Neural Machine Translation²
CMPT 413/825, Fall 2019

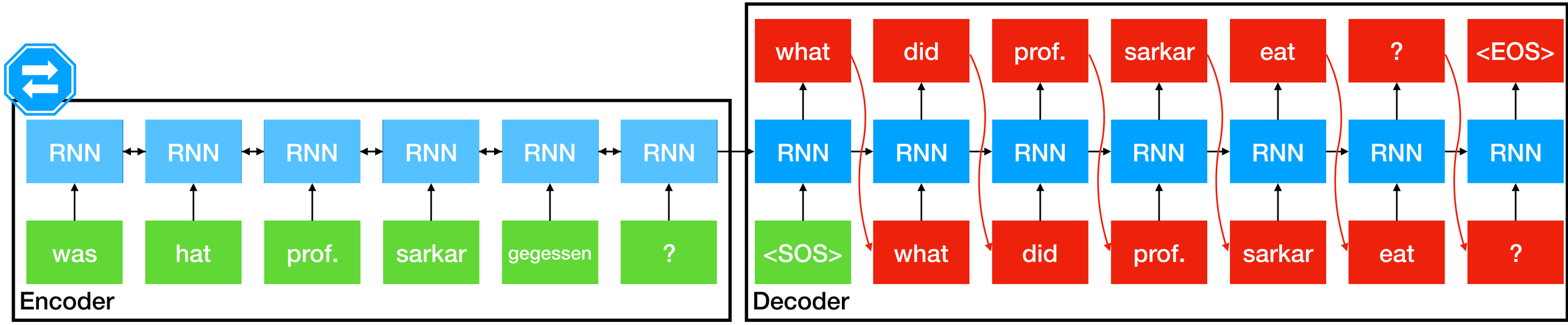
Jetic Gū
School of Computing Science
12 Nov. 2019

Overview

- Focus: Neural Machine Translation
- Architecture: Encoder-Decoder Neural Network
- Main Story:
  - Extension to Seq2Seq: Copy Mechanism
  - Extension to Seq2Seq: Ensemble
  - Extension to Seq2Seq: BeamSearch
  - [Extra] Beyond Seq2Seq: Attention is all you need
  - [Extra] Beyond NMT

P0
Review

Sequence-to-Sequence



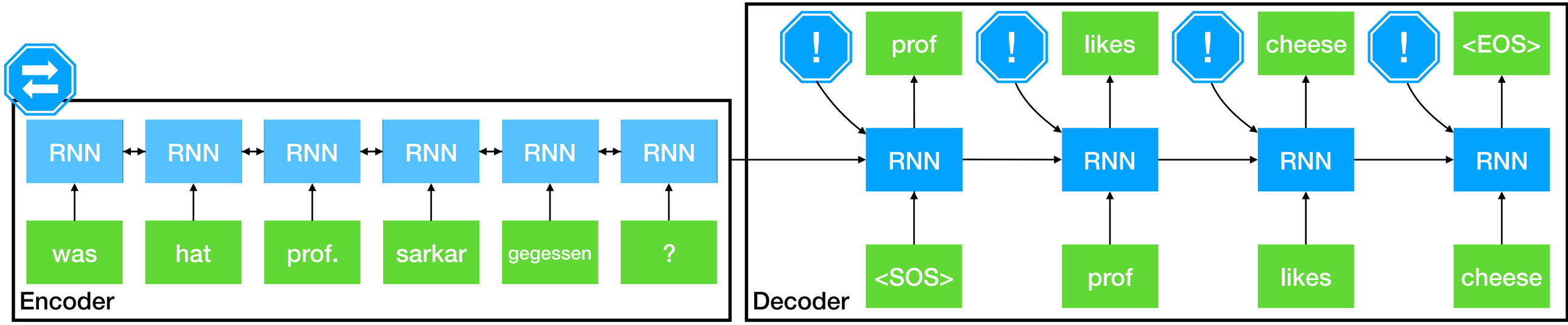
$$Pr(E \mid F) = \Pi_t Pr(e' = e_t \mid F, e_{<t})$$

CLM

Review

P0
Review

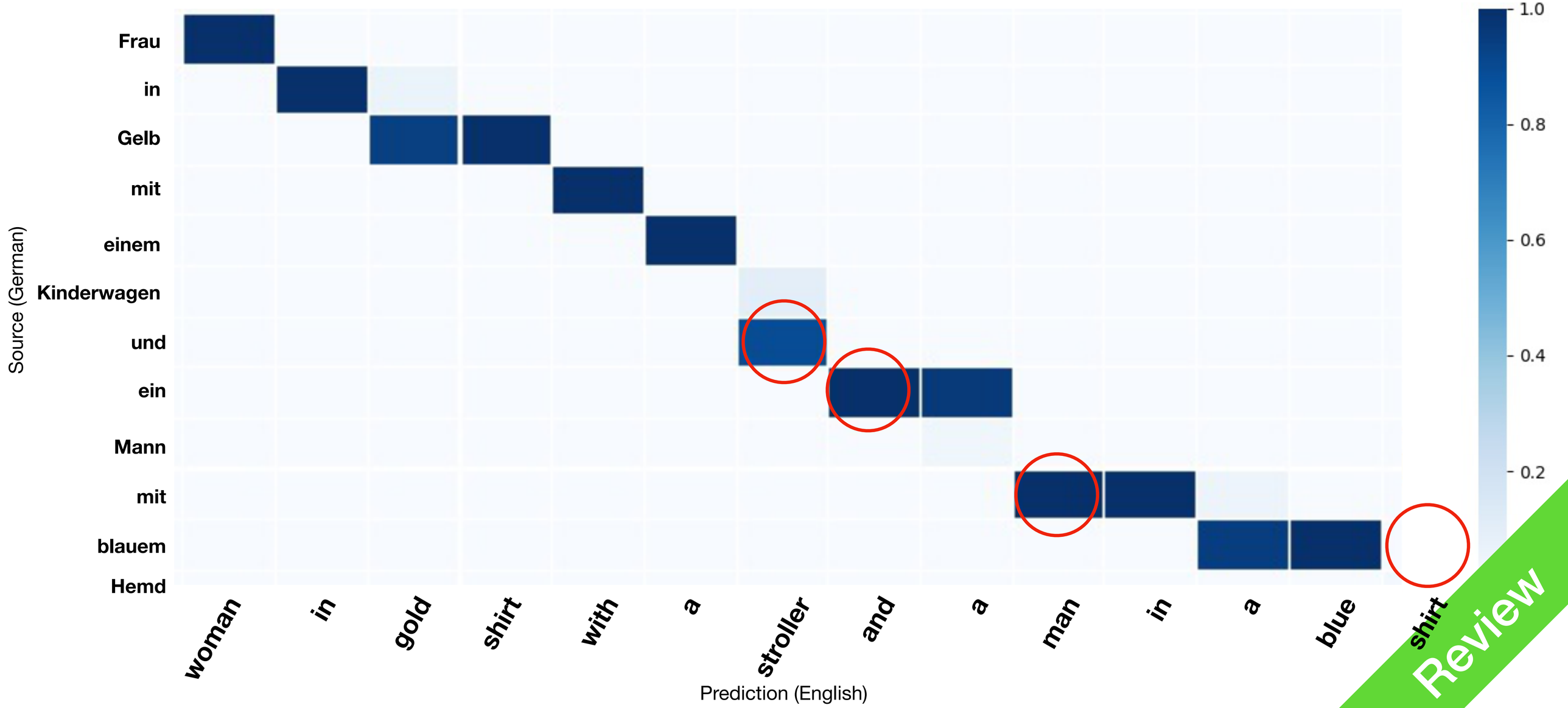
Attention



Review

P0
Review

Attention



Review

P0
Review

Attention

- Why does Attention work?
- What is in the context vector?
- What is in α?
  - ~~It's alignment~~[1] !
  - learns to refer to useful information in src[2]
  - similar to human attention: we pay attention to whatever is needed

$$score_{t,i} = f(h_i^{enc}, h_t^{dec})$$

$$\alpha_t = \text{softmax}(score_{t_i})$$

$$context_t = \sum_i \alpha_{t,i} h_i^{enc}$$

Review

1. CL2015008 [Bahdanau et al.] Neural Machine Translation by Jointly Learning to Align and Translate
2. CL2017342 [Ghader et Monz] What does Attention in Neural Machine Translation Pay Attention to?

P0
Review

Common Problems of NMT

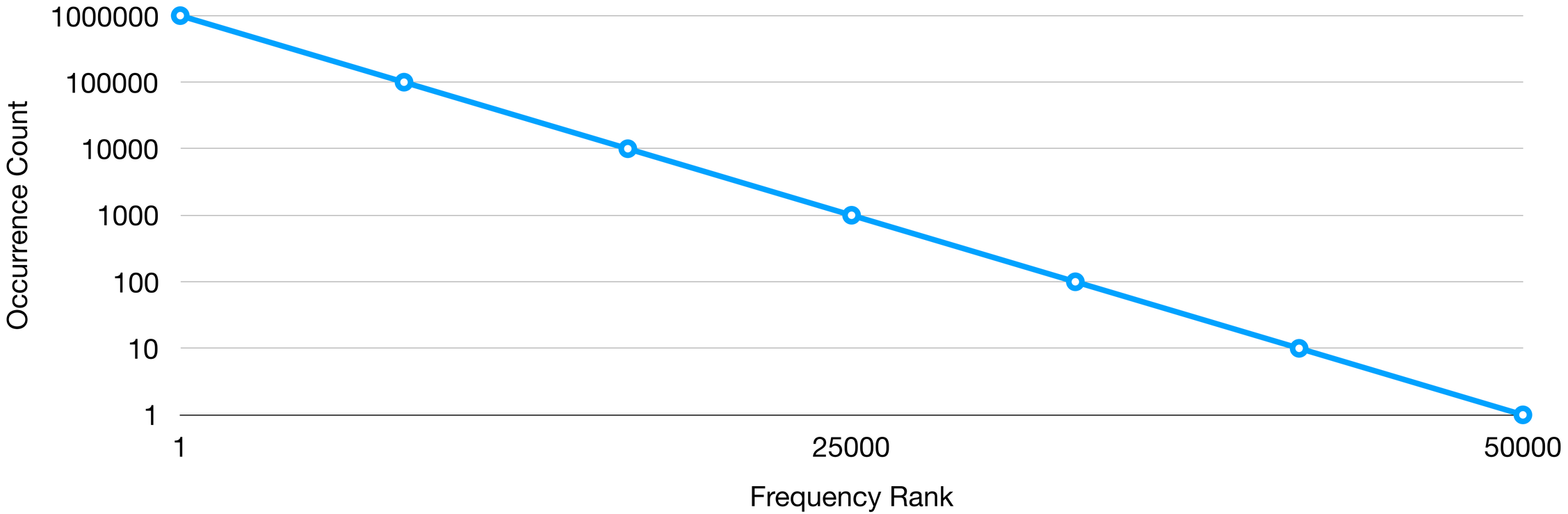
- Out-of-Vocabulary (OOV) Problem; Rare word problem
  - Frequent word are translated correctly, rare words are not
  - In any corpus, word frequencies are exponentially unbalanced

Review

P0
Review

OOV

Zipf's Law



- rare word are exponentially less frequent than frequent words
- e.g. in HW4, 45%|65% (src|tgt) of the unique words occur once

Review

1. https://en.wikipedia.org/wiki/Zipf%27s_law

P0
Review

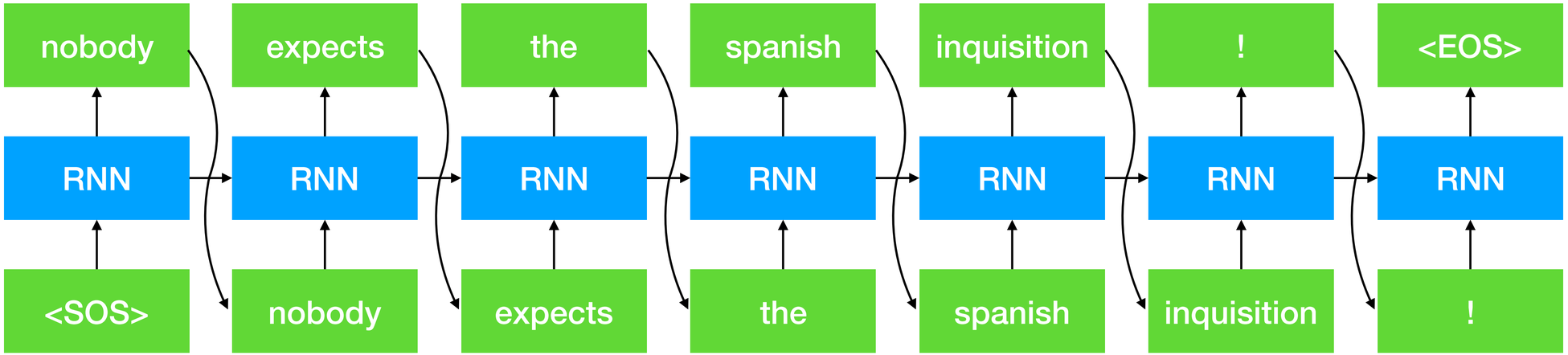
Common Problems of NMT

- Out-of-Vocabulary (OOV) Problem; Rare word problem
  - Frequent word are translated correctly, rare words are not
  - In any corpus, word frequencies are exponentially unbalanced
- Under translation
  - Crucial information are left untranslated; premature <EOS> generation

Review

P0
Review

Under-trans<EOS>

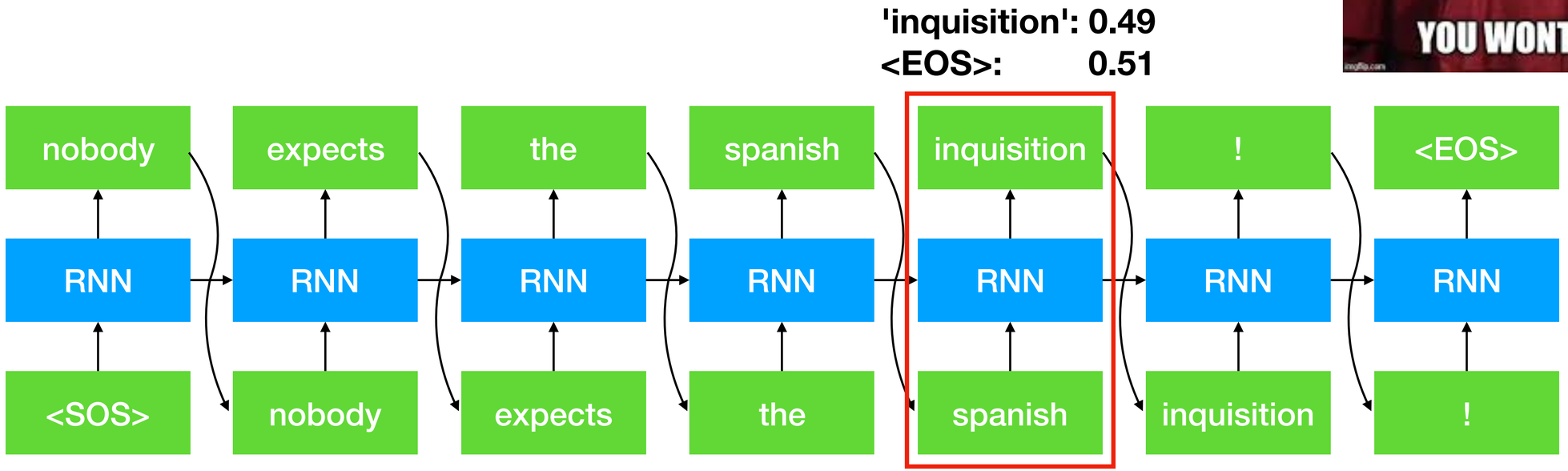


- Under translation
  - Crucial information are left untranslated; premature <EOS> generation

Review

P0
Review

Under-trans<EOS>

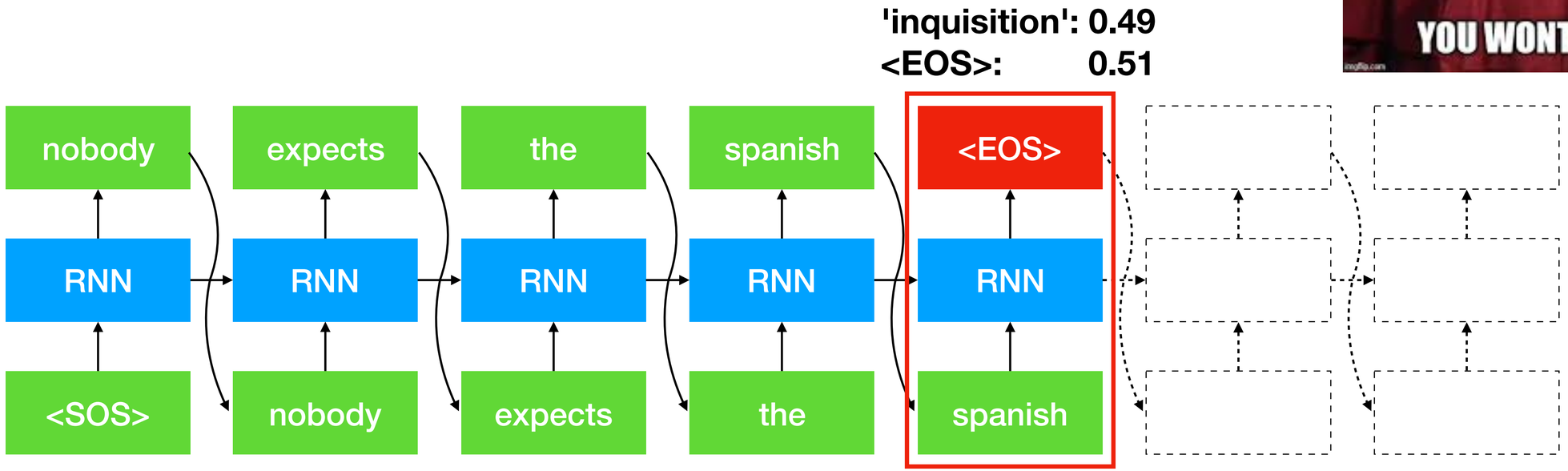


- Under translation
  - Crucial information are left untranslated; premature <EOS> generation

