A Decomposable Attention Model for Natural Language Inference

Ankur Parikh, Oscar Tackstrom, Dipanjan Das, Jakob Uszkoreit

Presented by: Xikun Zhang

University of Illinois, Urbana-Champaign

Natural Language Inference

- A key part of our understanding of natural language is the ability to understand sentence semantics.
- Semantic Entailment or, more popularly, the task of Natural Language Inference (NLI) is a core Natural Language Understanding task (NLU). While it poses as a classification task, it is uniquely well-positioned to serve as a benchmark task for research on NLU. It attempts to judge whether one sentence can be inferred from another.
- More specifically, it tries to identify the relationship between the meanings of a pair of sentences, called the premise and the hypothesis. The relationship could be one of the following:
  - Entailment: the hypothesis is a sentence with a similar meaning as the premise
  - Contradiction: the hypothesis is a sentence with a contradictory meaning
  - Neutral: the hypothesis is a sentence with mostly the same lexical items as the premise but a different meaning.

Natural Language Inference (Cont’d)

- Determine entailment/contradiction/neutral relationships between a premise and a hypothesis.

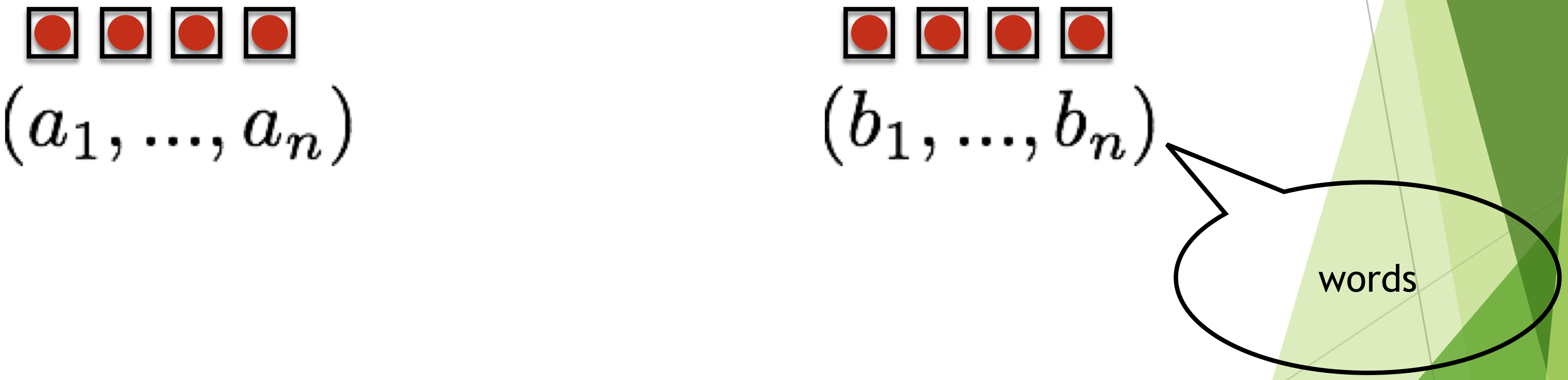
Premise *Bob is in his room, but because of the thunder and lightning outside, he cannot sleep.*

Hypothesis 1 *Bob is awake.* → **entailment**

Hypothesis 2 *It is sunny outside.* → **contradiction**

Hypothesis 3 *Bob has a big house.* → **neutral**

Recent Work (Sentence Encoding)

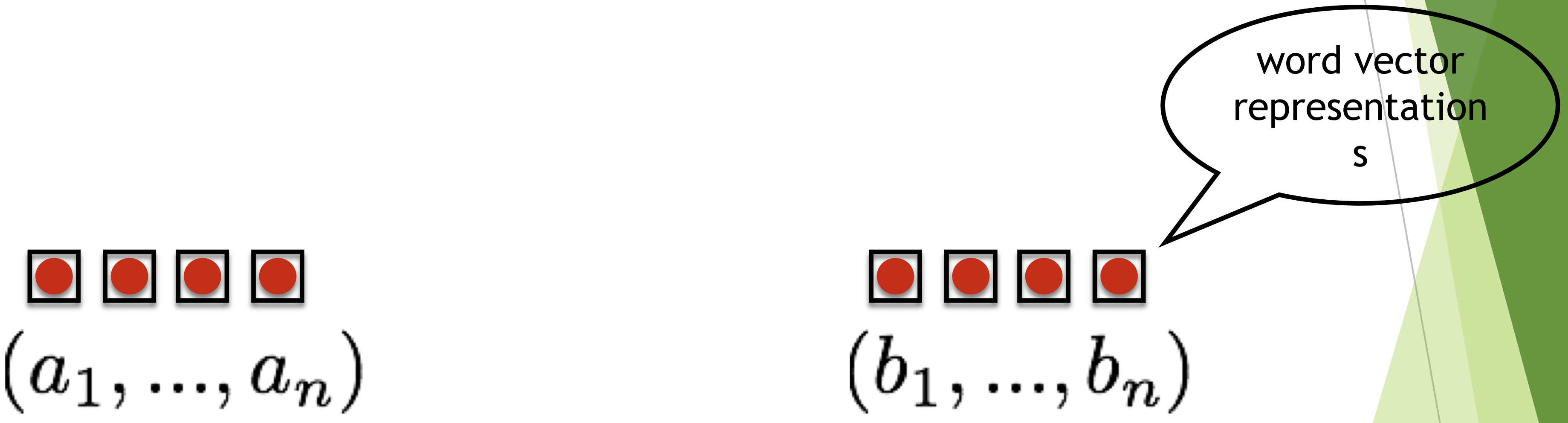


$(a_1, ..., a_n)$

$(b_1, ..., b_n)$

words

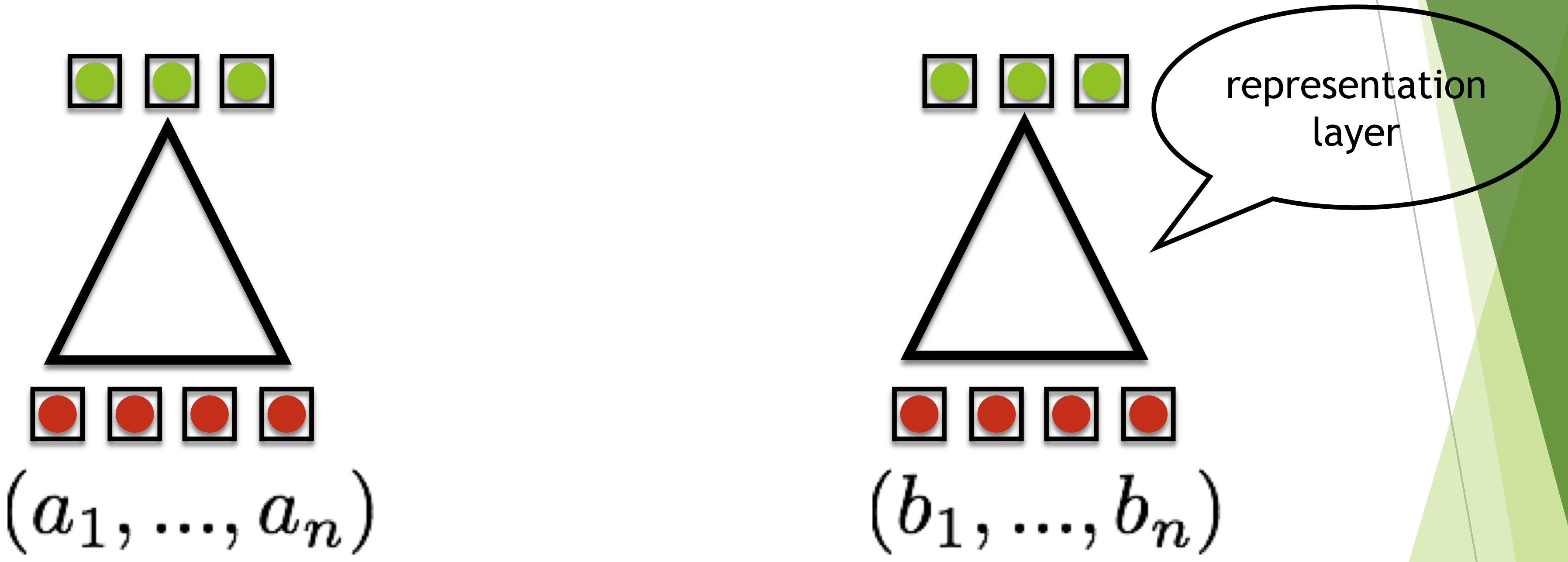
Recent Work (Sentence Encoding)



