COS 598C Advanced Topics in Computer Science:
Deep Learning for Natural Language Processing

Contextualized Word Embeddings

Winter 2020

Announcements

- Course website and presentation schedule up at https://www.cs.princeton.edu/courses/archive/spring20/cos598C/

Next lecture

- **Lecturers**: Haochen Li, Daniel Wang
- **Feedback providers**: Zexuan Zhong, Jace Lu, Jinyuan Qi
- **Papers**: OpenAI GPT/BERT

Overview

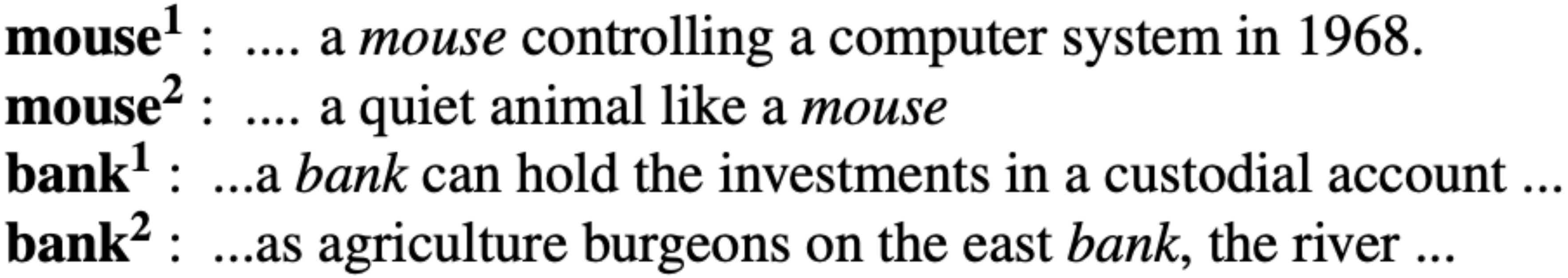
- (McCann et al, NIPS’2017) **Learned in translation: contextualized word vectors**
- (Peters et al, NAACL’2018) **Deep contextualized word representations**

Limitations of word2vec

- One vector for each word type

$$v(\text{bank}) = \begin{pmatrix} -0.224 \\ 0.130 \\ -0.290 \\ 0.276 \end{pmatrix}$$

- Polysemous words, e.g., bank, mouse



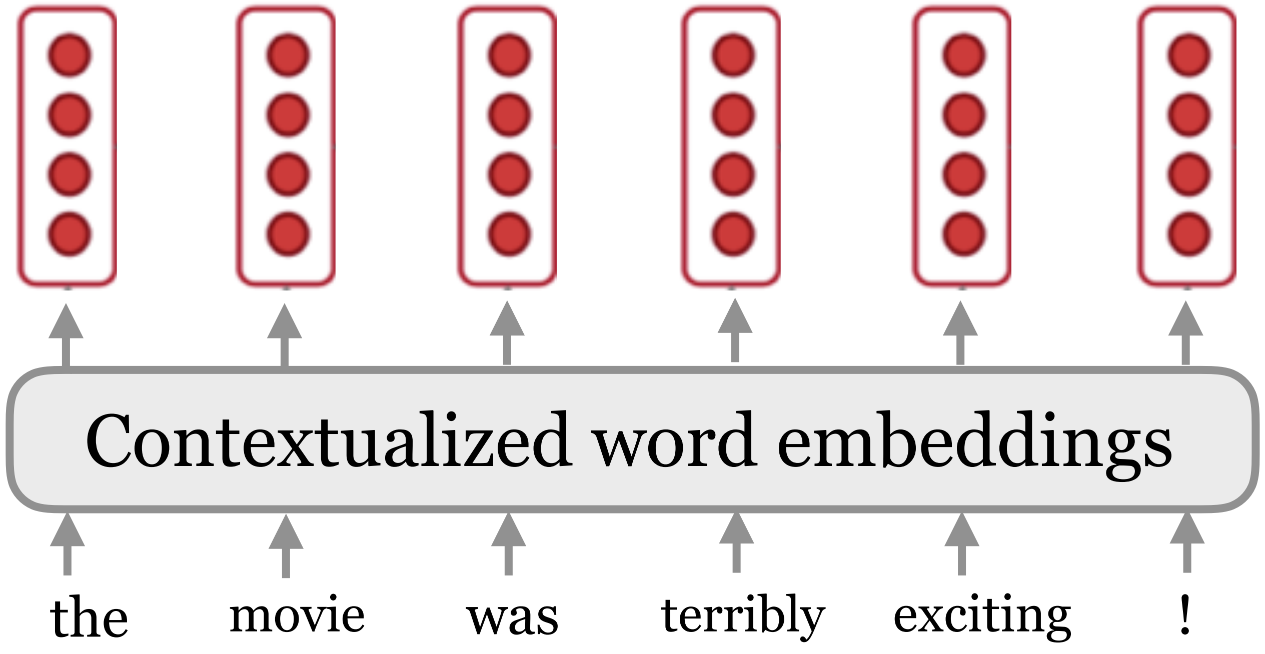
mouse^{1} : .... a *mouse* controlling a computer system in 1968.
mouse^{2} : .... a quiet animal like a *mouse*
bank^{1} : ...a *bank* can hold the investments in a custodial account ...
bank^{2} : ...as agriculture burgeons on the east *bank*, the river ...

