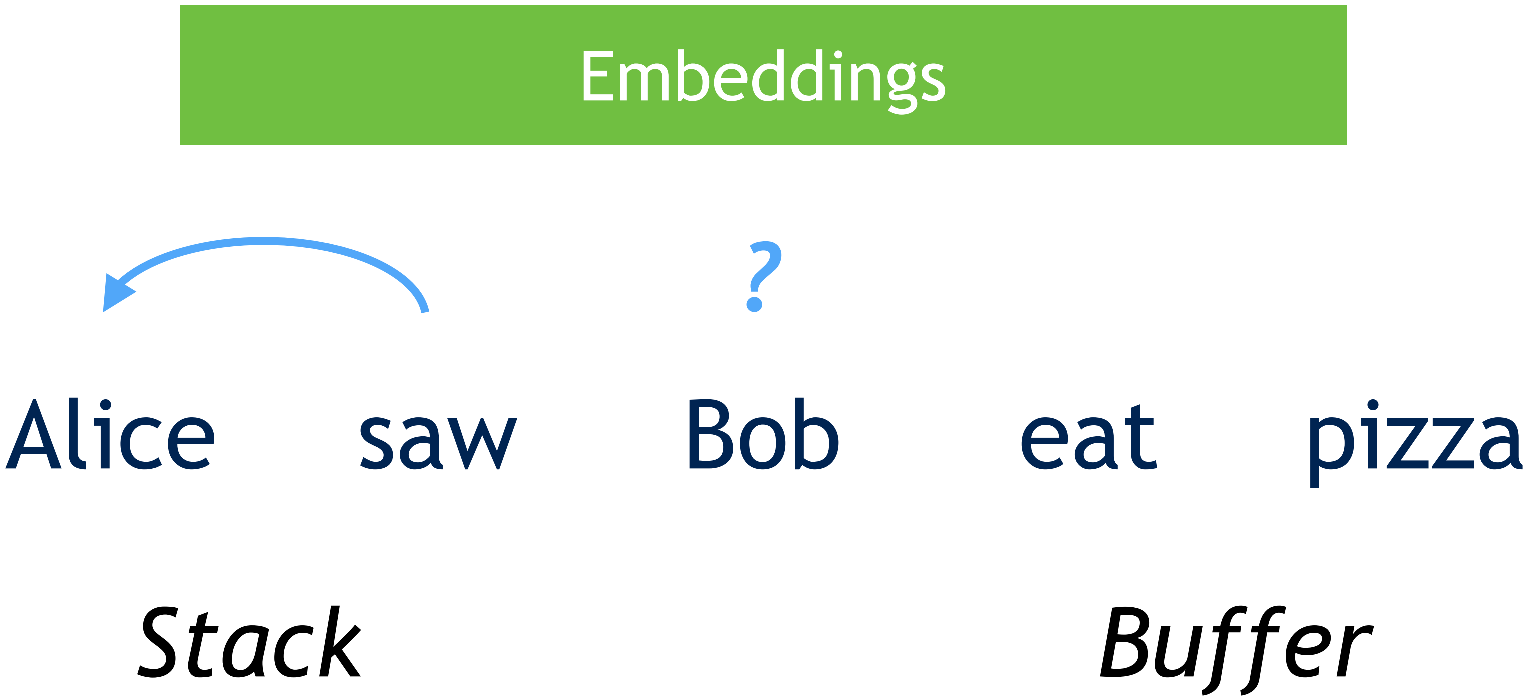
Alice saw Bob eat pizza

Stack *Buffer*

Transition-Based Neural Networks



Transition-Based Neural Networks



Embeddings

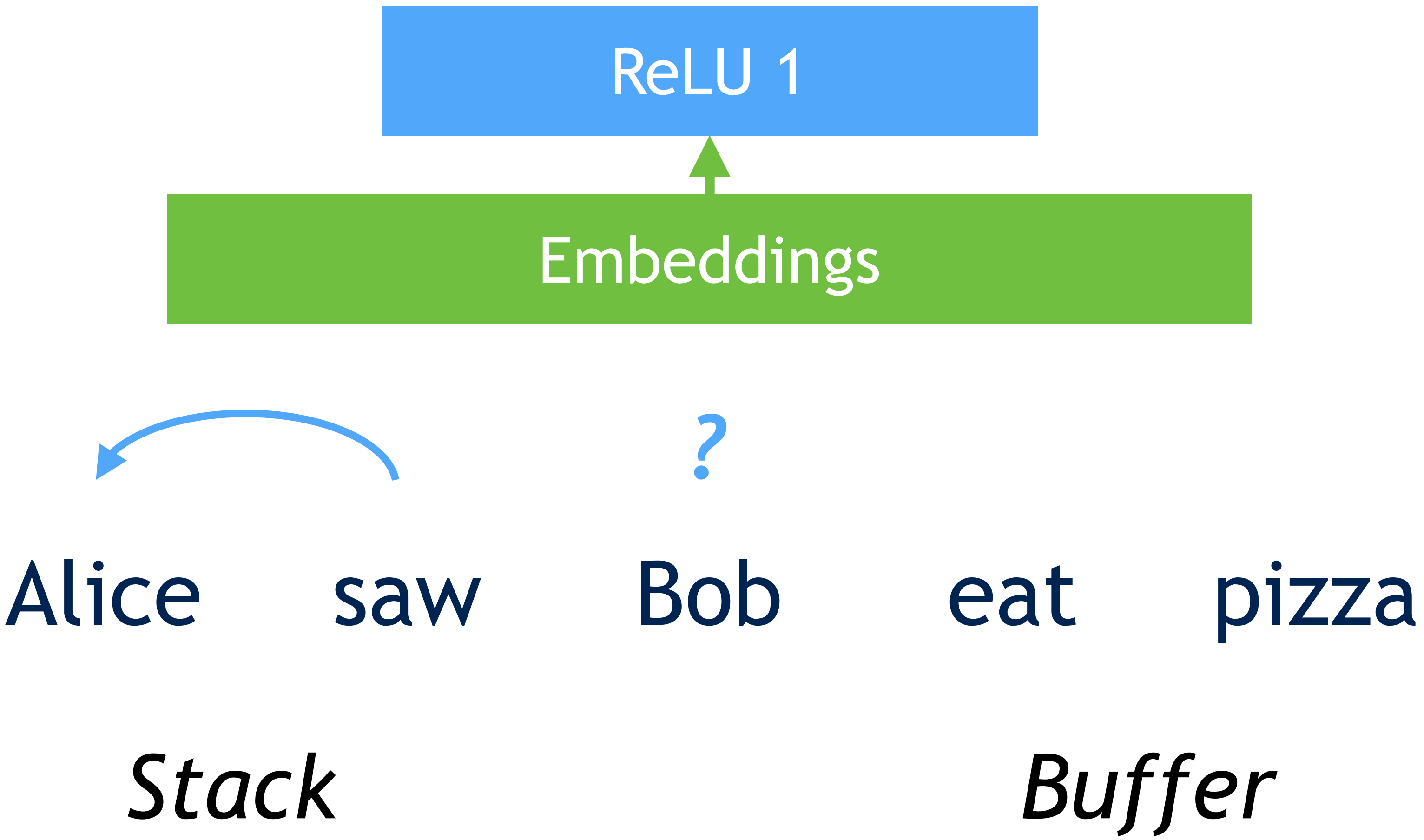
?

Alice saw Bob eat pizza

Stack

Buffer

Transition-Based Neural Networks



ReLU 1

Embeddings

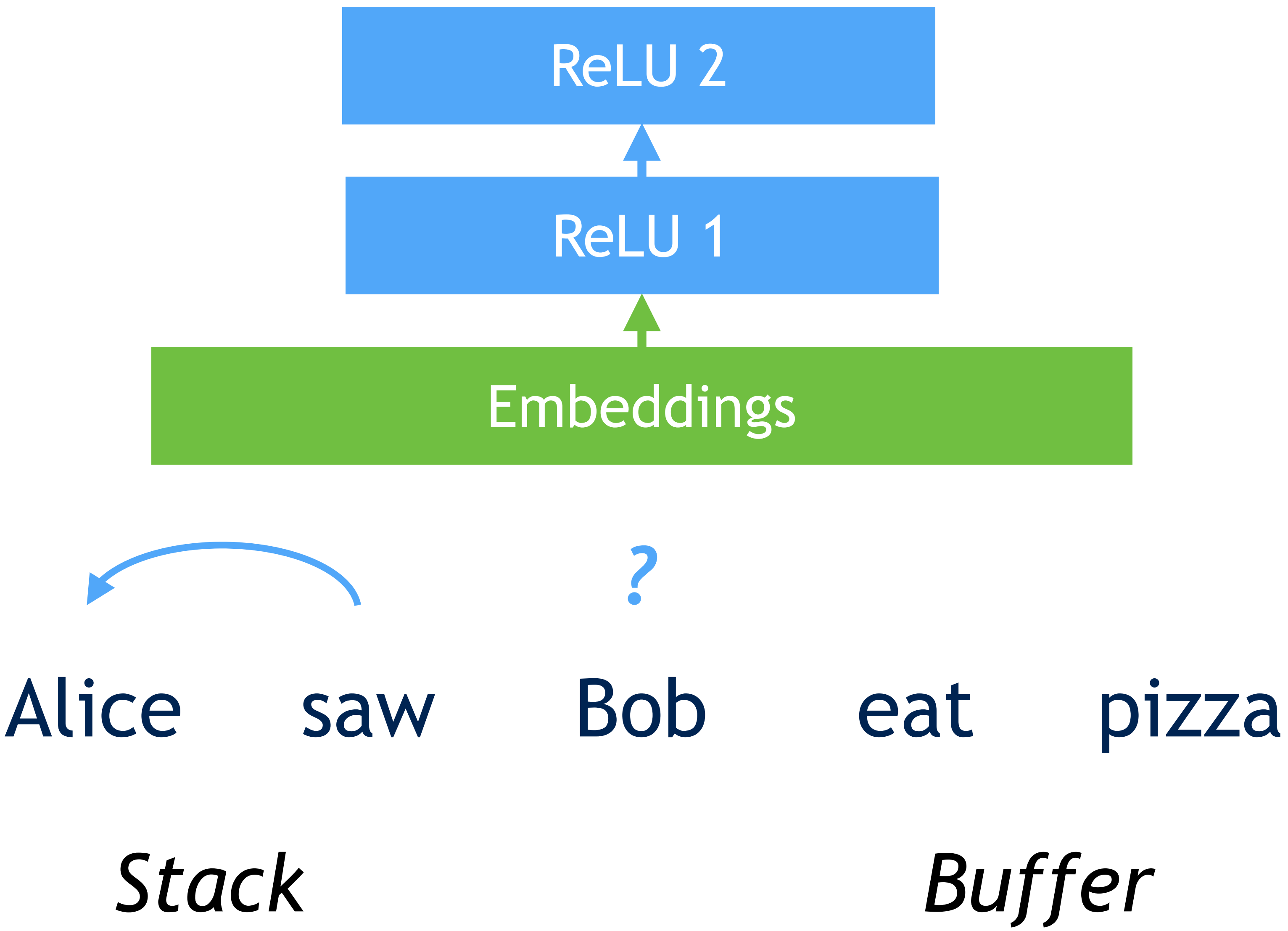
?

Alice saw Bob eat pizza

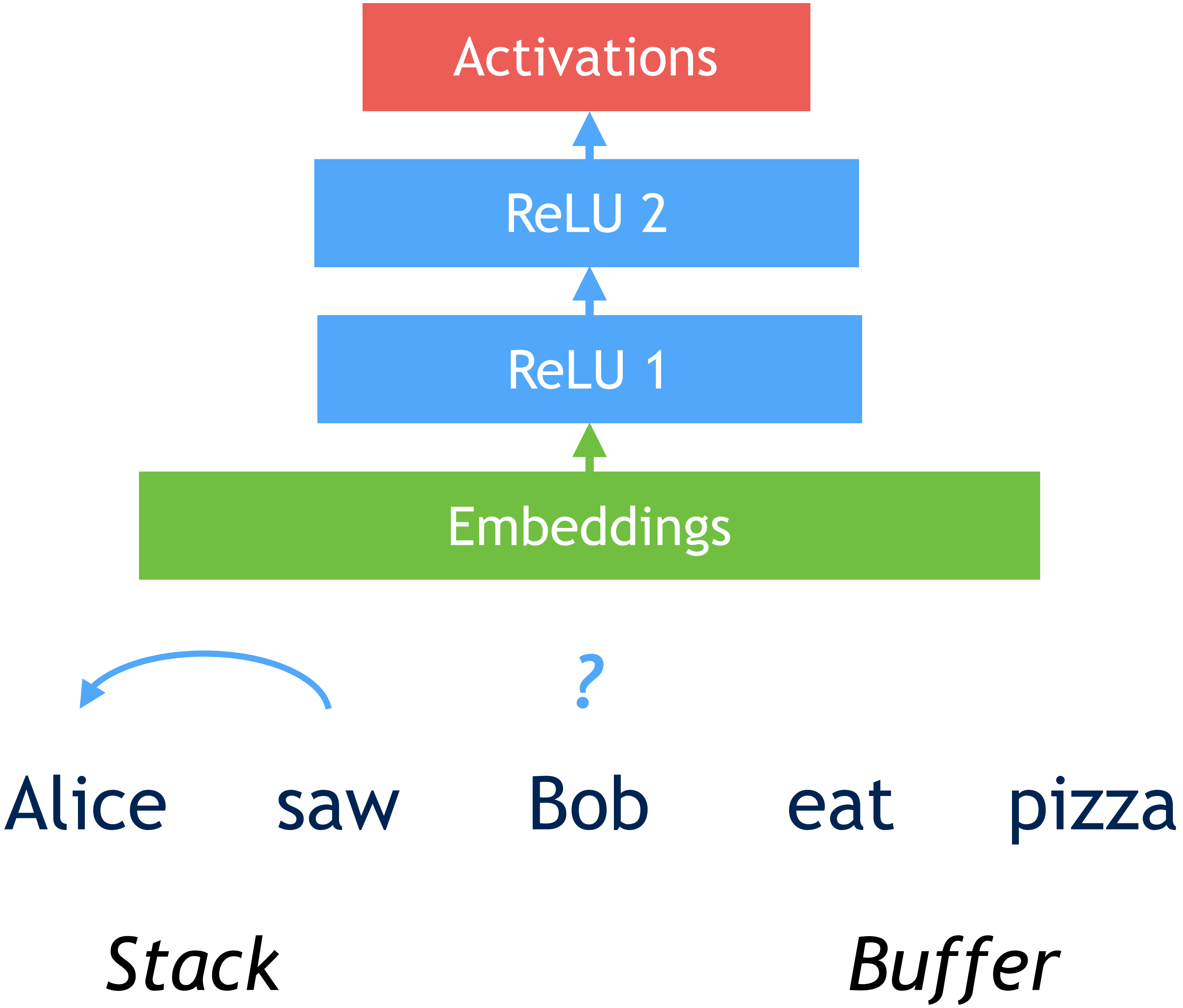
Stack

Buffer

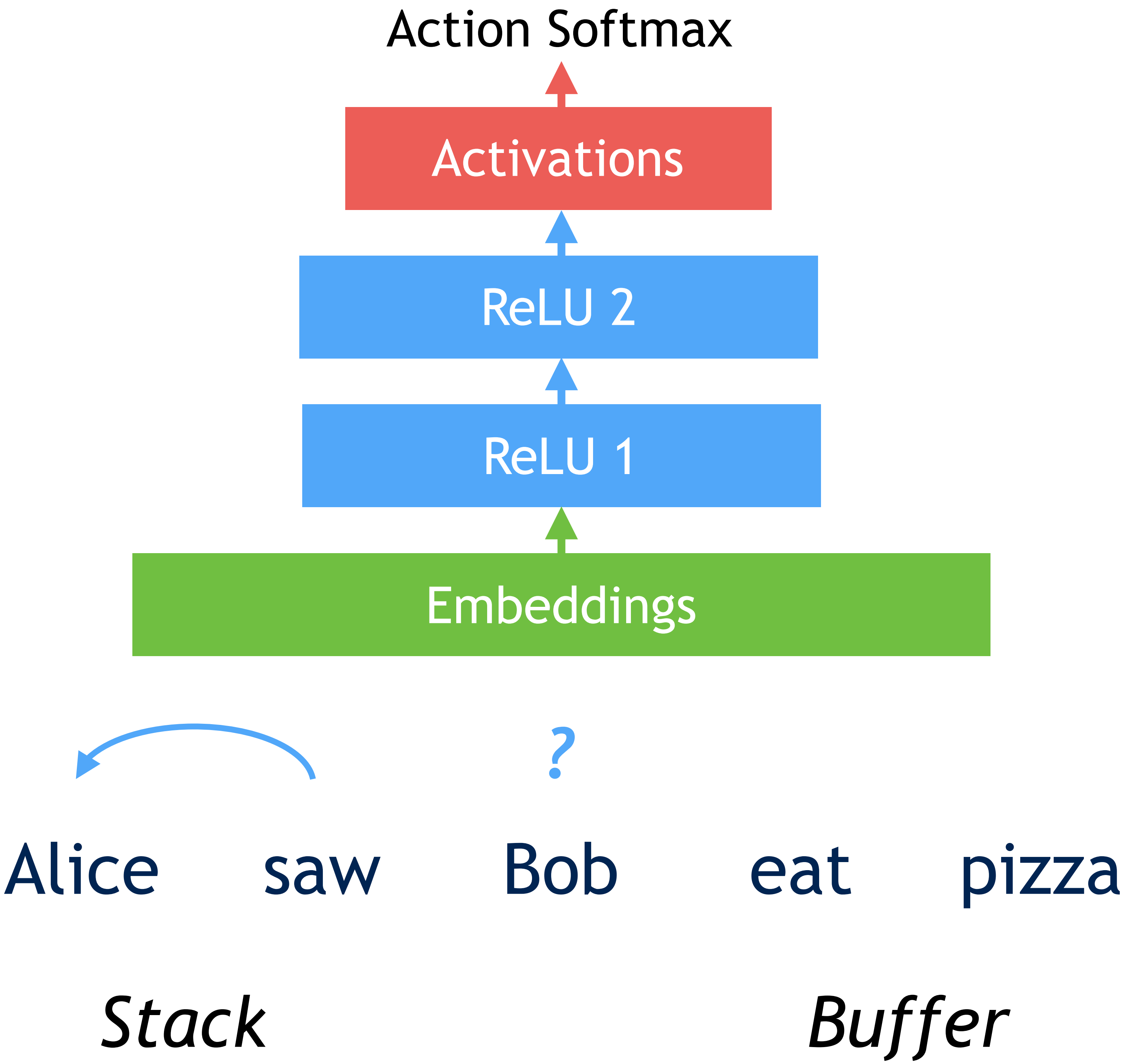
Transition-Based Neural Networks



Transition-Based Neural Networks



Transition-Based Neural Networks



Action Softmax

Activations

ReLU 2

ReLU 1

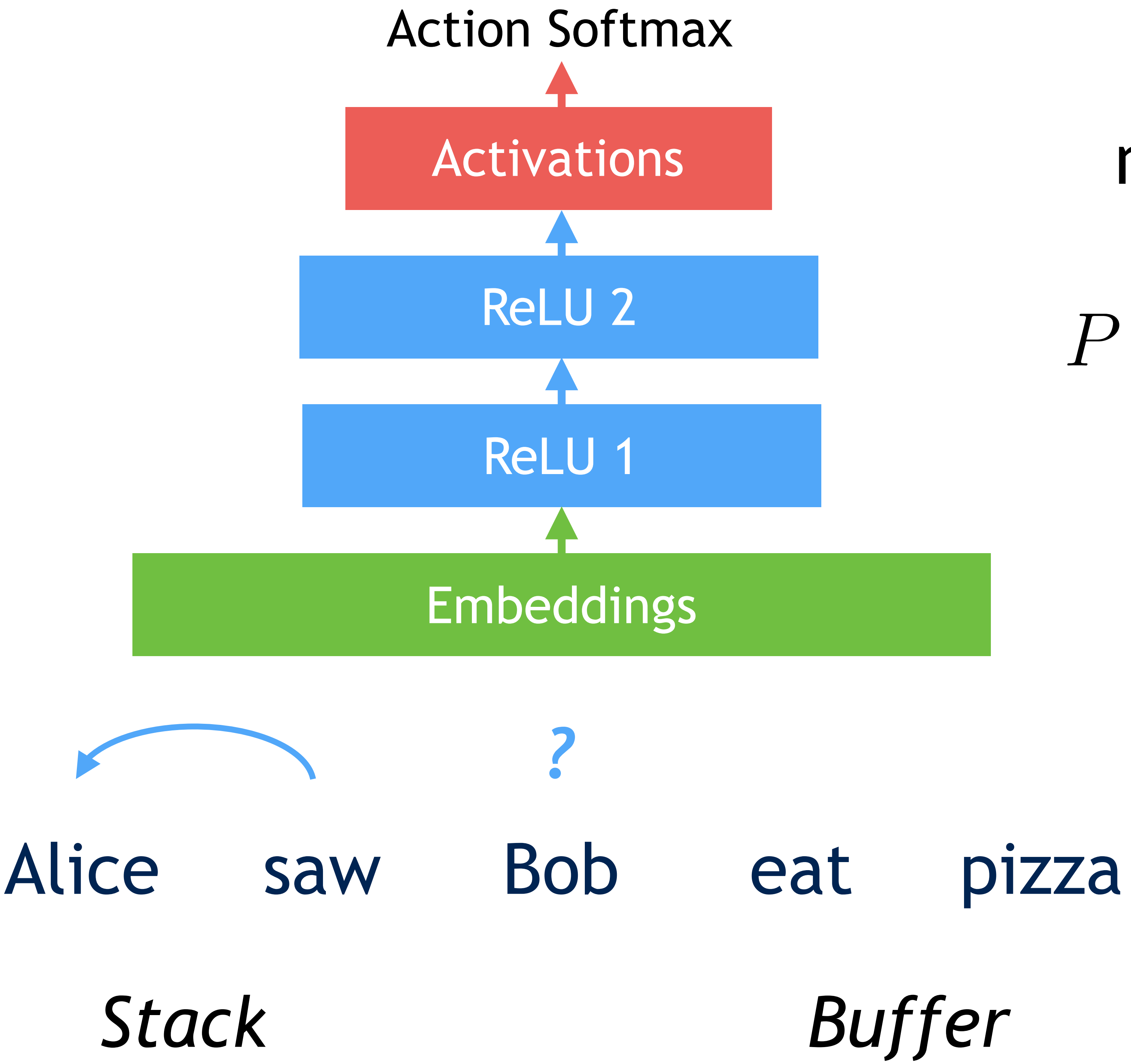
Embeddings

?