Review

P0
Review

Under-trans<EOS>



- Under translation
  - Crucial information are left untranslated; premature <EOS> generation

Review

P0
Review

Common Problems of NMT

- Out-of-Vocabulary (OOV) Problem; Rare word problem
  - Frequent word are translated correctly, rare words are not
  - In any corpus, word frequencies are exponentially unbalanced
- Under translation
  - Crucial information are left untranslated; premature <EOS> generation

Review

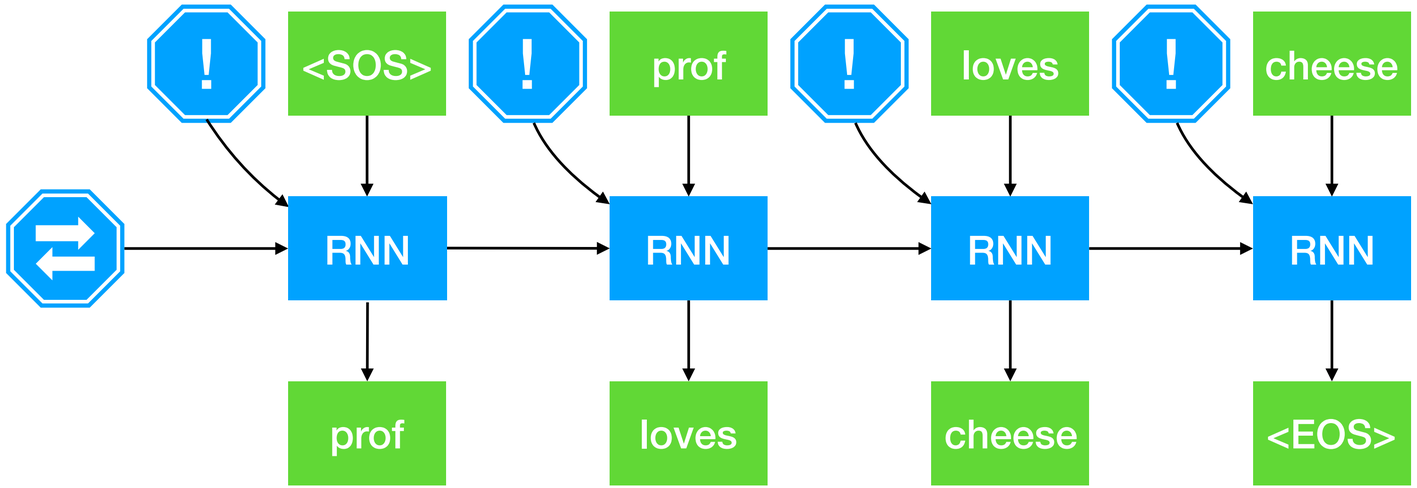
P0
Review

Common Problems of NMT

- Out-of-Vocabulary (OOV) Problem; Rare word problem
  - Copy Mechanisms
  - Char-level Encoder (oops for logogram, e.g. Chinese)
- Under translation
  - Ensemble
  - Beam search
  - Coverage models

Review

Copy Mechanism

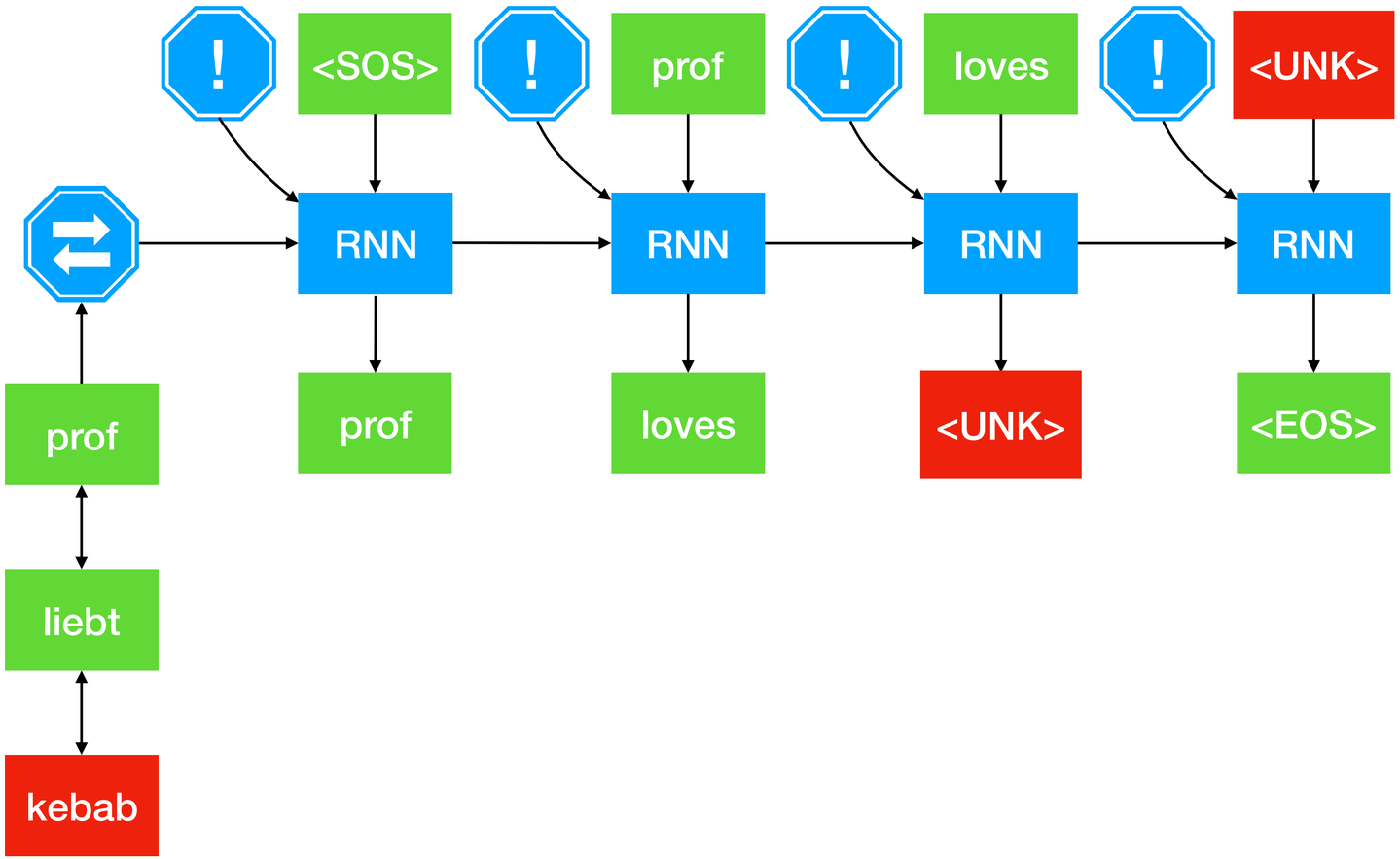


1. CL2015015 [Luong et al.] Addressing the Rare Word Problem in Neural Machine Translation

Concept

P1
Copy

Copy Mechanism

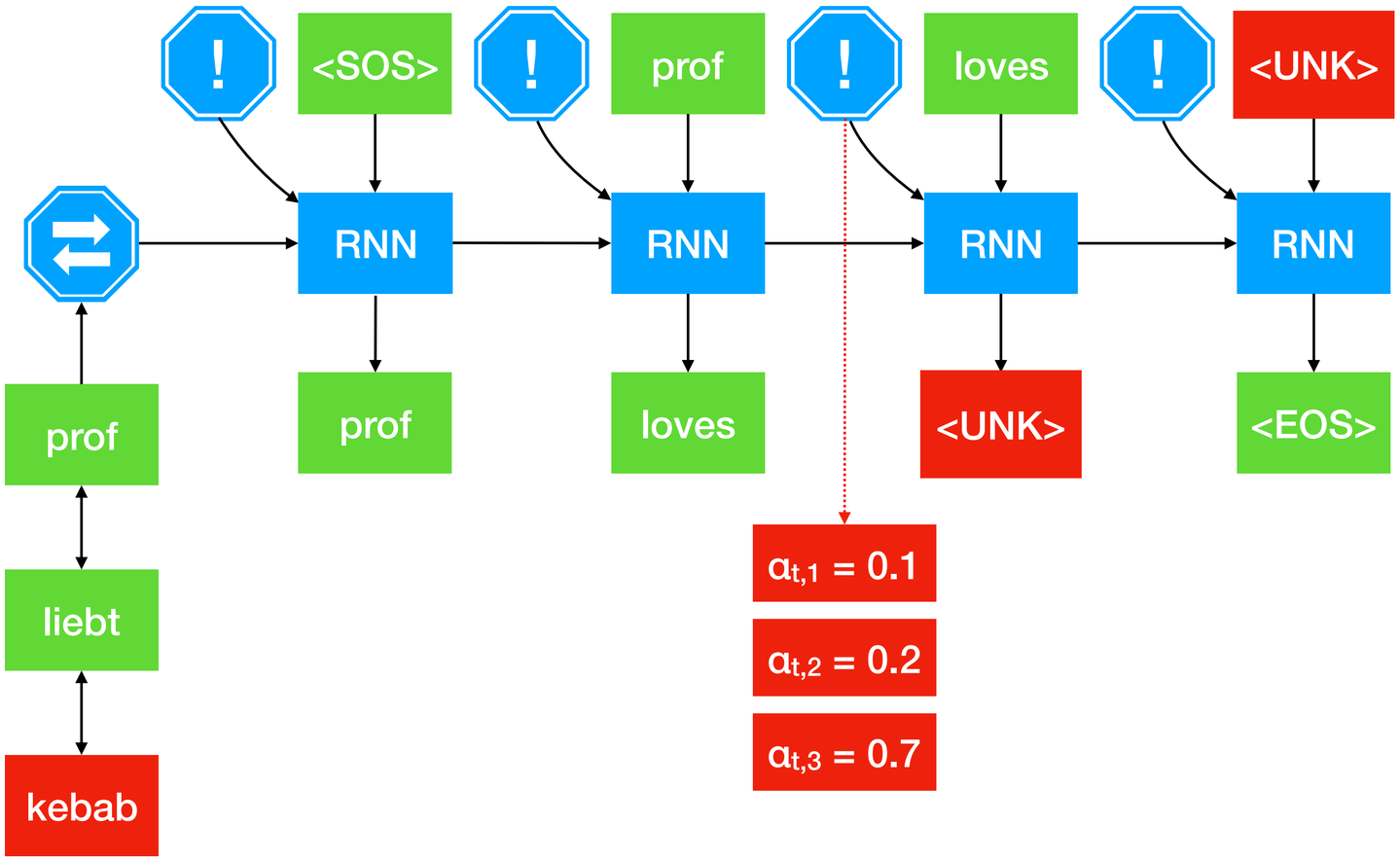


1. CL2015015 [Luong et al.] Addressing the Rare Word Problem in Neural Machine Translation

Concept

P1
Copy

Copy Mechanism

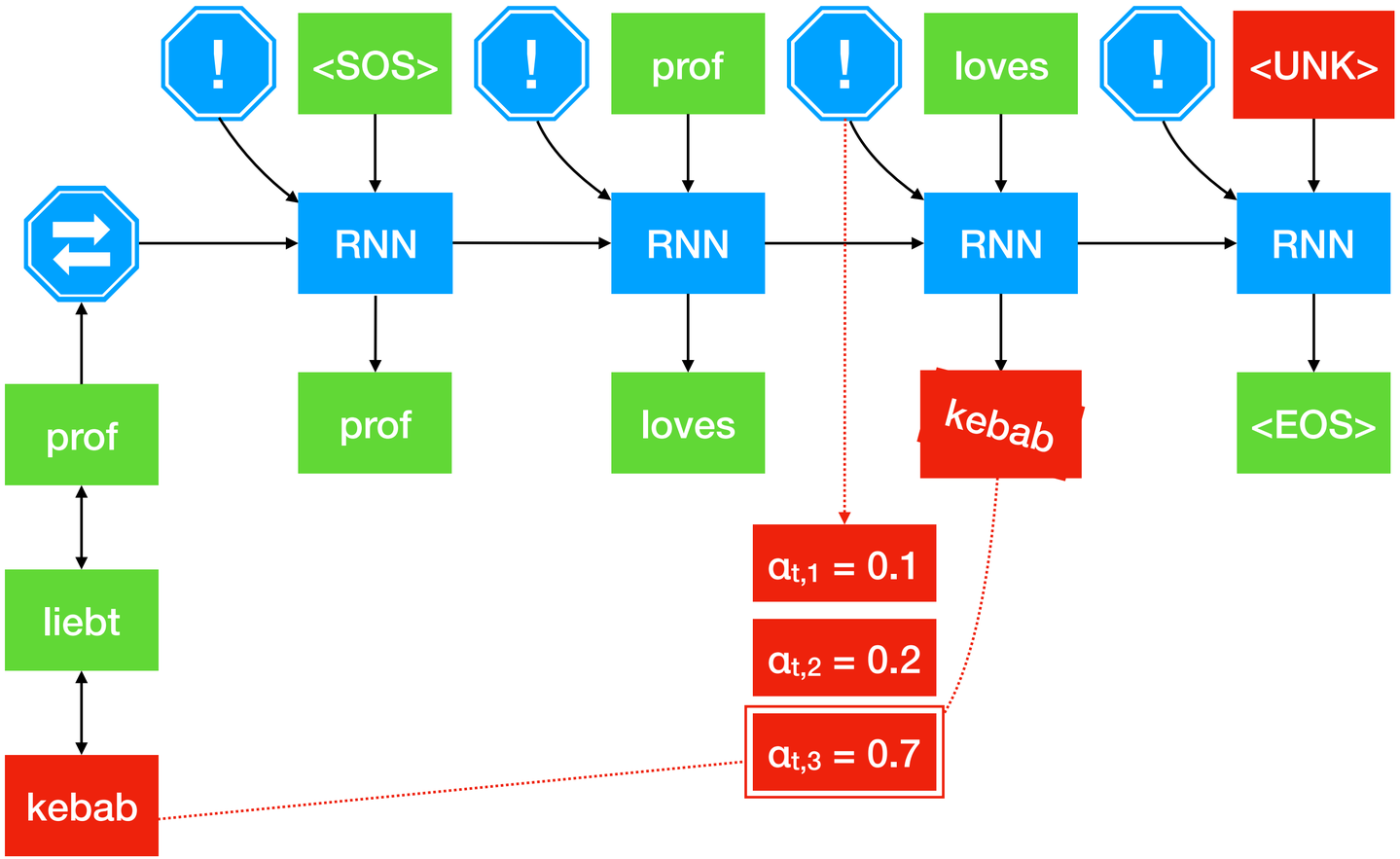


Concept

1. CL2015015 [Luong et al.] Addressing the Rare Word Problem in Neural Machine Translation

P1
Copy

Copy Mechanism

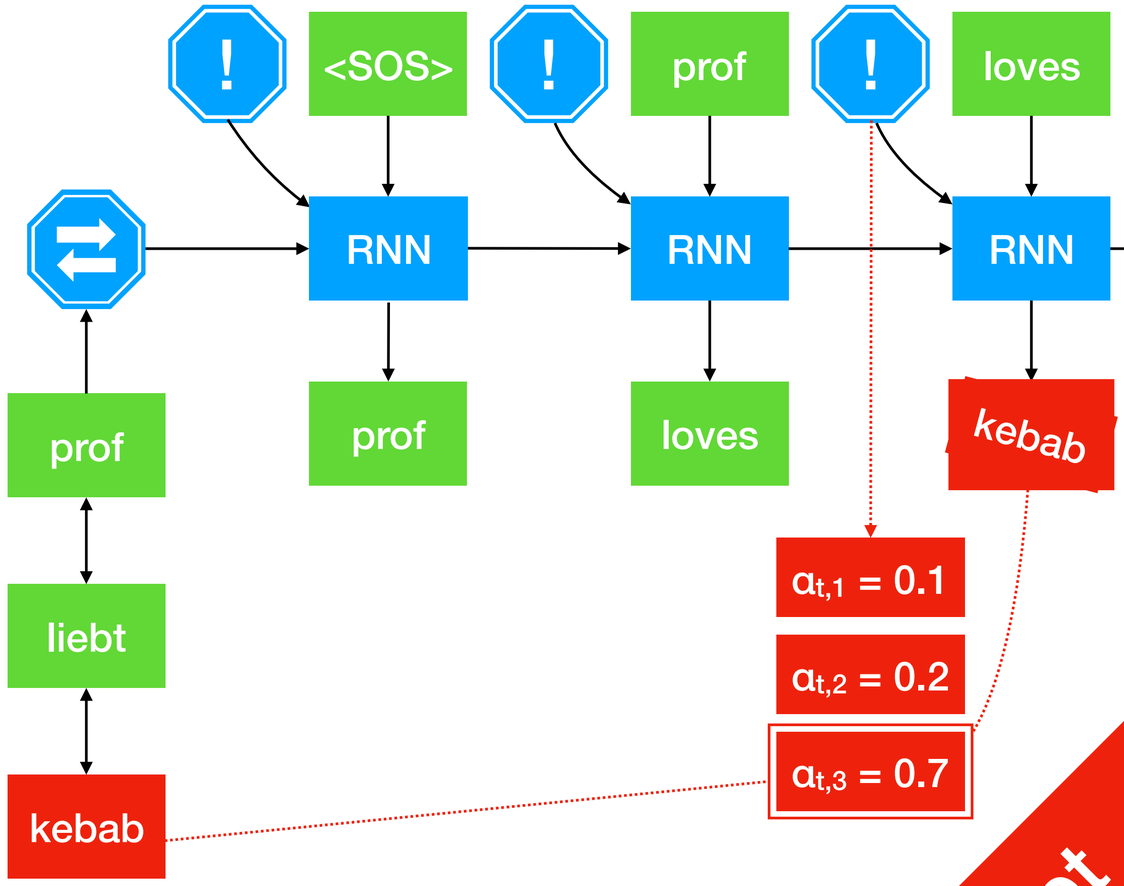


Concept

1. CL2015015 [Luong et al.] Addressing the Rare Word Problem in Neural Machine Translation

Copy Mechanism

- Sees <UNK> at step t
  - looks at attention weight α_t
  - replace <UNK> with the source word $f_{\text{argmax}_i \alpha_{t,i}}$