- Words don’t appear in isolation. The word use (e.g., syntax and semantics) depends on its context. Why not learn the representations for each word in its context?

Contextualized word embeddings

Build a vector for each word conditioned on its **context**!

= representation for each token is a function of the entire input sentence



$$g : (w_1, w_2, \ldots, w_n) \longrightarrow \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$$

Contextualized word embeddings

Compute contextual vector:

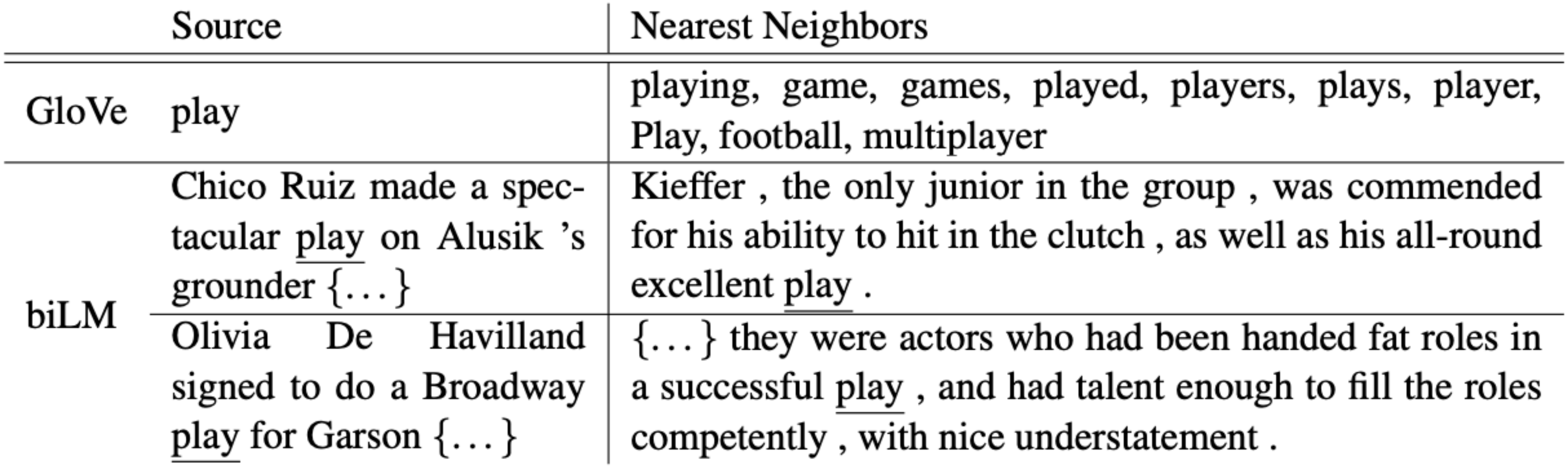
$$\mathbf{c}_k = f(w_k \mid w_1, w_2, \ldots, w_n) \in \mathbb{R}^d$$

$$f(\text{play} \mid \text{Elmo and Cookie Monster play a game})$$

$$\neq$$

$$f(\text{play} \mid \text{The Broadway play premiered yesterday})$$

Contextualized word embeddings



	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {. . . }	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {. . . }	{. . . } they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