Alice saw Bob eat pizza

Stack *Buffer*

Transition-Based Neural Networks



Locally normalized model:

$$P\left(\text{action}|\text{context}\right)$$

Transition-Based Neural Networks

Action Softmax

- Locally normalized models are often easy to train
- Globally normalized models using the same #params can be much more accurate
- Applies to multiple tasks

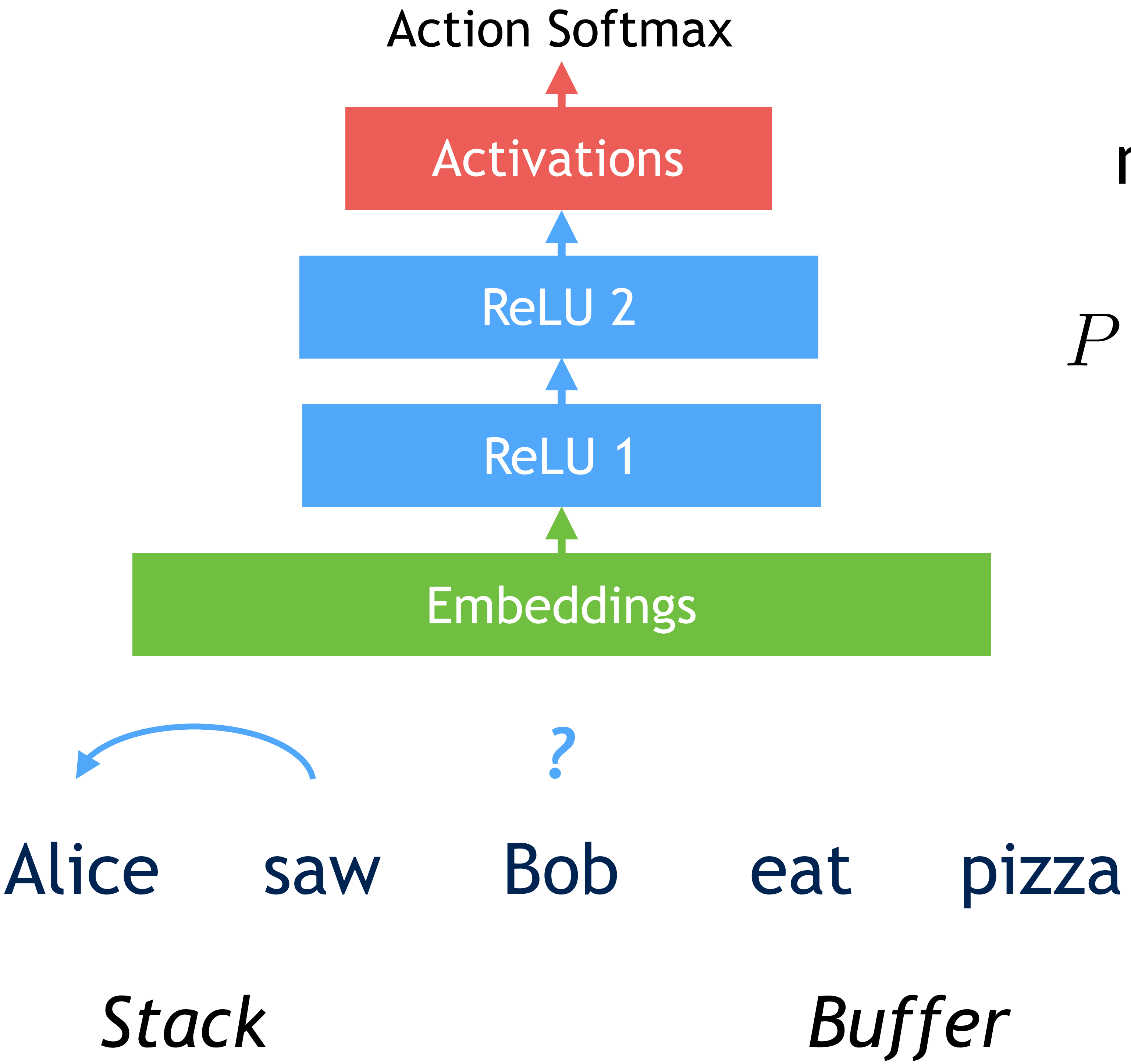
Locally
nalized
odel:
$\mathrm{tion}|\mathrm{context})$

Al

Stack

Buffer

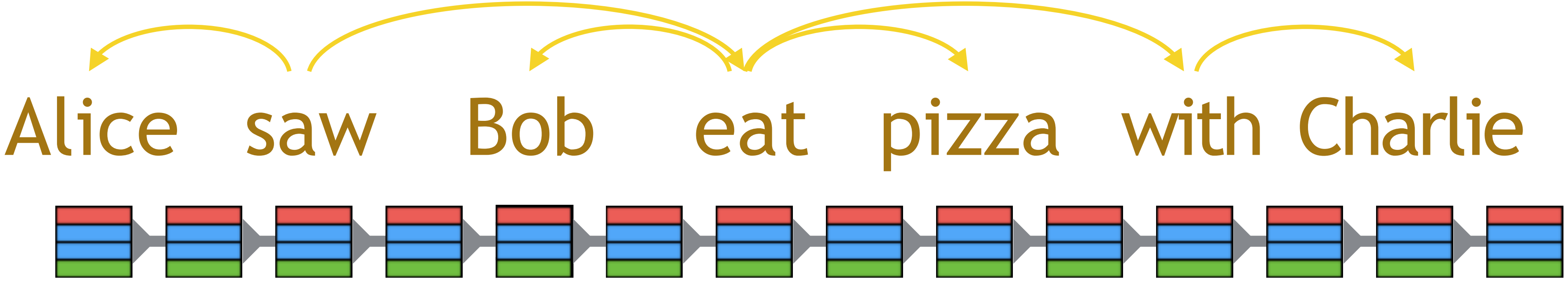
Transition-Based Neural Networks



Locally normalized model:

$P\left(\text{action}|\text{context}\right)$

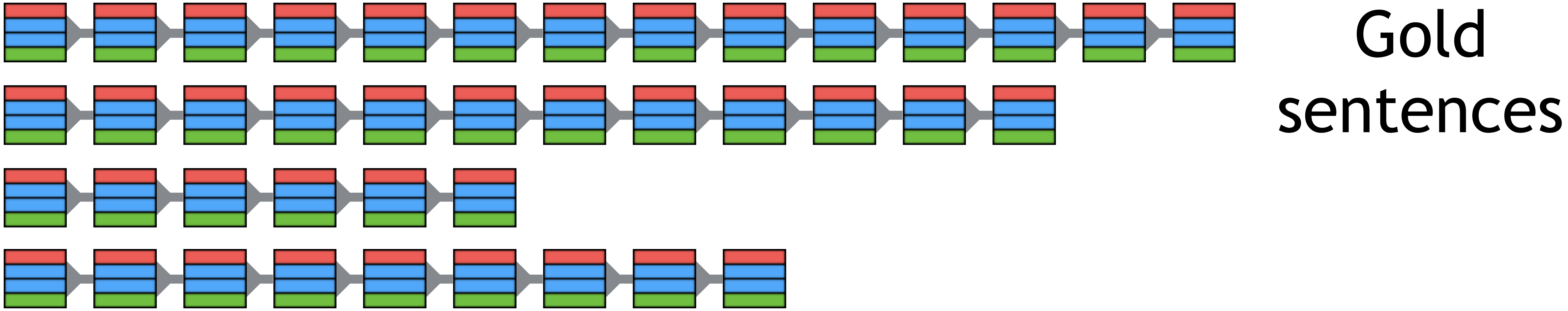
Locally Normalized Training



[Chen & Manning ’14, Weiss et al. ’15]

Locally Normalized Training

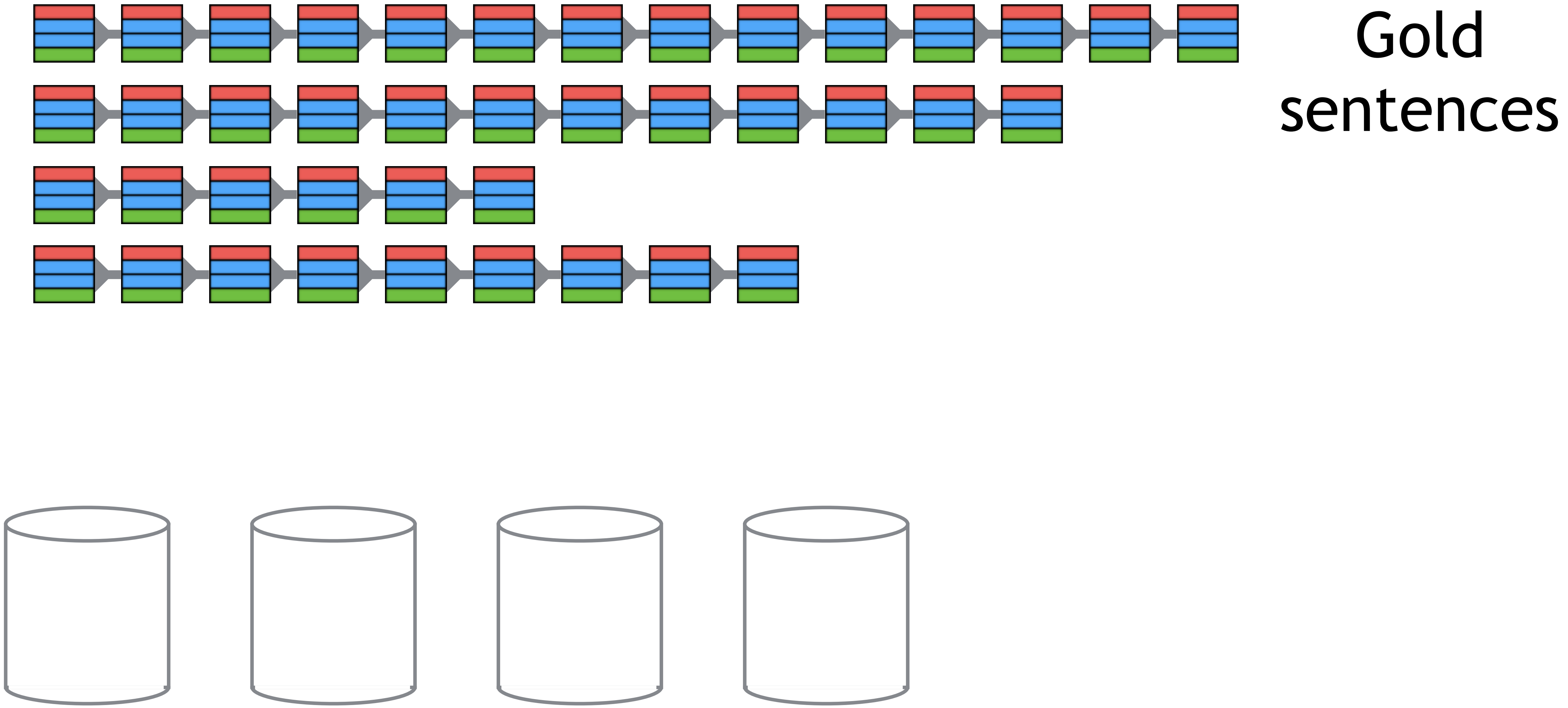
Oracle maps gold structures to gold action sequences:



[Chen & Manning ’14, Weiss et al. ’15]

Locally Normalized Training

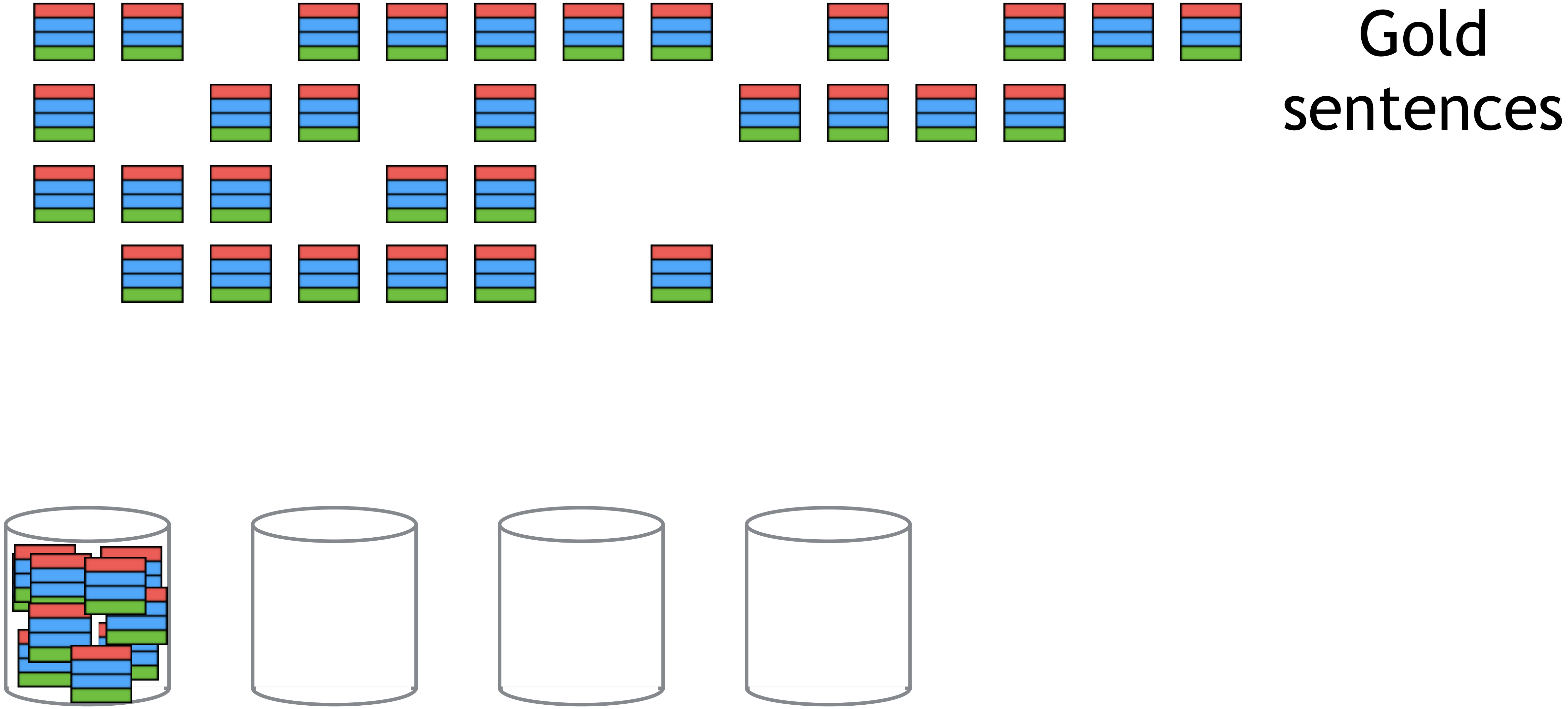
Oracle maps gold structures to gold action sequences:



[Chen & Manning ’14, Weiss et al. ’15]

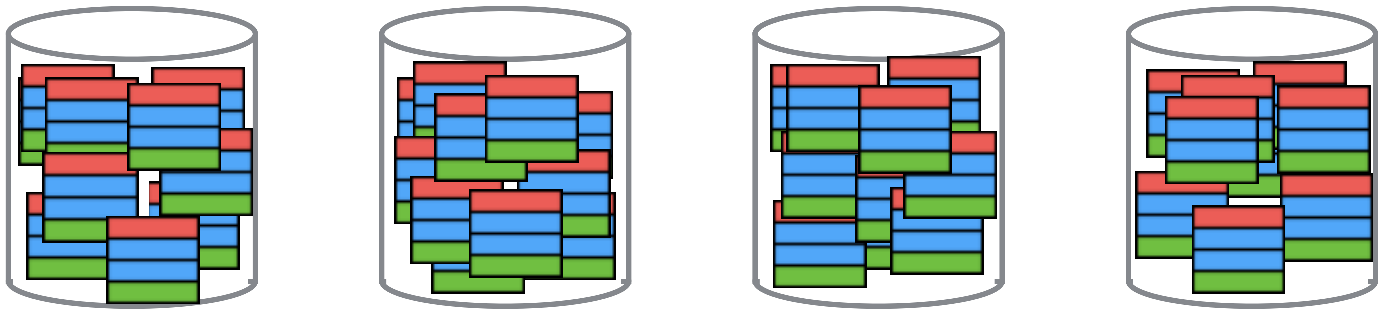
Locally Normalized Training

Oracle maps gold structures to gold action sequences:



[Chen & Manning ’14, Weiss et al. ’15]

Locally Normalized Training



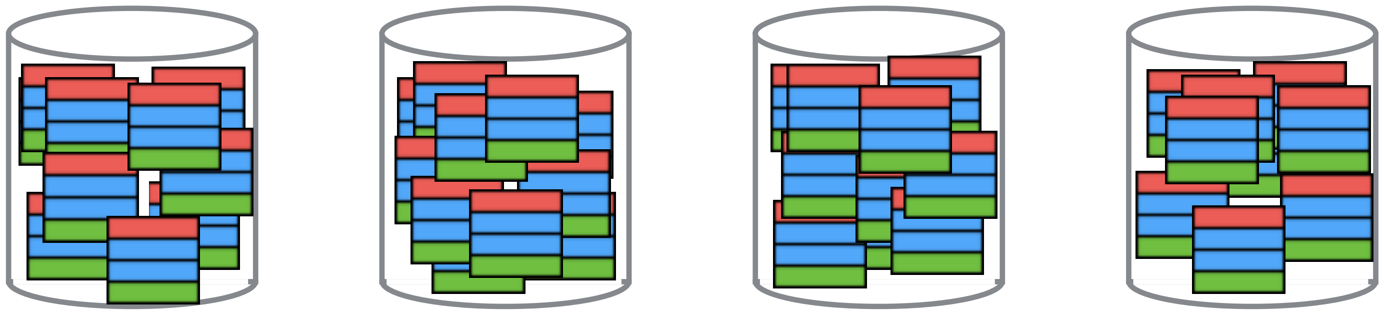
Mini-batches

[Chen & Manning ’14, Weiss et al. ’15]

Locally Normalized Training

Some advantages:

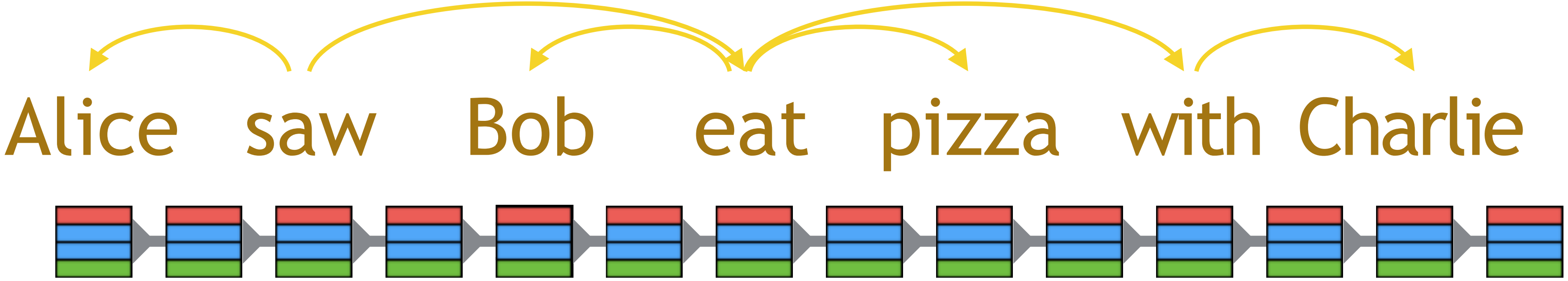
- Trivially Parallelizable
- SGD Training recipes
- Standard NN Packages



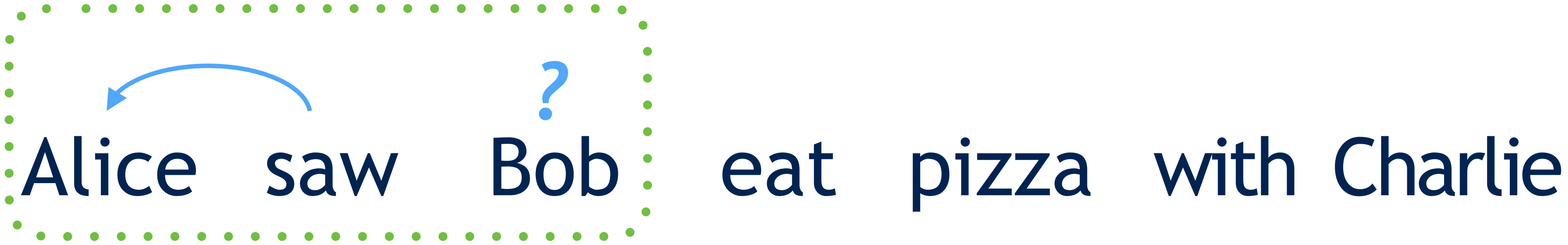
Mini-batches

[Chen & Manning ’14, Weiss et al. ’15]

Locally Normalized Inference

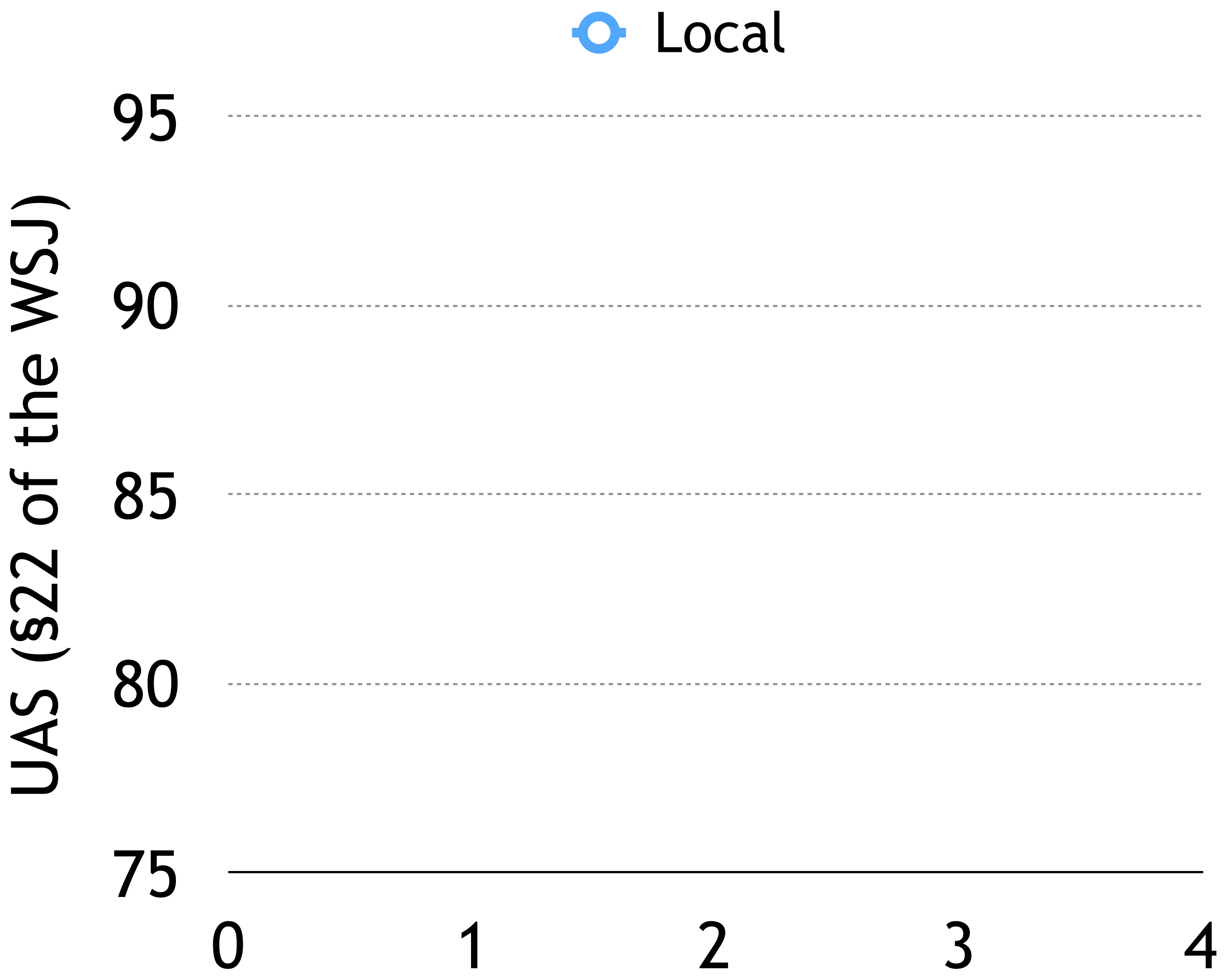


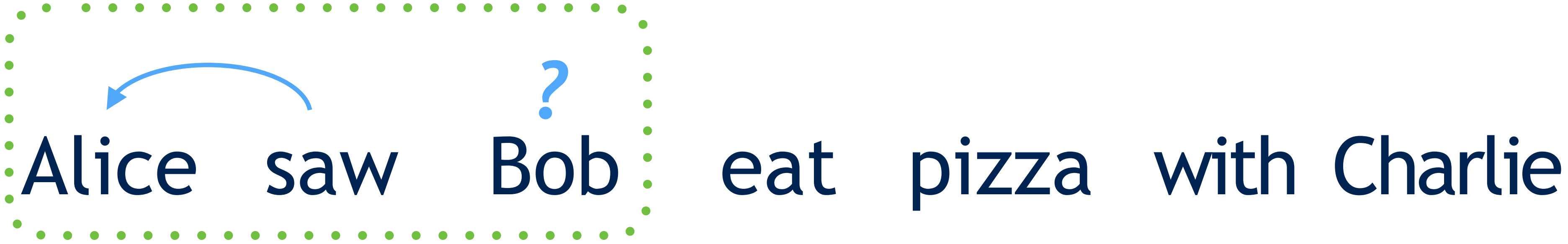
How Important is Lookahead?



Alice saw Bob eat pizza with Charlie

How Important is Lookahead?

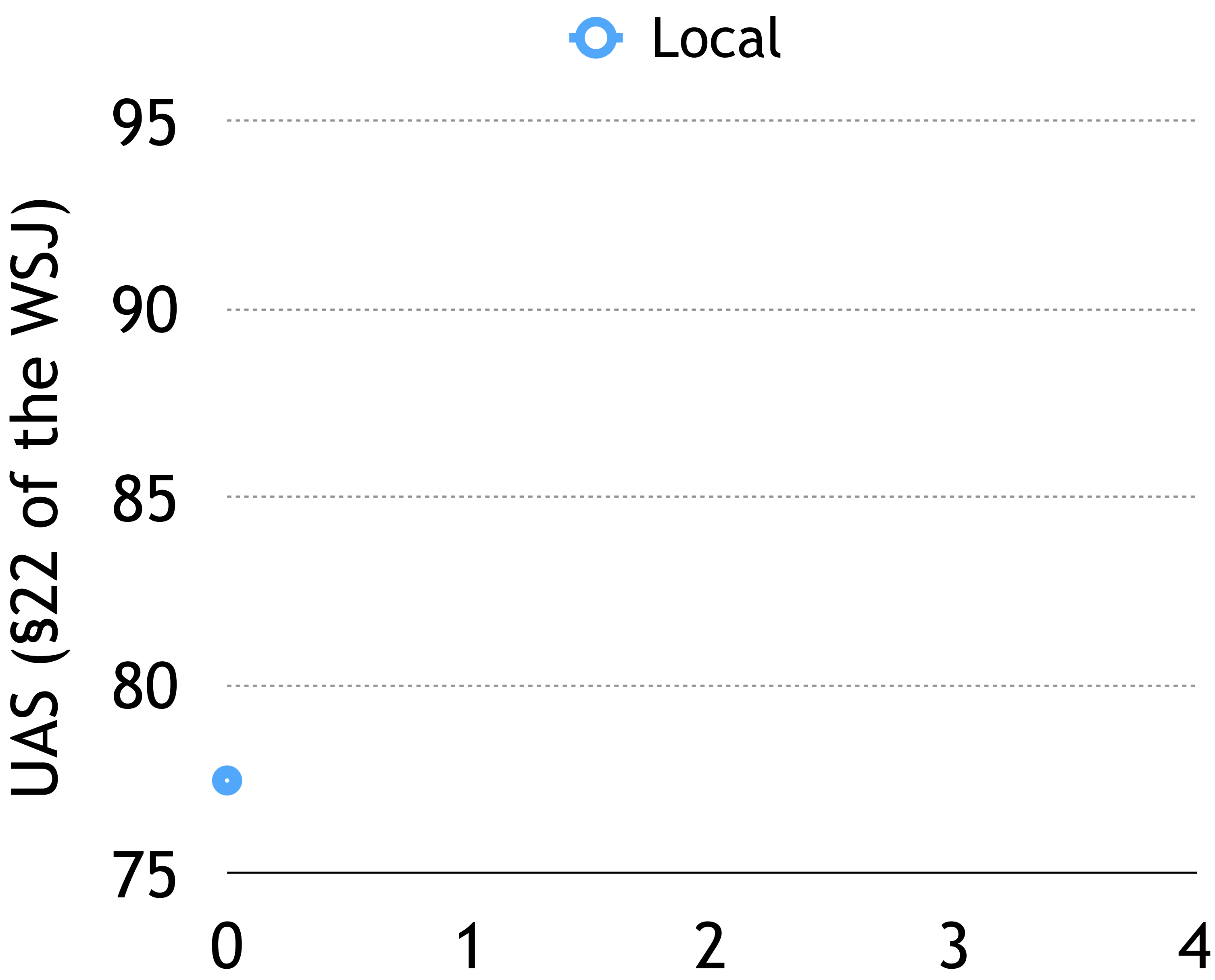


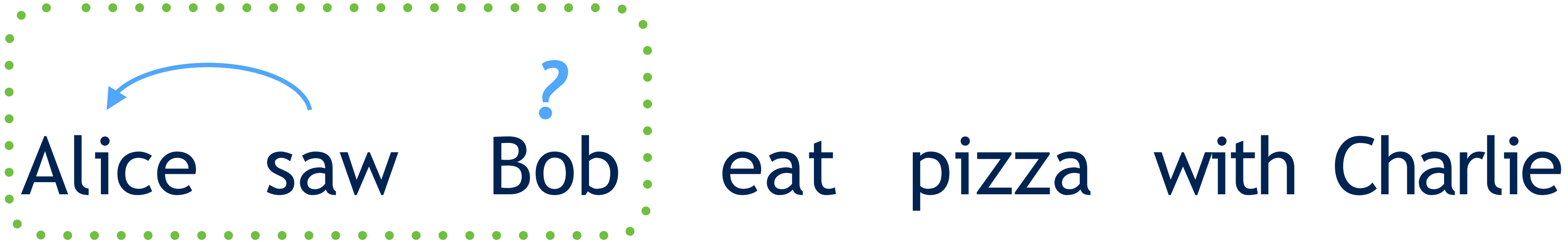


?