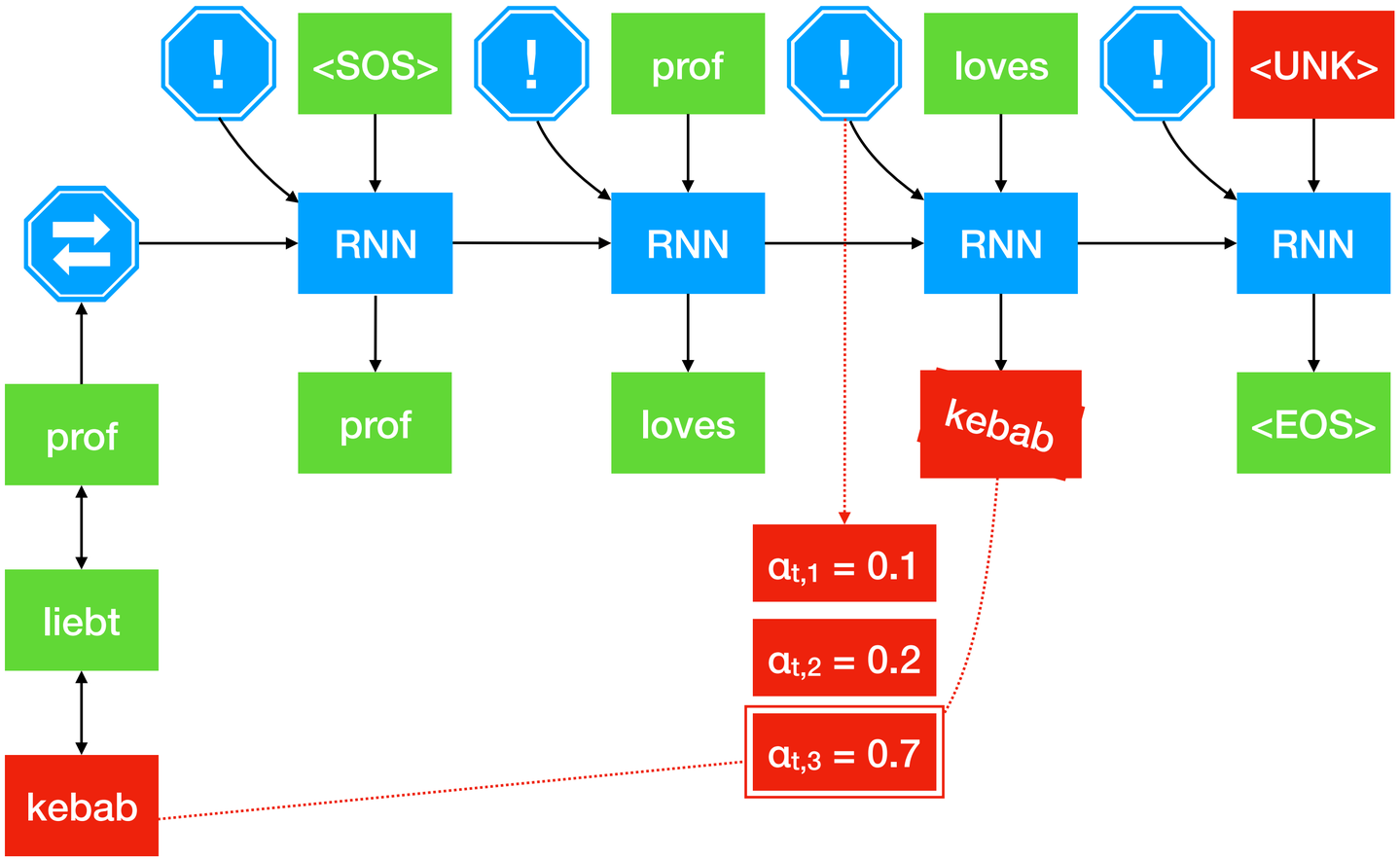


Concept

1. CL2015015 [Luong et al.] Addressing the Rare Word Problem in Neural Machine Translation

P1
Copy

*Pointer-Generator Copy Mechanism

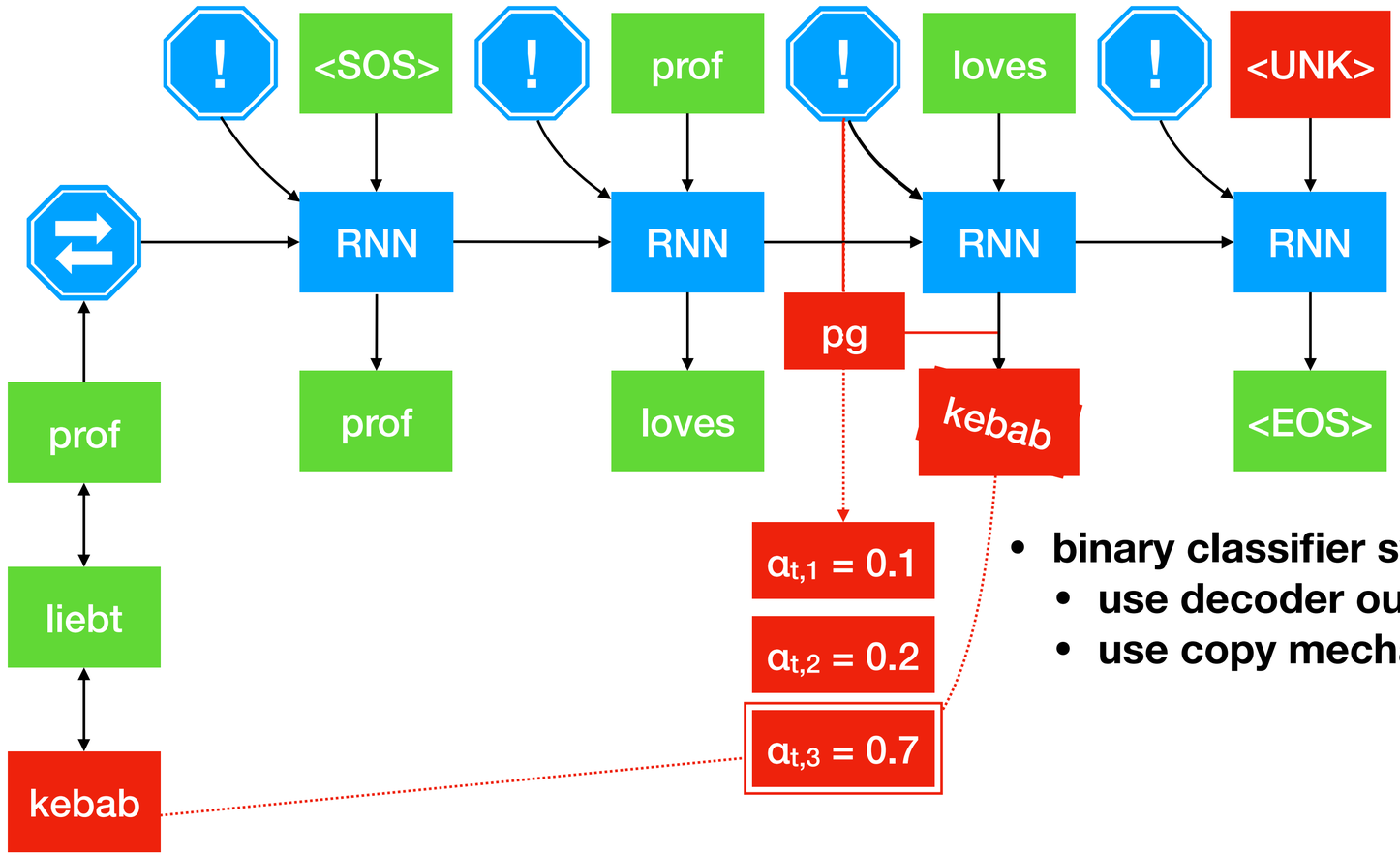


Concept

1. CL2016032 [Gulcehre et al.] Pointing the Unknown Words

P1
Copy

*Pointer-Generator Copy Mechanism



- **binary classifier switch**
  - **use decoder output**
  - **use copy mechanism**

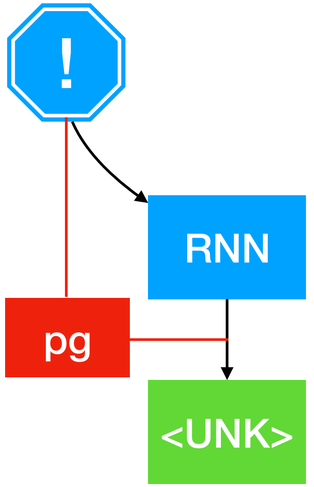
Concept

1. CL2016032 [Gulcehre et al.] Pointing the Unknown Words

P1
Copy

*Pointer-Generator Copy Mechanism

- Sees <UNK> at step t, or $pg([h_t^{dec}; context_t]) \leq 0.5$
  - looks at attention weight α_t
  - replace <UNK> with the source word $f_{\text{argmax}_i \alpha_{t,i}}$



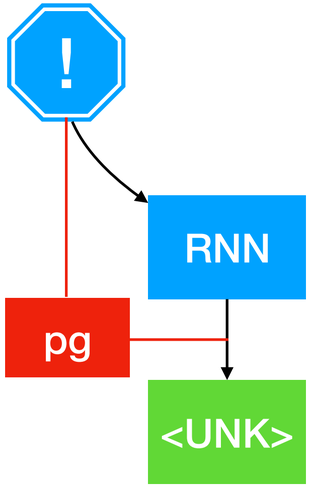
Concept

1. CL2016032 [Gulcehre et al.] Pointing the Unknown Words

P1
Copy

*Pointer-Based Dictionary Fusion

- Sees <UNK> at step t, or $pg([h_t^{dec}; context_t]) \leq 0.5$
  - looks at attention weight α_t
  - replace <UNK> with translation of the source word $\text{dict}(f_{\text{argmax}_i \alpha_{t,i}})$



Concept

1. CL2019331 [Gū et al.] Pointer-based Fusion of Bilingual Lexicons into Neural Machine Translation

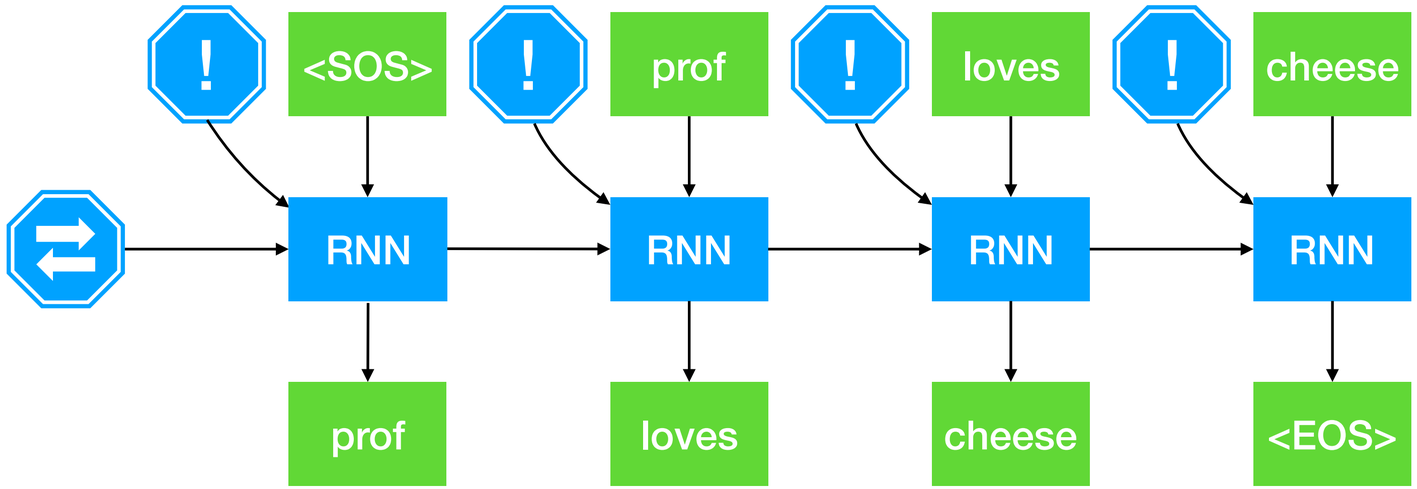
Ensemble

- Similar to voting mechanism, but with probabilities
  - multiple models of different parameters (usually from different checkpoints of the same training instance)
  - use the output with the highest probability across all models