(Peters et al, 2018): Deep contextualized word representations

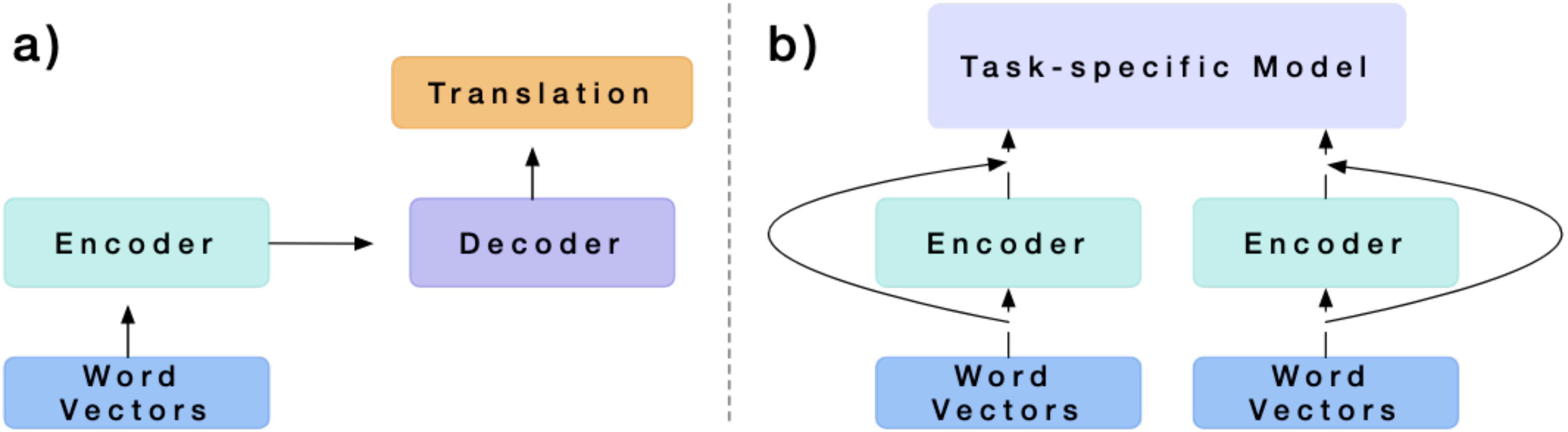
Noah Smith. 2019.
Contextual Word Representations: A Contextual Introduction

“With hindsight, we can now see that by representing word types independent of context, we were solving a problem that was harder than it needed to be. Because words mean different things in different contexts, we were requiring that type representations capture all of the possibilities. Moving to word token vectors simplifies things, asking the word token representation to capture only what a word means in this context. For the same reasons that the collection of contexts a word type is found in provide clues about its meaning(s), a particular token’s context provides clues about its specific meaning.”