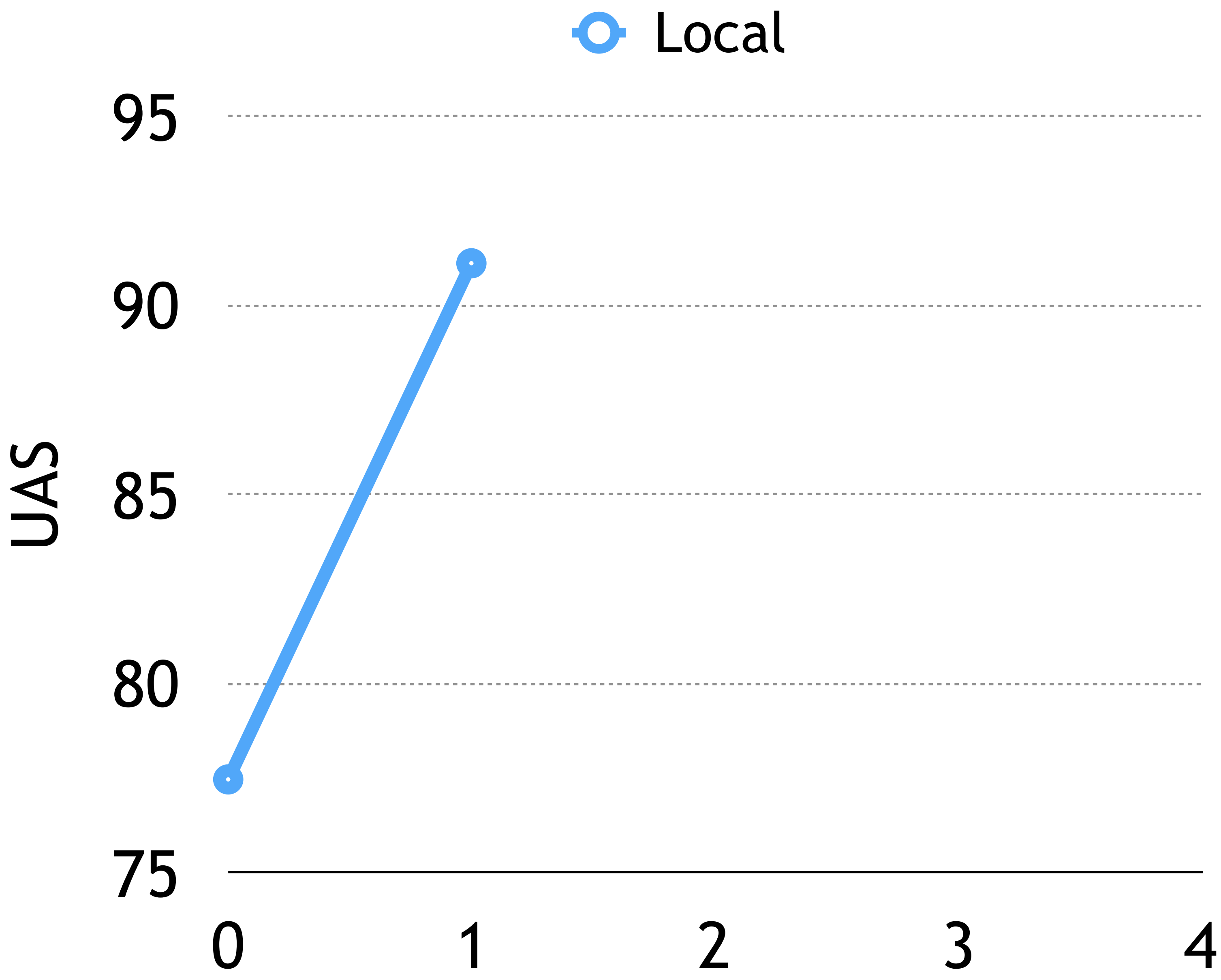
Alice saw Bob eat pizza with Charlie

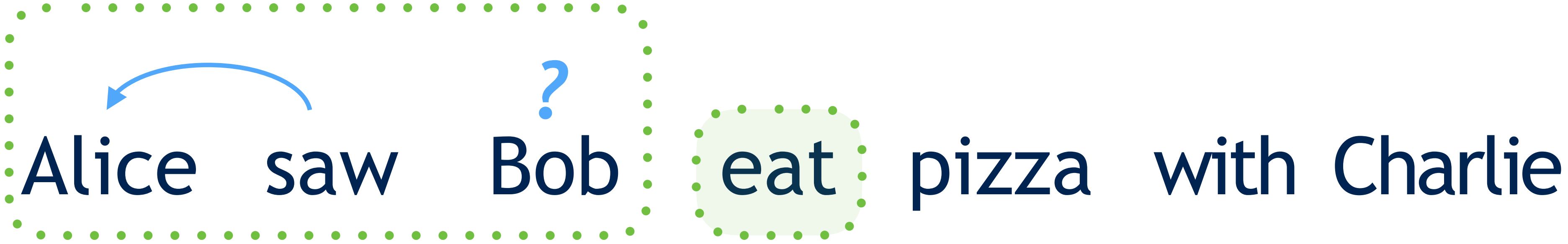
How Important is Lookahead?





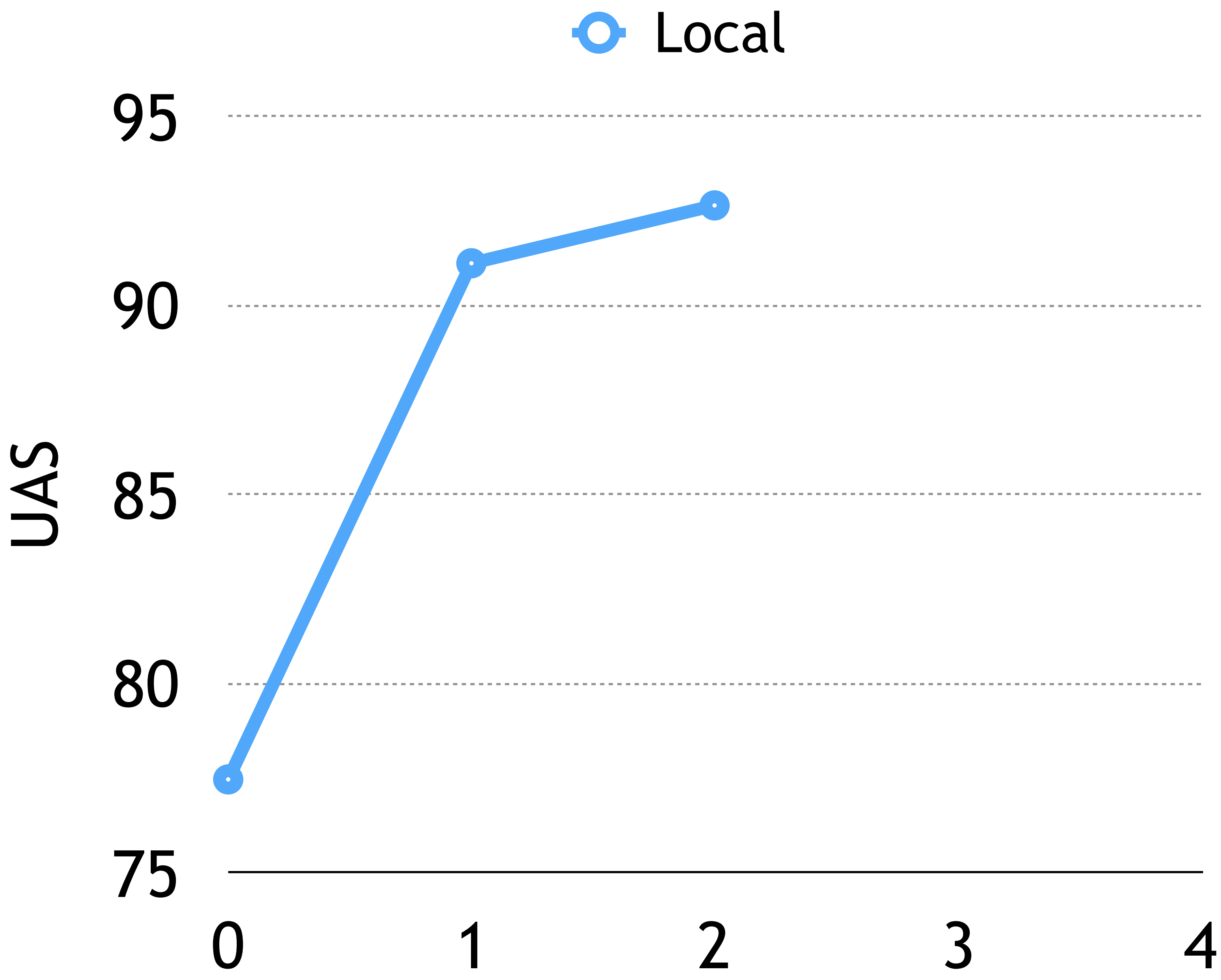
How Important is Lookahead?

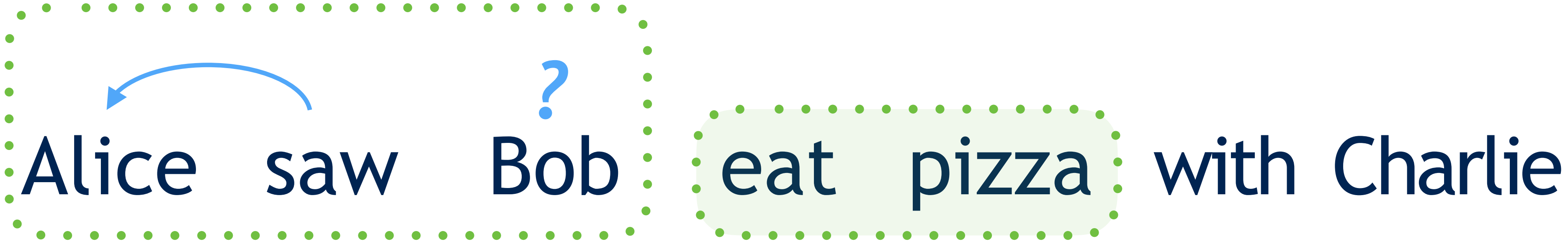




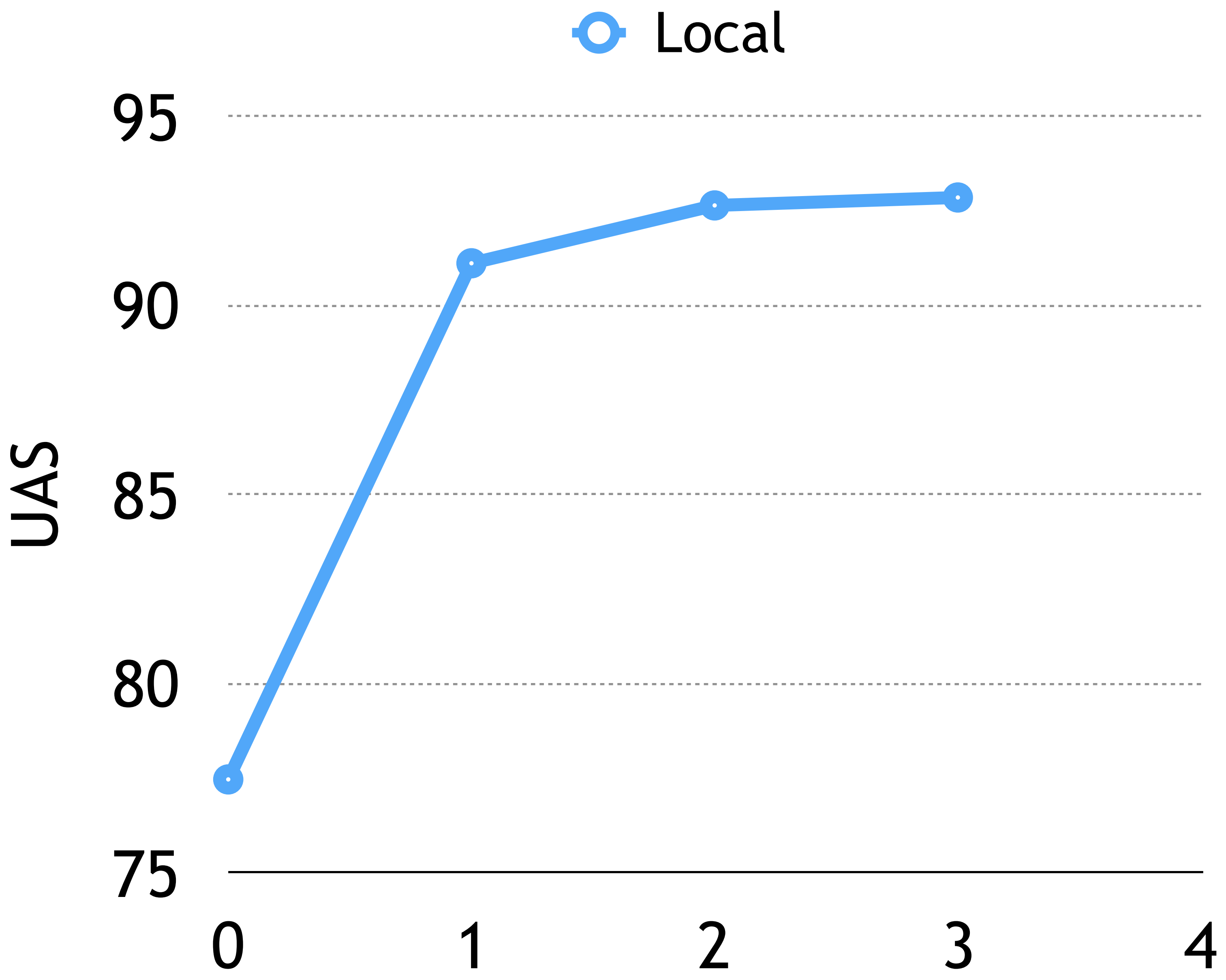
Alice saw Bob eat pizza with Charlie

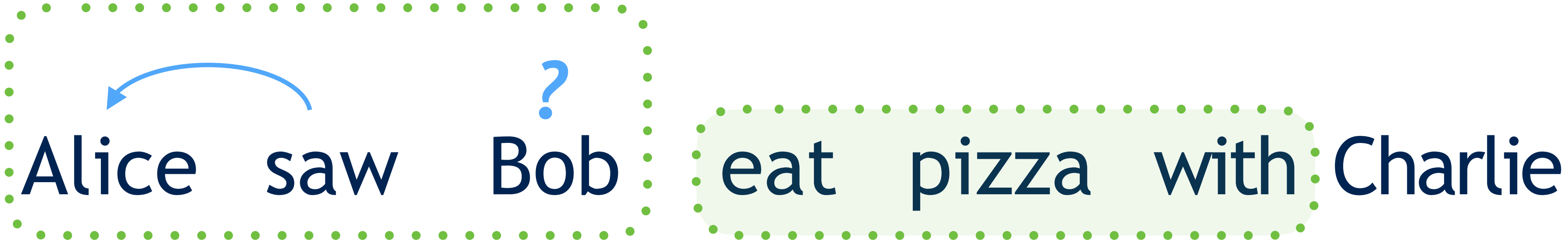
How Important is Lookahead?





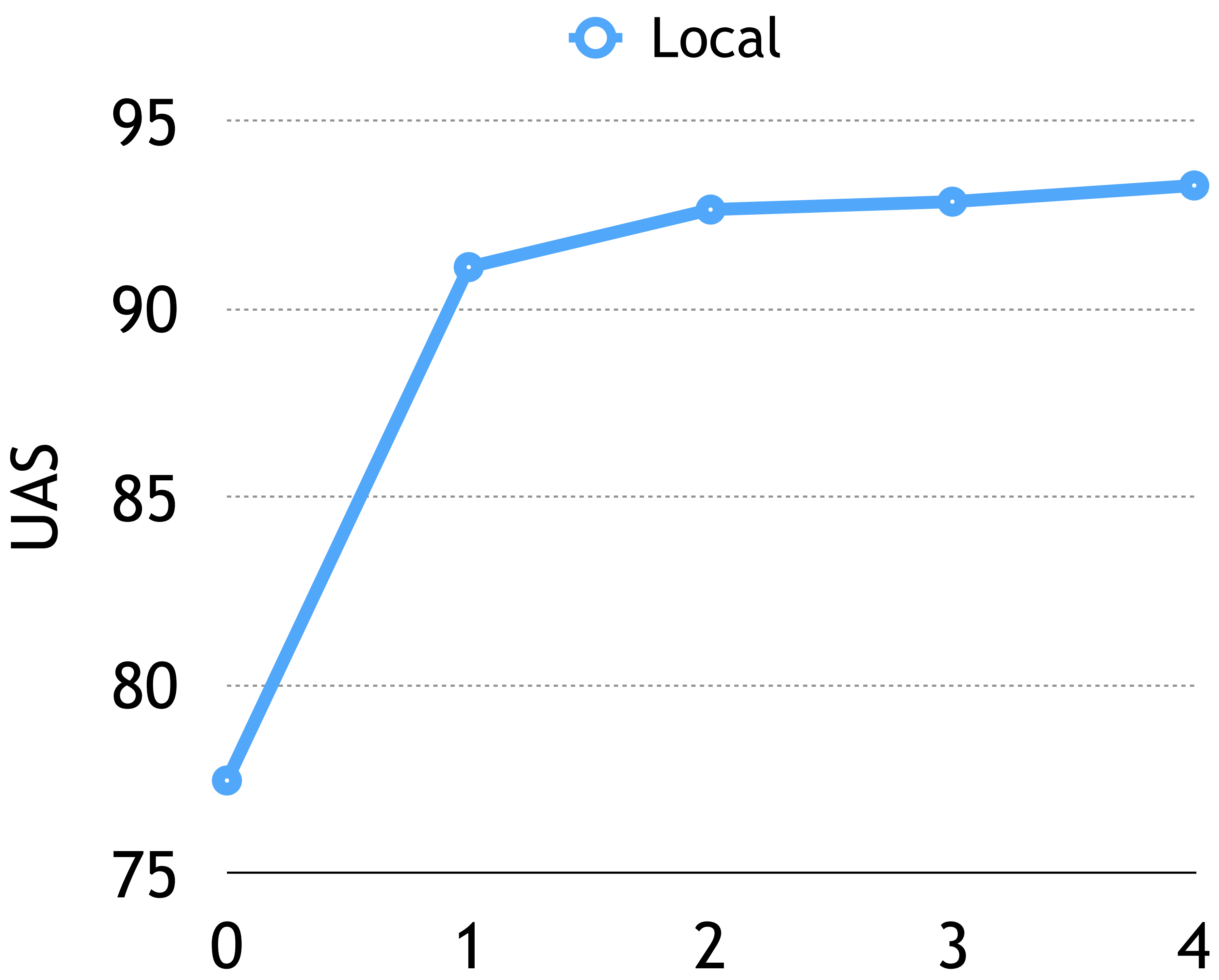
How Important is Lookahead?

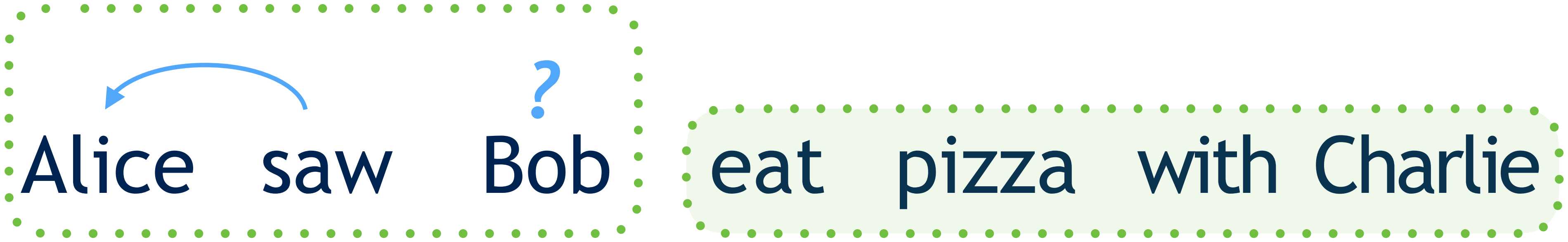




Alice saw Bob eat pizza with Charlie

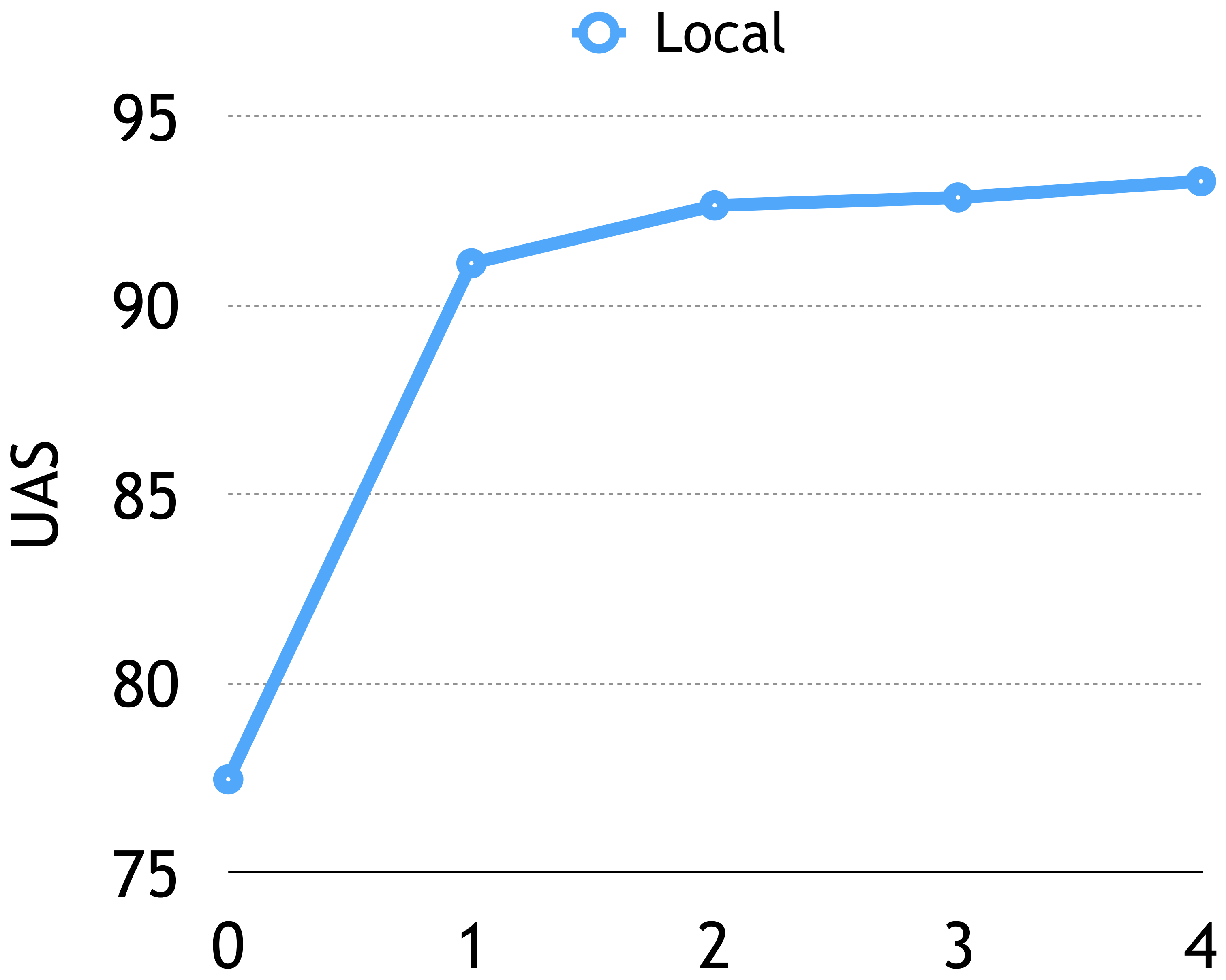
How Important is Lookahead?