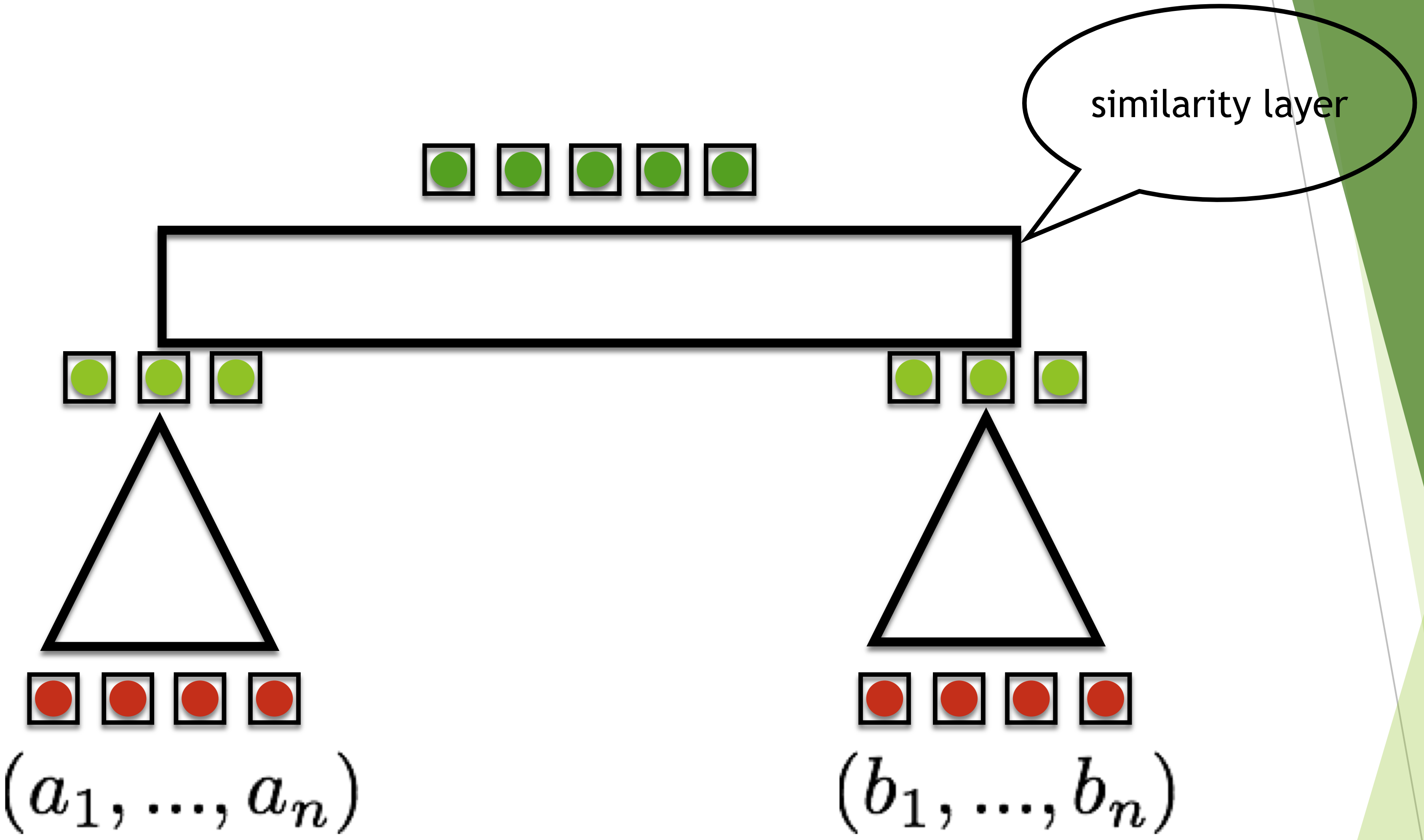
$(a_1, ..., a_n)$ $(b_1, ..., b_n)$

word vector representations

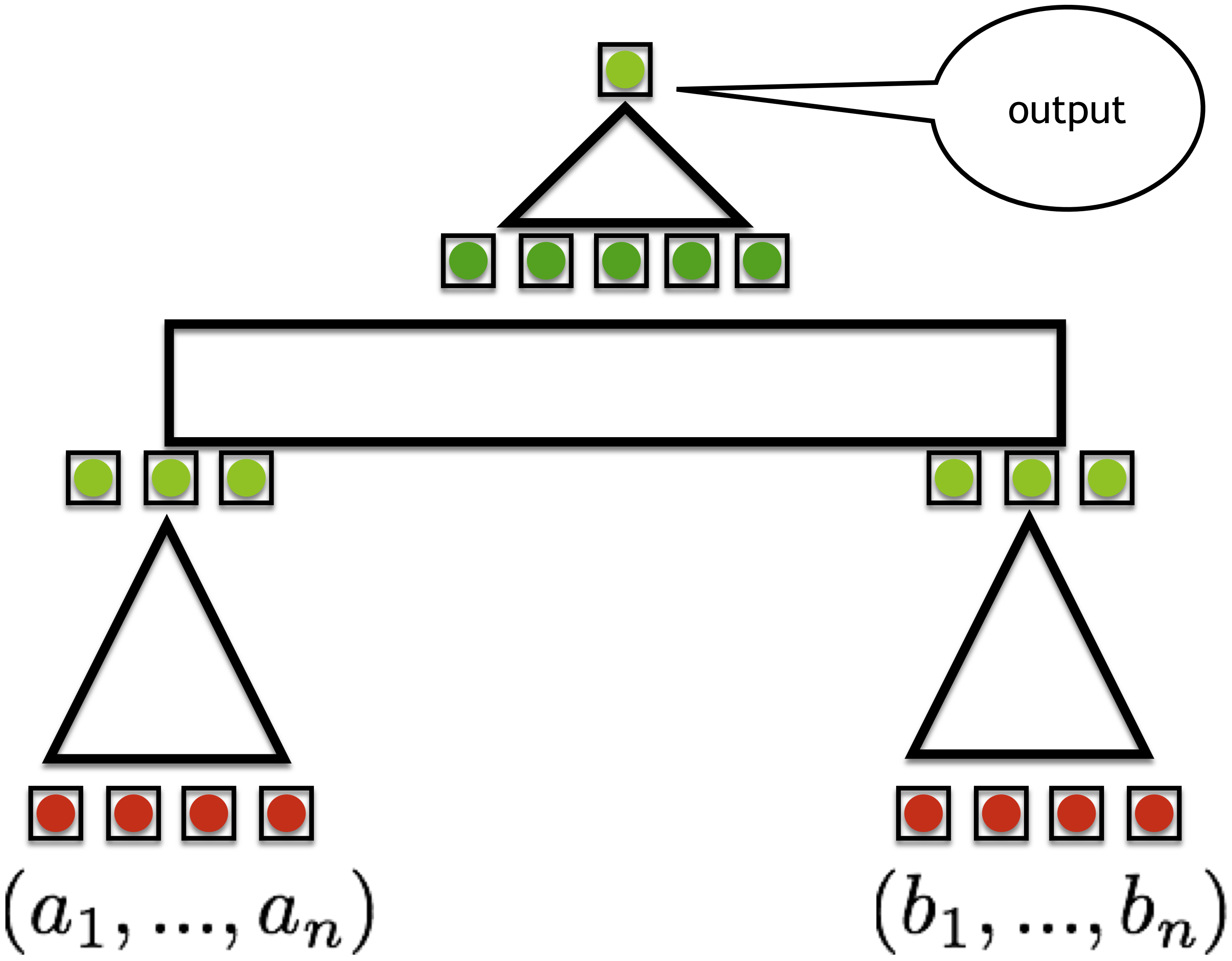
Recent Work (Sentence Encoding)