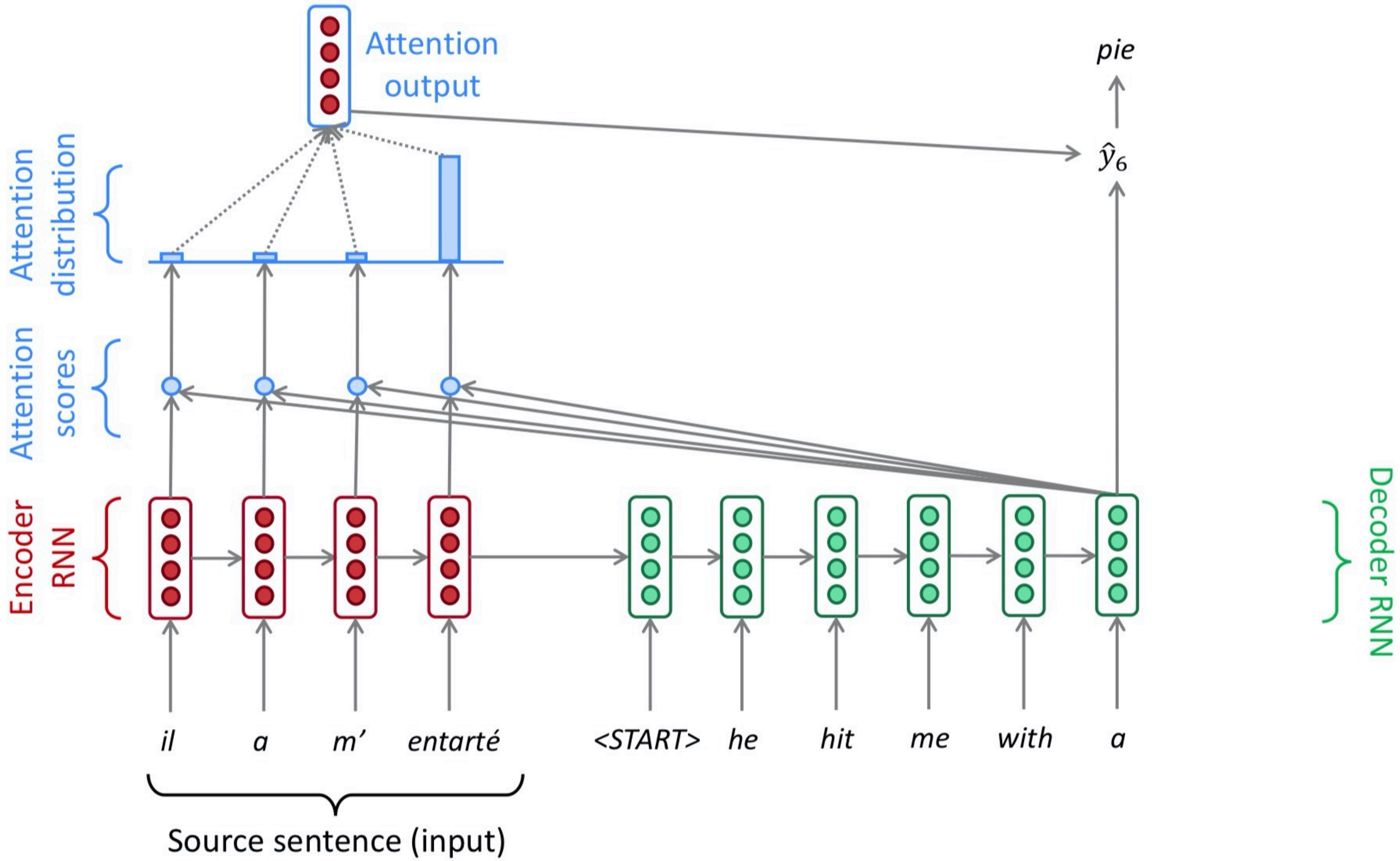
CoVe

Key idea:

- Train a standard sequence-to-sequence (with attention) model for English-to-German translation.
- Take the encoder directly as contextualized word embeddings.