Concept

P2
Ensemble

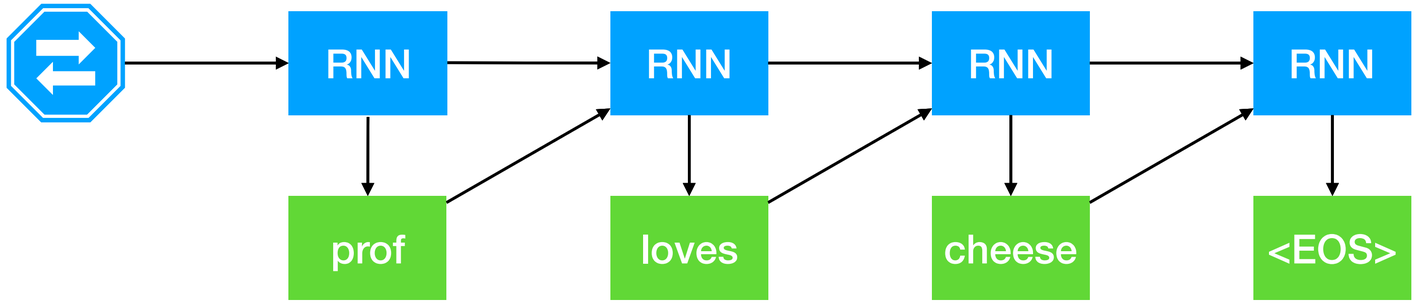
Ensemble



Demo

P2
Ensemble

Ensemble

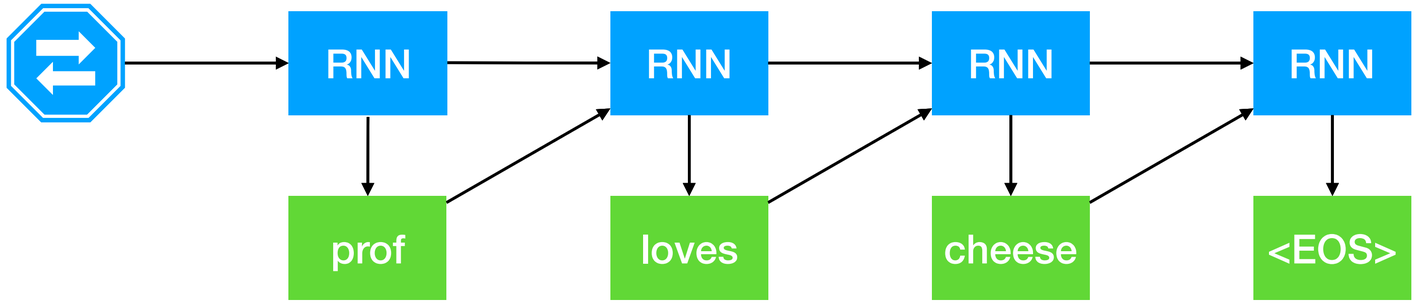


Demo

P2
Ensemble

Ensemble

seq2seq_E049.pt



Demo

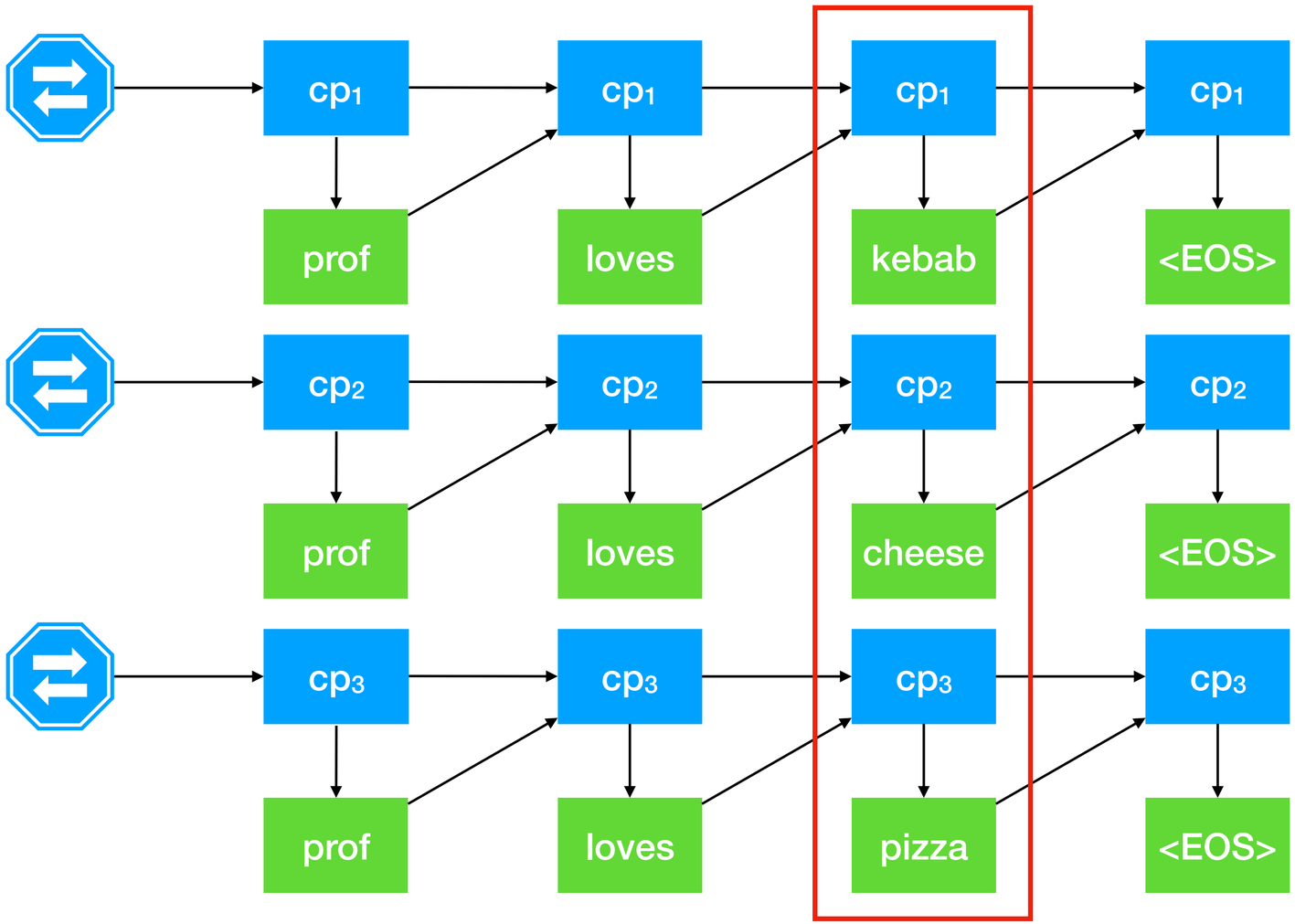
P2
Ensemble

Ensemble

seq2seq_E049.pt

seq2seq_E048.pt

seq2seq_E047.pt



Demo

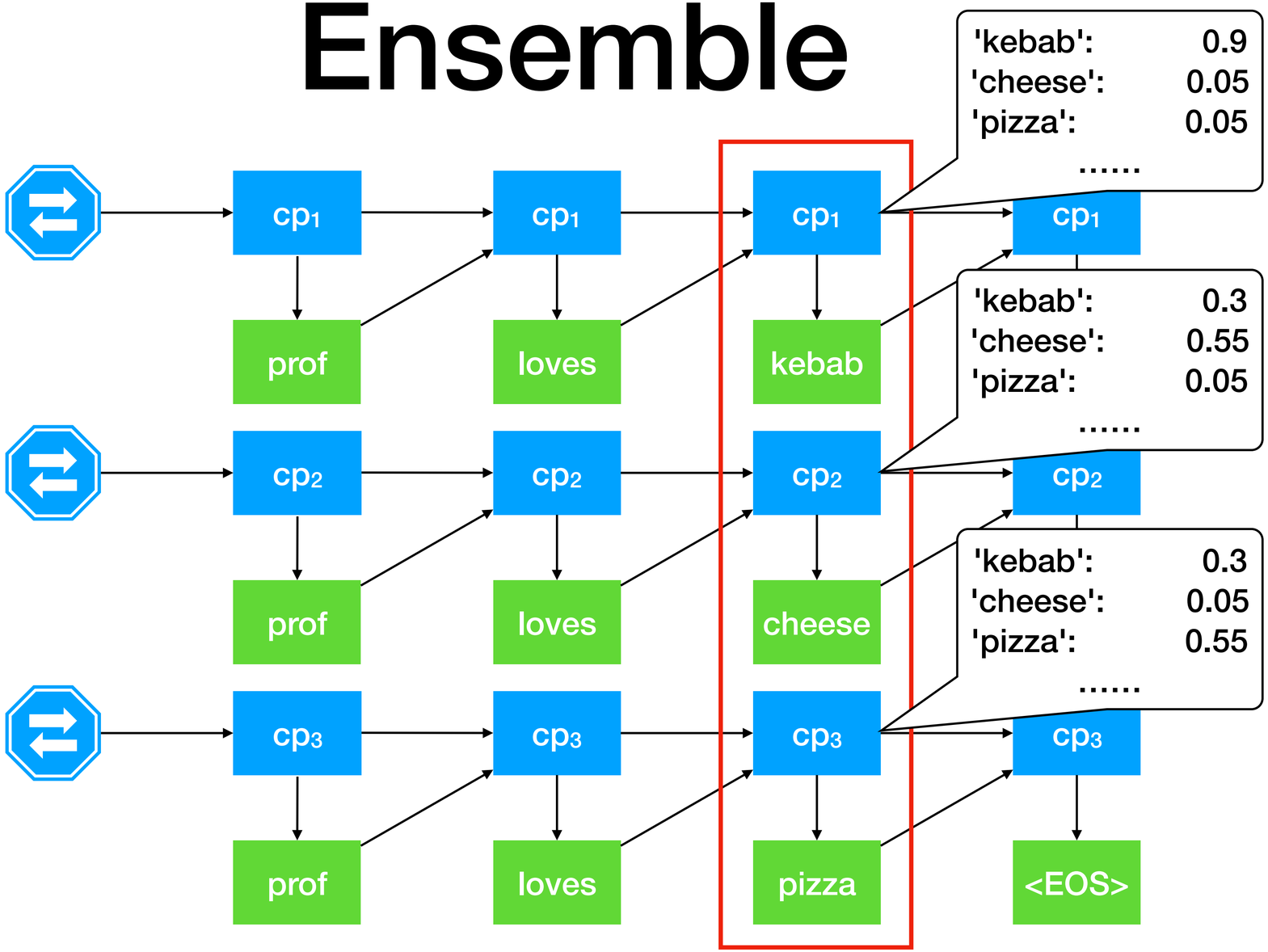
P2
Ensemble

Ensemble

seq2seq_E049.pt

seq2seq_E048.pt

seq2seq_E047.pt



Demo

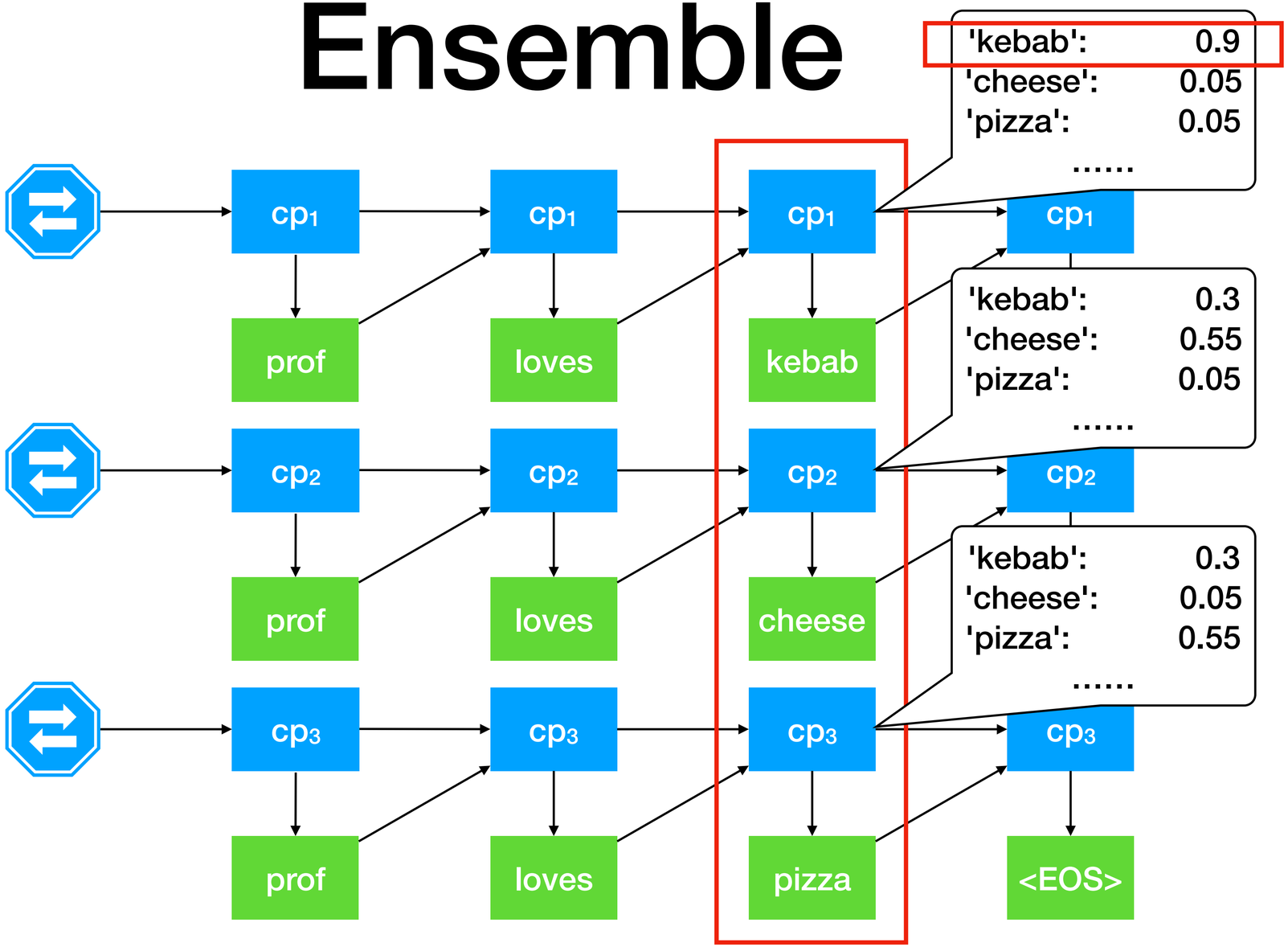
P2
Ensemble

Ensemble

seq2seq_E049.pt

seq2seq_E048.pt

seq2seq_E047.pt



Demo

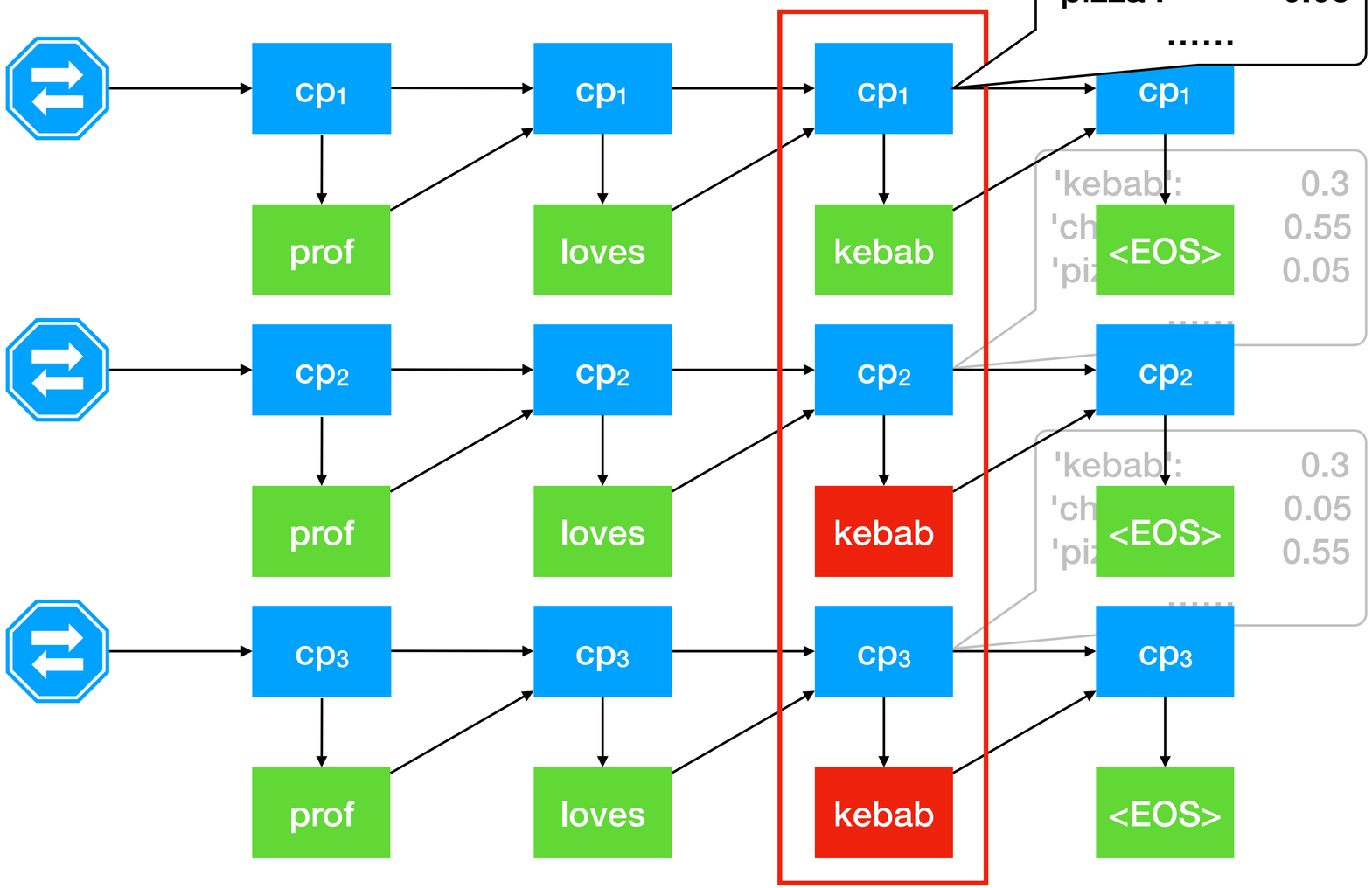
P2
Ensemble

Ensemble

seq2seq_E049.pt

seq2seq_E048.pt

seq2seq_E047.pt



Demo

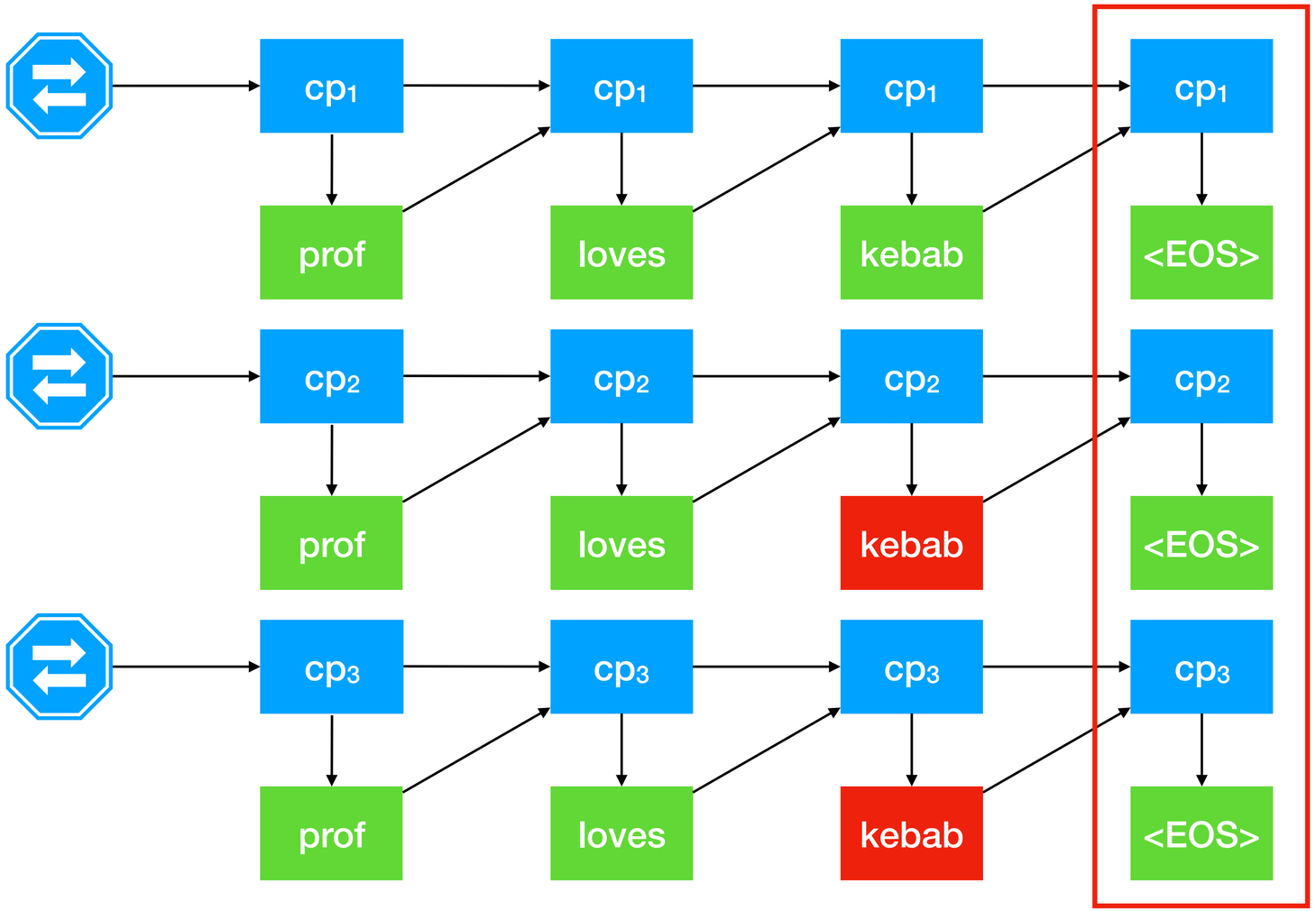
P2