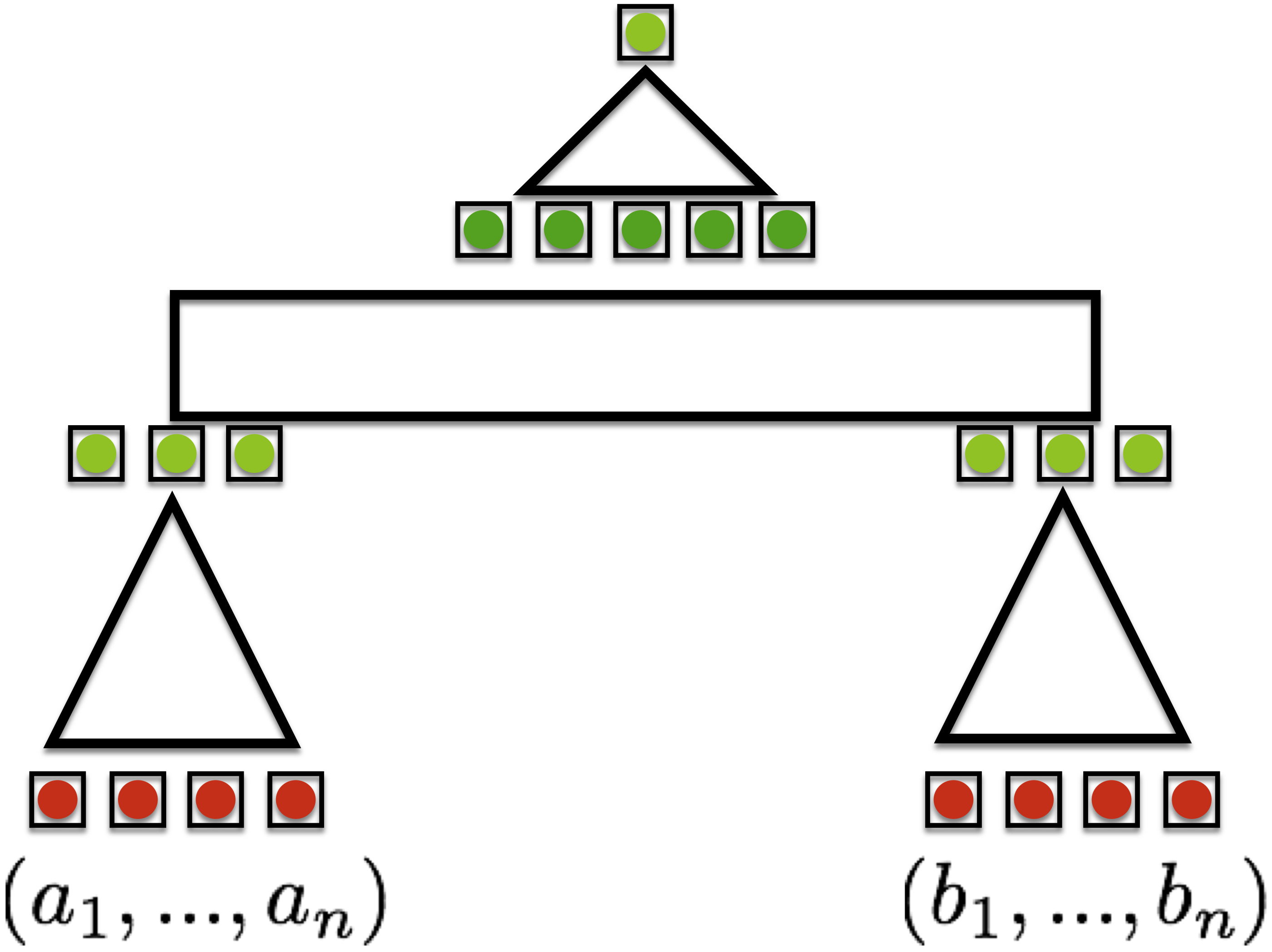
Recent Work (Sentence Encoding)



Recent Work (Sentence Encoding)



Recent Work (Sentence Encoding)

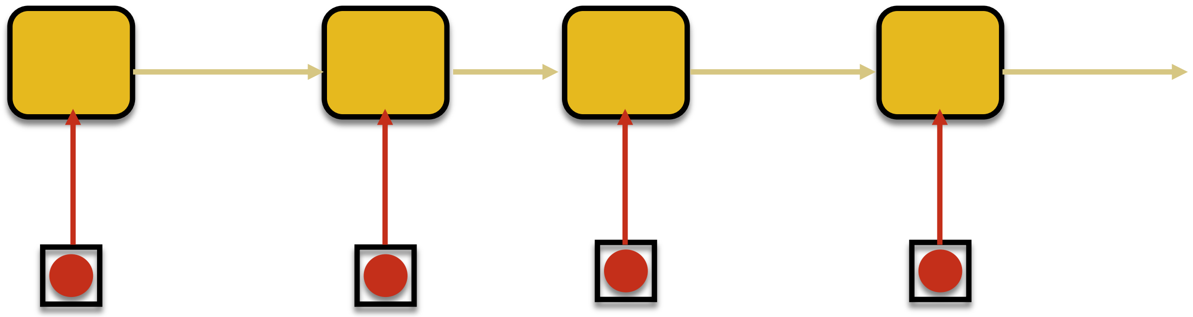


Lot of papers using this family of neural architectures:

Hu et al. (2014)
Bowman et al. (2015)
He et al. (2015)

Recent Work (Seq2Seq)