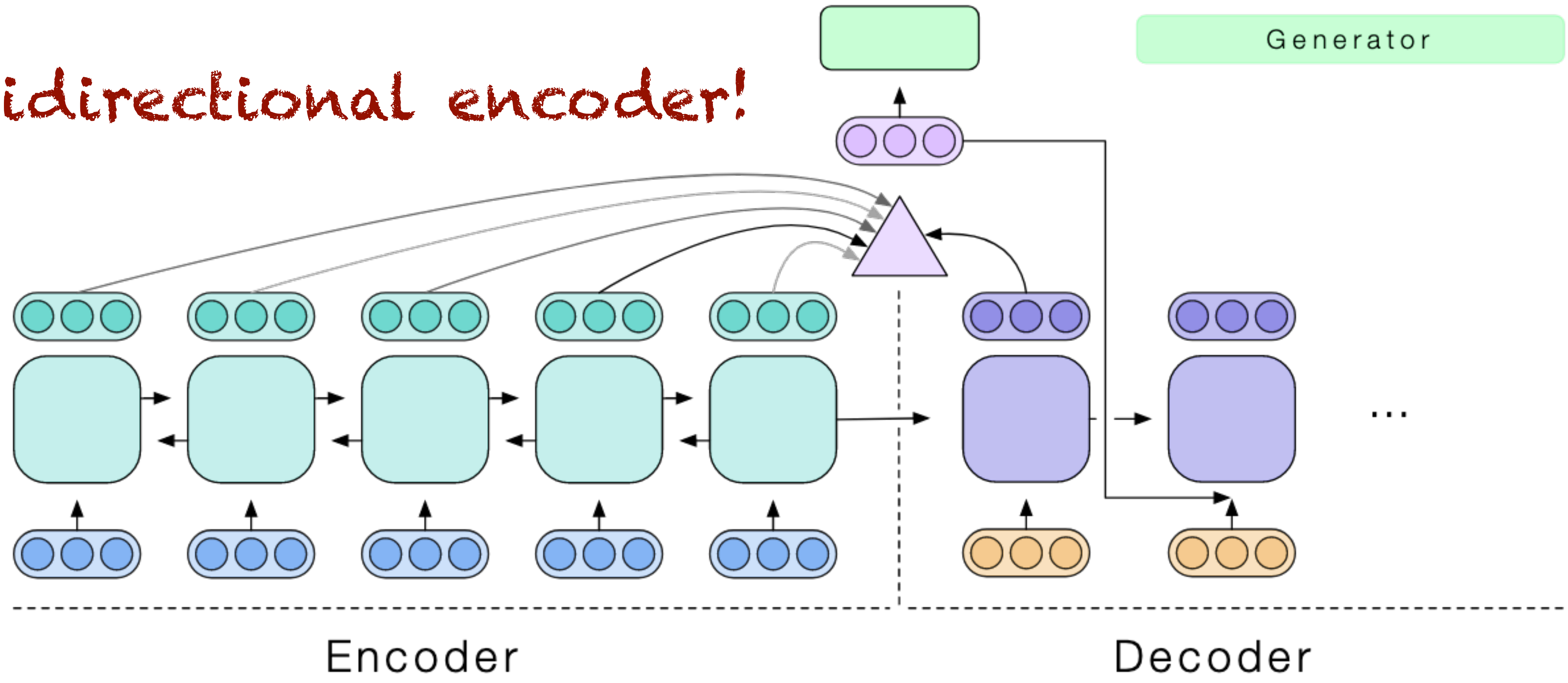
Sequence-to-sequence with attention



(slide credit: Abigail See)

CoVe

Bidirectional encoder!