Ensemble

Ensemble

seq2seq_E049.pt

seq2seq_E048.pt

seq2seq_E047.pt



Demo

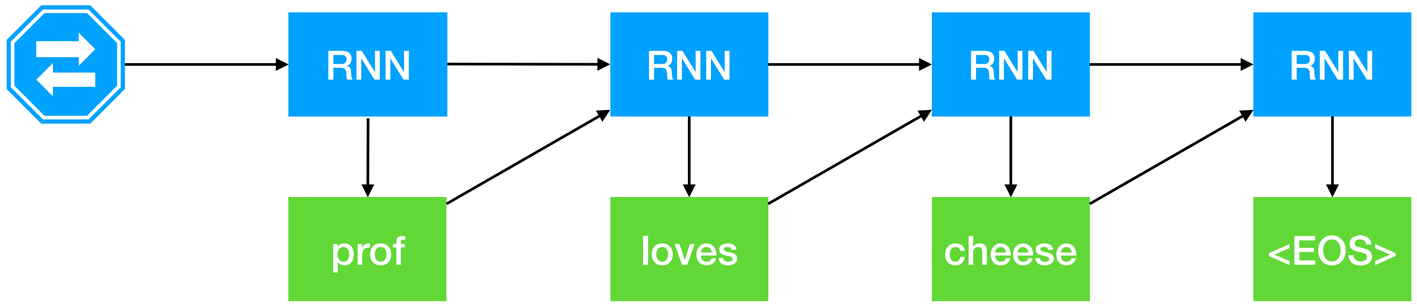
BeamSearch

- Consider multiple hypothesis at each step t, formulate decoding as a search programme on a tree

Concept

P3
BeamSearch

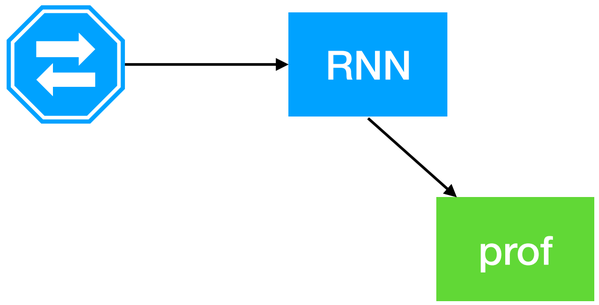
BeamSearch



Demo

P3
BeamSearch

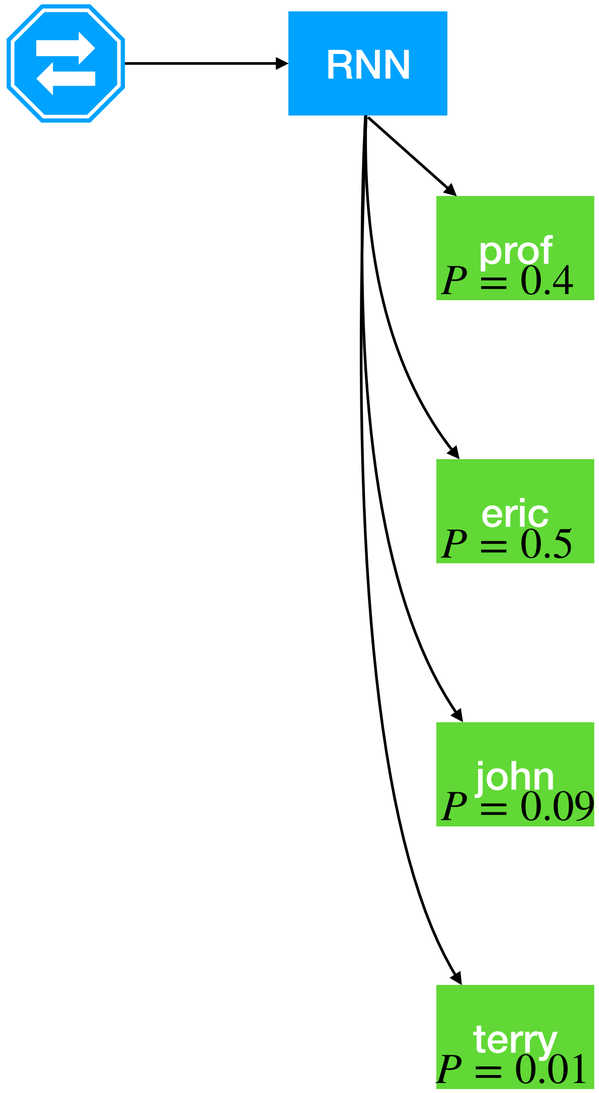
BeamSearch (Size=3)



Demo

P3
BeamSearch

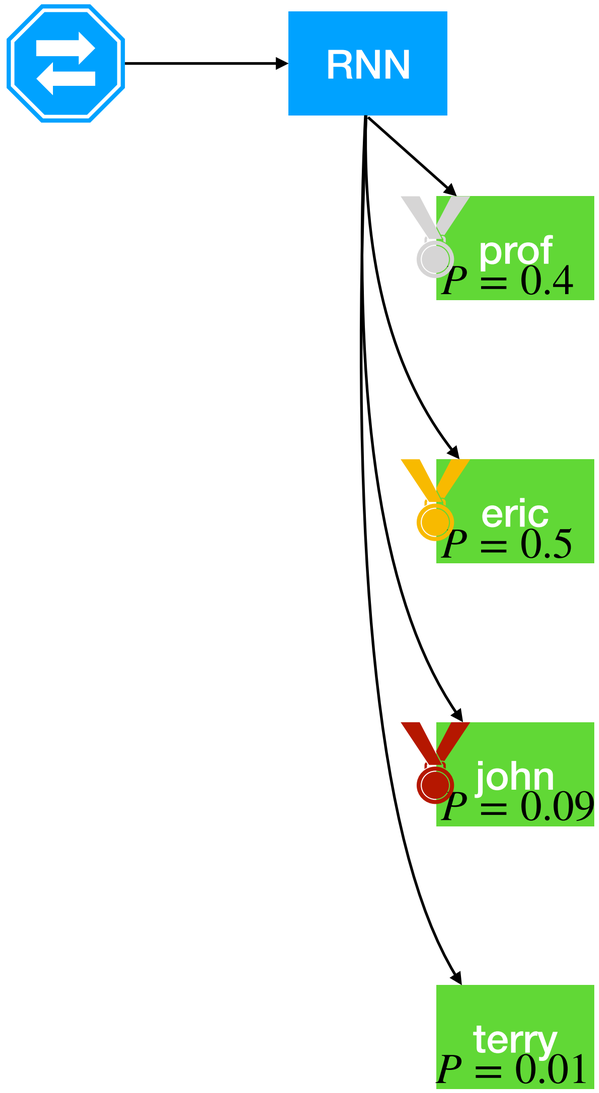
BeamSearch (Size=3)



Demo

P3
BeamSearch

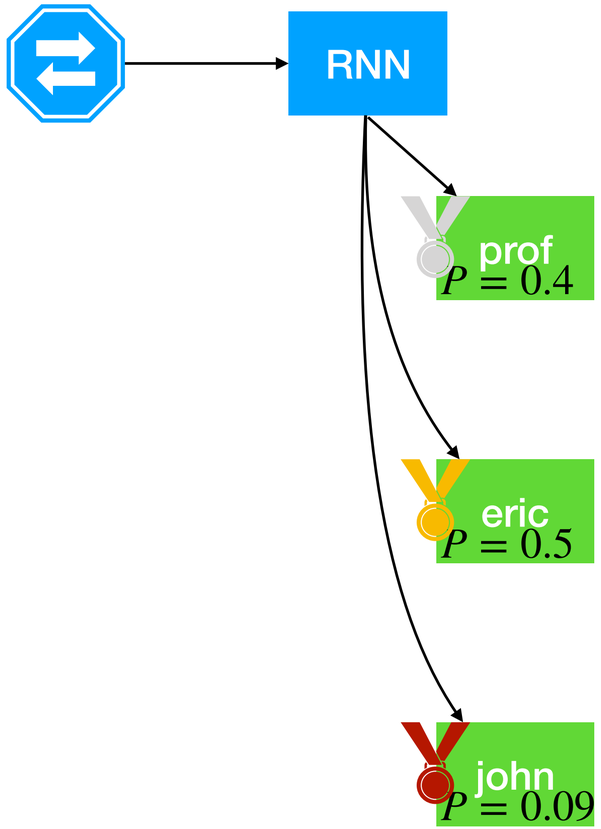
BeamSearch (Size=3)



Demo

P3
BeamSearch

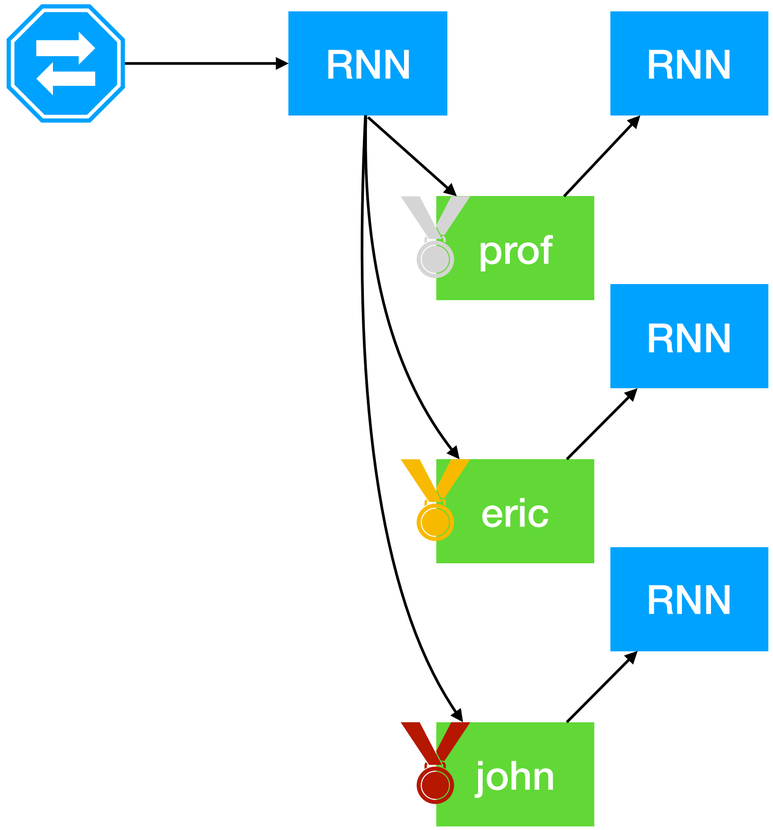
BeamSearch (Size=3)



Demo

P3
BeamSearch

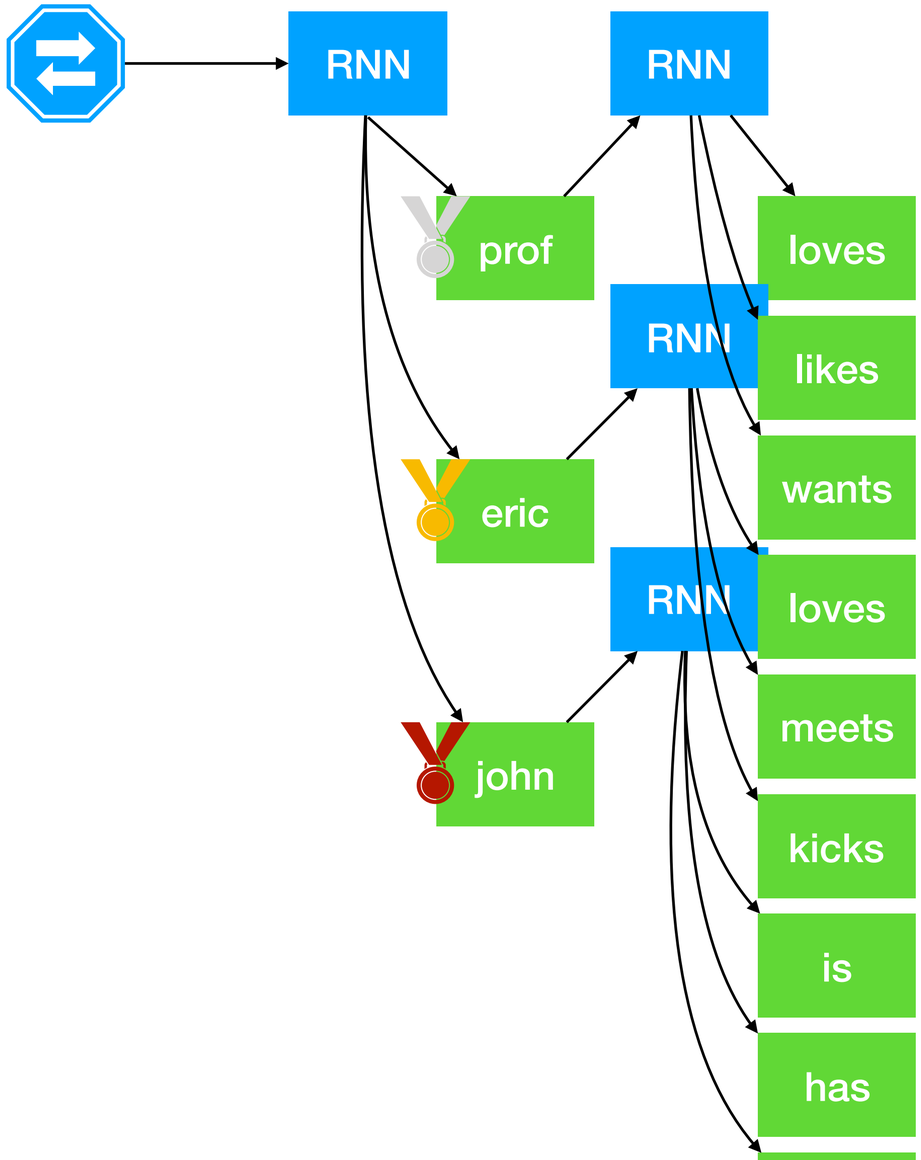
BeamSearch (Size=3)