Alice saw Bob ? eat pizza with Charlie

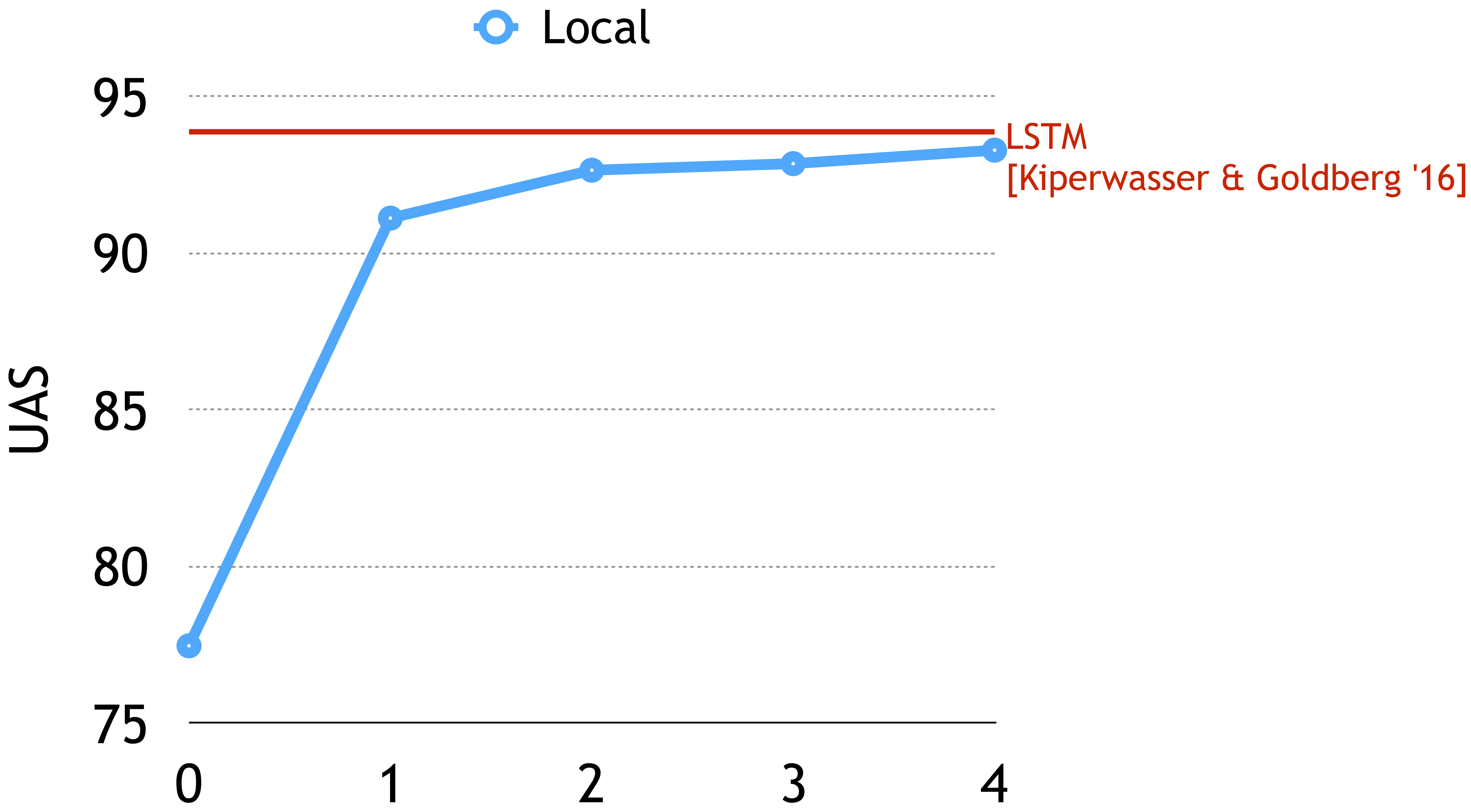
How Important is Lookahead?

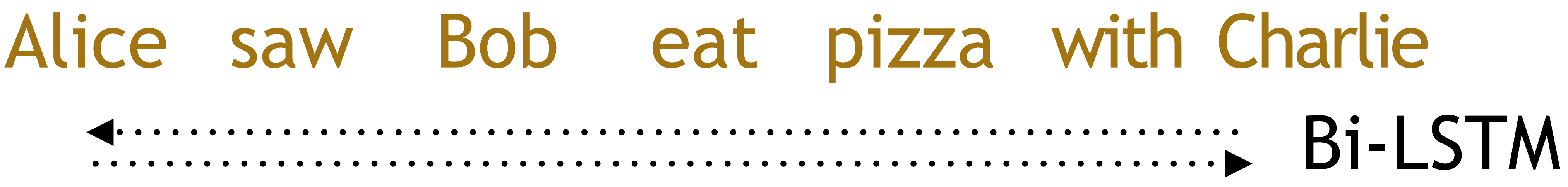


Alice saw Bob eat pizza with Charlie

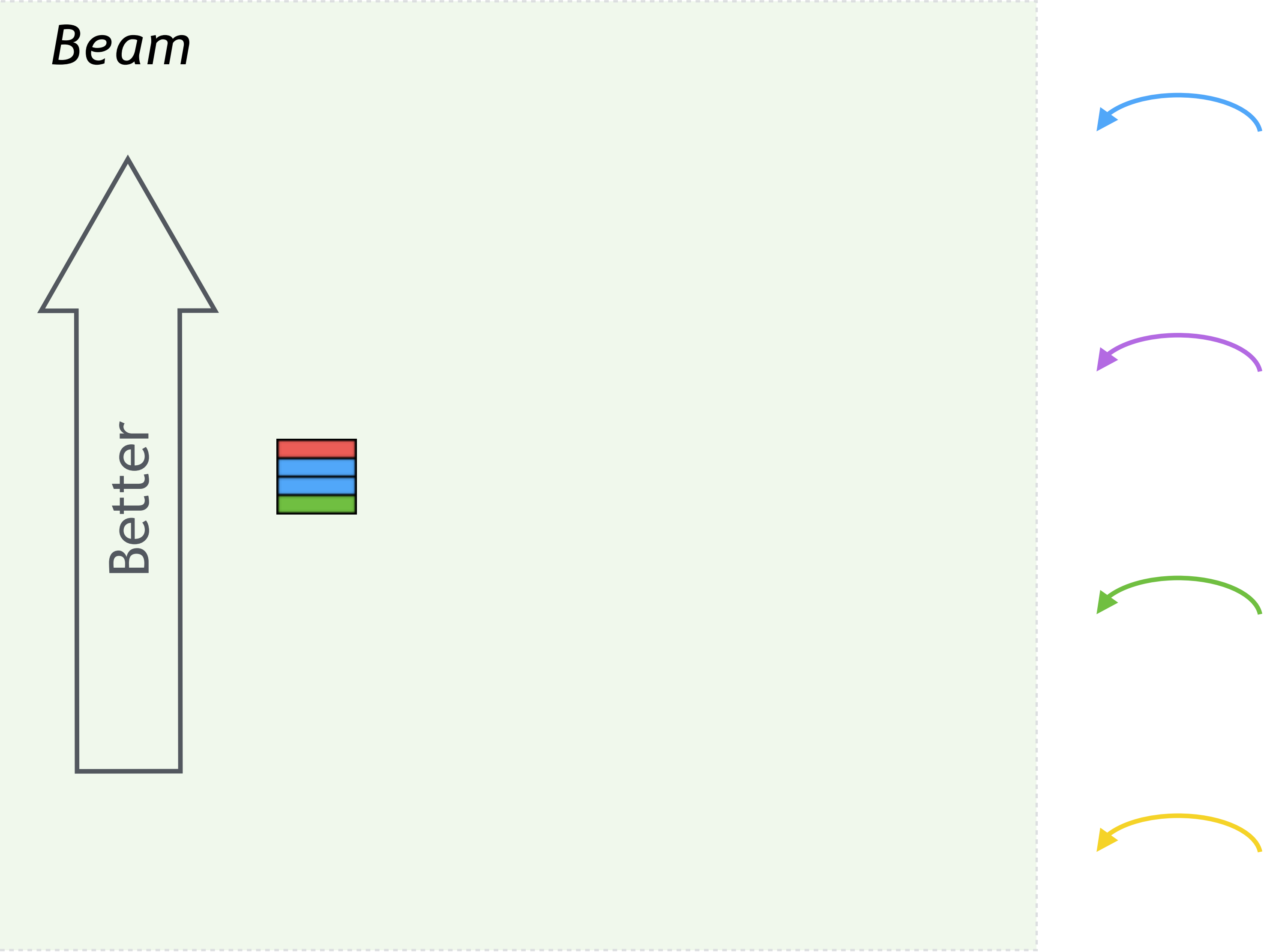
Bi-LSTM

How Important is Lookahead?





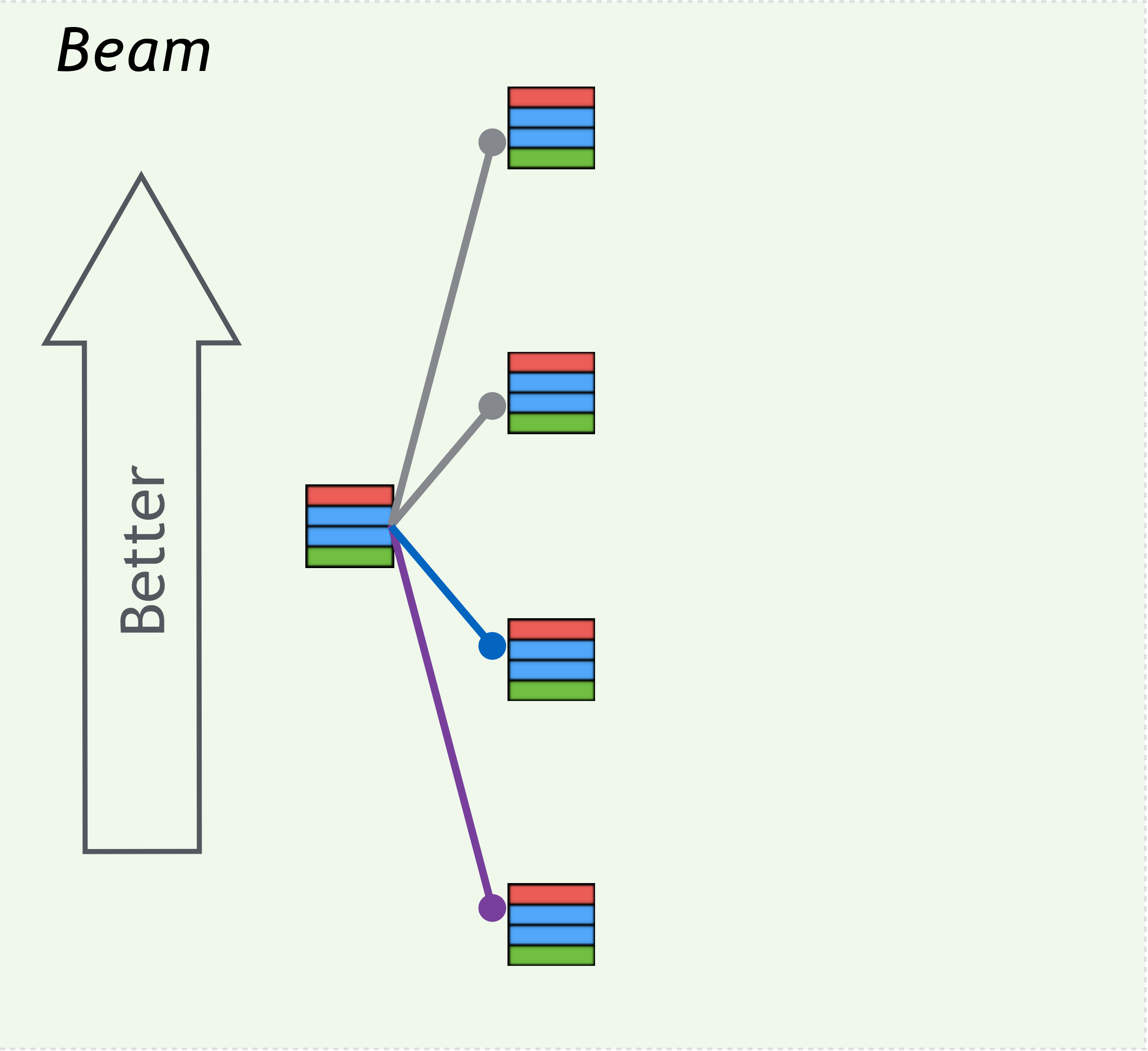
Beam Search with Local Model



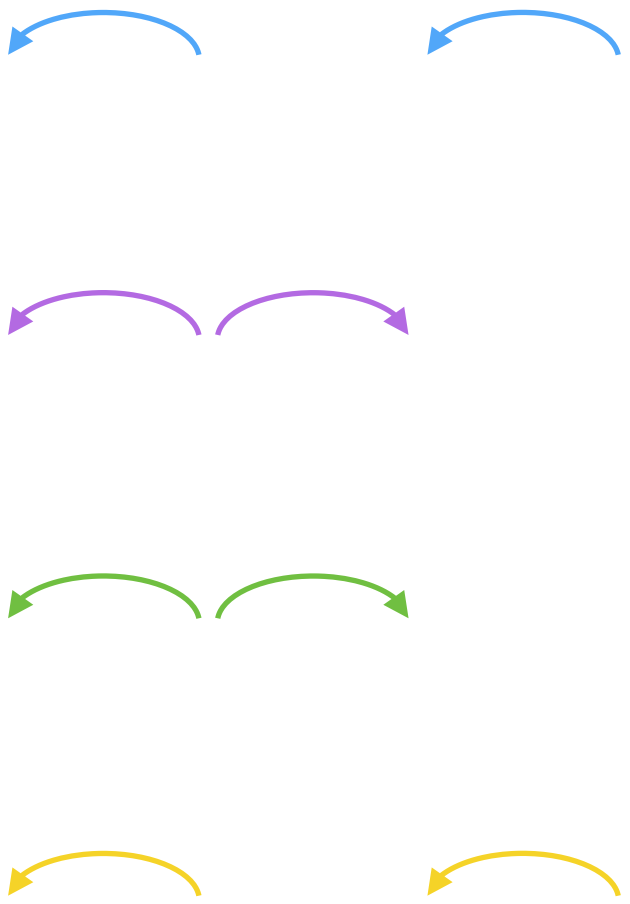
(Schematic)

Alice saw Bob eat pizza with Charlie

Beam Search with Local Model

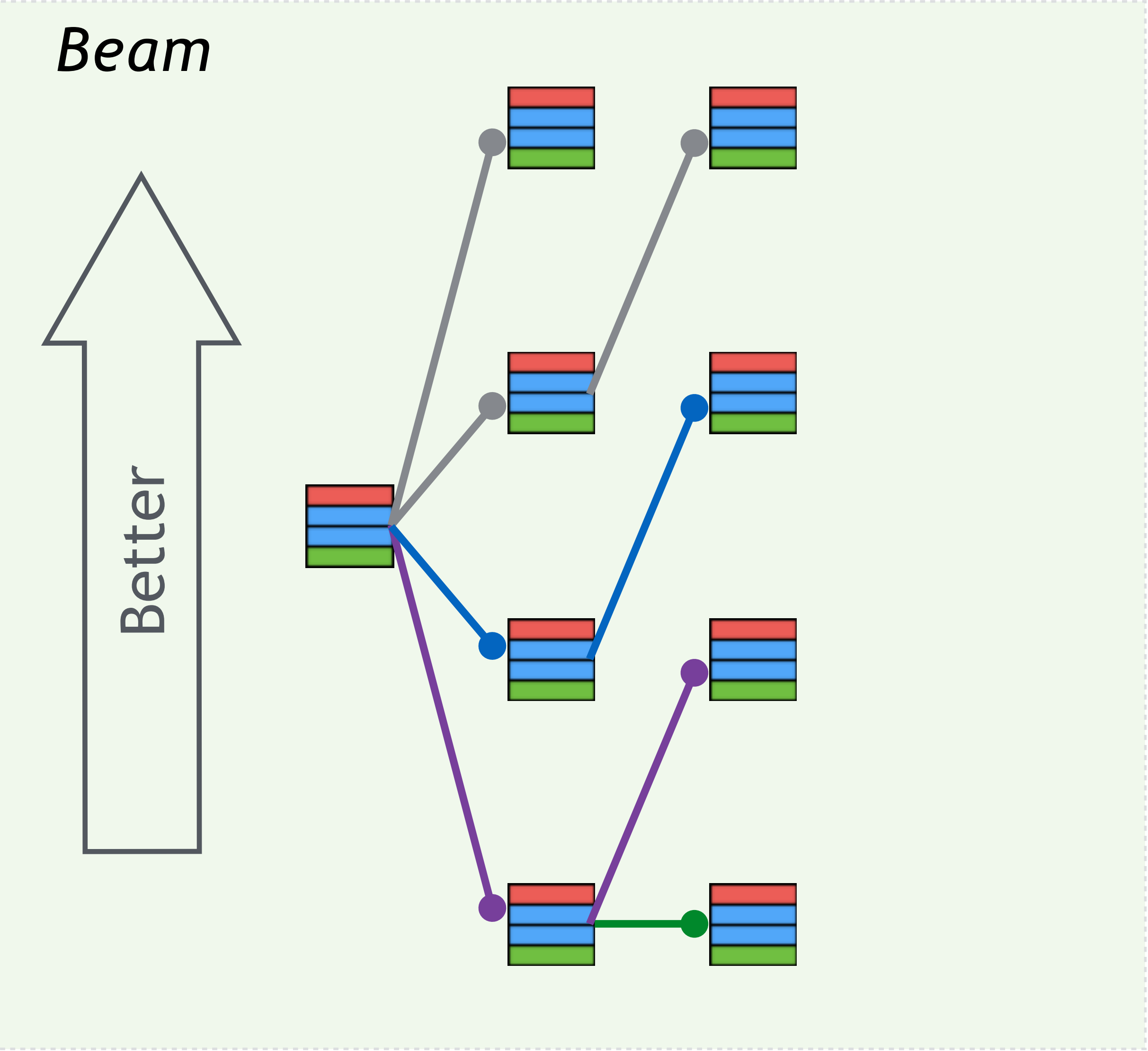


(Schematic)

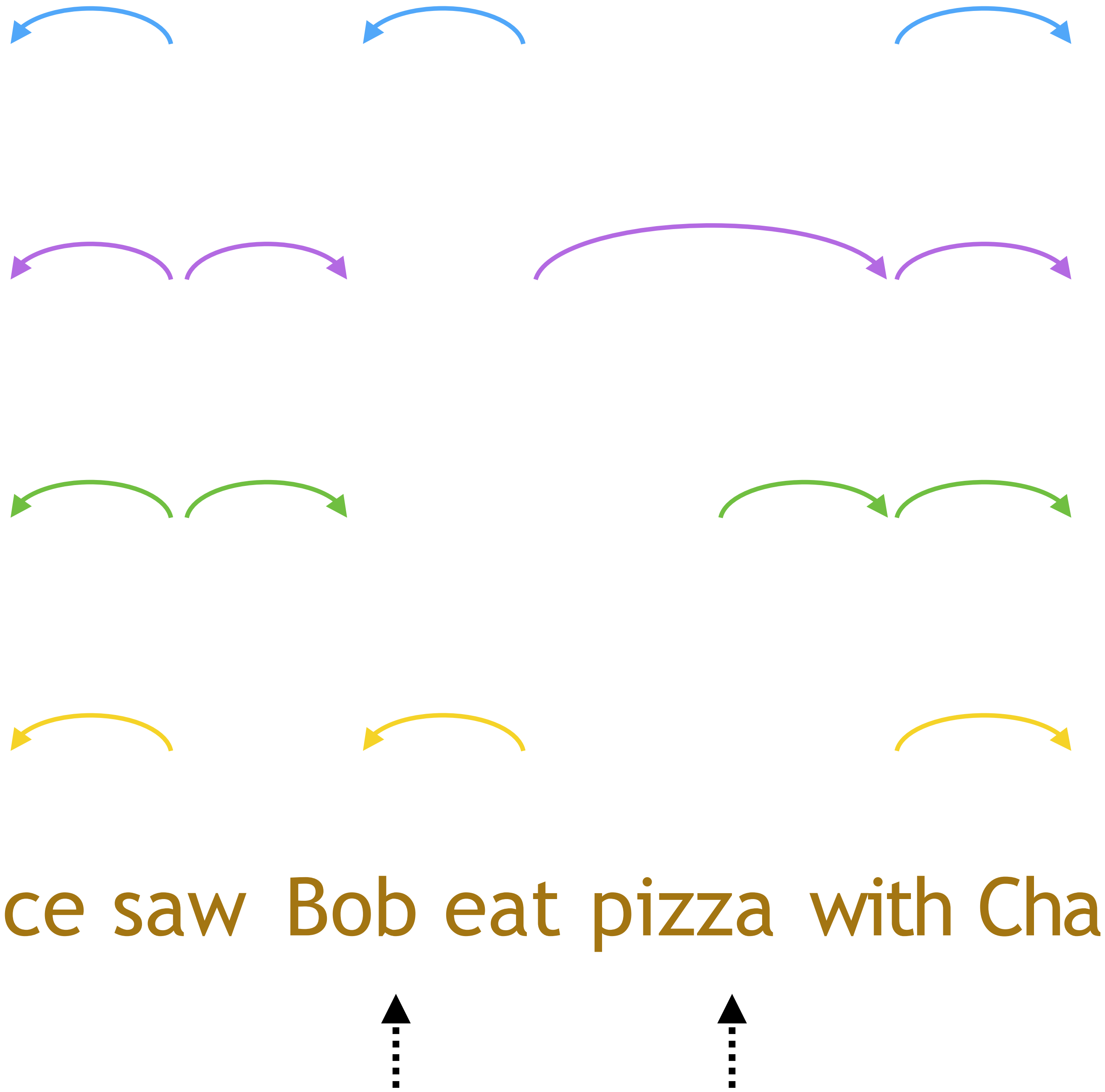


Alice saw Bob eat pizza with Charlie

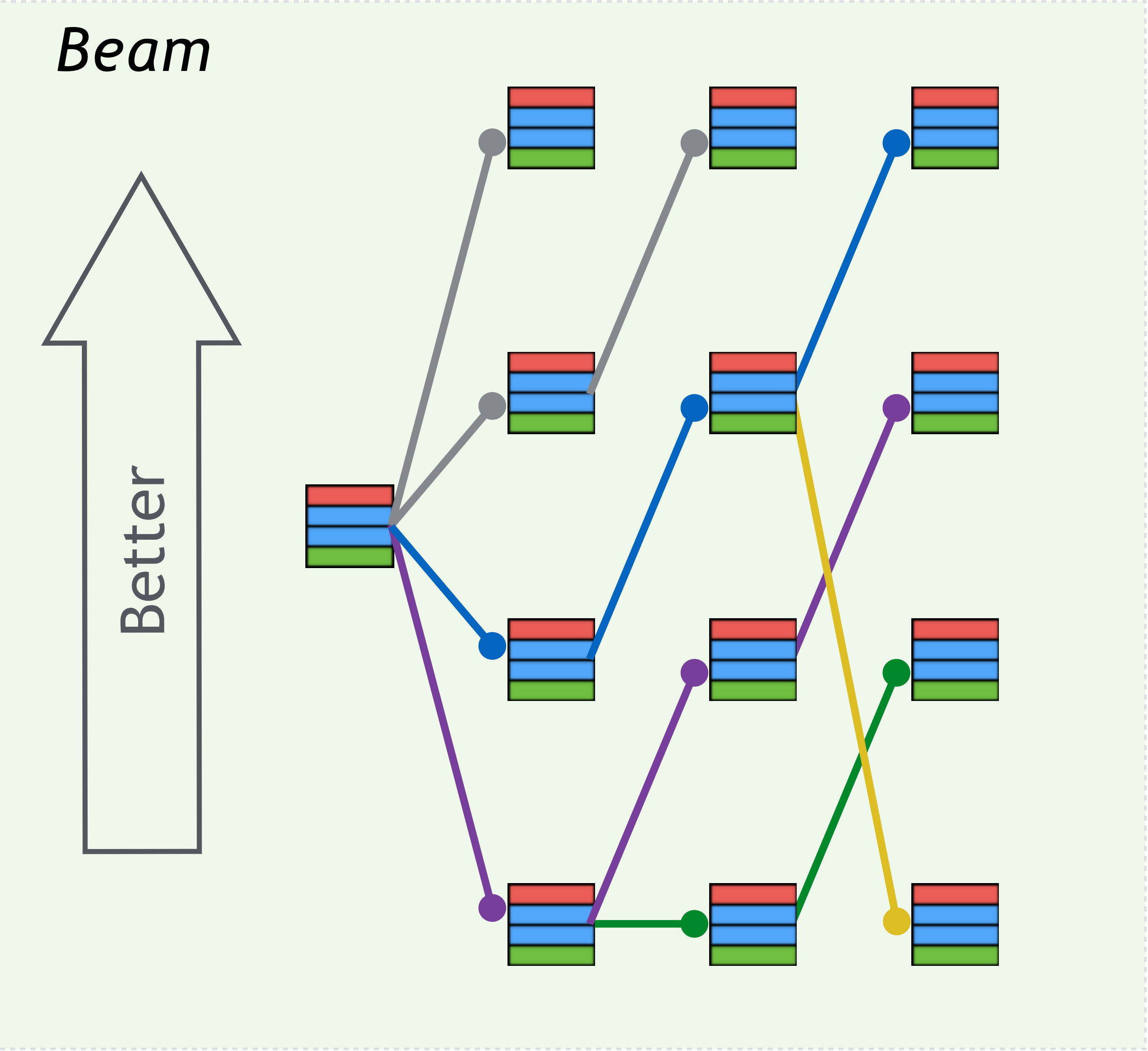
Beam Search with Local Model



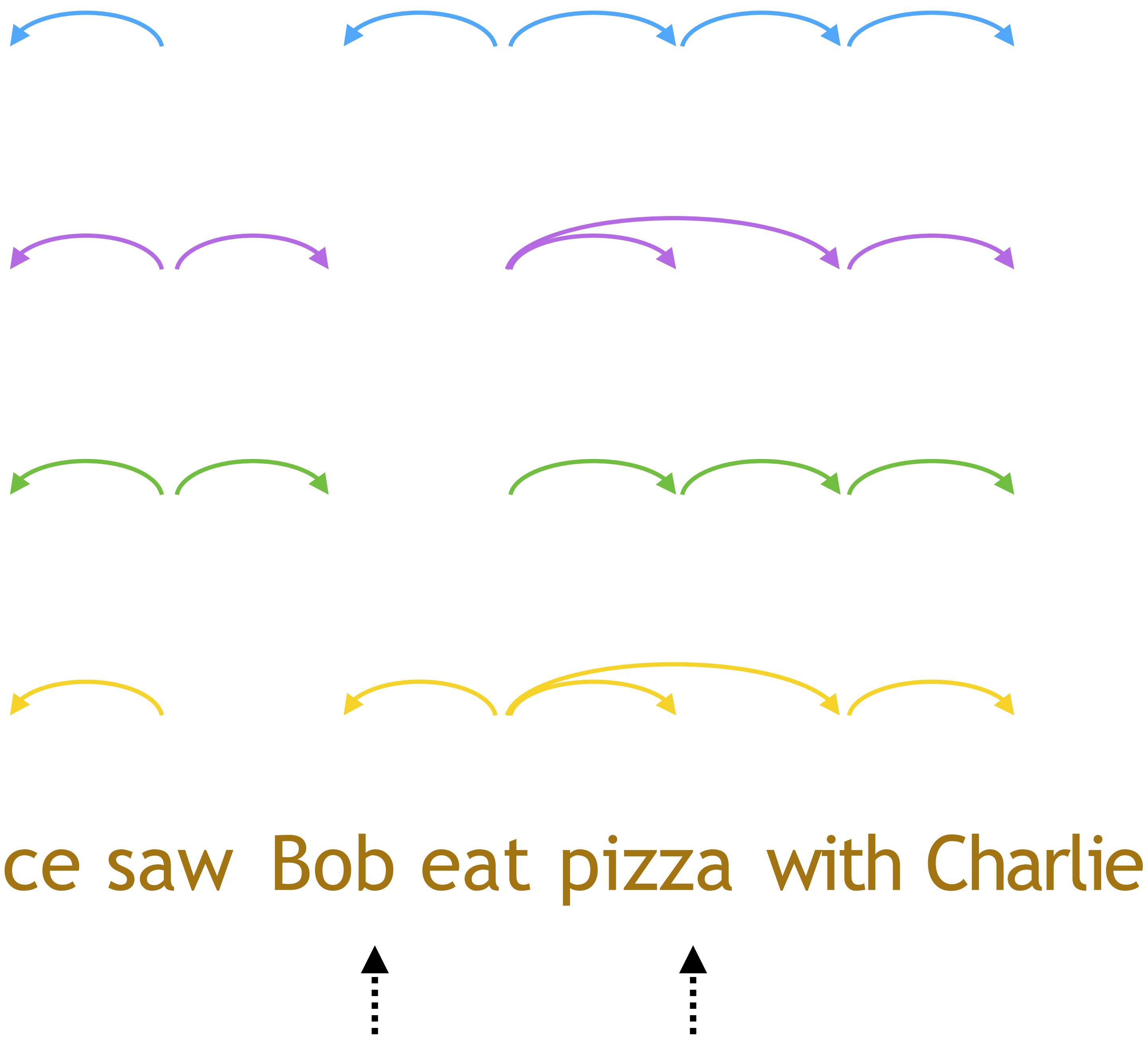
(Schematic)