$\mathrm{CoVe}(w) = \text{MT-LSTM}(\mathrm{GloVe}(w))$

$\tilde{w} = [\mathrm{GloVe}(w); \mathrm{CoVe}(w)]$

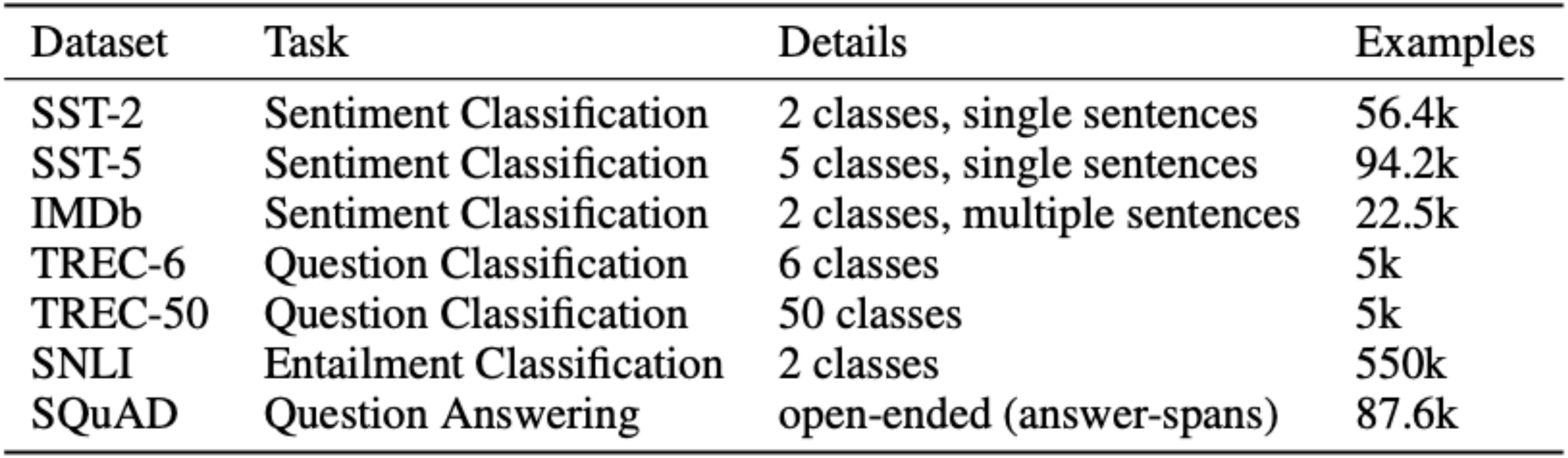
Training details

Machine Translation datasets

- Small: 30,000 sentence pairs (Flickr captions)
- Medium: 209,772 (TED presentations)
- Large: 7M (Web crawl data, news, European Parliament proceedings…)

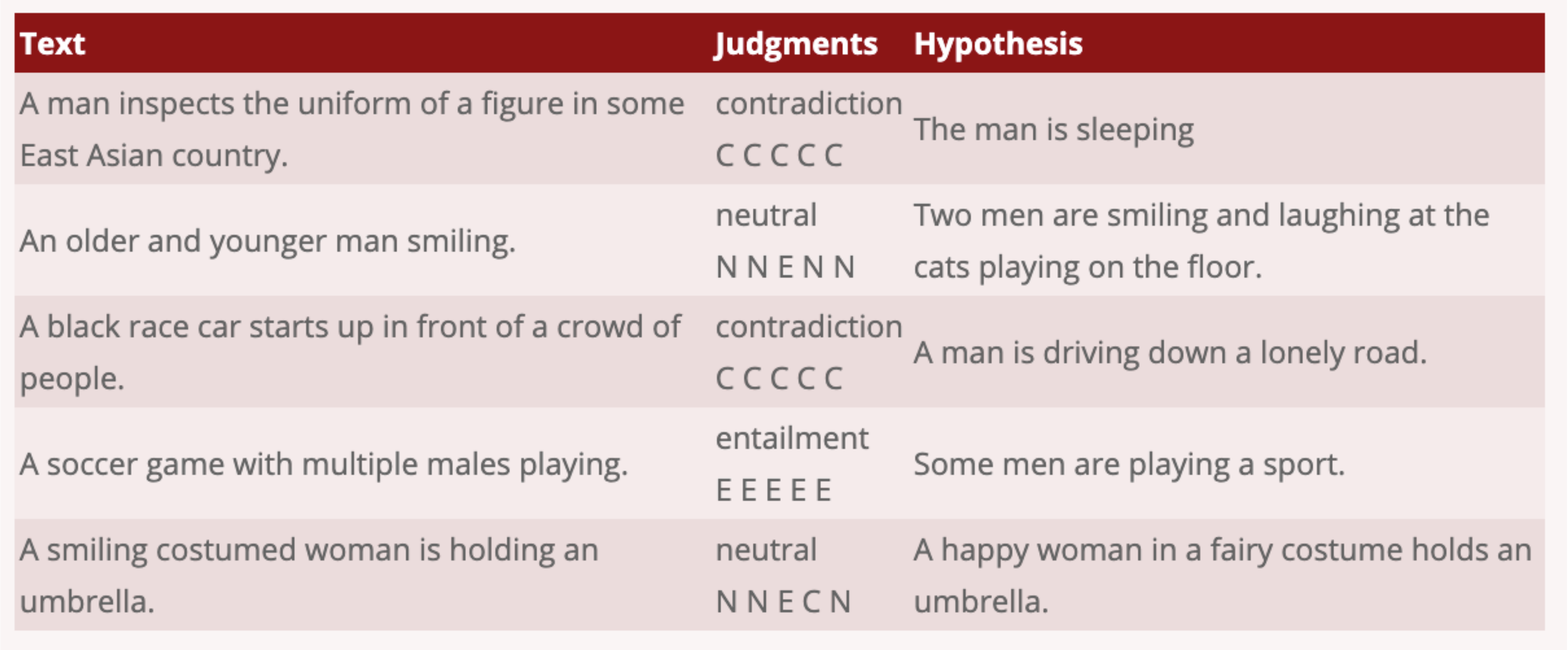
- Encoder: two-layer **bidirectional** LSTMs
- Decoder: two-layer unidirectional LSTMs