Demo

P3
BeamSearch

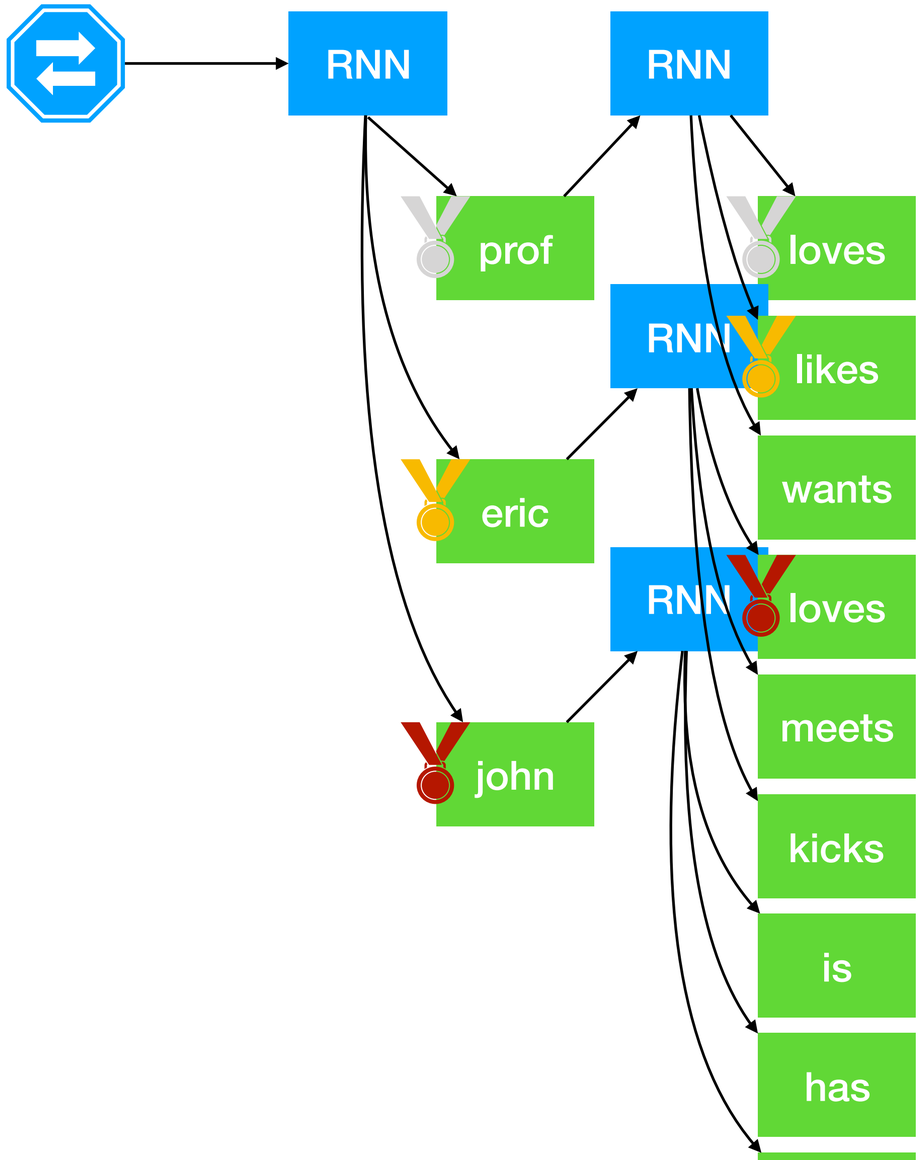
BeamSearch (Size=3)



Demo

P3
BeamSearch

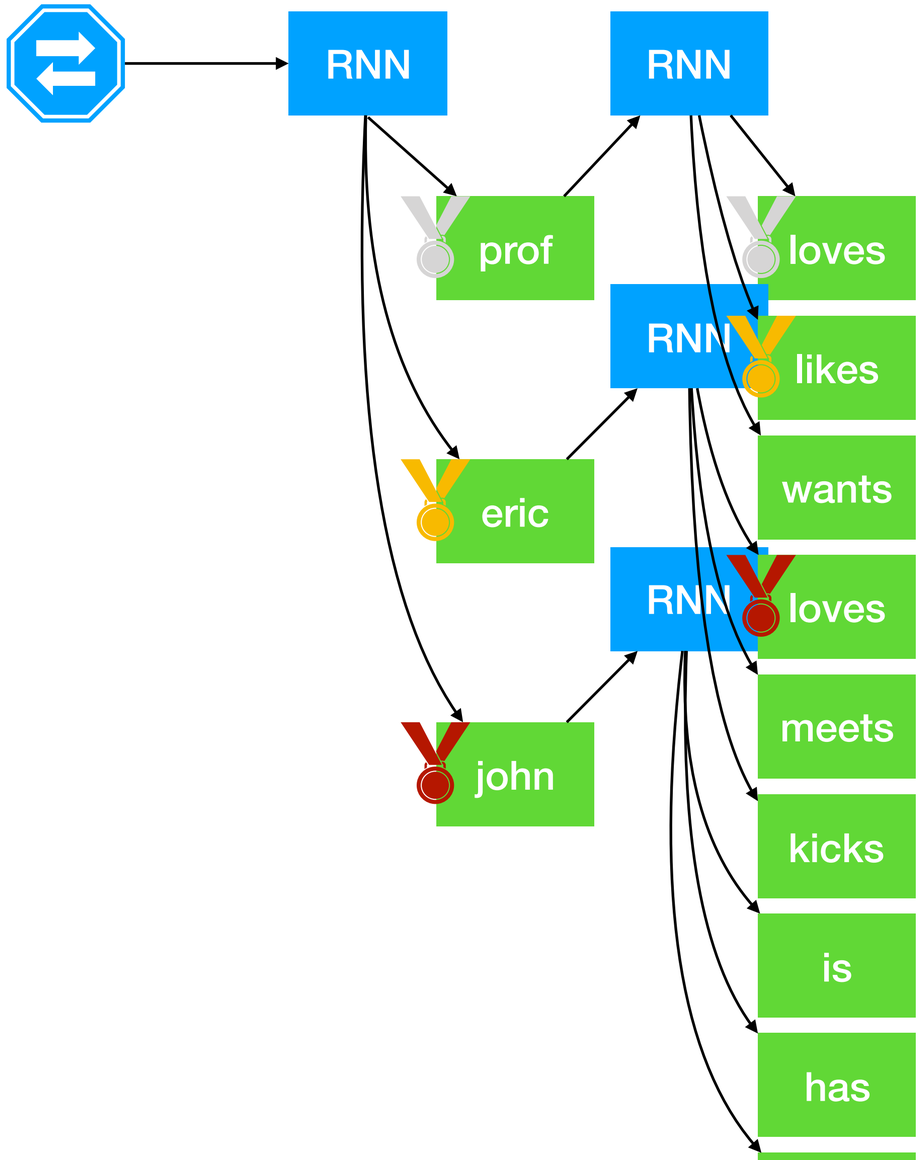
BeamSearch (Size=3)



Demo

P3
BeamSearch

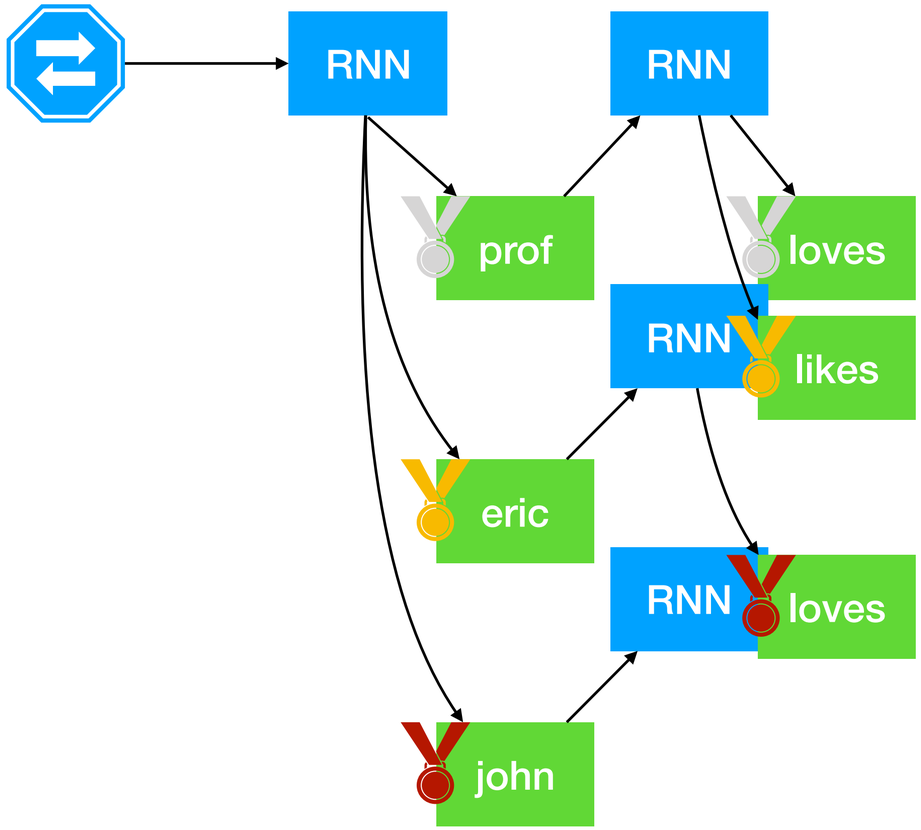
BeamSearch (Size=3)



Demo

P3
BeamSearch

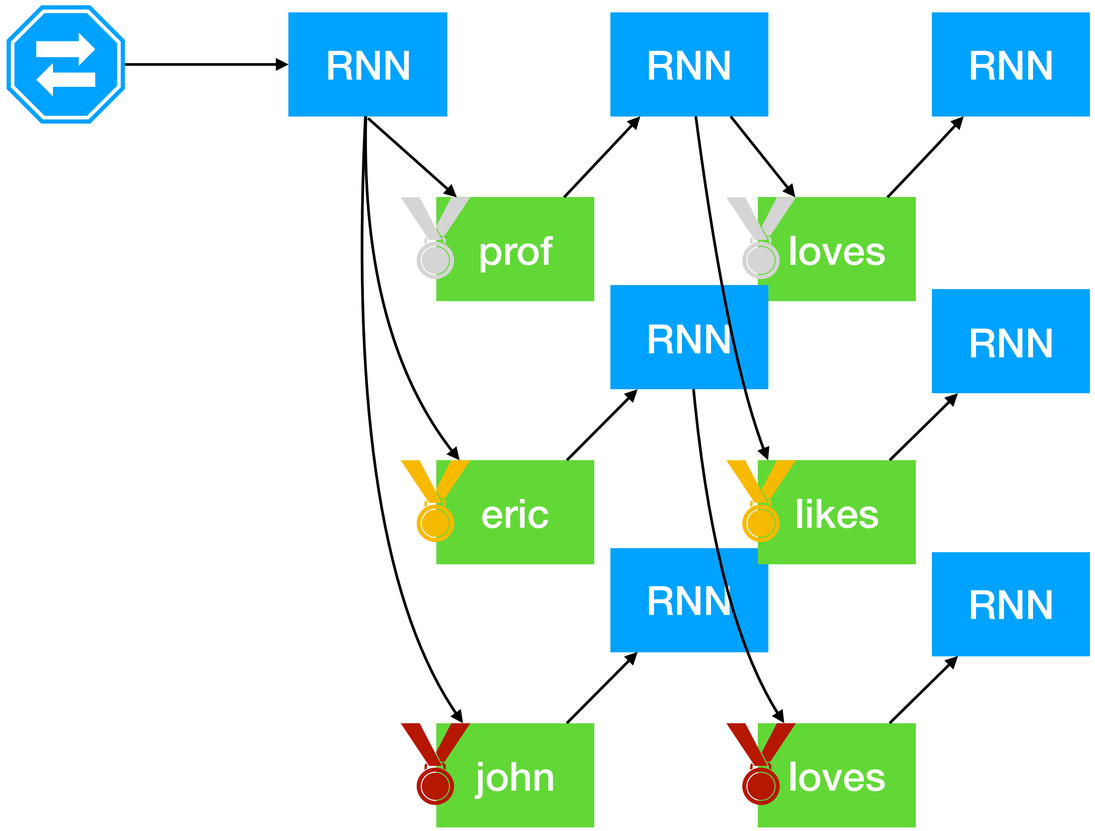
BeamSearch (Size=3)



Demo

P3
BeamSearch

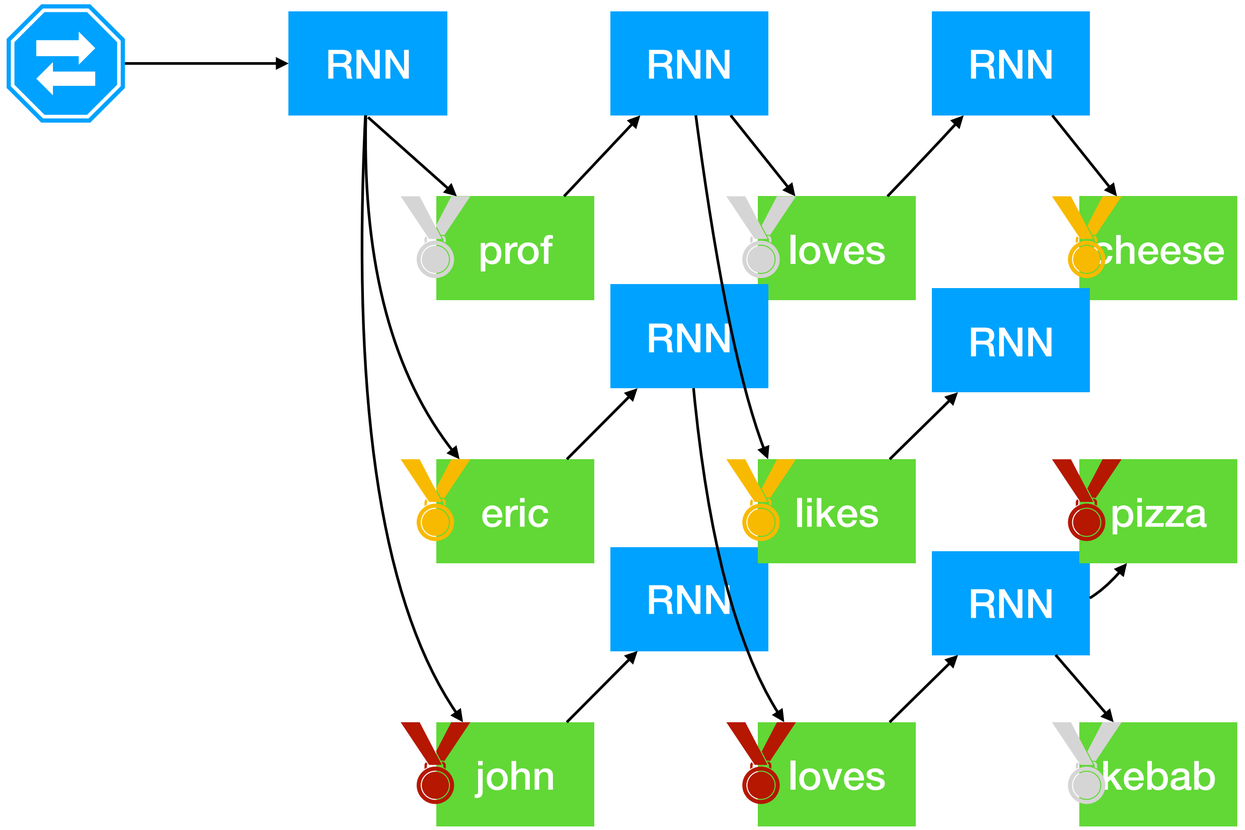
BeamSearch (Size=3)



Demo

P3
BeamSearch

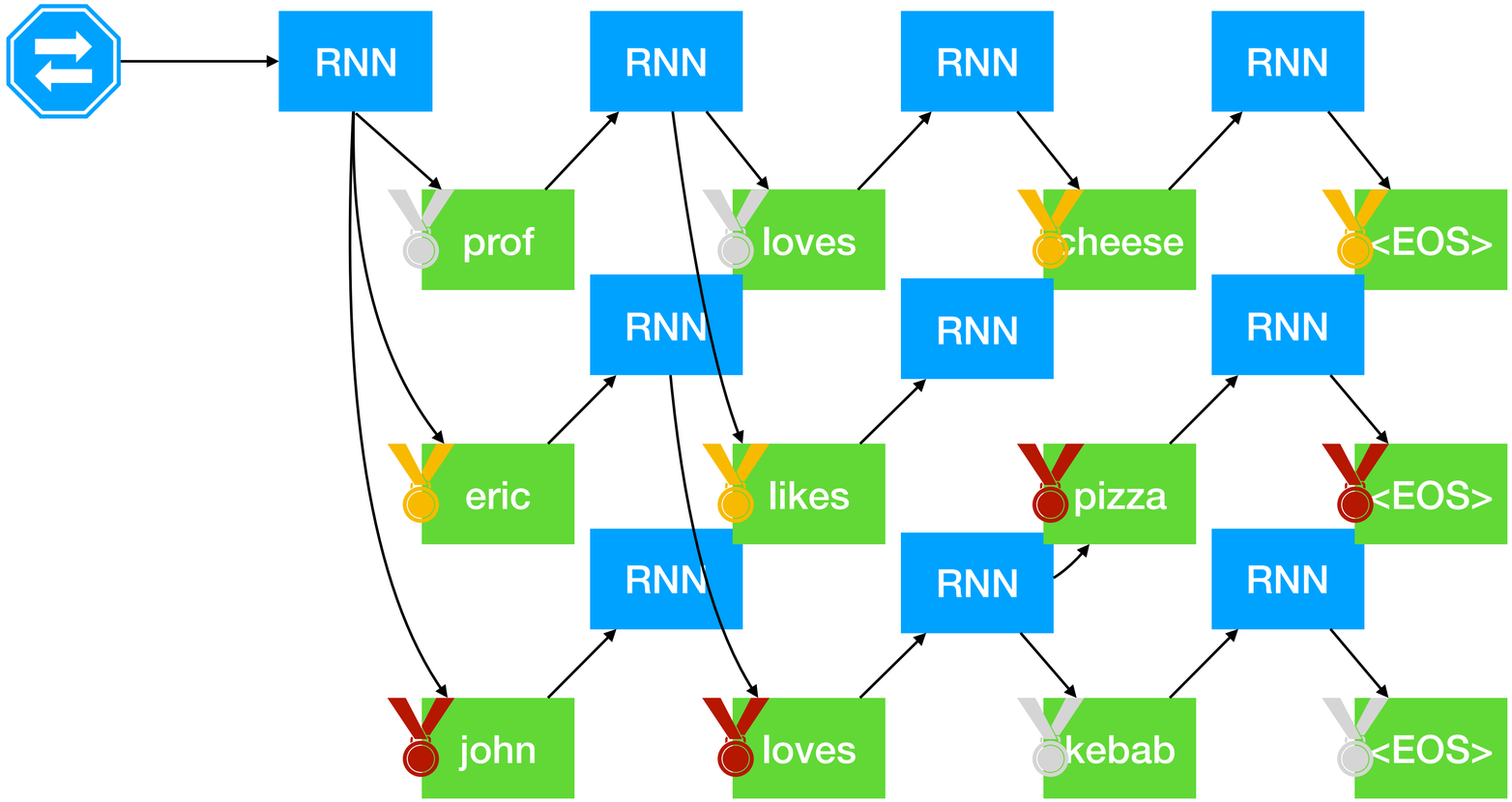
BeamSearch (Size=3)