encoder recurrent neural network

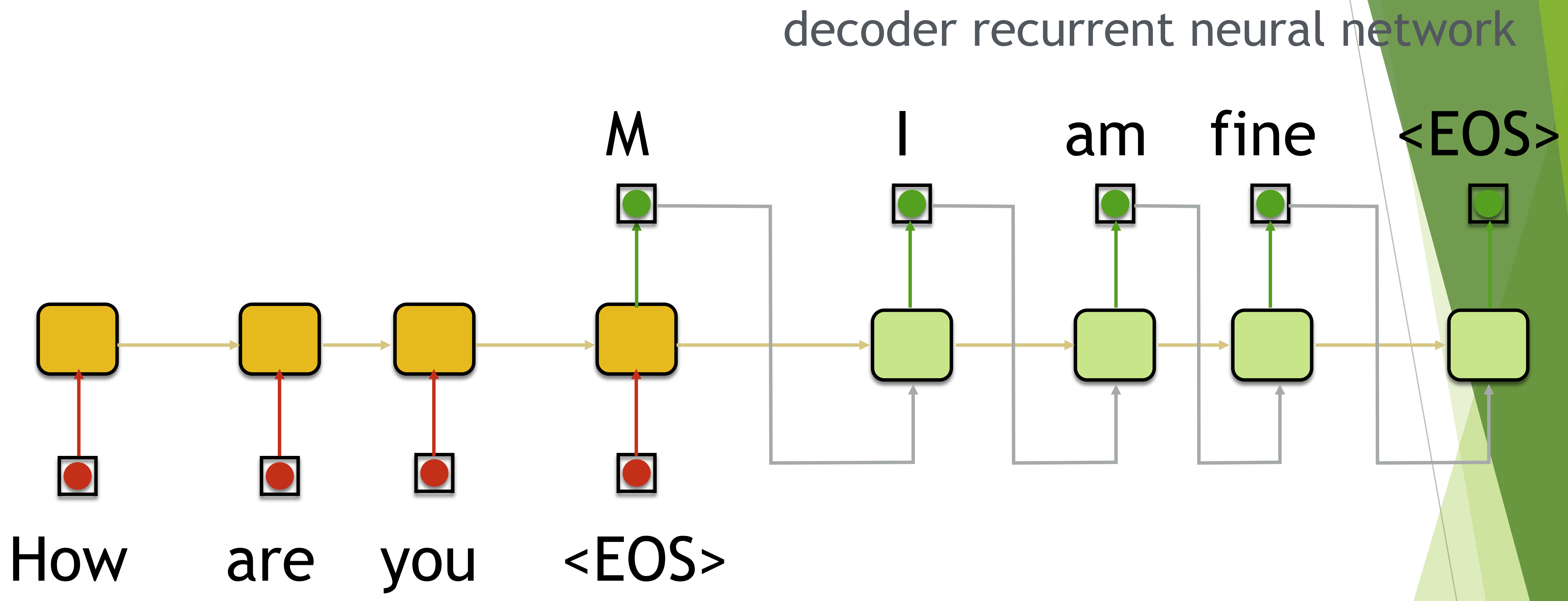


How are you <EOS>

model for machine translation

(Sutskever et al. 2014, Cho et al. 2014)

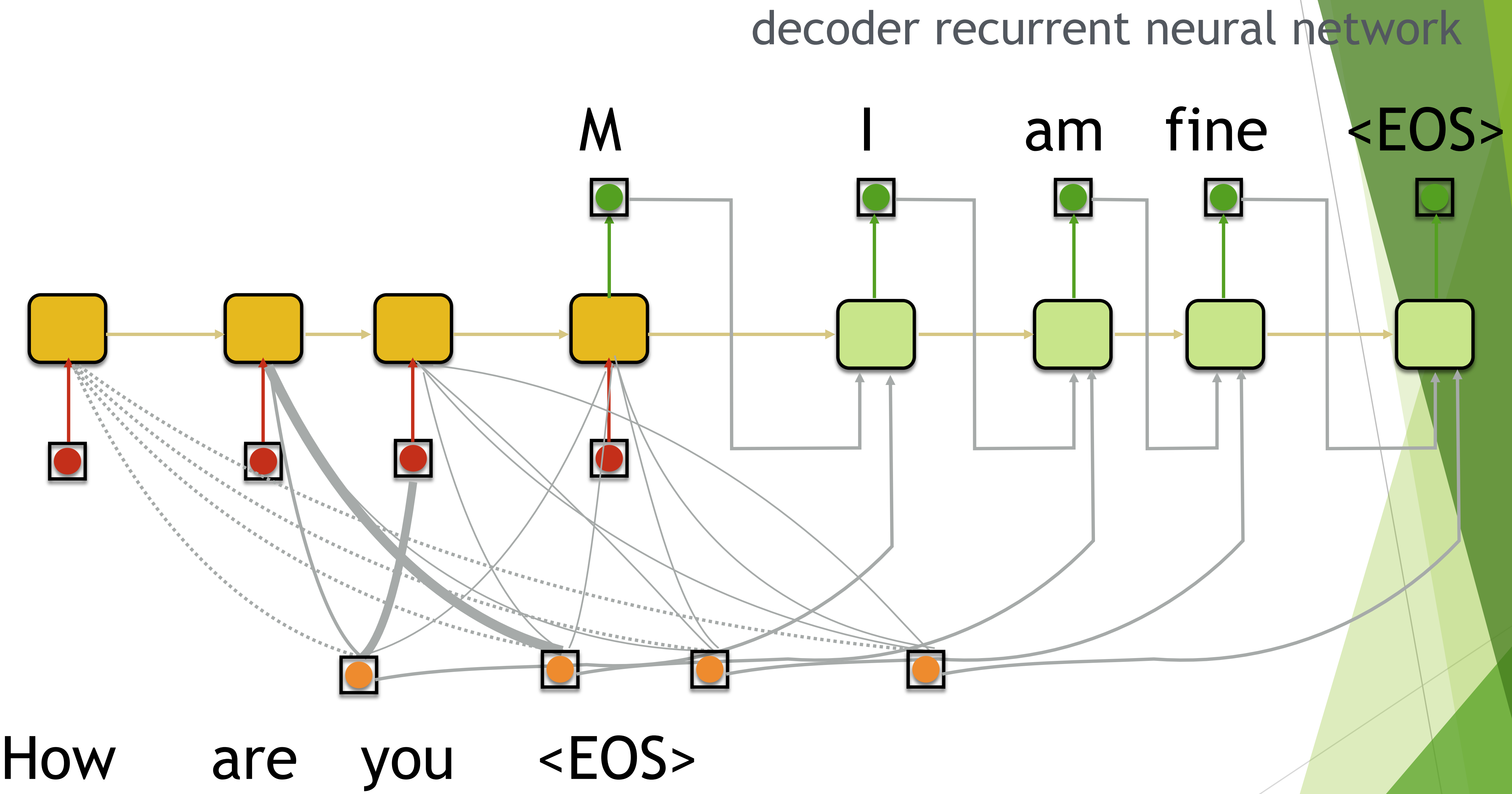
Recent Work (Seq2Seq)



model for machine translation

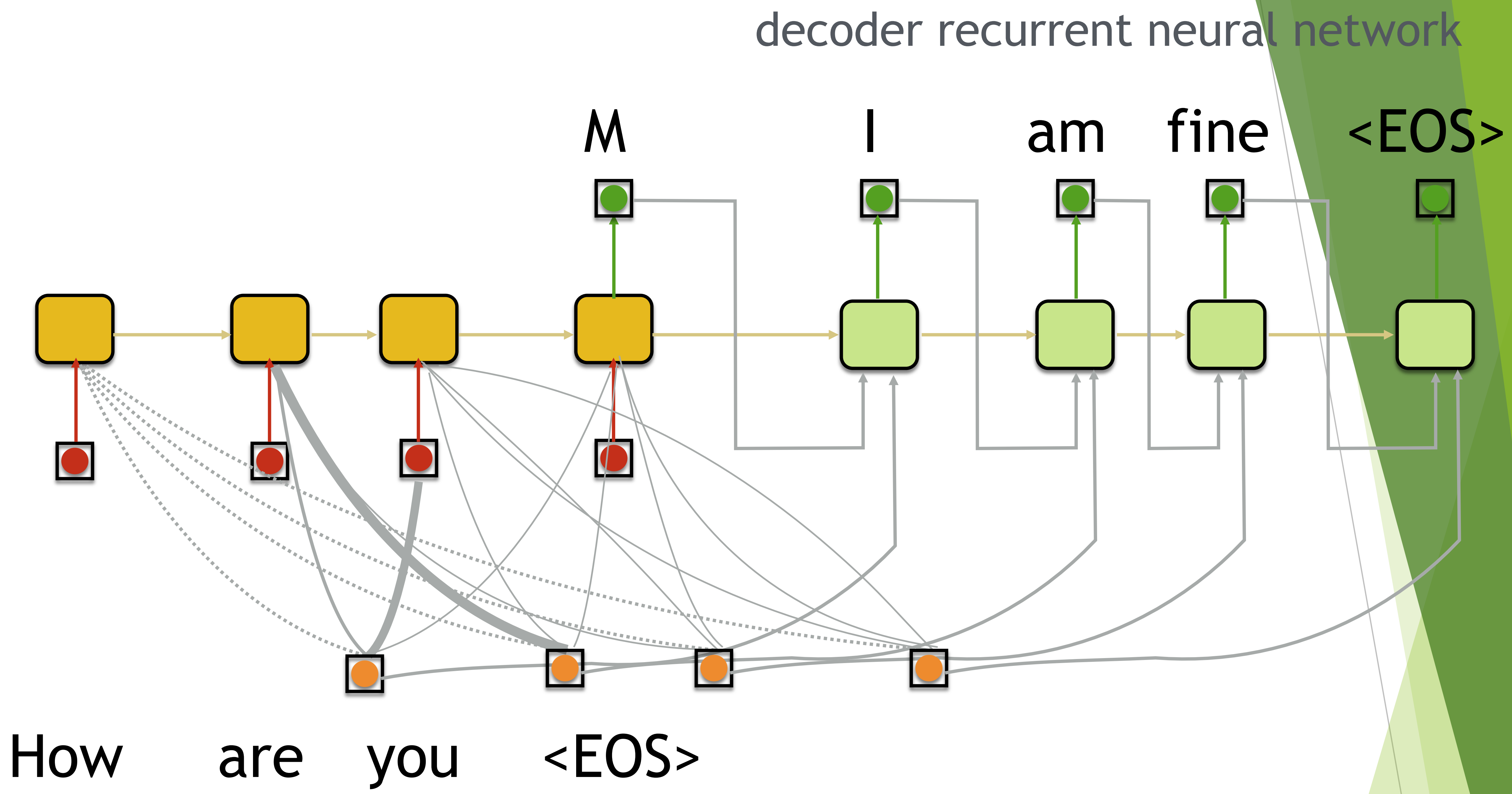
(Sutskever et al. 2014, Cho et al. 2014)