Evaluation



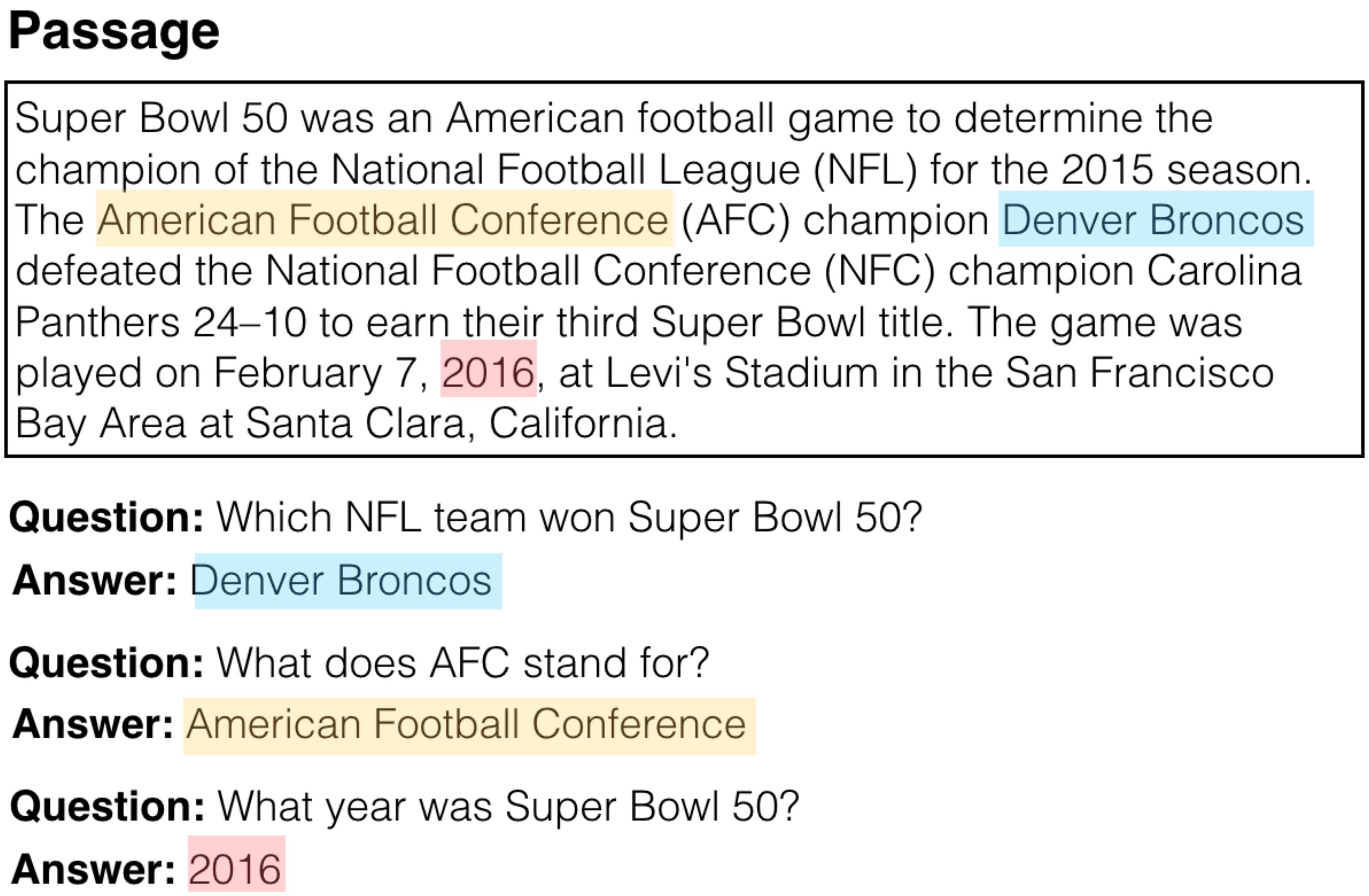
Dataset	Task	Details	Examples
SST-2	Sentiment Classification	2 classes, single sentences	56.4k
SST-5	Sentiment Classification	5 classes, single sentences	94.2k
IMDb	Sentiment Classification	2 classes, multiple sentences	22.5k
TREC-6	Question Classification	6 classes	5k
TREC-50	Question Classification	50 classes	5k
SNLI	Entailment Classification	2 classes	550k
SQuAD	Question Answering	open-ended (answer-spans)	87.6k

Evaluation: 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.