Beam Search with Local Model

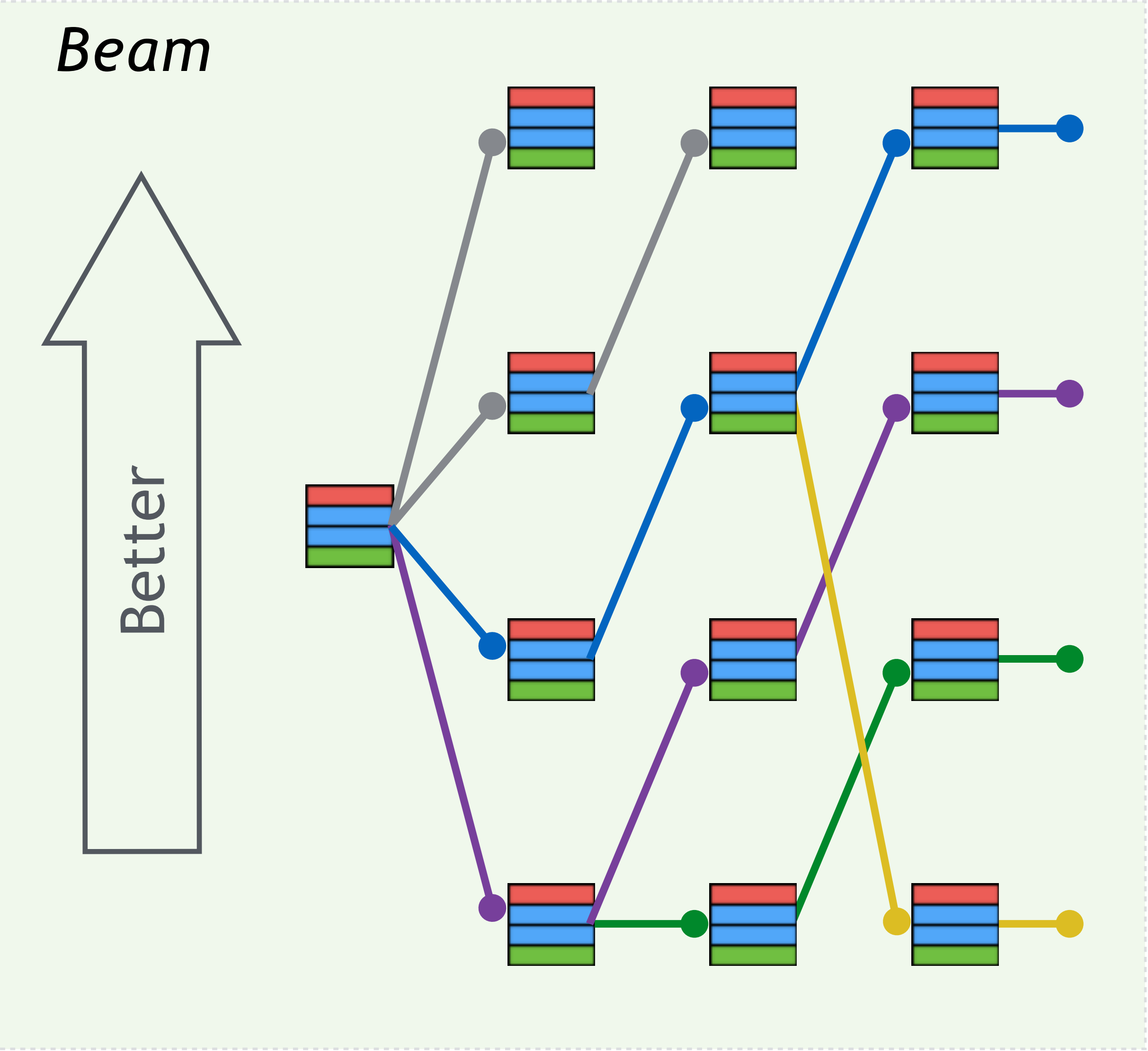


(Schematic)

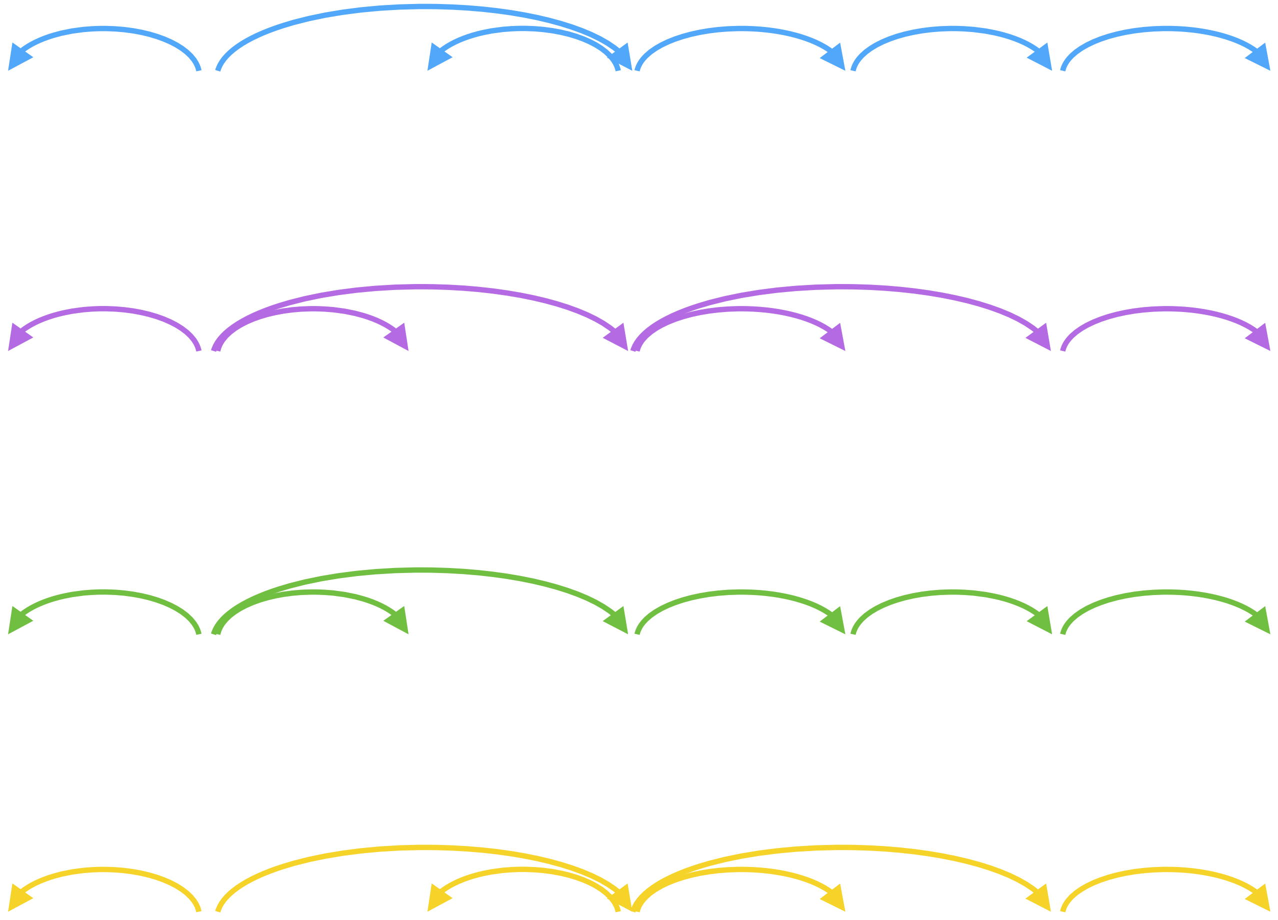


Alice saw Bob eat pizza with Charlie

Beam Search with Local Model

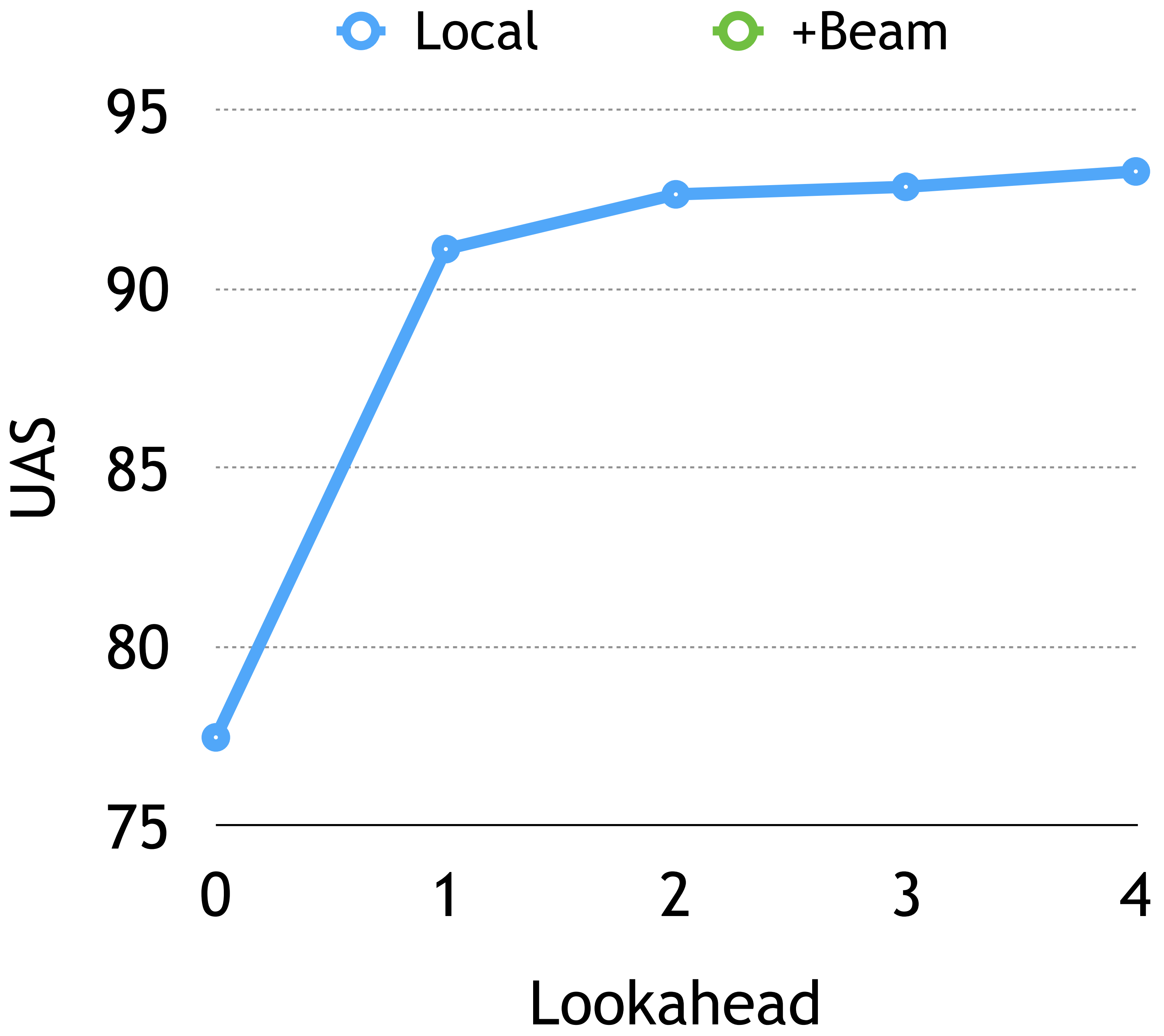


(Schematic)

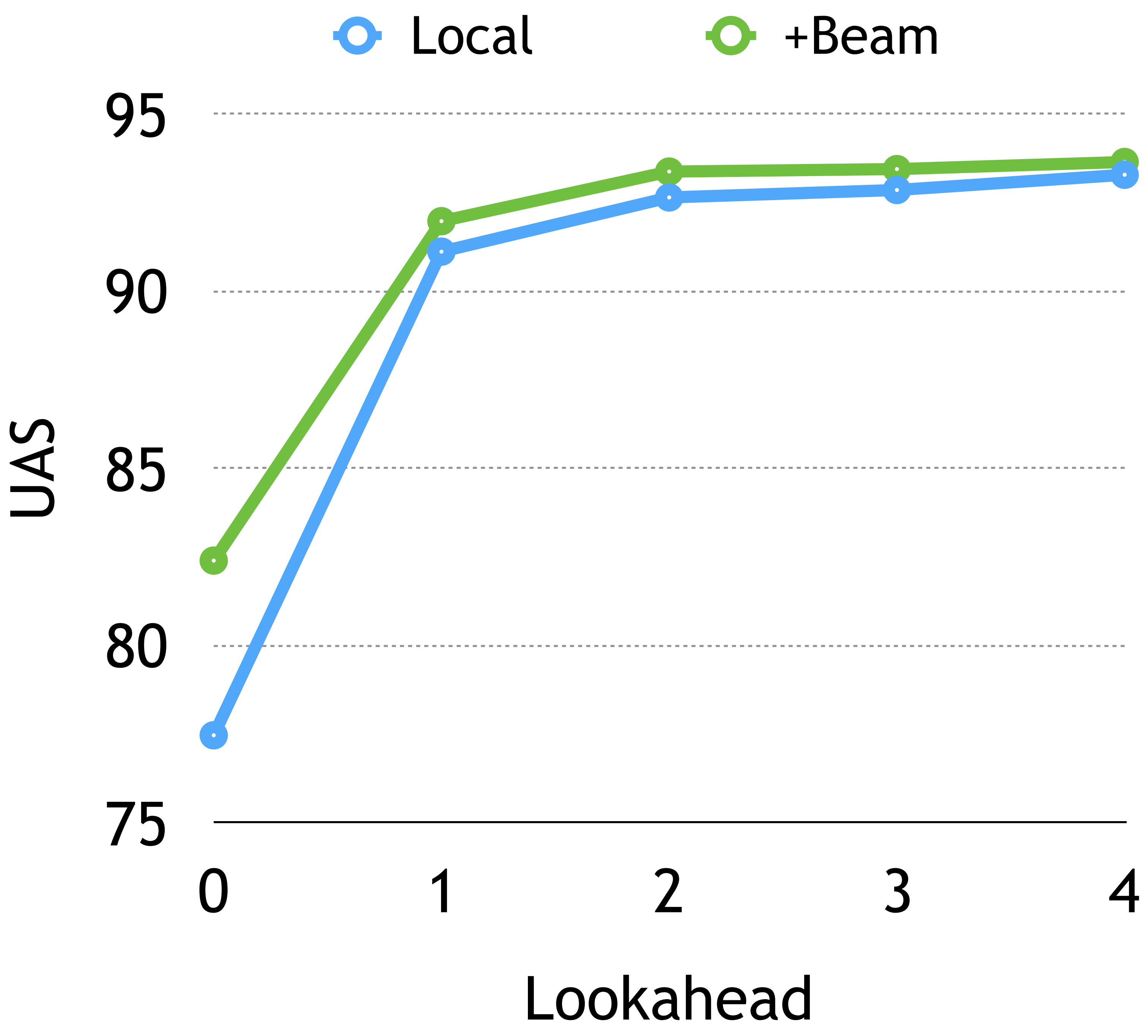


Alice saw Bob eat pizza with Charlie

Beam Search with Local Model



Beam Search with Local Model

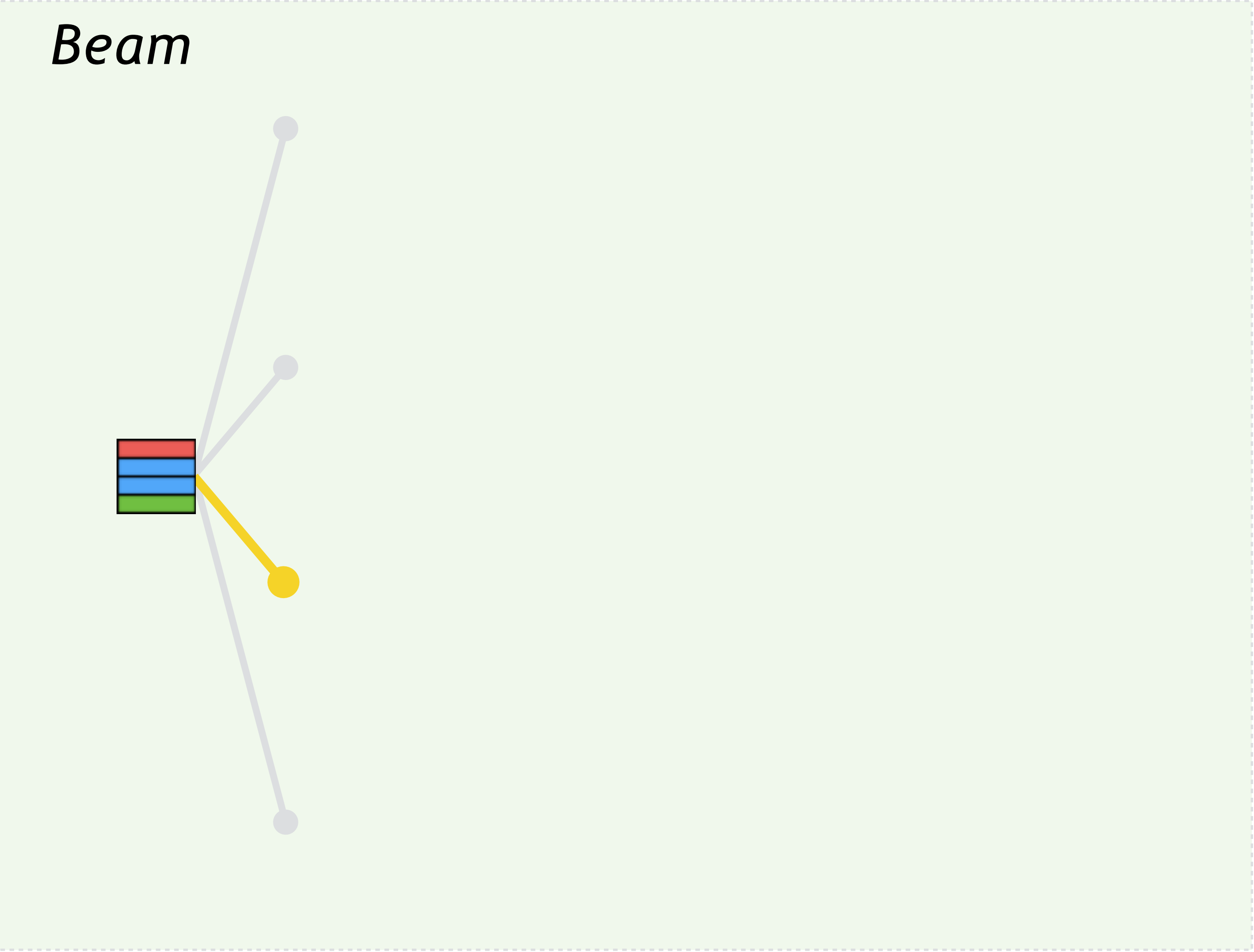


Training with Early Updates

Beam

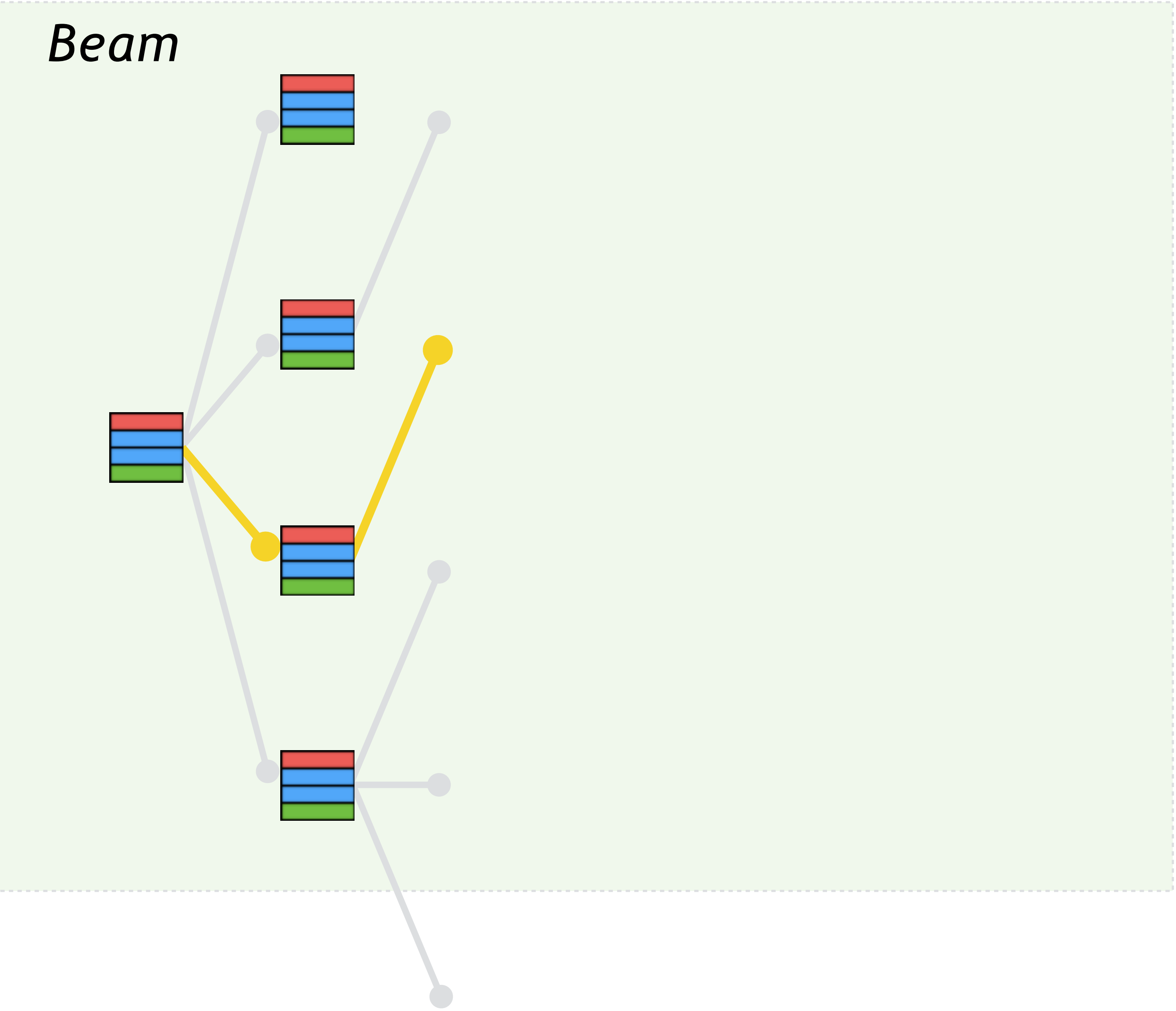
[Collins and Roark ’04, Zhou et al.’15]

Training with Early Updates



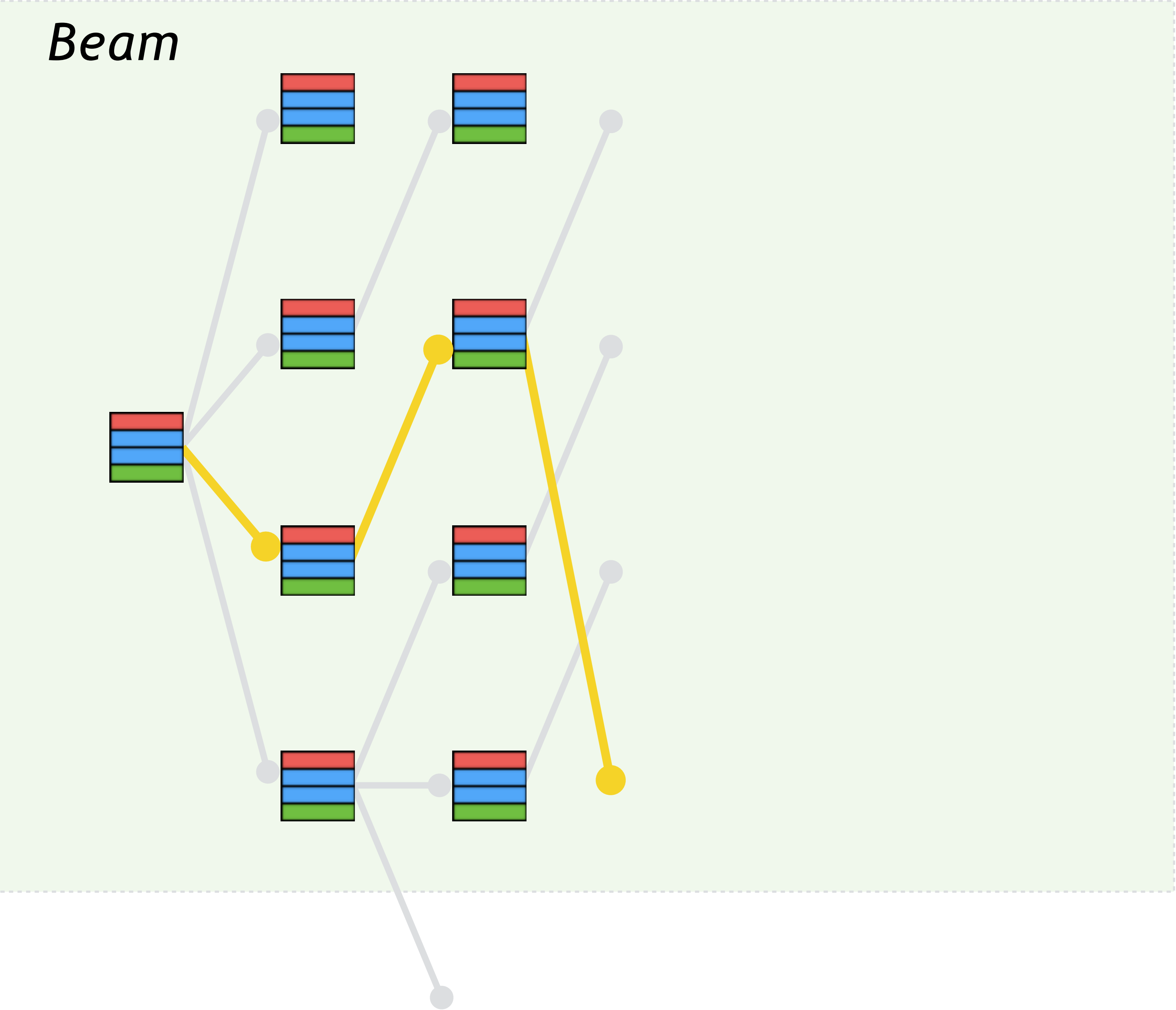
[Collins and Roark ’04, Zhou et al.’15]