Recent Work



sequence to sequence model with attention

(Bahdanau et al. 2014)



machine translation

(Bahdanau et al. 2014)

reading comprehension

(Hermann et al. 2015)

sentence similarity/entailment

(Rocktaschel et al. 2015, Wang and Jiang 2015, Cheng et al. 2016)

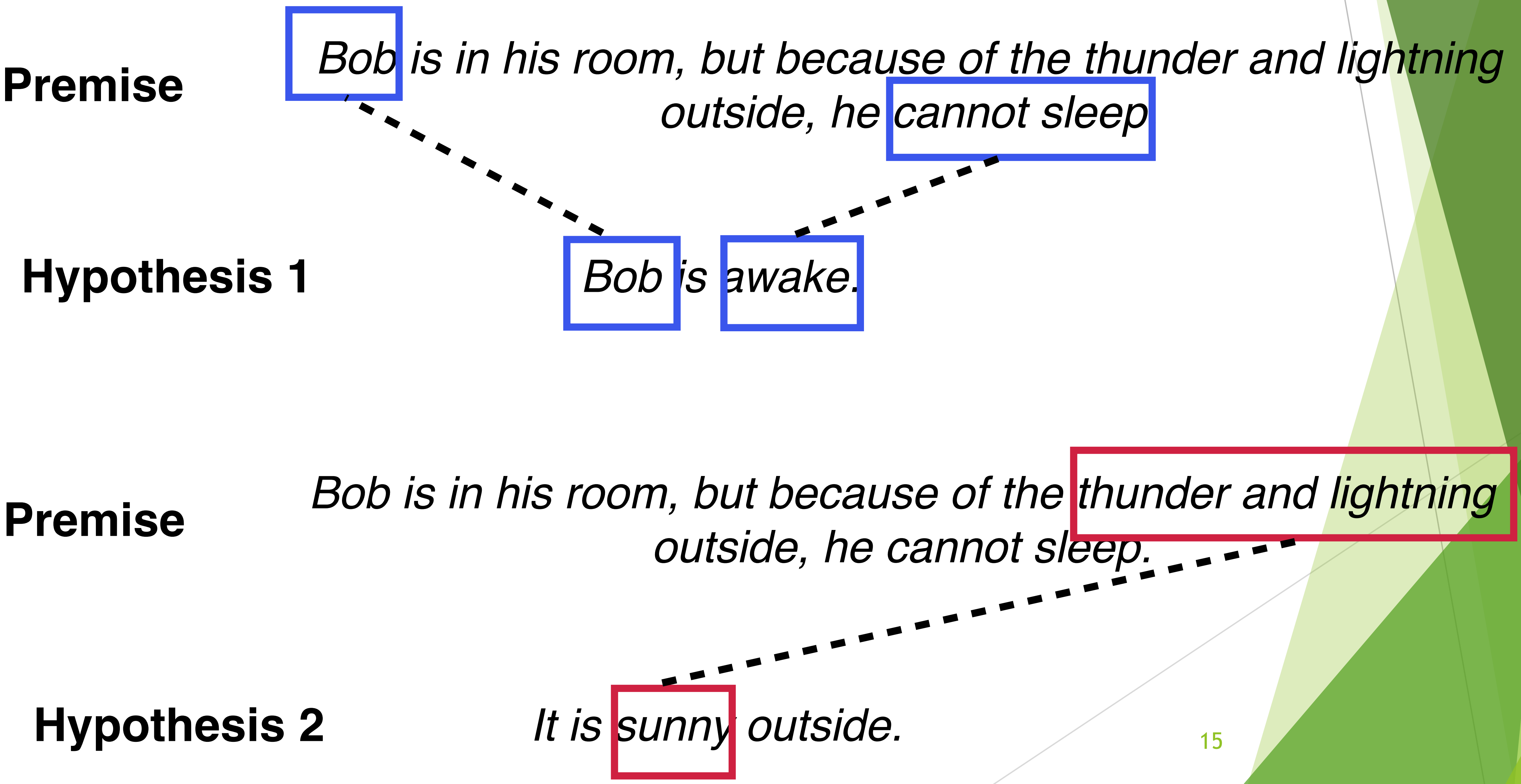
Motivation for this Work

- Alignment plays key role in many NLP tasks:
  - Machine translation [Koehn, 2009]
  - Sentence Similarity [Haghighi et al., 2005; Koehn, 2009; Das and Smith, 2009, Chang et al., 2010; Fader et al., 2013]
  - Natural Language Inference [Marsi and Krahmer, 2005; McCartney et al., 2006; Hickl and Bensley, 2007; McCartney et al., 2008]
  - Semantic Parsing [Andreas et al., 2013]
- Attention is the neural counterpart to alignment [Bahdanau et al. 2014]

Motivation for this Work

How well can we do with just alignment/attention,

without building complex sentence representations?



Premise *Bob is in his room, but because of the thunder and lightning outside, he cannot sleep*

Hypothesis 1 *Bob is awake.*

Premise *Bob is in his room, but because of the thunder and lightning outside, he cannot sleep.*