Demo

P3
BeamSearch

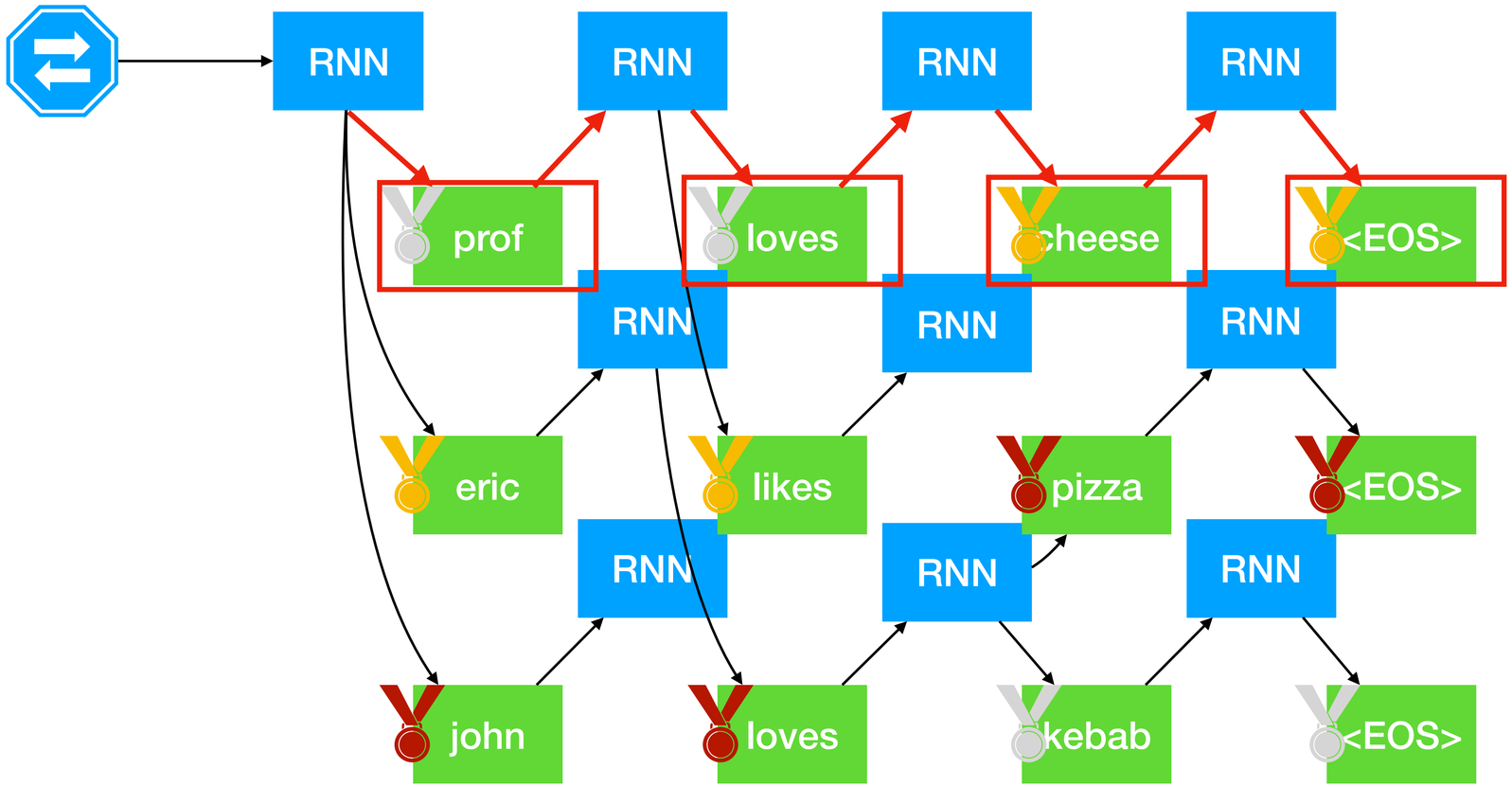
BeamSearch (Size=3)



Demo

P3
BeamSearch

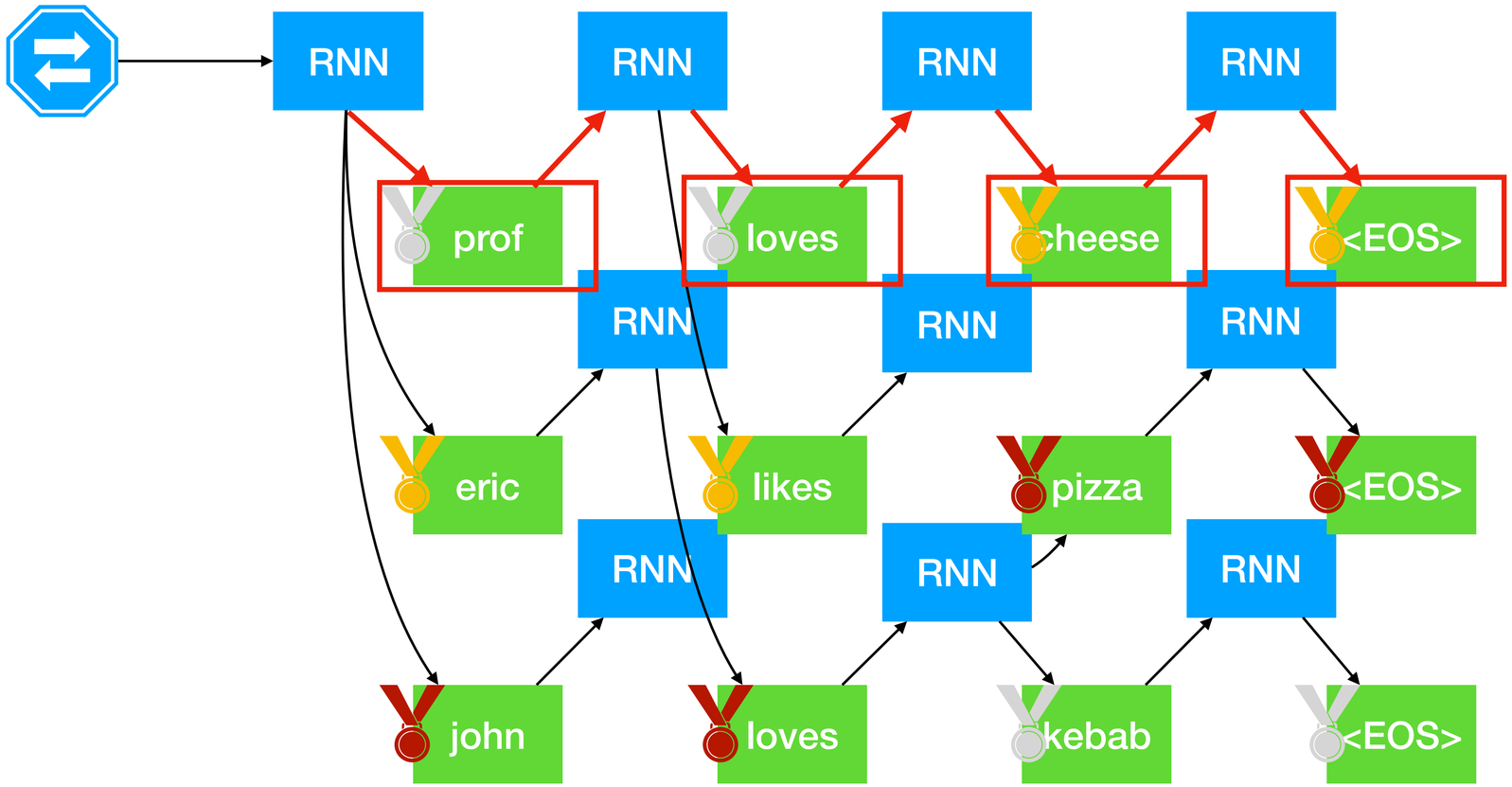
BeamSearch (Size=3)



Demo

P3
BeamSearch

BeamSearch (Size=3)

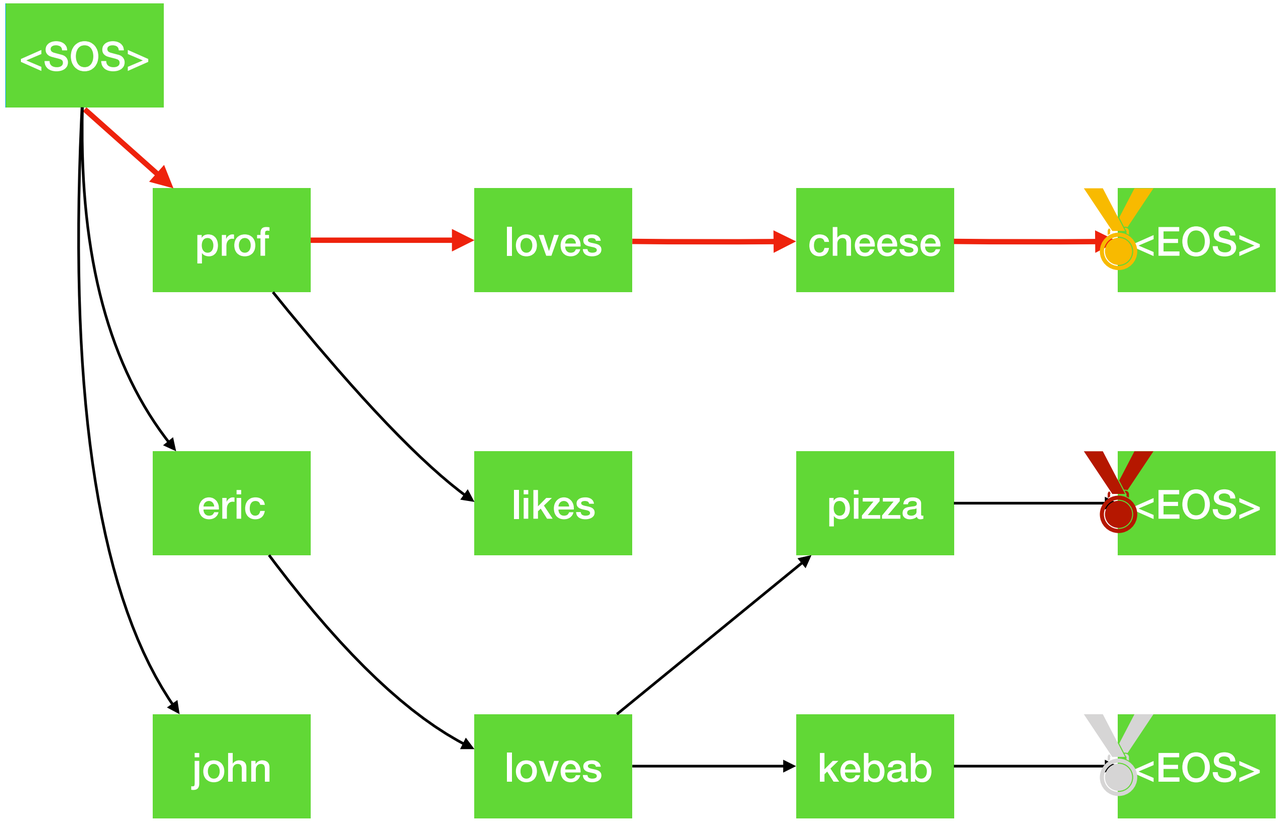


Output: "prof loves cheese"

Demo

P3
BeamSearch

BeamSearch (Size=3)

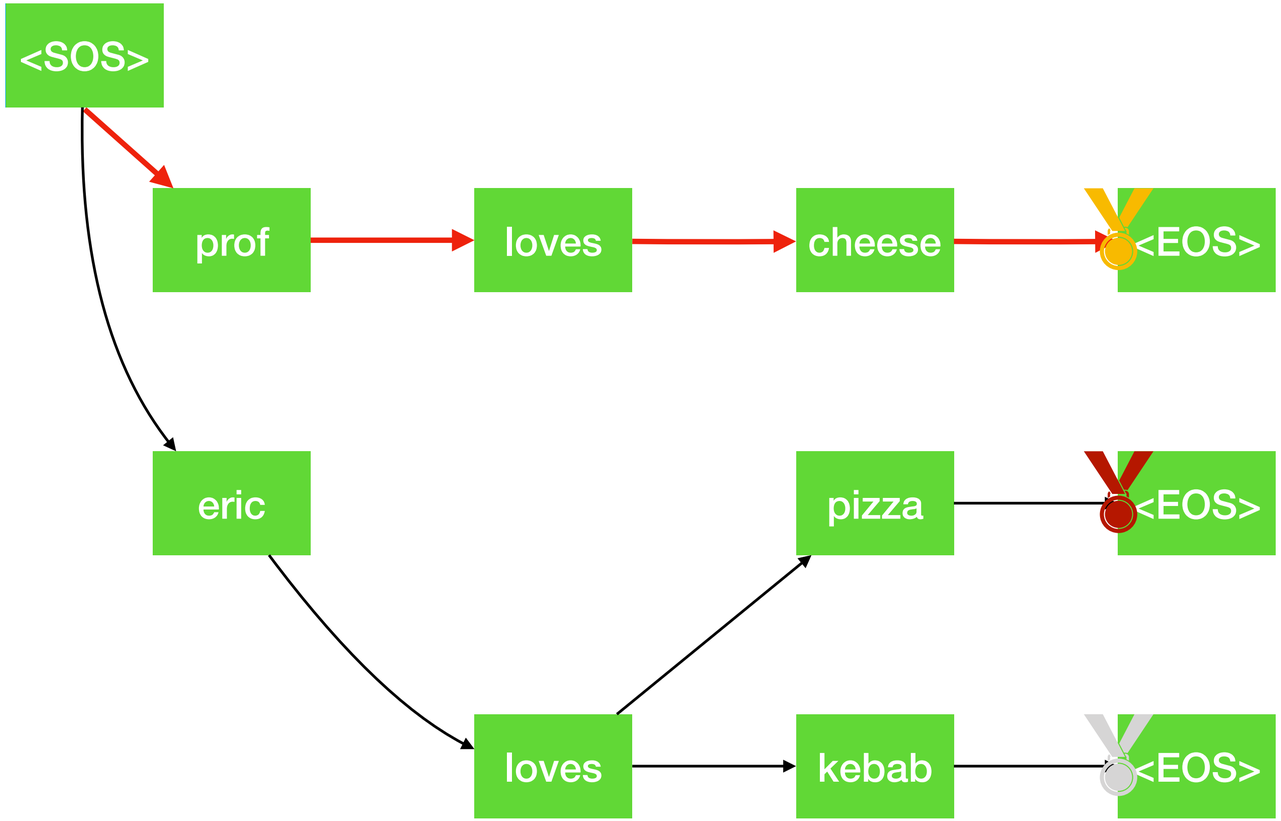


Output: "prof loves cheese"

Demo

P3
BeamSearch

BeamSearch (Size=3)