Training with Early Updates



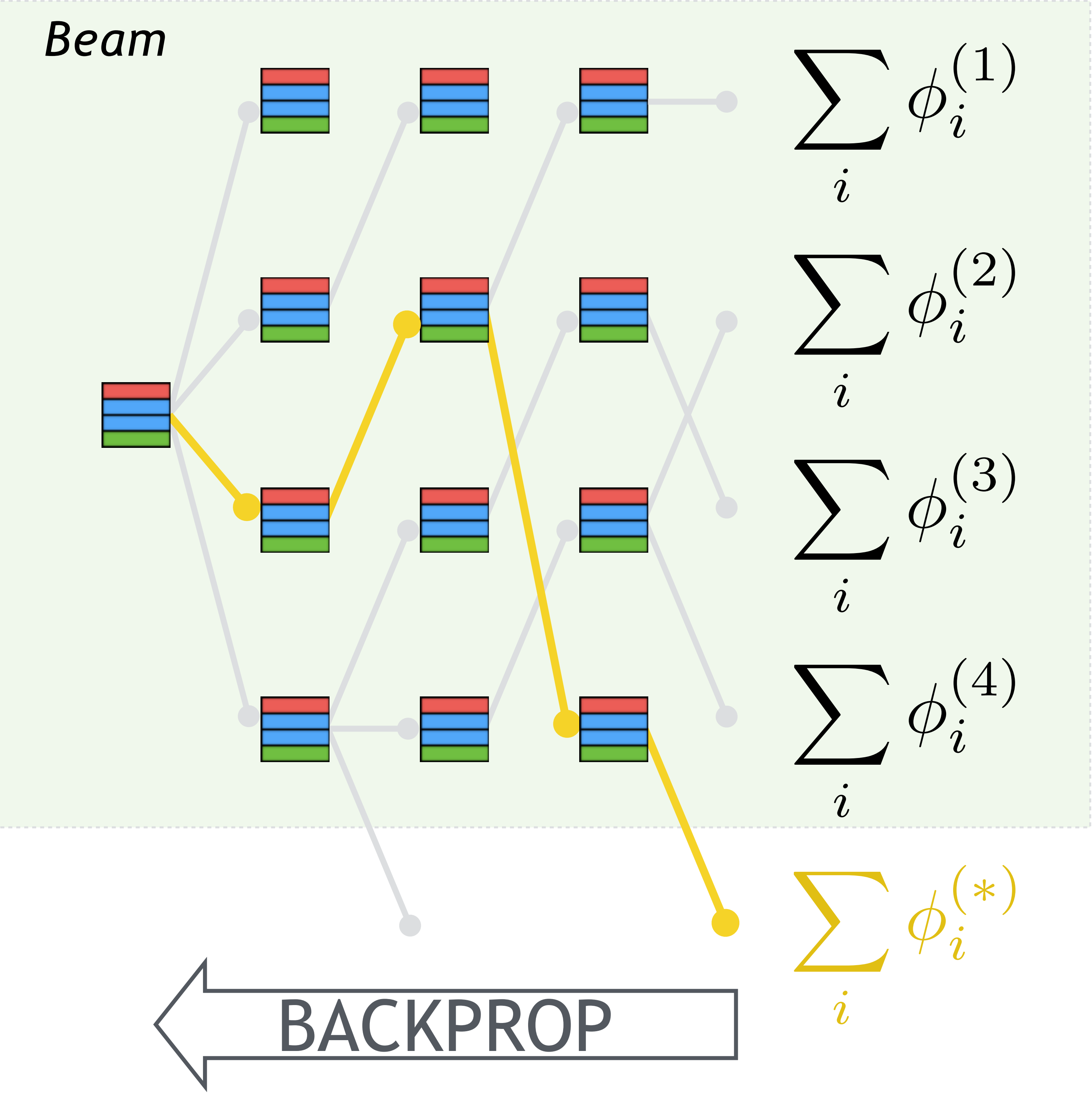
[Collins and Roark ’04, Zhou et al.’15]

Training with Early Updates



[Collins and Roark ’04, Zhou et al.’15]

Training with Early Updates



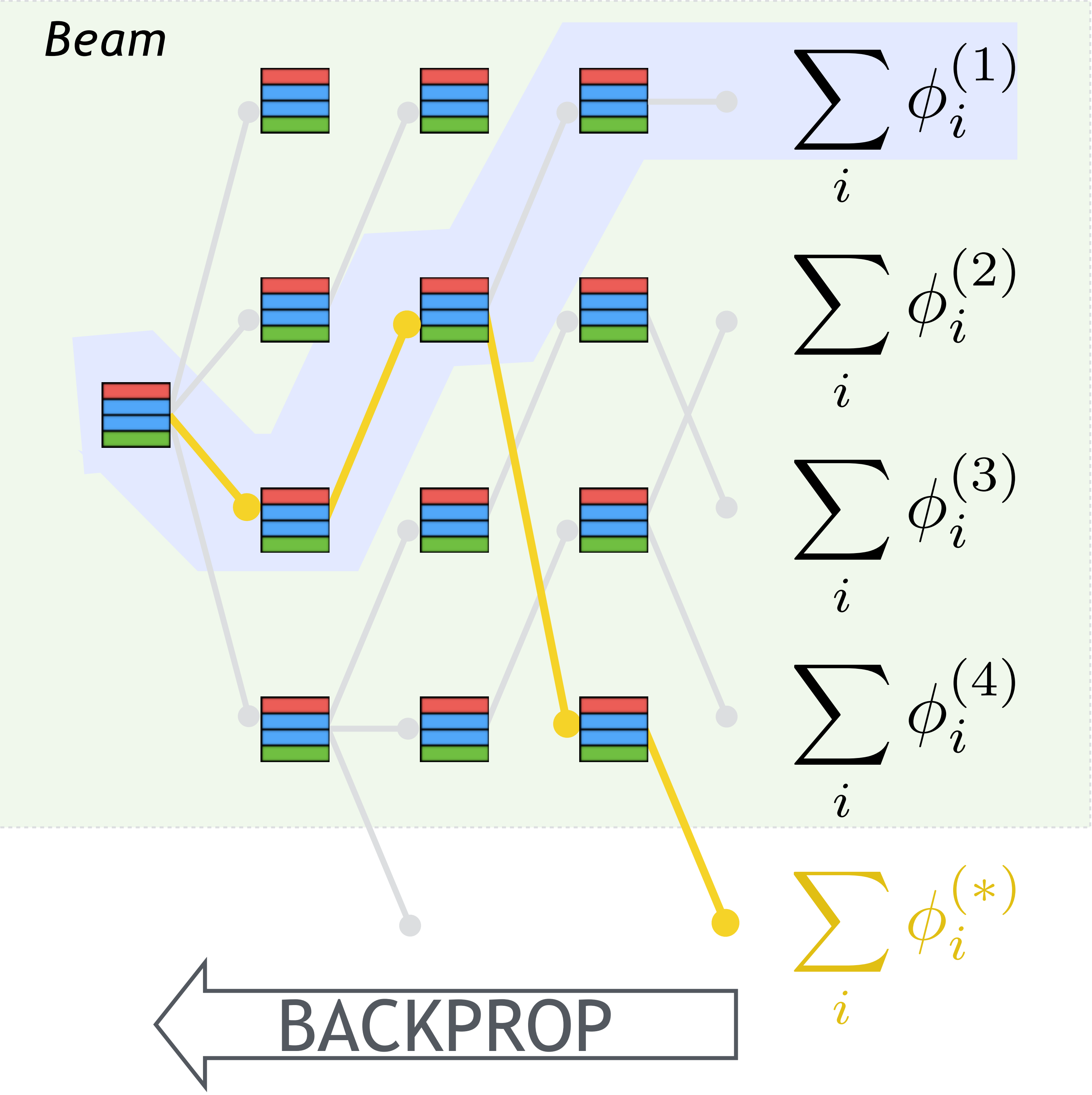
Globally normalized with respect to the beam:

$$\frac{\exp \sum_i \phi_i^{(*)}}{\sum_{j=1}^{|\text{Beam}|} \exp \sum_i \phi_i^{(j)}}$$

Backpropagate through all steps, paths, and layers

[Collins and Roark ’04, Zhou et al.’15]

Training with Early Updates



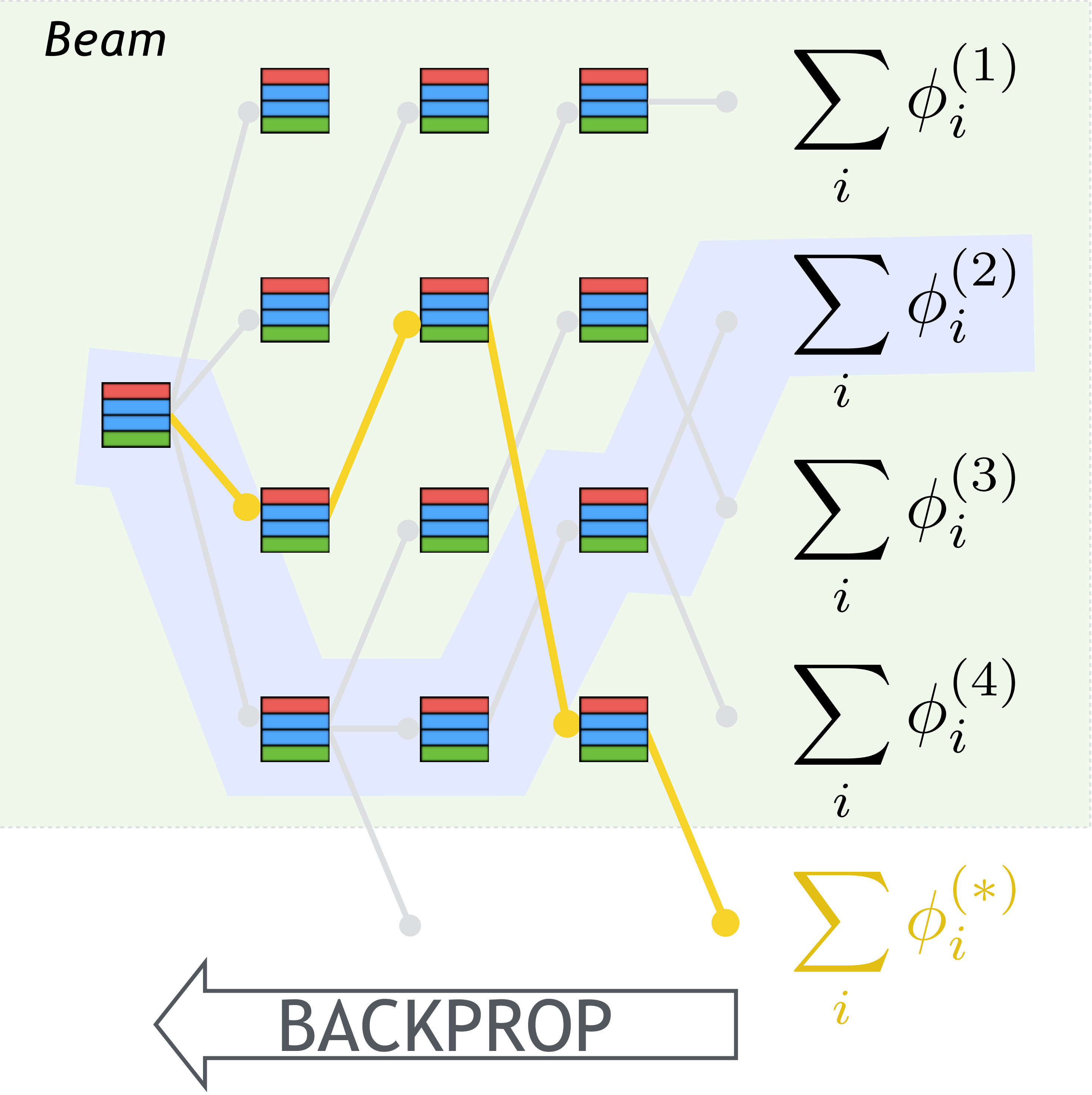
Globally normalized with respect to the beam:

$$\frac{\exp \sum_i \phi_i^{(*)}}{\sum_{j=1}^{|\text{Beam}|} \exp \sum_i \phi_i^{(j)}}$$

Backpropagate through all steps, paths, and layers

[Collins and Roark ’04, Zhou et al.’15]

Training with Early Updates



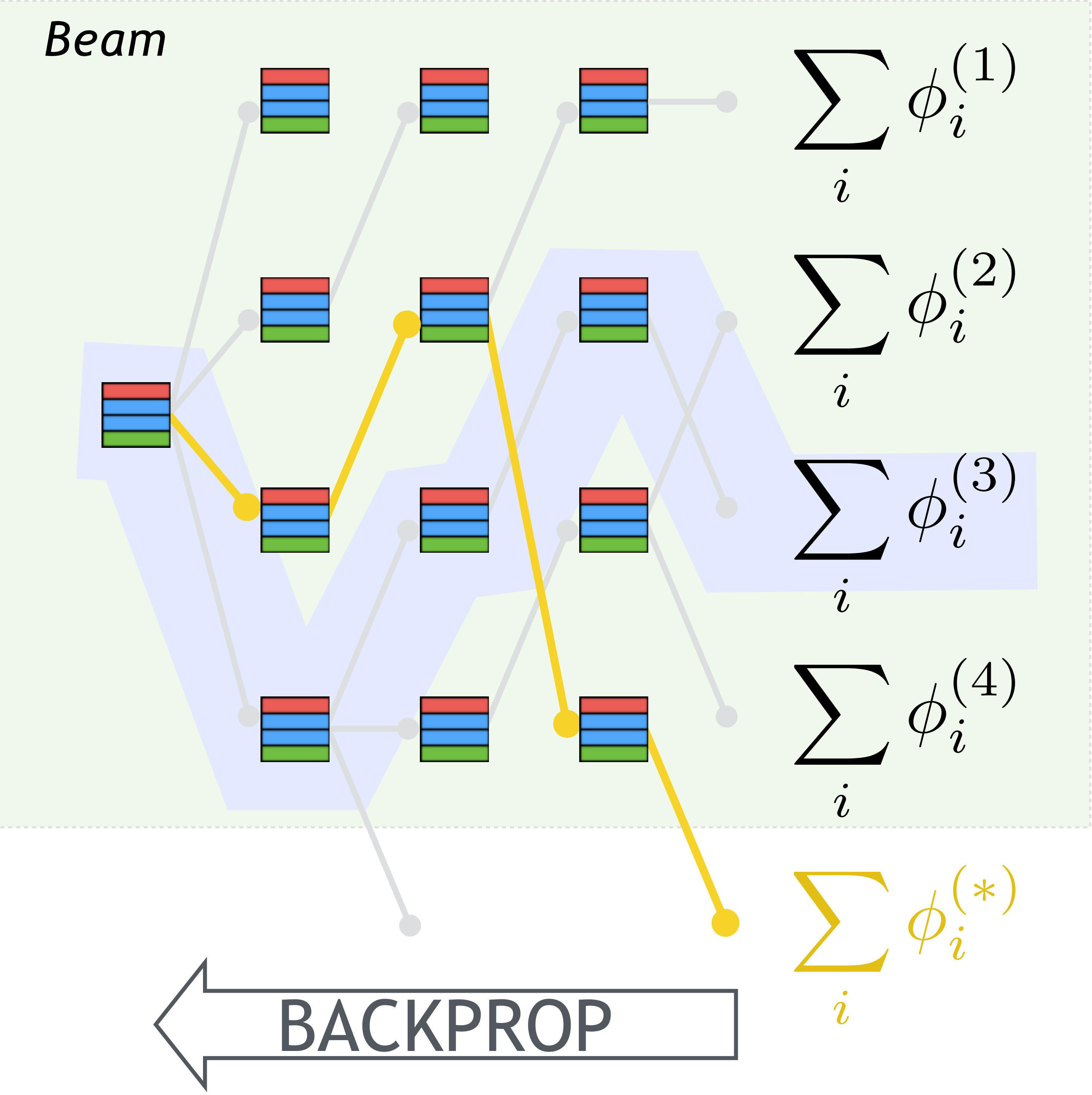
Globally normalized with respect to the beam:

$$\frac{\exp \sum_i \phi_i^{(*)}}{\sum_{j=1}^{|\mathrm{Beam}|} \exp \sum_i \phi_i^{(j)}}$$

Backpropagate through all steps, paths, and layers

[Collins and Roark ’04, Zhou et al.’15]

Training with Early Updates



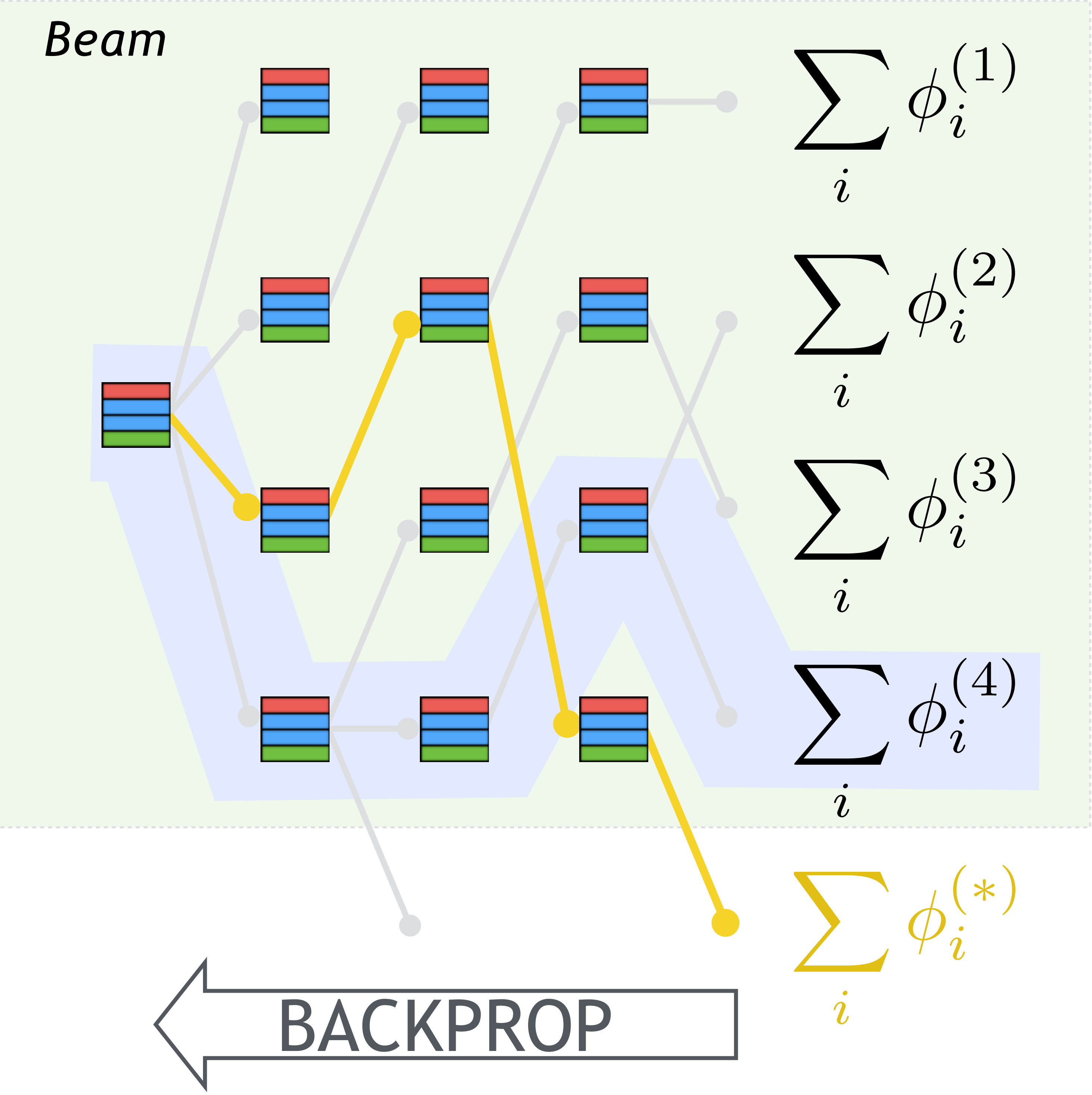
Globally normalized with respect to the beam:

$$\frac{\exp \sum_i \phi_i^{(*)}}{\sum_{j=1}^{|\text{Beam}|} \exp \sum_i \phi_i^{(j)}}$$

Backpropagate through all steps, paths, and layers

[Collins and Roark ’04, Zhou et al.’15]

Training with Early Updates



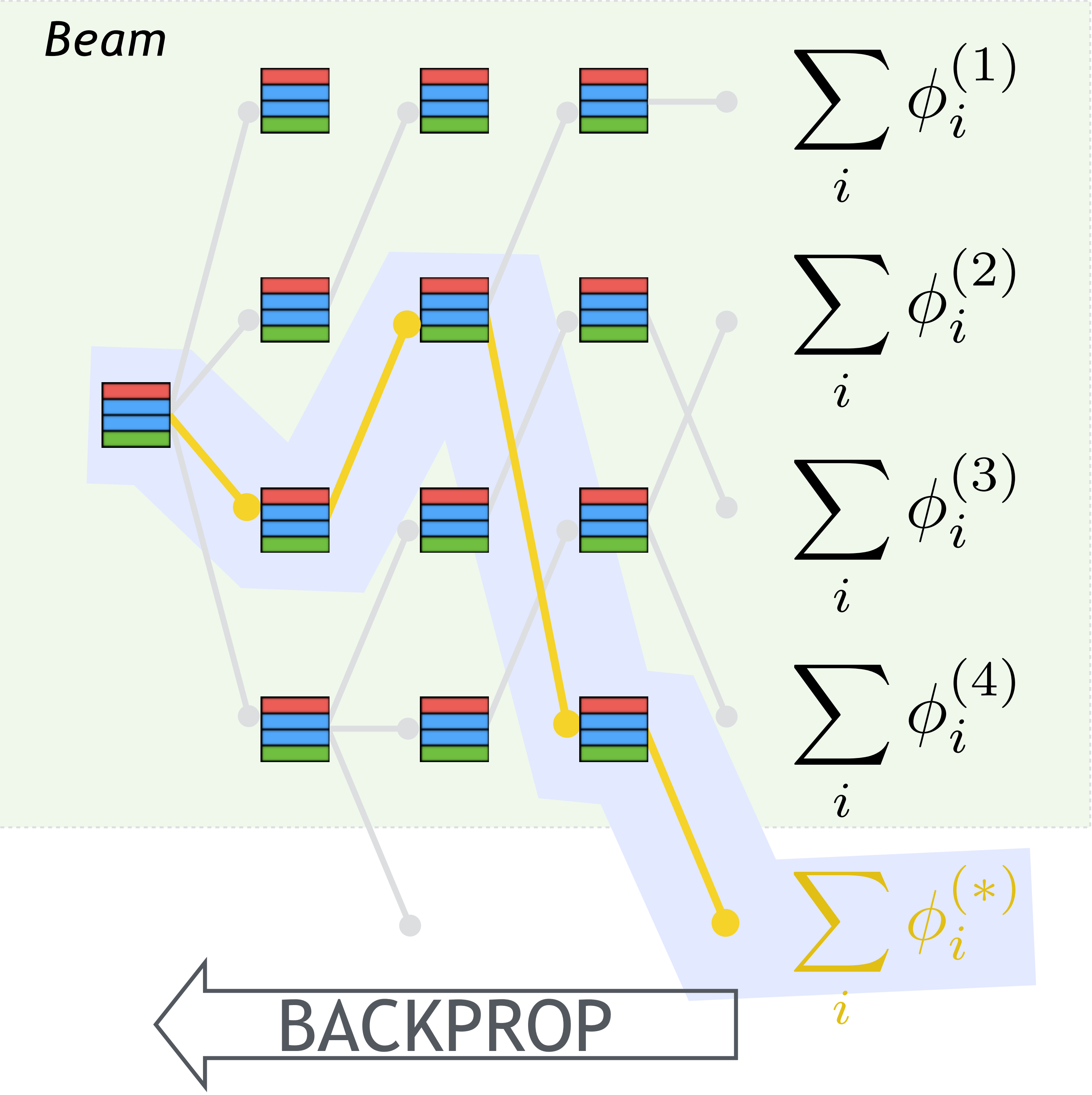
Globally normalized with respect to the beam:

$$\frac{\exp \sum_i \phi_i^{(*)}}{\sum_{j=1}^{|\mathrm{Beam}|} \exp \sum_i \phi_i^{(j)}}$$