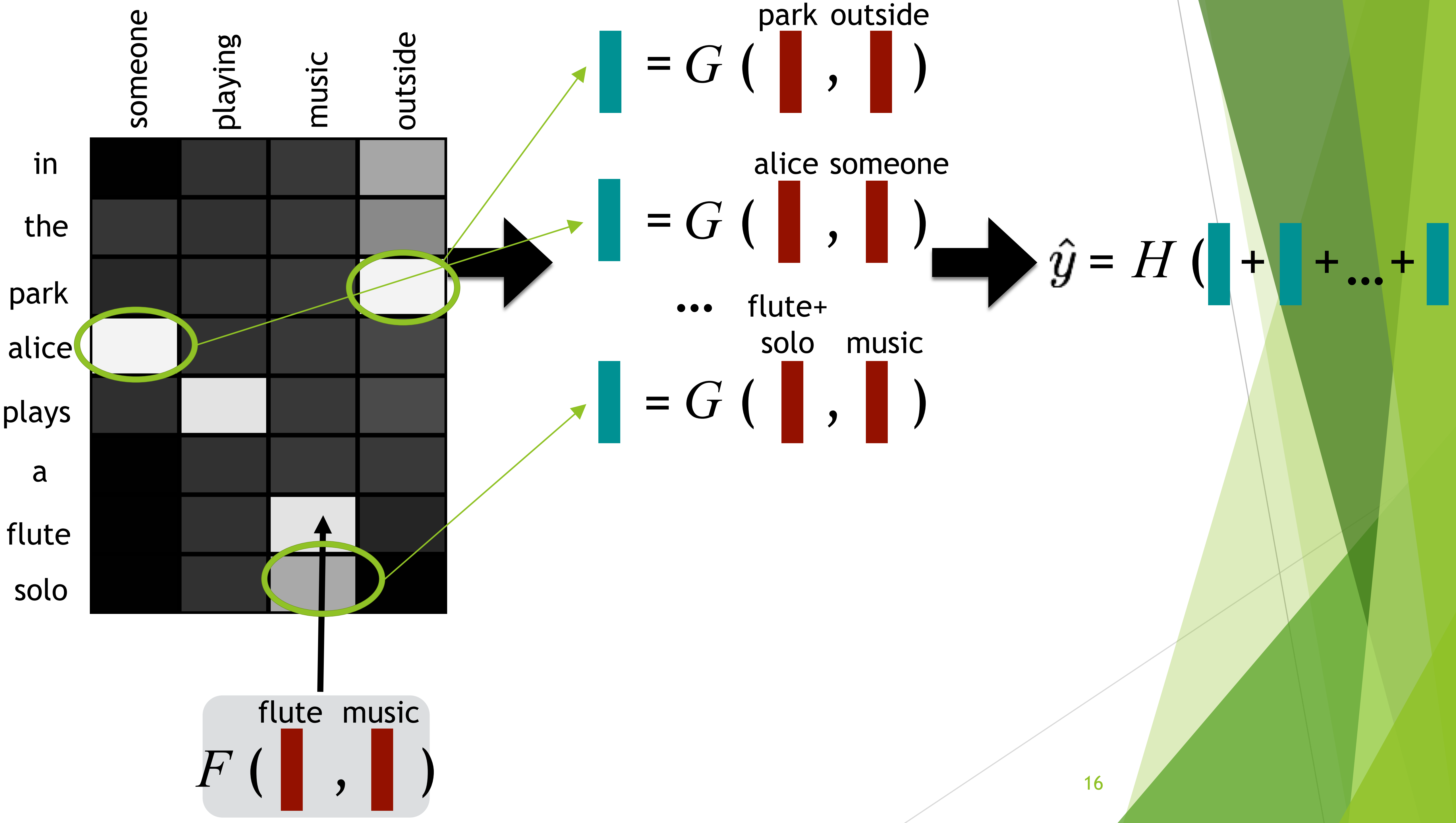
Hypothesis 2 *It is sunny outside.*

Decomposable Attention

1. Attend

2. Compare

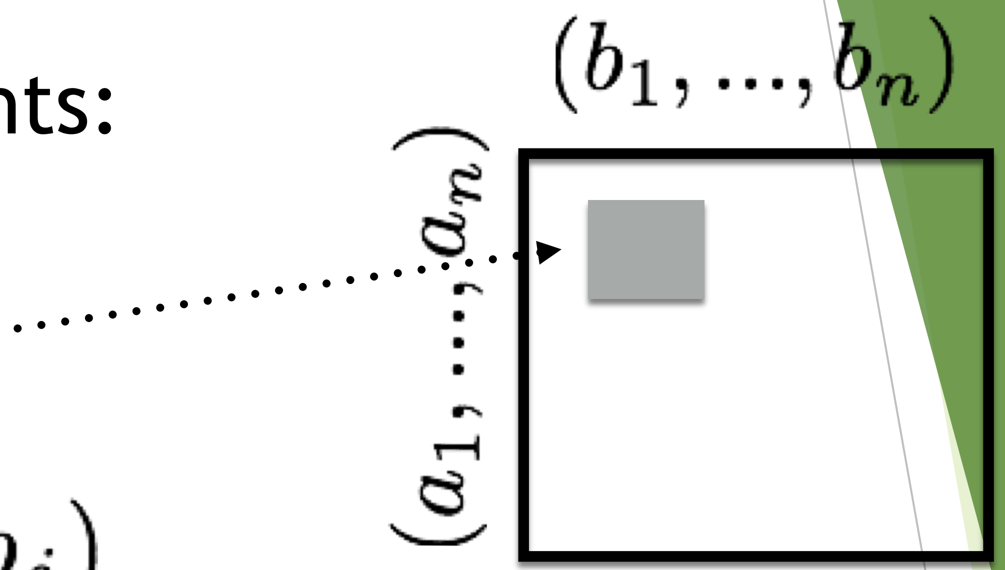
3. Aggregate



Step 1: Attend

Unnormalized attention weights:

$$e_{ij} = F^*(a_i, b_j)$$



In practice, $$e_{ij} = F(a_i)^\top F(b_j)$$

$$\alpha_j = \sum_{i=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{kj})} a_i$$

$$\beta_i = \sum_{j=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})} b_j$$

sub-phrase in
sentence 1 aligned to b_j

sub-phrase in
sentence 2 aligned to a_i

Attend 2: Compare

Separately compare aligned subphrases:

$$\mathbf{v}_{1,i} := G([a_i, \beta_i]) \quad \forall i \in [1, \ldots, n]$$

$$\mathbf{v}_{2,j} := G([b_j, \alpha_j]) \quad \forall j \in [1, \ldots, n]$$

G is a feed forward network

Step 3: Aggregate

- Combine results and classify.

$$\mathbf{v}_1 = \sum_{i=1}^{n} \mathbf{v}_{1,i} \qquad \mathbf{v}_2 = \sum_{j=1}^{n} \mathbf{v}_{2,j}$$

$$\hat{\mathbf{y}} = H([\mathbf{v}_1, \mathbf{v}_2])$$

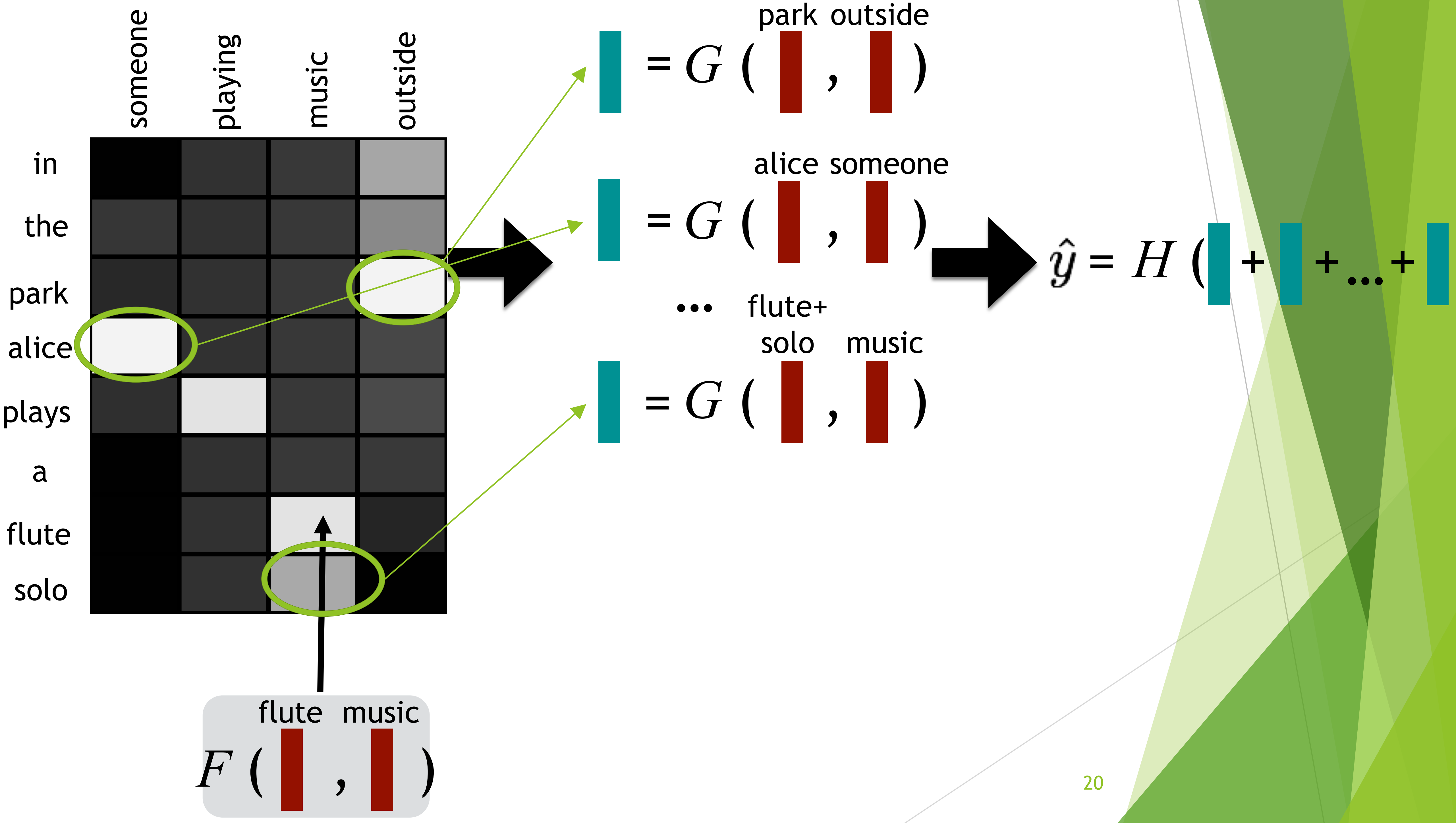
In practice, H is a feed forward neural network + linear layer + sigmoid