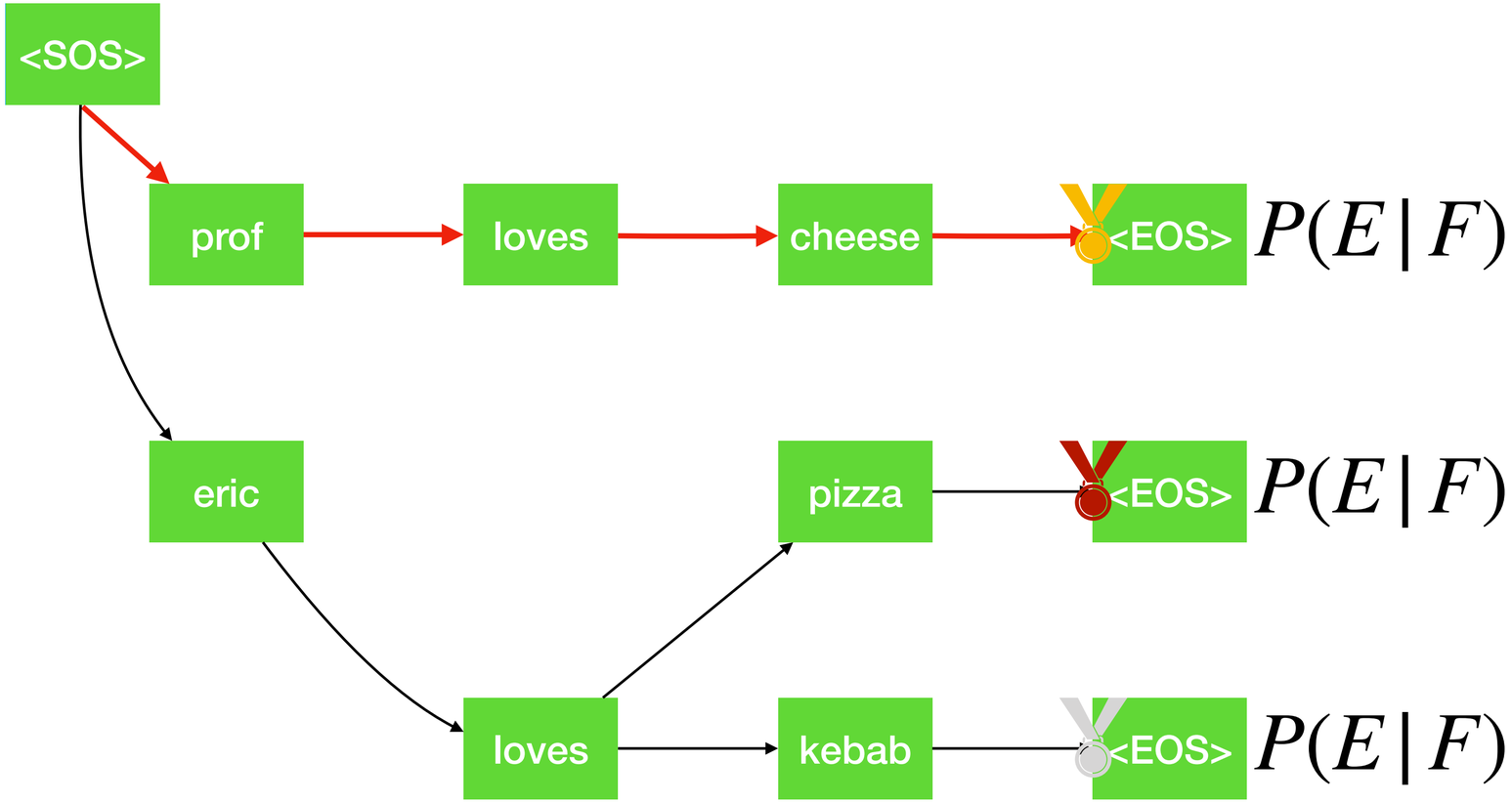


Output: "prof loves cheese"

Demo

P3
BeamSearch

BeamSearch (Size=3)

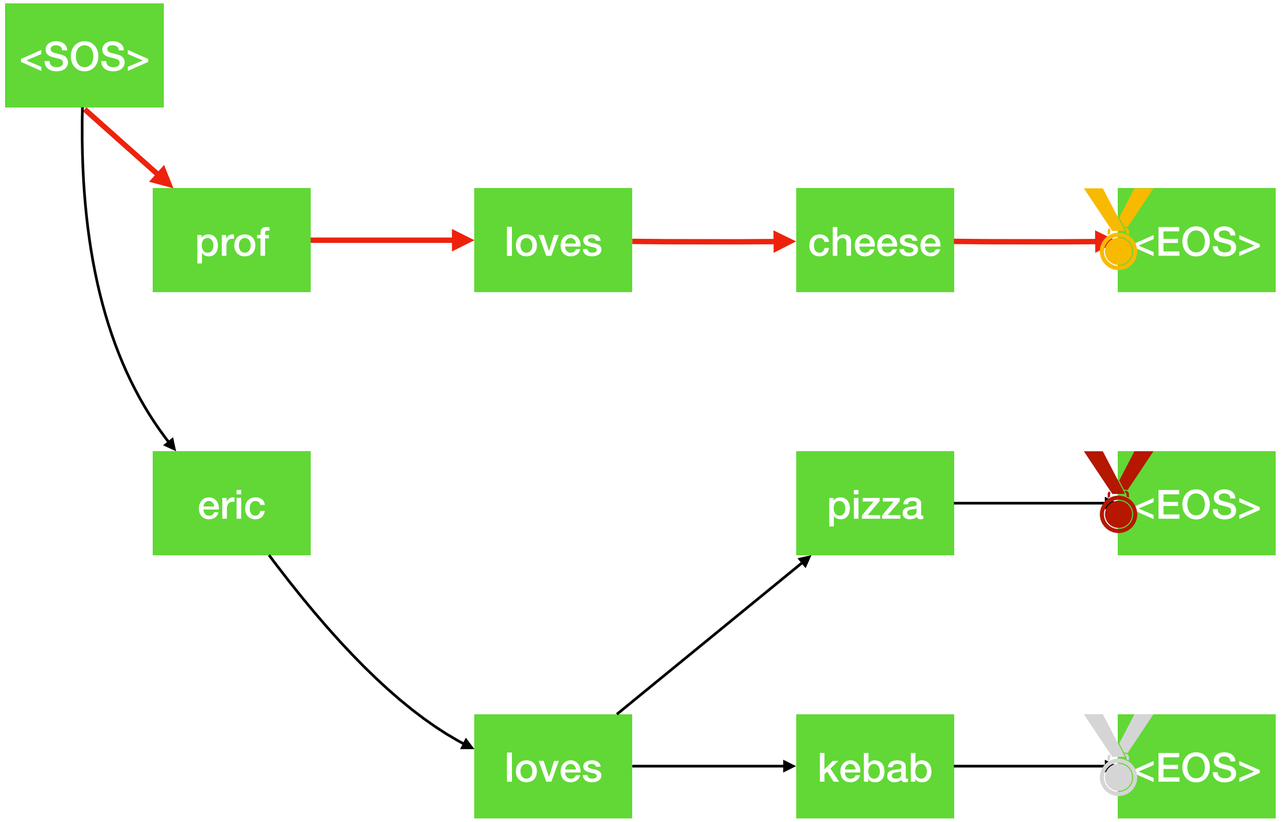


Output: "prof loves cheese"

Demo

P3
BeamSearch

BeamSearch (Size=3)



$P(\text{'prof loves cheese <EOS>'} \mid F)$

$P(\text{'eric loves pizza <EOS>'} \mid F)$

$P(\text{'eric loves kebab <EOS>'} \mid F)$

Output: "prof loves cheese"

Demo

BeamSearch

- Consider multiple hypotheses at each step t, formulate decoding as a search programme on a tree
- Consider multiple complete translation hypotheses, but reduces the total search space from $|V_E|^m$ to

Concept

BeamSearch

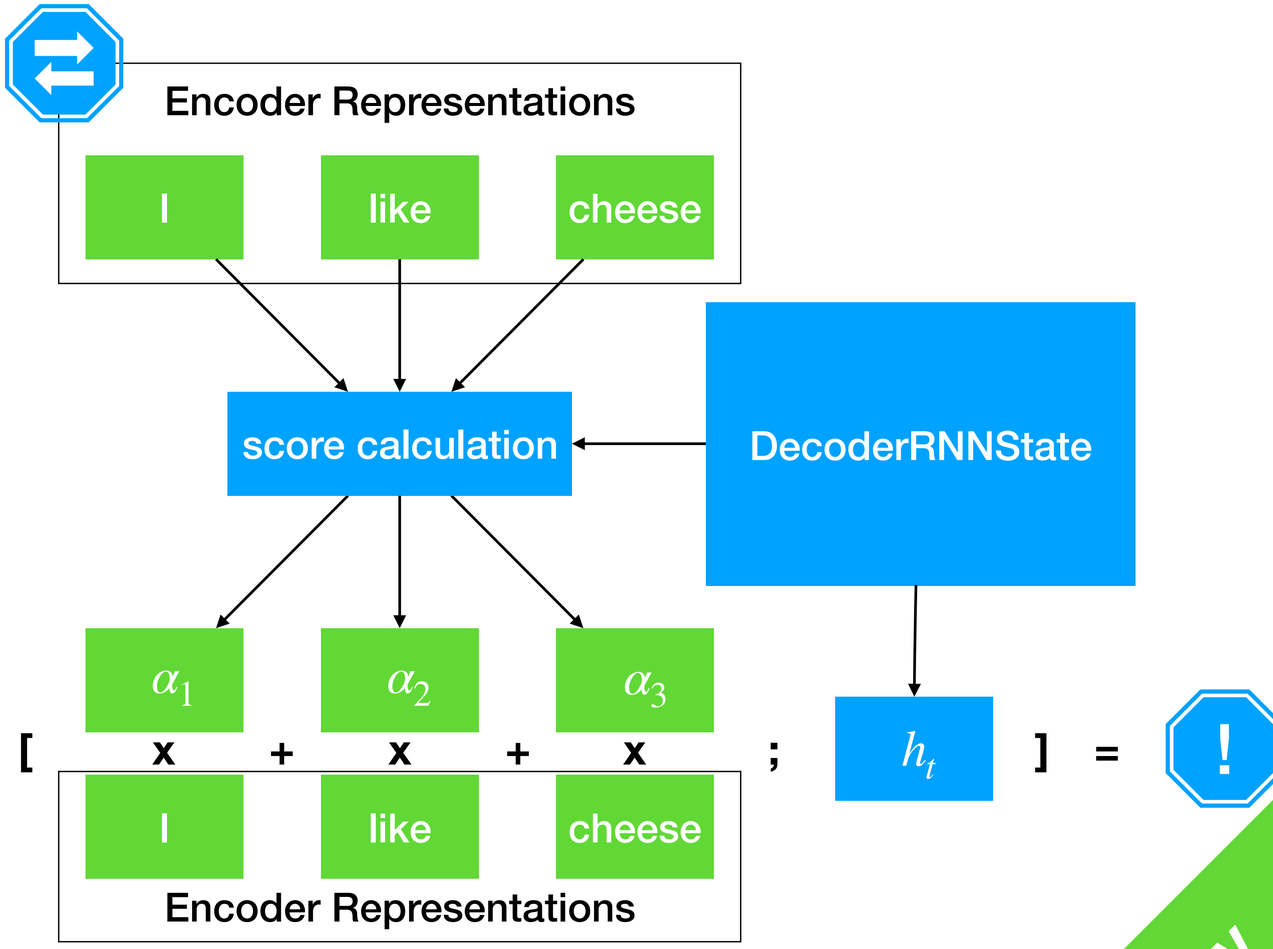
- Consider multiple hypotheses at each step t, formulate decoding as a search programme on a tree
- Consider multiple complete translation hypotheses, but reduces the total search space from $|V_E|^m$ to $\text{batchSize} \times m$

Concept

Extra1
Transformer

Self-Attentive NMT

- Aggregating src information $h^{enc}_{1:|F|}$, with h^{dec}_t as query



Review

Extra1
Transformer

Self-Attentive NMT

- Aggregating src information $h^{enc}_{1:|F|}$, with h^{dec}_t as query
- $\overrightarrow{h}^{enc}$: aggregating information $f_{<i}$, with f_i as query
- $\overleftarrow{h}^{enc}$: aggregating information $f_{>i}$, with f_i as query
- Can we replace the RNN with attention?

Review

Extra1
Transformer

Self-Attentive NMT

- Transformer[1]
  - Encoder-Decoder
  - No RNN -> all RNNs in Seq2Seq replaced by identical self-attention blocks
  - Multi-headed attention blocks*
    - read the paper for more information

Concept

1. CL2017244 [Vaswani et al.] Attention Is All You Need

Extra1
Transformer

Self-Attentive NMT

- Does everything Seq2Seq does
- BERT: Bidirectional Encoder Representation of Transformer

1. CL2017244 [Vaswani et al.] Attention Is All You Need
2. CL2018460 [Devlin et al.] BERT Pre-training of Deep Bidirectional Transformers for Language Understanding

Concept

Extra2
Beyond NMT

Beyond NMT

- NMT weaknesses
  - Longer src sentences -> SMT does it better, augment attention?
  - Growing lexicon/Terminologies -> Pointer-based Dictionary Fusion
  - Sensitive domain -> offline computing, human-involved system to ensure accuracy
  - Mobile Platform deployment -> model compression

Concept

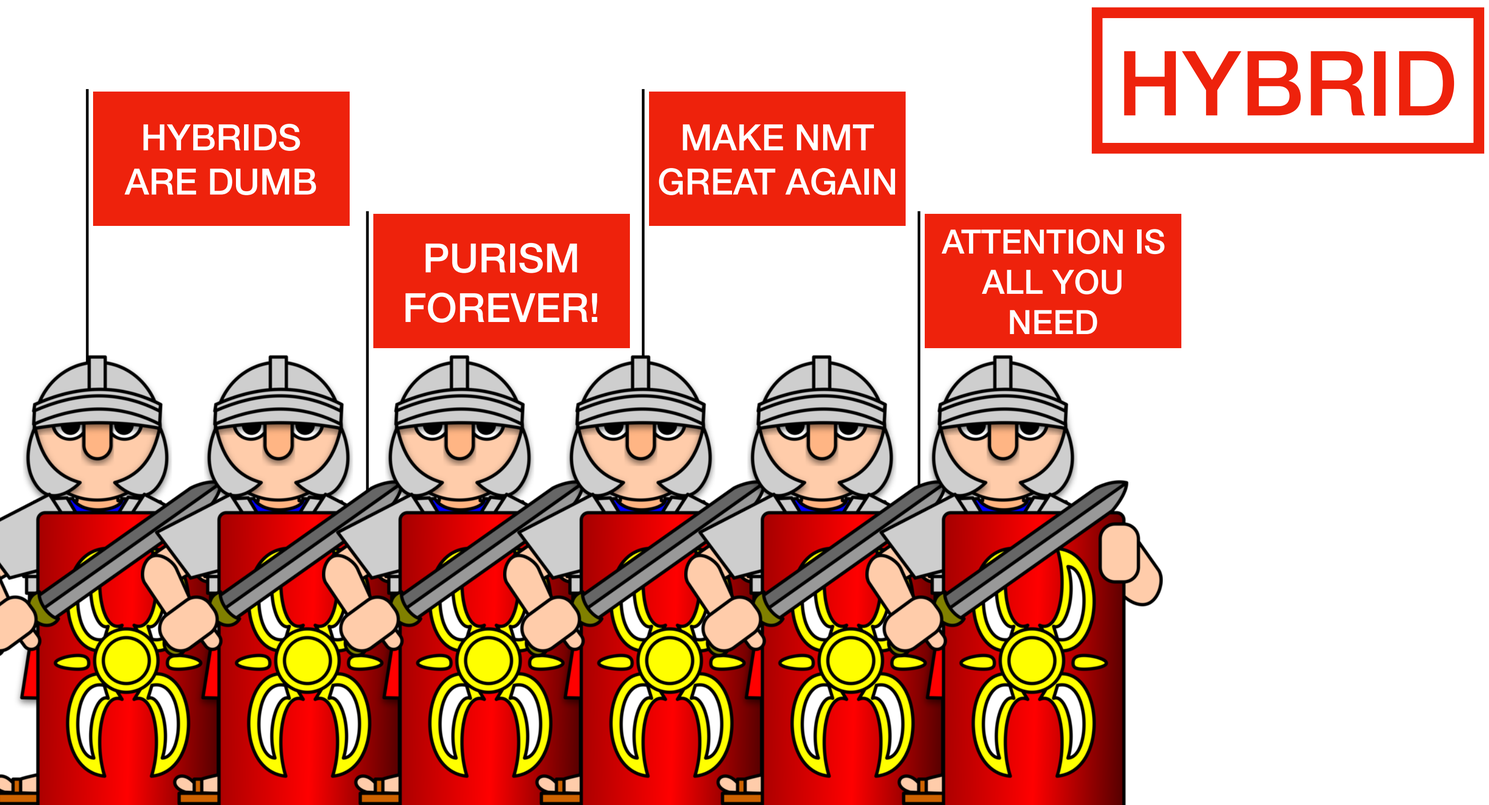
Extra2
Beyond NMT

Beyond NMT

- NMT weaknesses
  - Longer src sentences -> SMT does it better, augment attention?
  - Growing lexicon/Terminolog HYBRID based Dictionary Fusion
  - Sensitive domain -> offline computing, human-involved system to ensure accuracy
  - Mobile Platform deployment -> model compression

Concept

Beyond NMT



Concept

My work here is done, thank you.