Backpropagate through all steps, paths, and layers

[Collins and Roark ’04, Zhou et al.’15]

Training with Early Updates



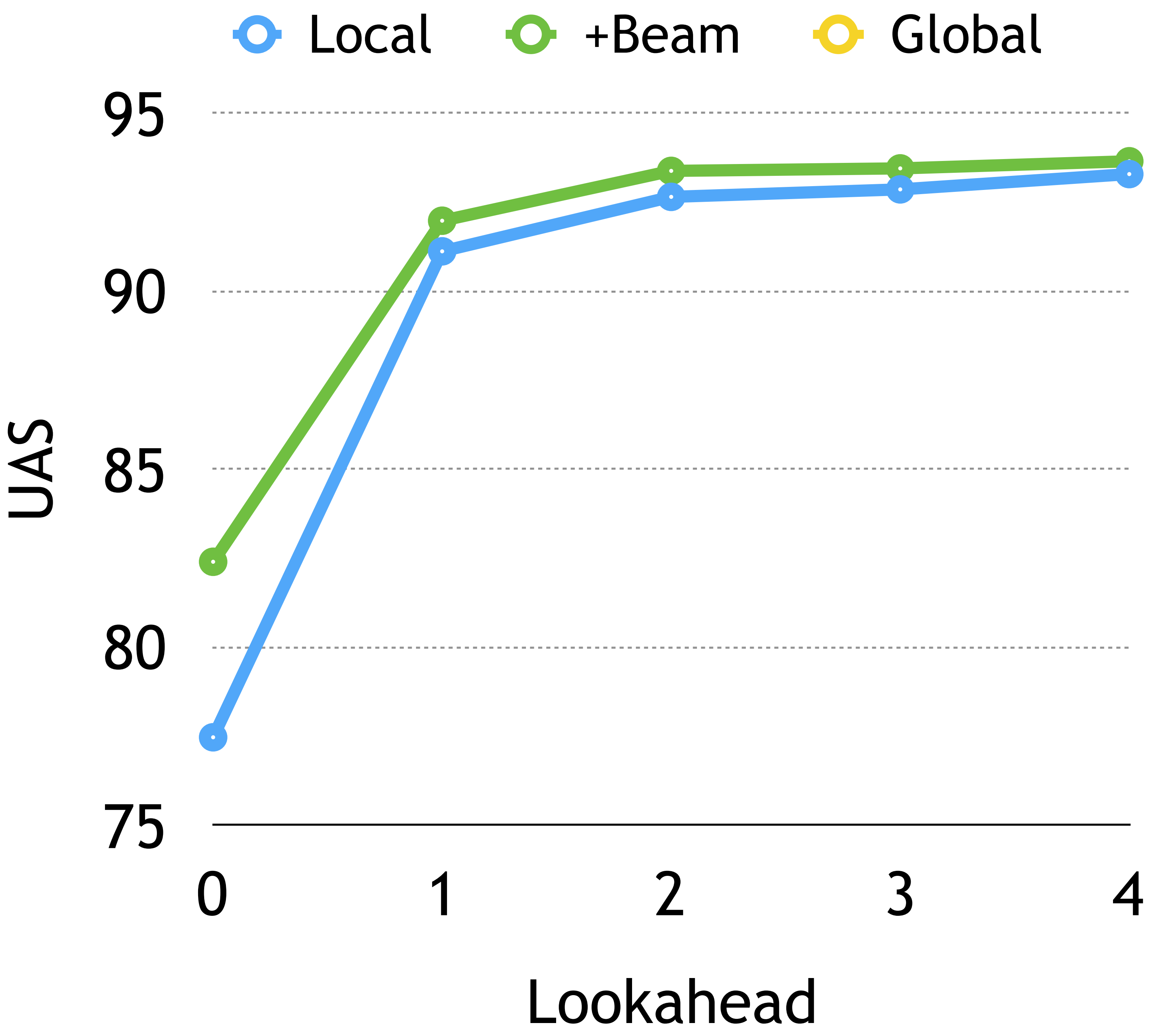
Globally normalized with respect to the beam:

$$\frac{\exp \sum_i \phi_i^{(*)}}{\sum_{j=1}^{|\text{Beam}|} \exp \sum_i \phi_i^{(j)}}$$

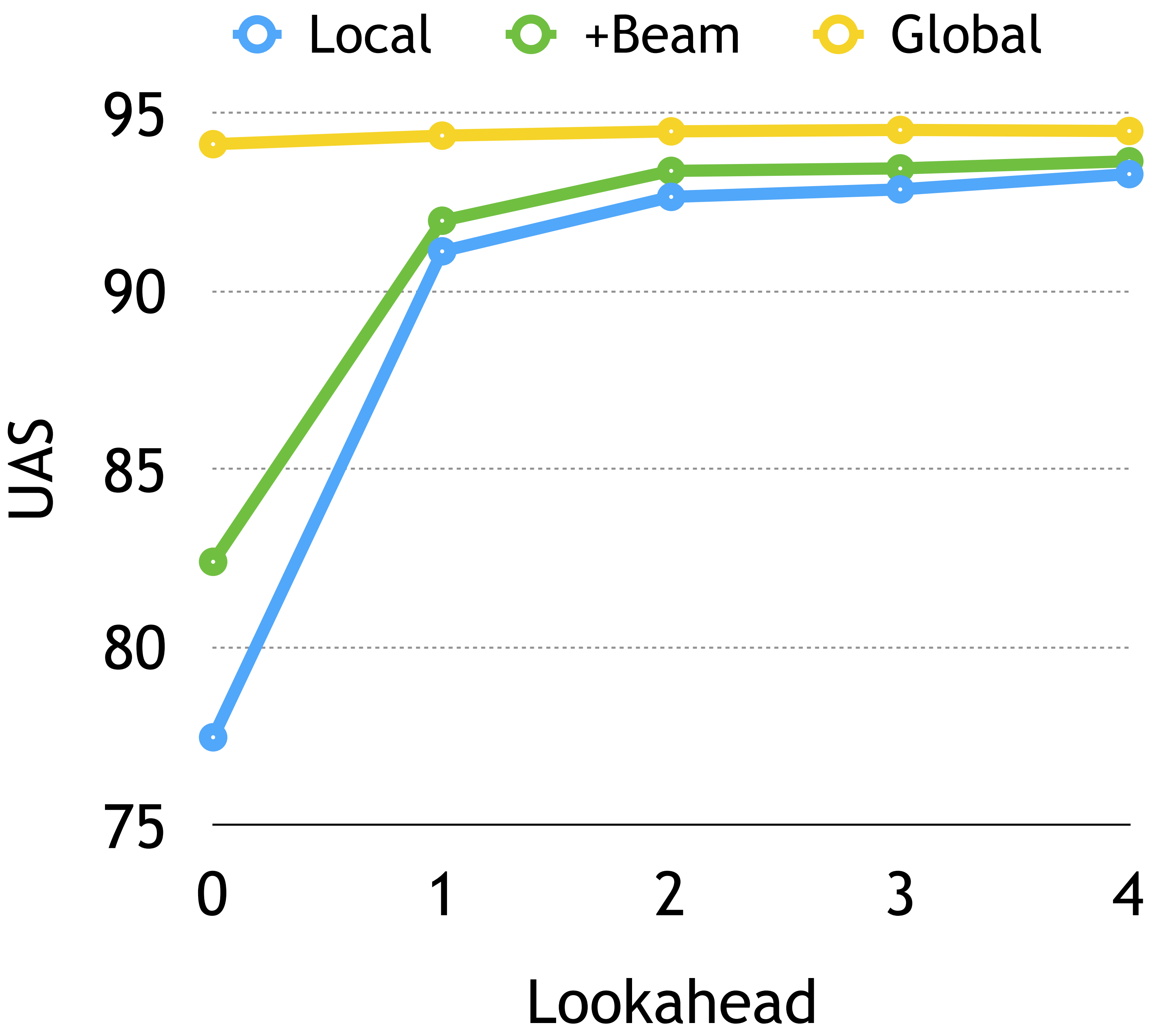
Backpropagate through all steps, paths, and layers

[Collins and Roark ’04, Zhou et al.’15]

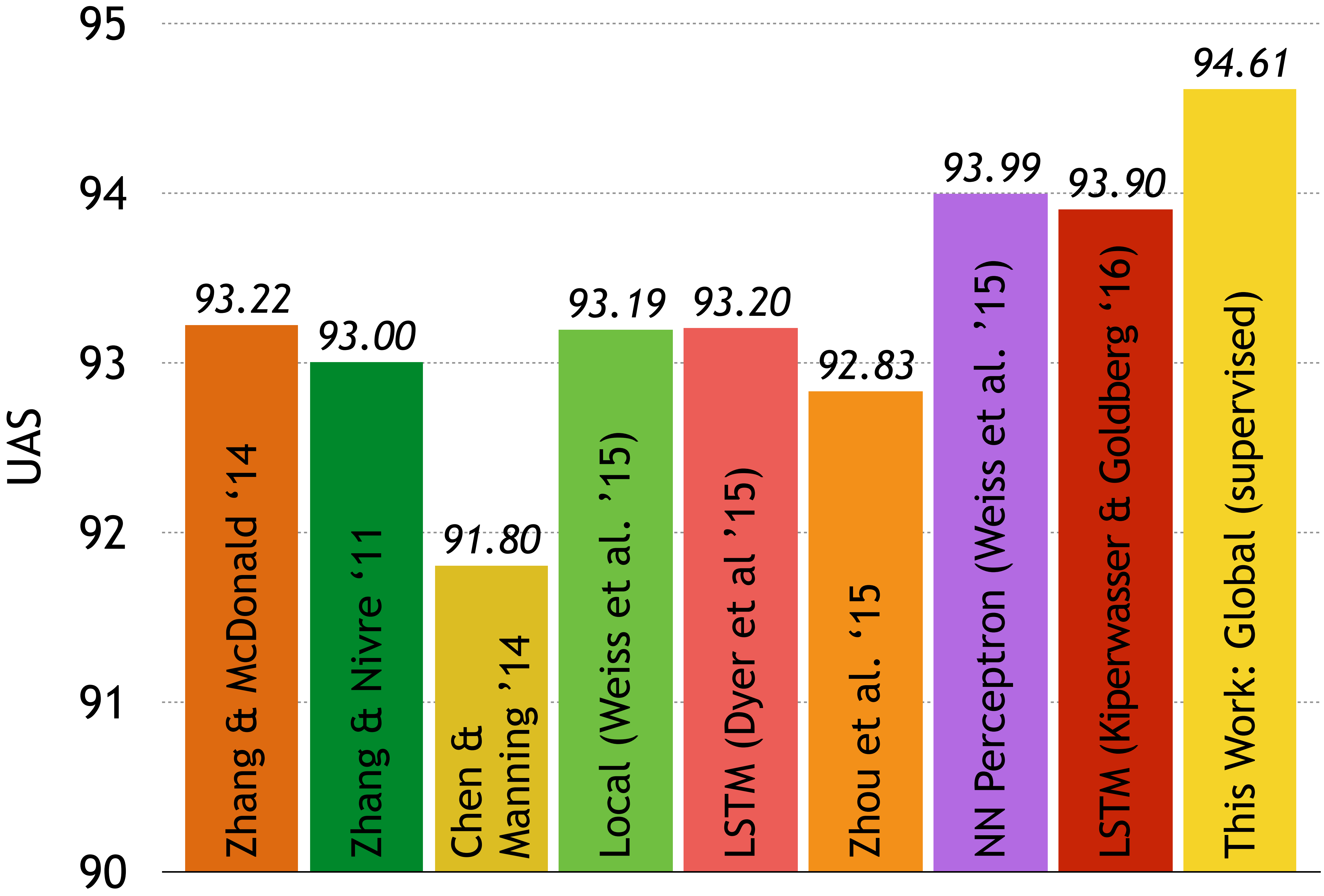
Globally Normalized Model



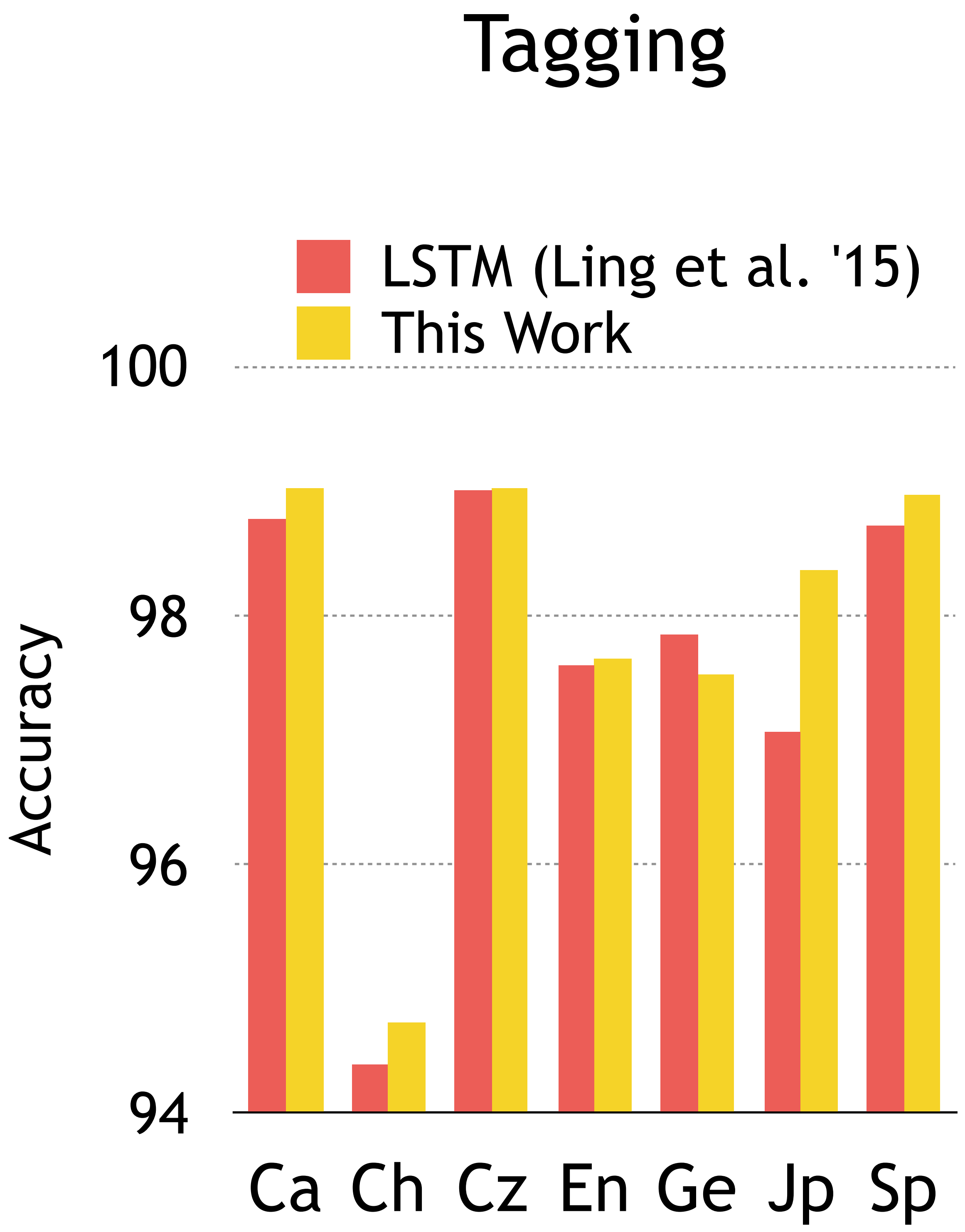
Globally Normalized Model

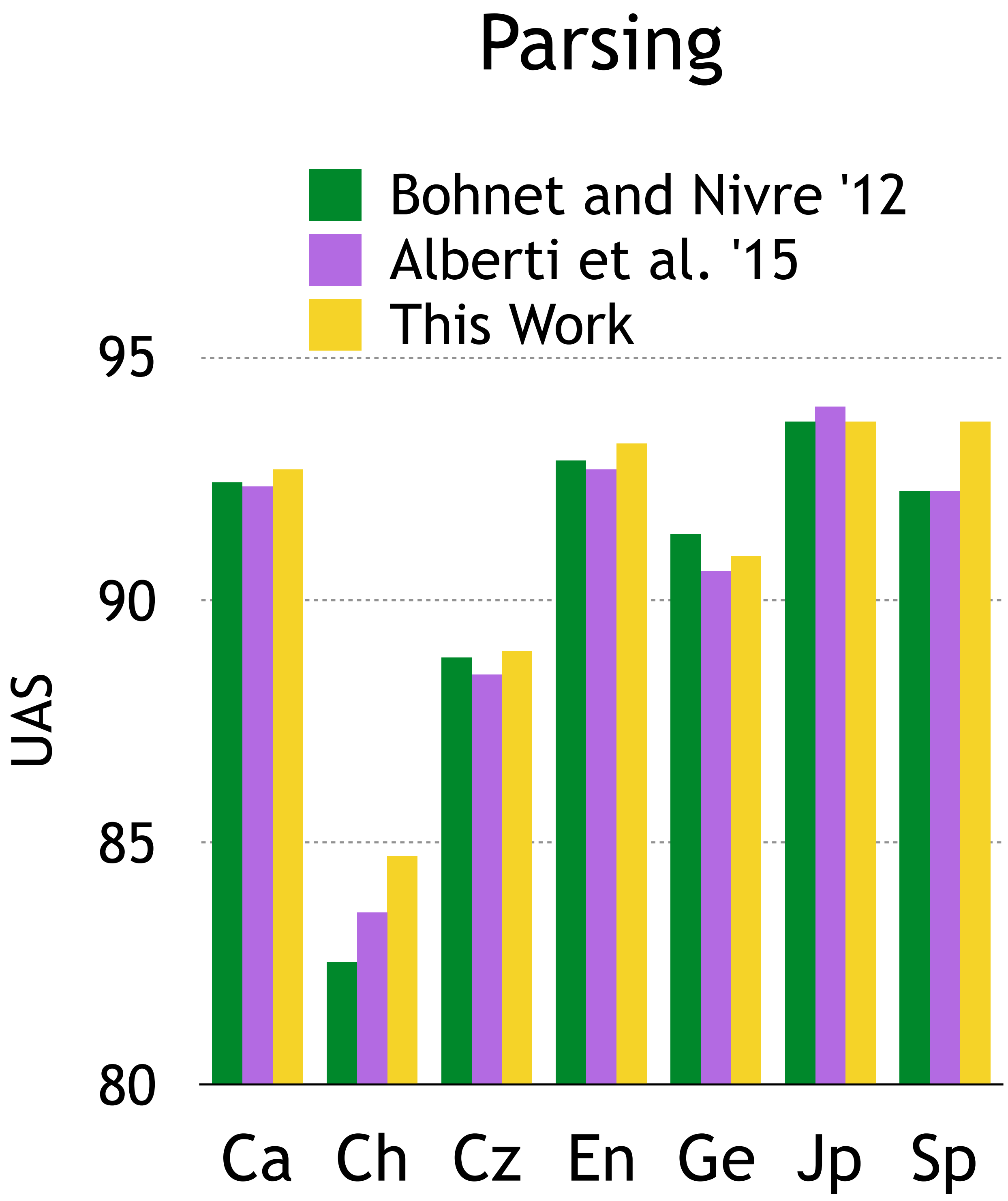


English WSJ Results



CoNLL’09 POS Tagging and Parsing Results





Sentence Compression Results

In Pakistan, former leader Pervez Musharraf has appeared in court for the first time, on treason charges.

Sentence Compression Results

In Pakistan, former leader Pervez Musharraf has appeared in court for the first time, on treason charges.

Transition System decides to **KEEP** or **DROP** words

Sentence Compression Results

In Pakistan, former leader **Pervez Musharraf has appeared in court** for the first time, **on treason charges.**

Transition System decides to **KEEP** or **DROP** words

Sentence Compression Results

Pervez Musharraf has appeared in court on treason charges.

Sentence Compression Results

`Pervez Musharraf has appeared in court on treason charges.`

	Seq2seq LSTM (Filippova et al. ’15)	Global model (This work)
Whole-sentence test accuracy	35.36	35.16
Human eval rating	4.66	4.67
Relative throughput	1x	100x

Sentence Compression Results

`Pervez Musharraf has appeared in court on treason charges.`

	Seq2seq LSTM (Filippova et al. ’15)	Global model (This work)
Whole-sentence test accuracy	35.36	35.16
Human eval rating	4.66	4.67
Relative throughput	1x	100x

Sentence Compression: Label Bias

	Predicted compression	Sequence probability under	
		Local	Global
Local	In Pakistan, former leader **Pervez Musharraf has appeared in court** for the first time, on treason charges**.**	0.13	0.05
+Beam	In Pakistan, former leader Pervez Musharraf has appeared in court for the first time, on treason charges**.**	0.16	<10^{-4}
Global	In Pakistan, former leader **Pervez Musharraf has appeared in court** for the first time, **on treason charges.**	0.06	0.07

Why does it work?

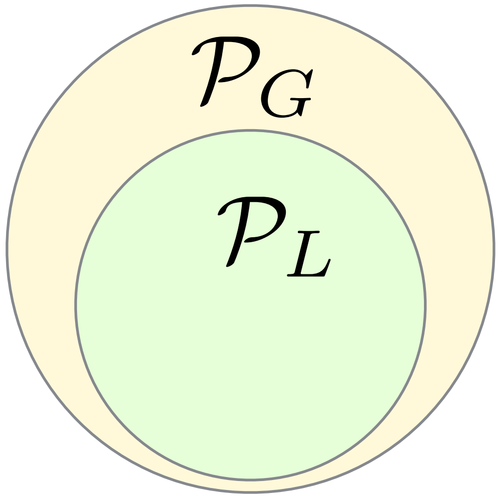
1. Global Models are More Expressive

Let

- $\mathcal{P}_L$ set of distributions under a Local model
- $\mathcal{P}_G$ set of distributions under a Global model

Theorem: $\mathcal{P}_L \subsetneq \mathcal{P}_G$

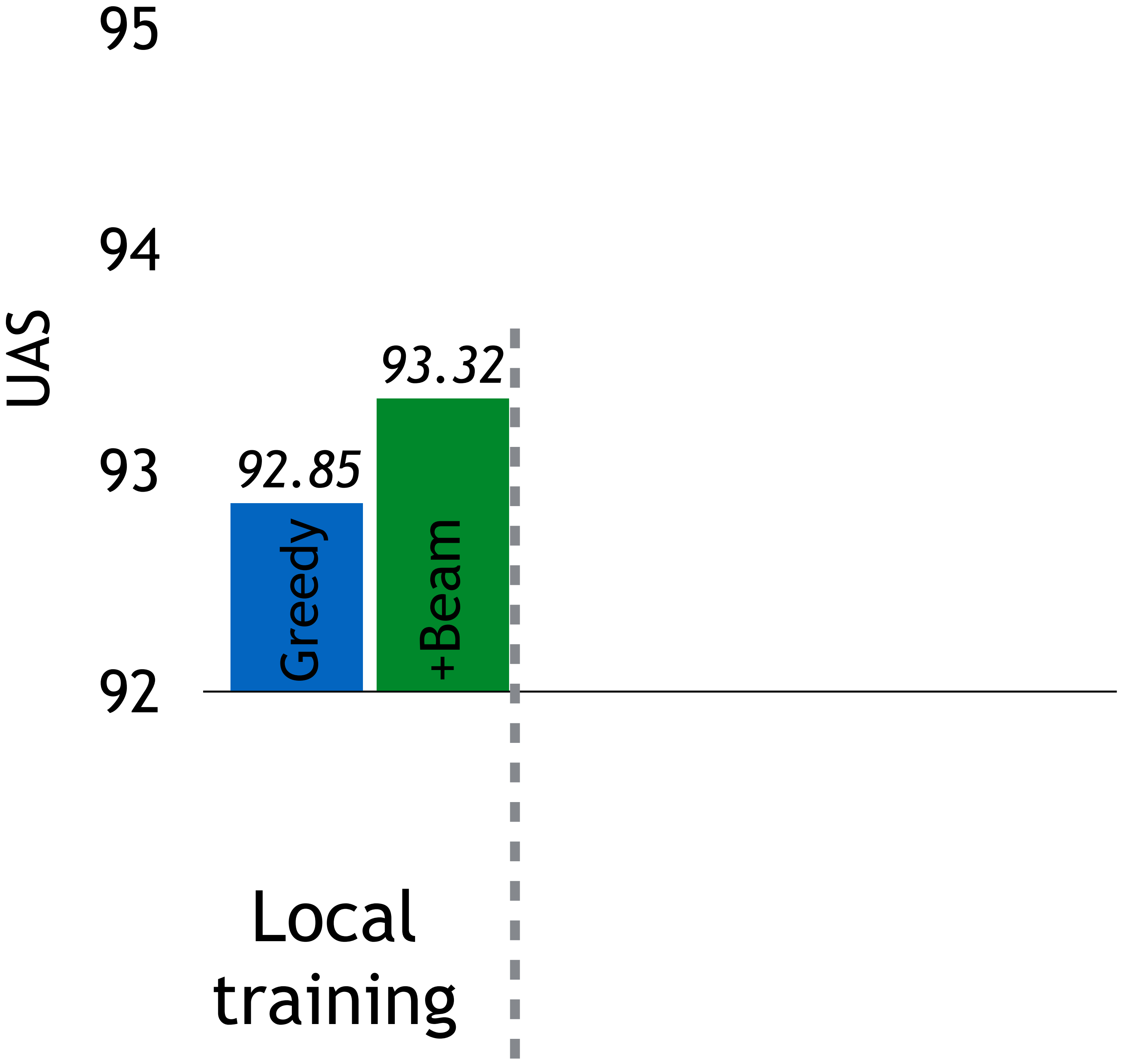
Therefore there are some distributions over sequences that cannot be captured in a finite-lookahead locally-normalized model.