Decomposable Attention

1. Attend

2. Compare

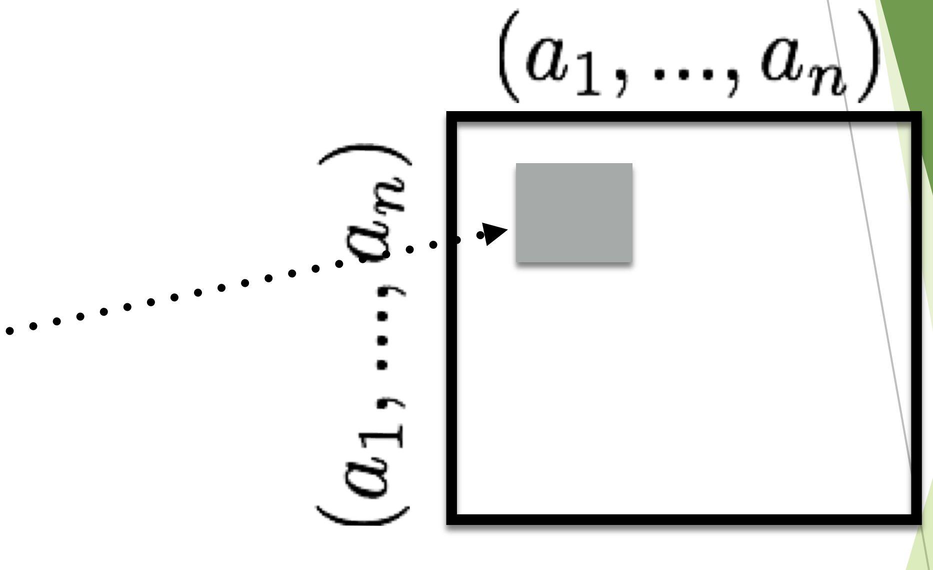
3. Aggregate



Beyond Unordered Words

- Intra-Attention - Construct a “context” using an extra attention layer
- Uses weak word order information via distance bias

$$f_{ij} = F_{\text{intra}}(a_i)^\top F_{\text{intra}}(a_j)$$



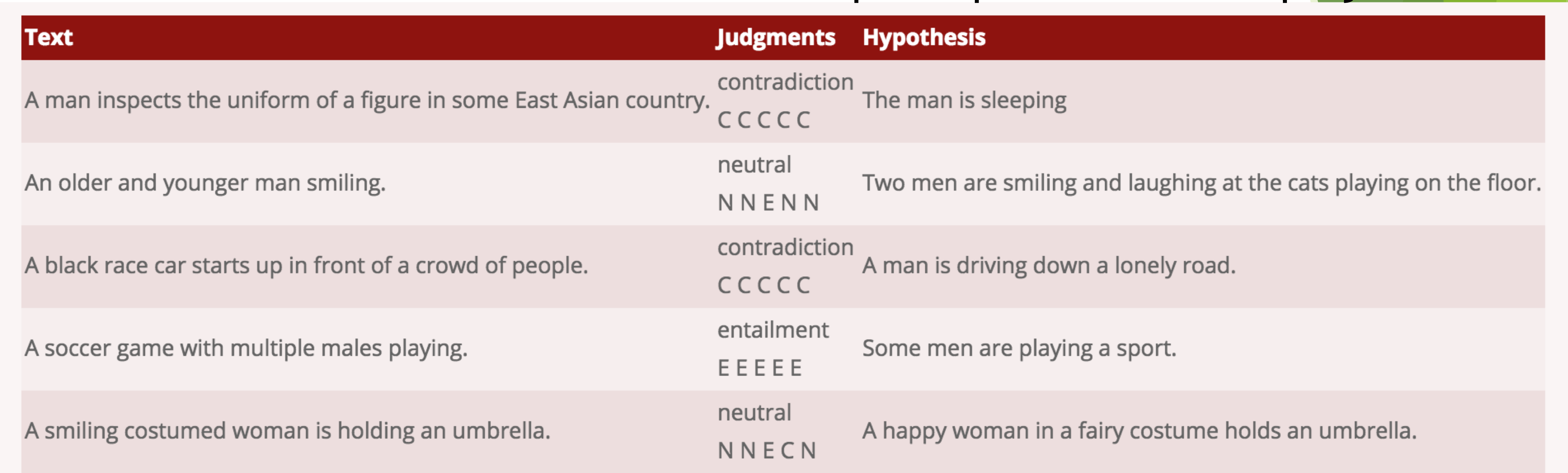
$$a'_i = \sum_{j=1}^{n} \frac{\exp(f_{ij} + d_{i-j})}{\sum_{k=1}^{n} \exp(f_{ik} + d_{i-j})} a_j \quad \Longrightarrow \quad \bar{a}_i = [a_i, a'_i]$$

The distance-sensitive bias terms $d_{i-j} \in \mathbb{R}$ provides the model with a minimal amount of sequence information, while remaining parallelizable. These terms are bucketed such that all distances greater than 10 words share the same bias.

Empirical Results

Dataset: Stanford Natural Language Inference Corpus (SNLI, Bowman et al. 2015)