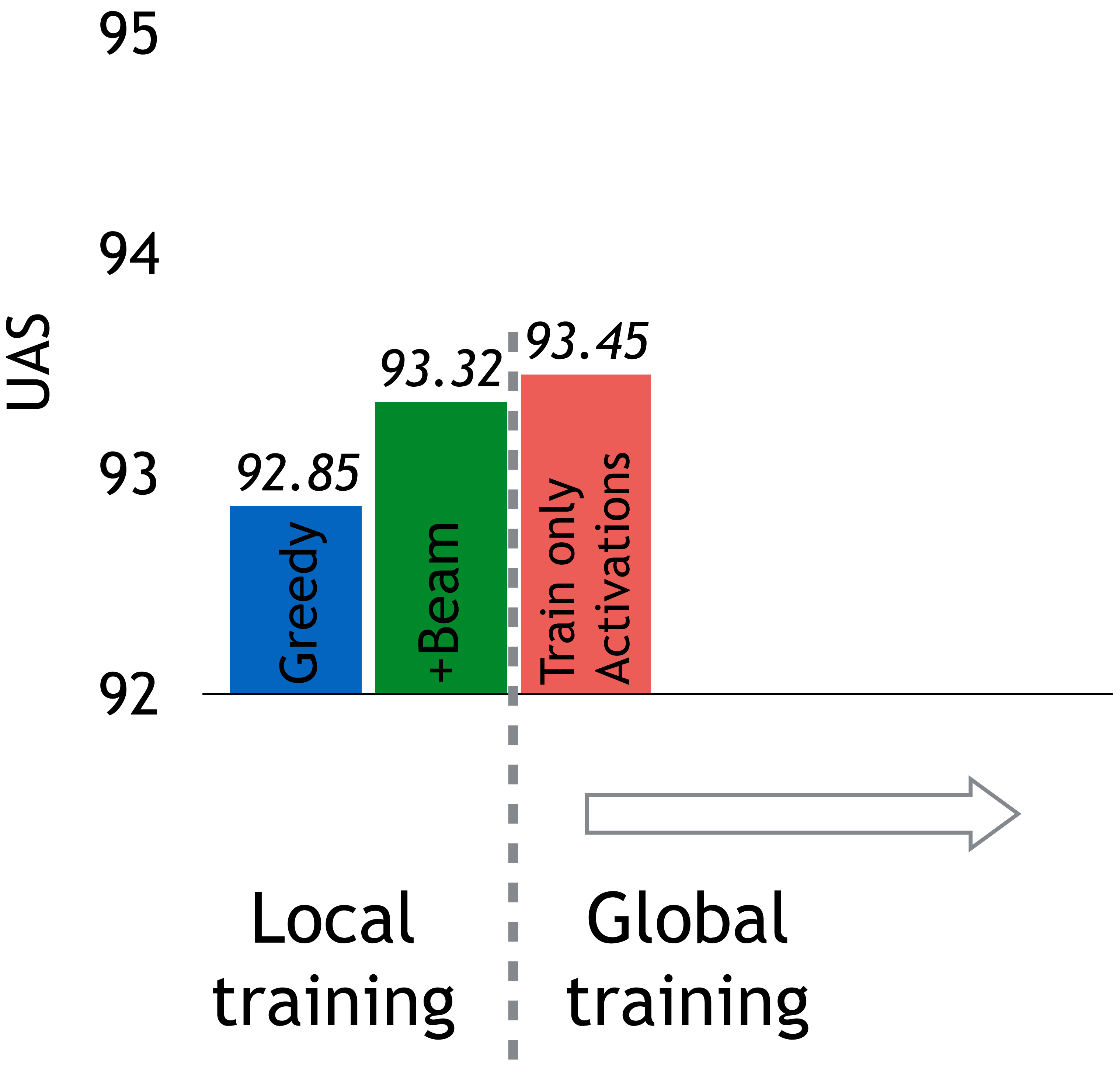
[This work, Smith and Johnson ’07]

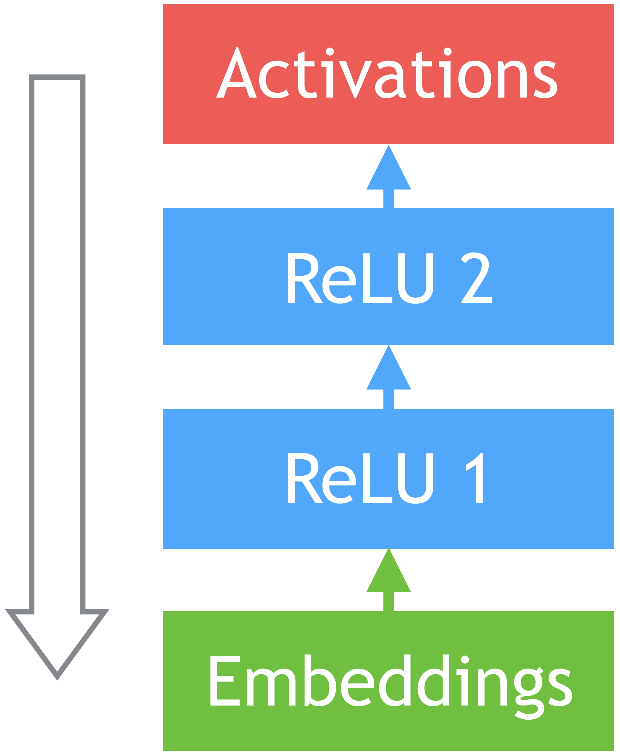
2. Backprop with a Beam

2. Backprop with a Beam

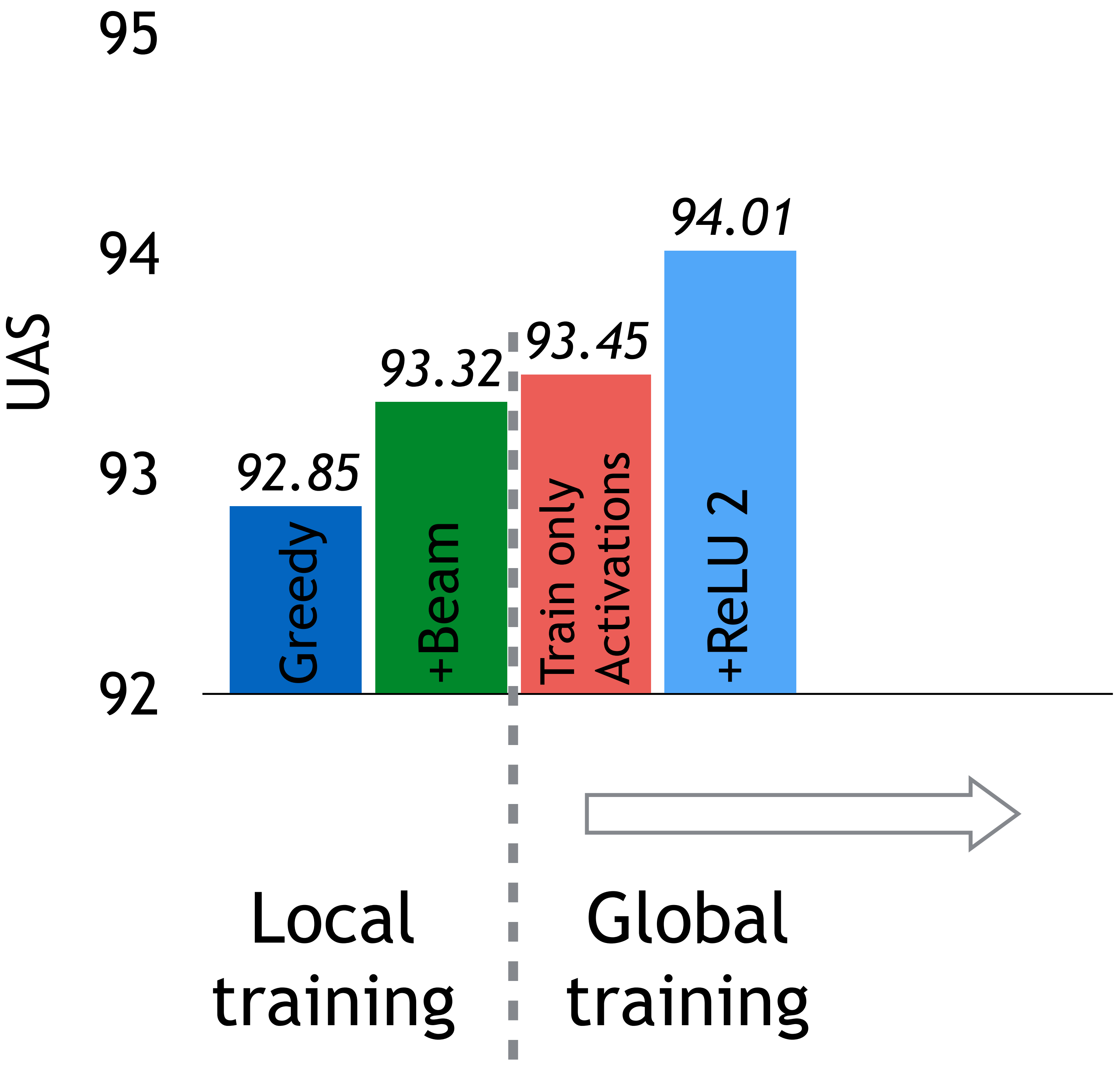


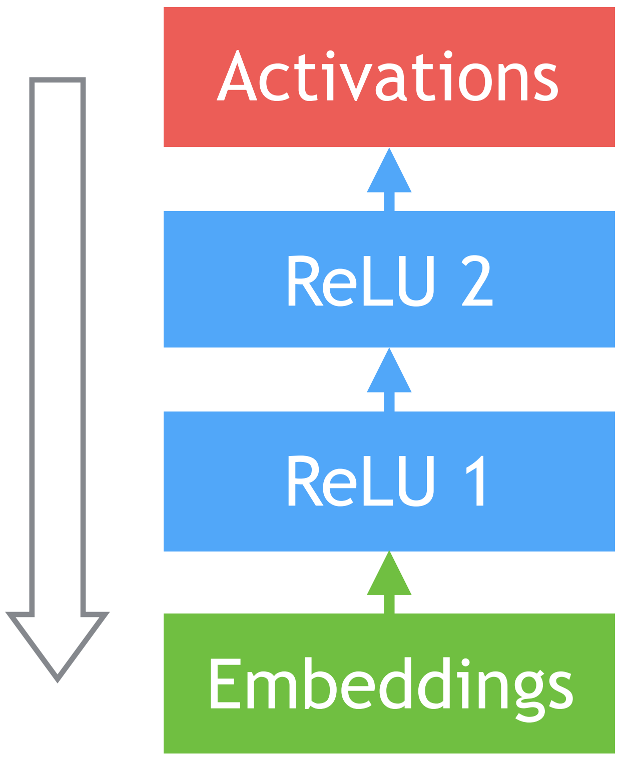
2. Backprop with a Beam



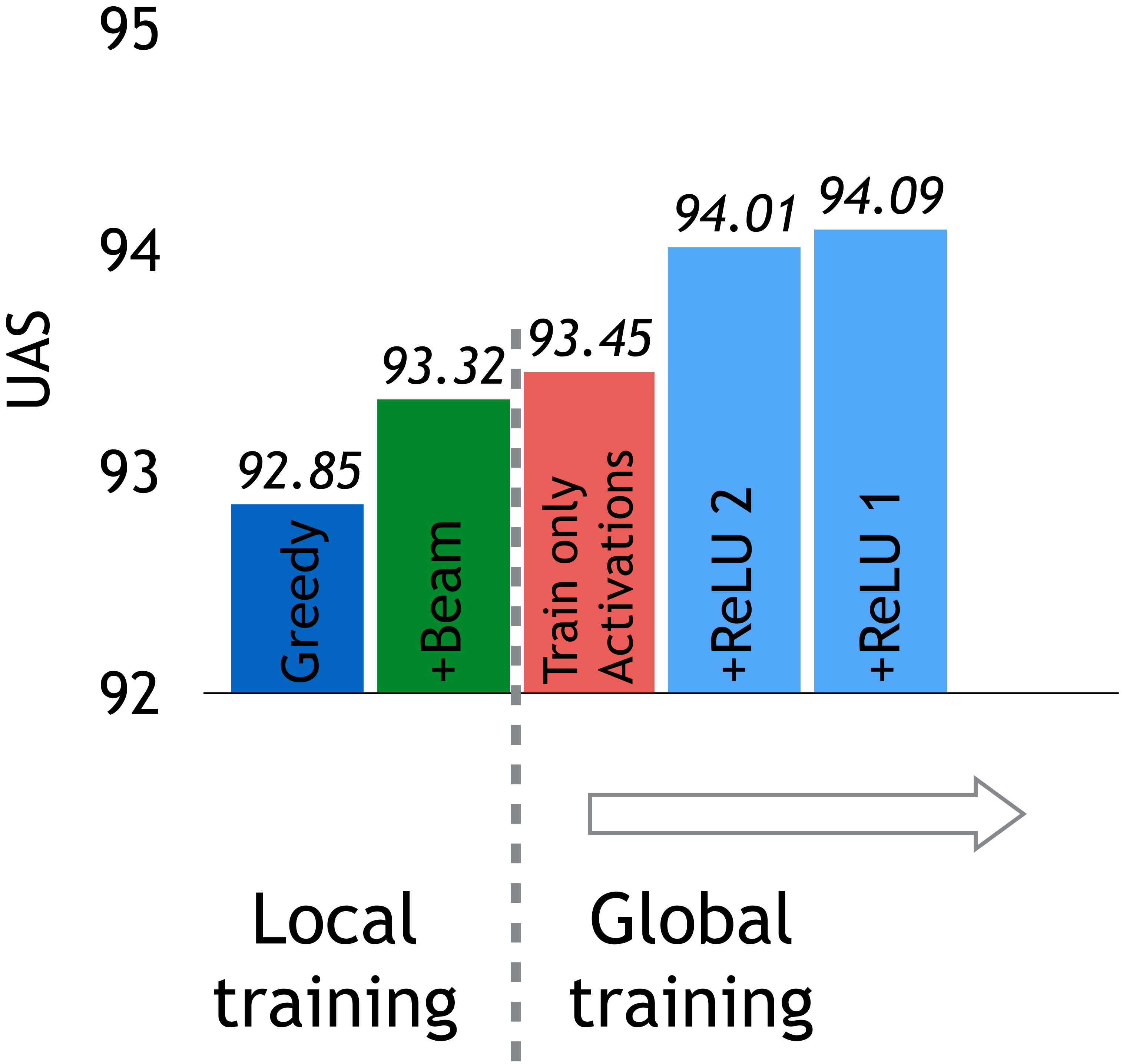


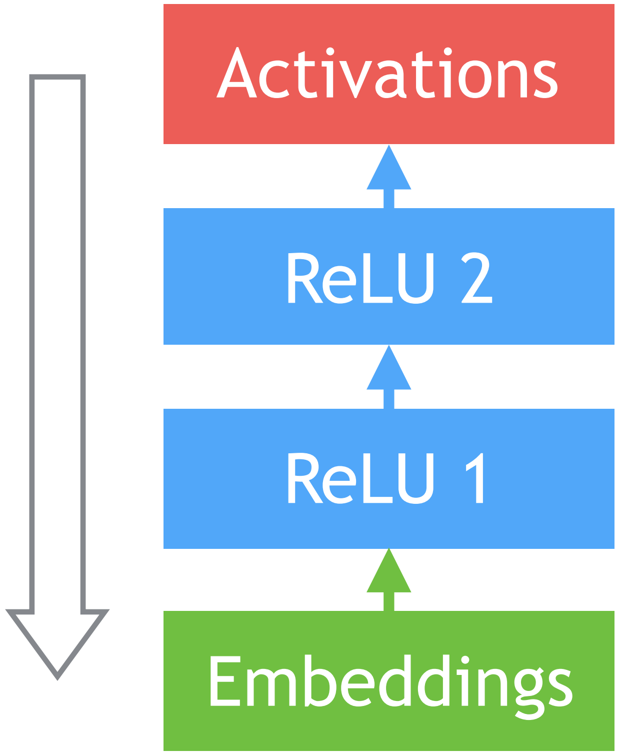
2. Backprop with a Beam



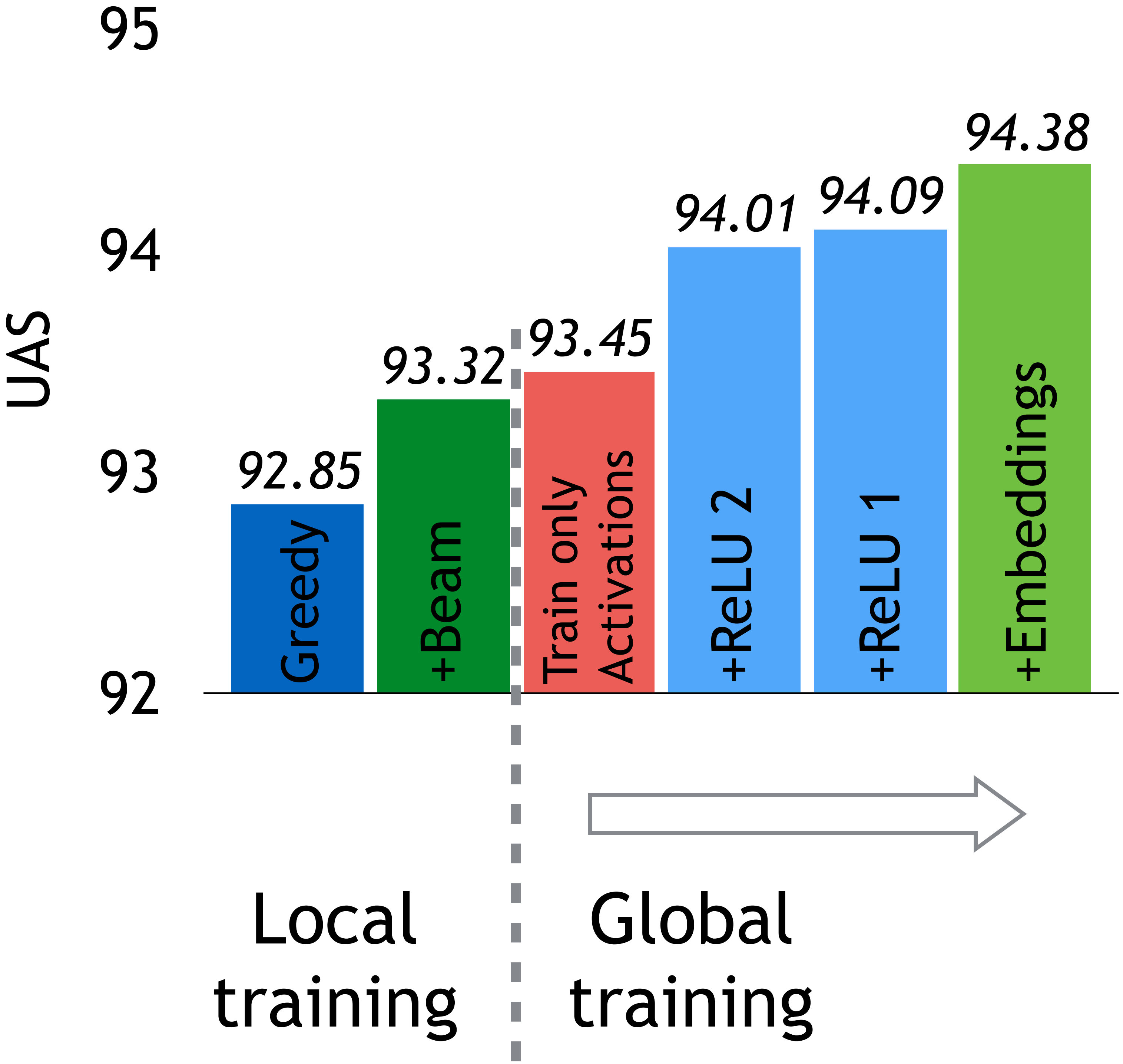


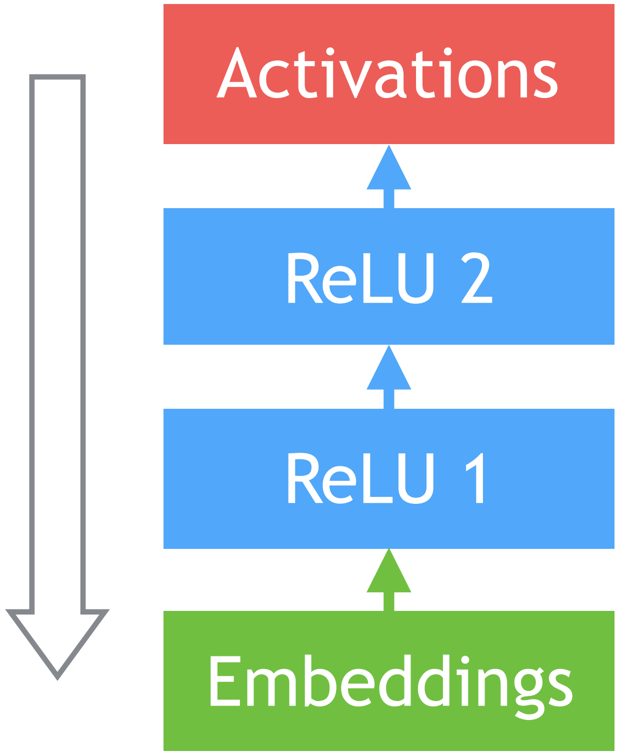
2. Backprop with a Beam





2. Backprop with a Beam





Conclusions

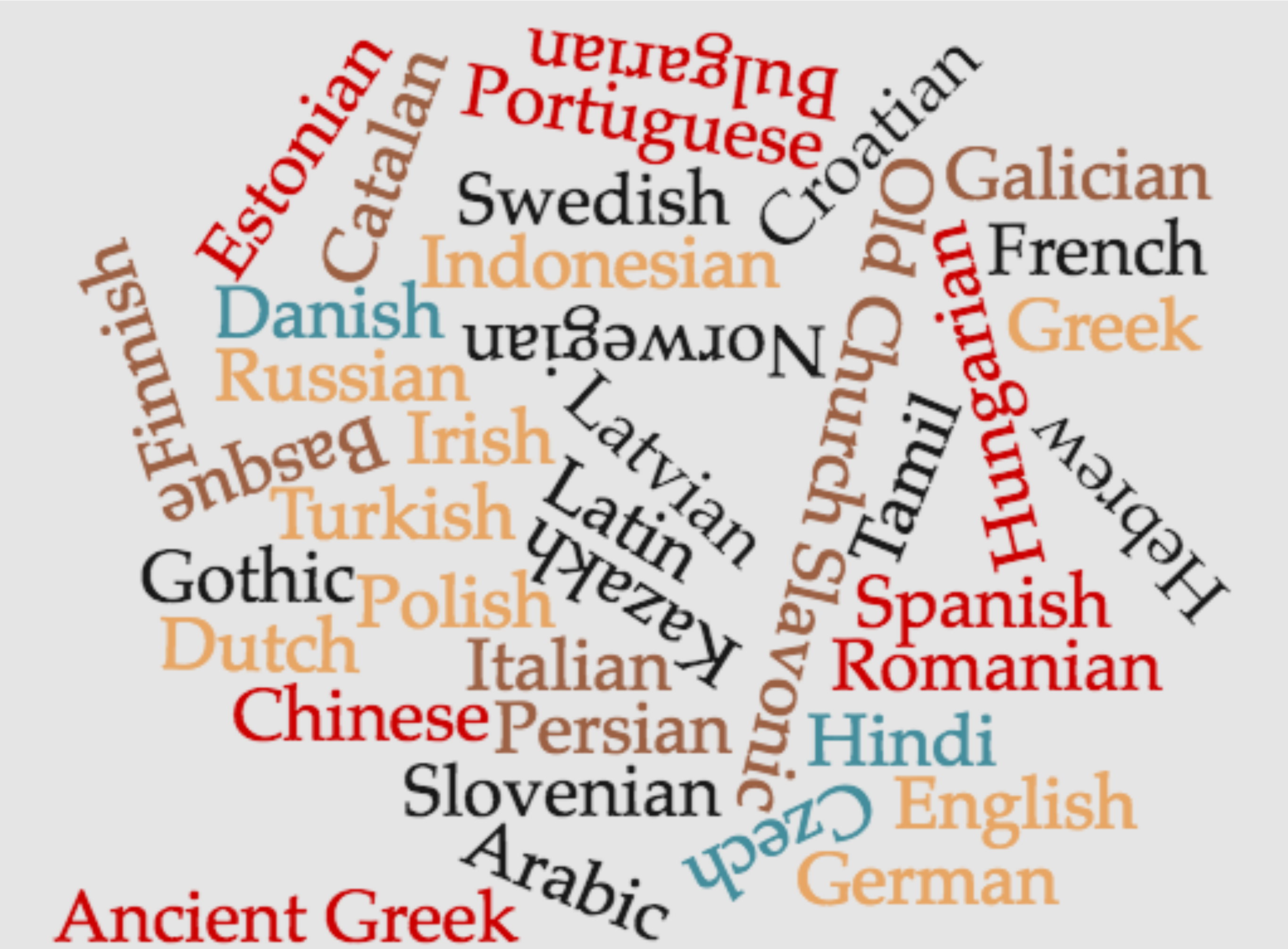
Global models:

- can be taught to do search better
- more accurate, in exchange for more training time
- same wicked fast decoding
- applicable to multiple tasks

Open Source: SyntaxNet

Parsey McParseface + 40 languages

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



ACL 2016 Google Booth

Come by for **demos**, info and swag

And check out the
Natural Language Understanding
team page: **`g.co/NLUTeam`**

Thank You!

[Do and Artires ’10]
[Filippova et al.’15]
[Goldberg and Nivre ’13]
[Hochreiter and Schmidhuber ’97]
[Huang et al.’15]

[Henderson ’03]
[Henderson ’04]
[Durrett and Klein ’15]
[Vinyals et al.’15]
[Watanabe and Sumita ’15]

[Bottou ’91]
[Bottou et al.’97]
[Lafferty et al.’01]
[Bottou and LeCun ’05]
[Le Cun et al.’98]

[Lei et al.’14]
[Ling et al.’15]
[Peng et al.’09]

[Collins and Roark ’04]
[Collins ’99]
[Liang et al.’08]
[Daume III et al.’09]

[Abney et al.’99]
[Chi ’99]
[Smith and Johnson ’07]

[Ross et al.’11]
[Yao et al.’14]
[Zheng et al.’15]
[Zhou and Xu’15]

[Nivre ’06]
[Nivre ’09]
[Bohnet and Nivre ’12]
[Martins et al.’13]
[Chen and Manning ’14]
[Zhang and McDonald ’14]
[Alberti et al.’15]
[Ballesteros et al.’15]
[Dyer et al.’15]
[Weiss et al.’15]
[Yazdani and Henderson ’15]
[Zhou et al.’15]
[Vaswani and Sagae ’16]

Appendix

Longer examples of ambiguity