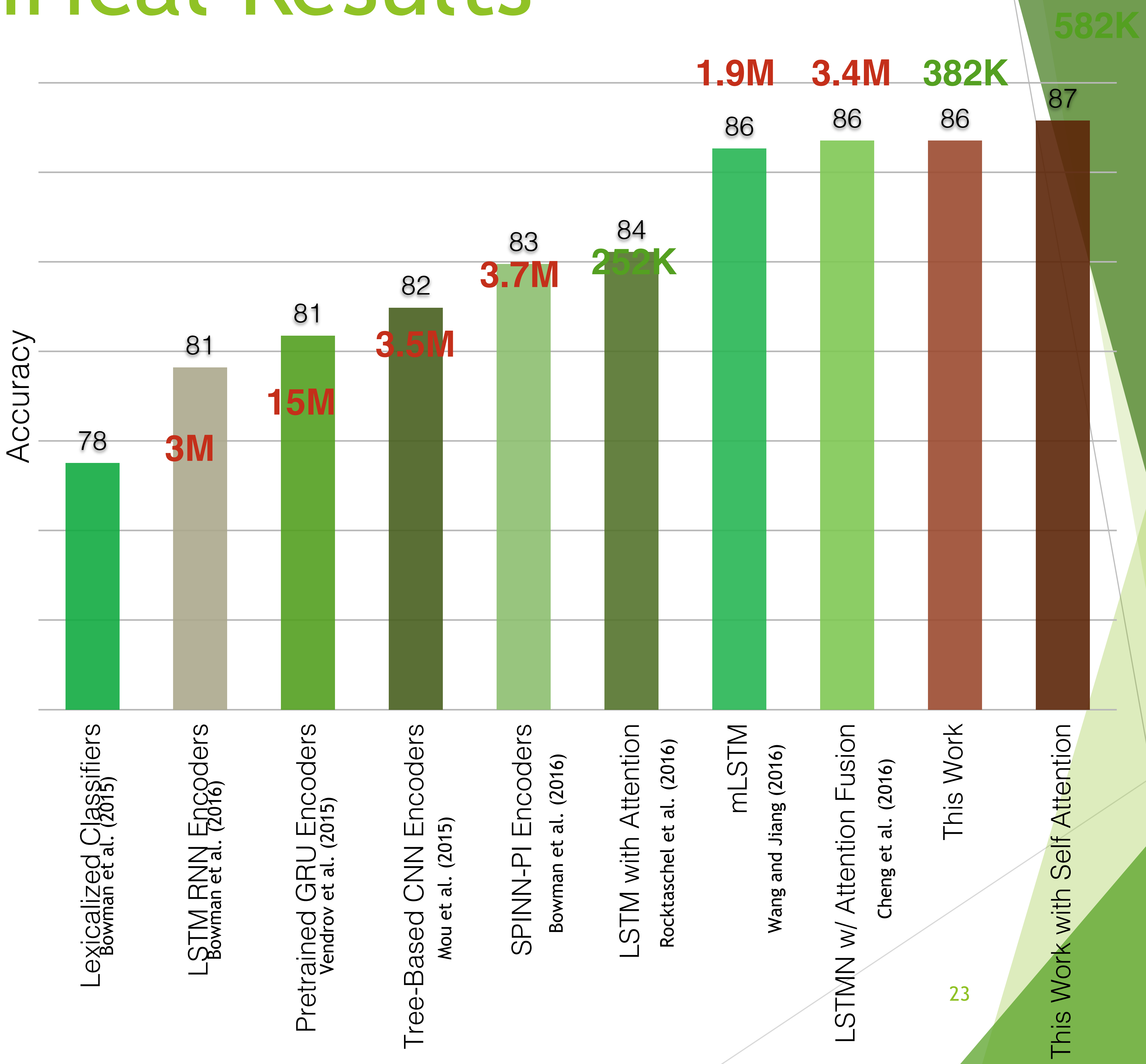
http://nlp.stanford.edu/projects/snli/



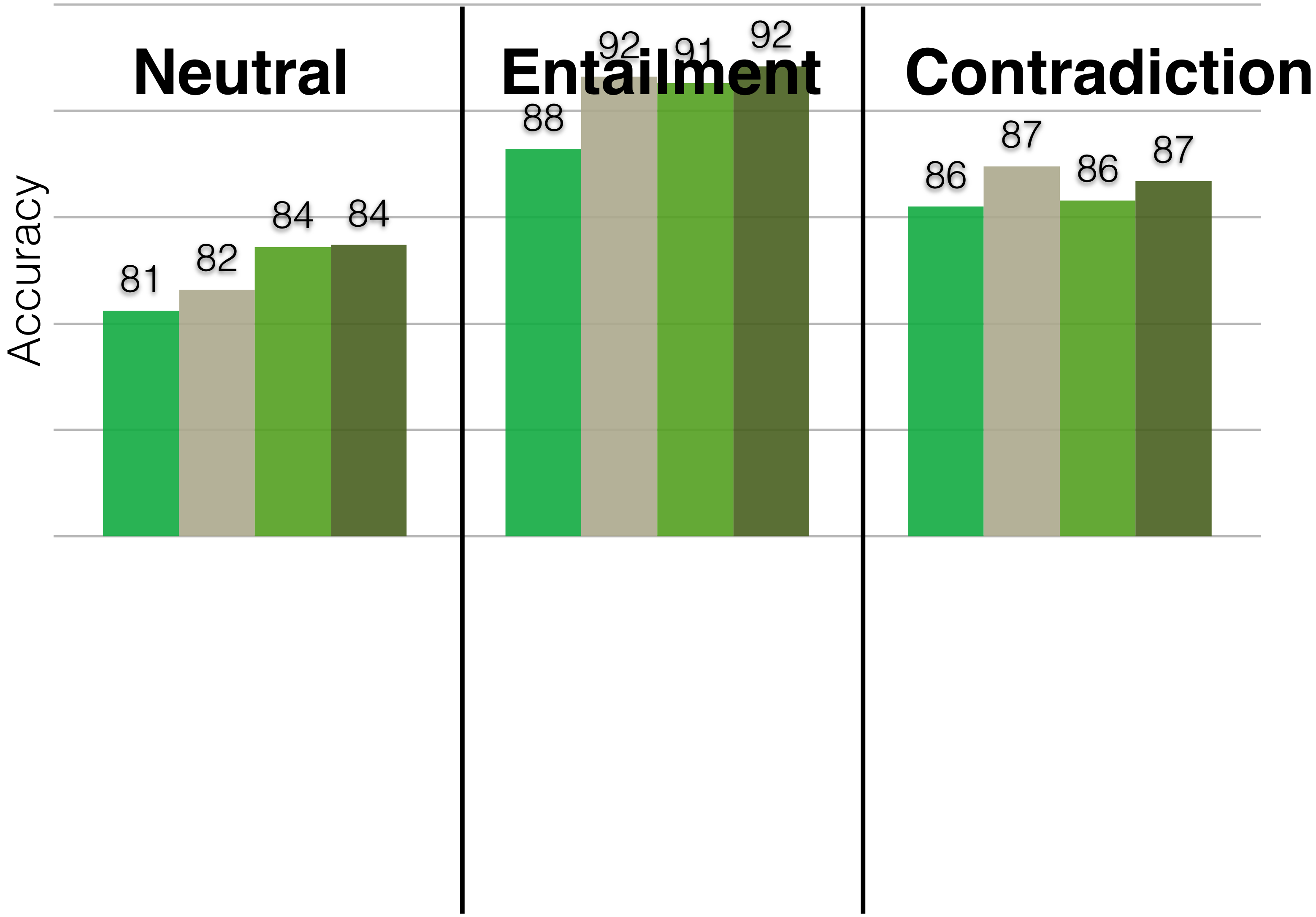
Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

549,367 sentence pairs for training
9,842 pairs for development
9,824 pairs for testing

Empirical Results



Empirical Results



Error Analysis - Wins

Sentence 1	Sentence 2	DA (vanilla)	DA (intra att.)	SPINN-PI	mLSTM	Gold
Two kids are standing in the ocean hugging each other.	Two kids enjoy their day at the beach.	N	N	E	E	N
A dancer in costumer performs on stage while a man watches.	the man is captivated	N	N	E	E	N
They are sitting on the edge of a fountain	The fountain is splashing the persons seated	N	N	C	C	N

Error Analysis - Losses

Sentence 1	Sentence 2	DA (vanilla)	DA (intra att.)	SPINN-PI	mLSTM	Gold
Two dogs play with tennis ball in field.	Dogs are watching a tennis match.	N	C	C	C	C
Two kids begin to make a snowman on a sunny winter day.	Two penguins making a snowman.	N	C	C	C	C
The horses pull the carriage, holding people and a dog, through the rain.	Horses ride in a carriage pulled by a dog.	E	E	C	C	C

Headroom

Sentence 1	Sentence 2	DA (vanilla)	DA (intra att.)	SPINN-PI	mLSTM	Gold
A woman closes her eyes as she plays her cello.	The woman has her eyes open	E	E	E	E	C
Two women having drinks and smoking cigarettes at the bar.	Three women are at a bar.	E	E	E	E	C
A band playing with fans watching.	A band watches the fans play	E	E	E	E	C

Conclusion

- We presented a simple attention-based approach to text similarity that is trivially parallelizable.
- Our results suggest that for at least the SNLI task pairwise comparisons are relatively more important than global sentence-level representations

Thank You