Evaluation: SQuAD



Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

Question: Which NFL team won Super Bowl 50?

Answer: Denver Broncos

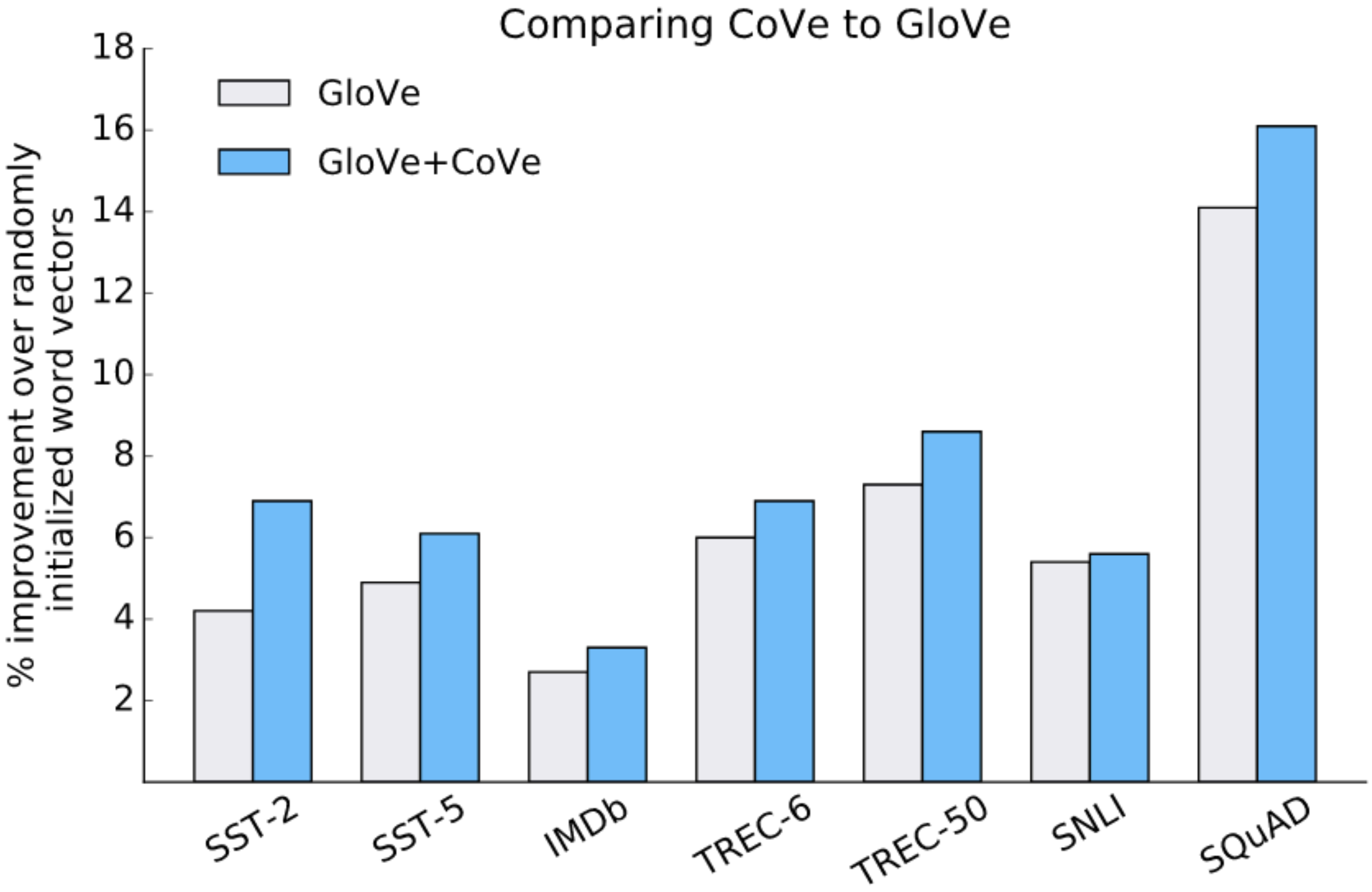
Question: What does AFC stand for?

Answer: American Football Conference

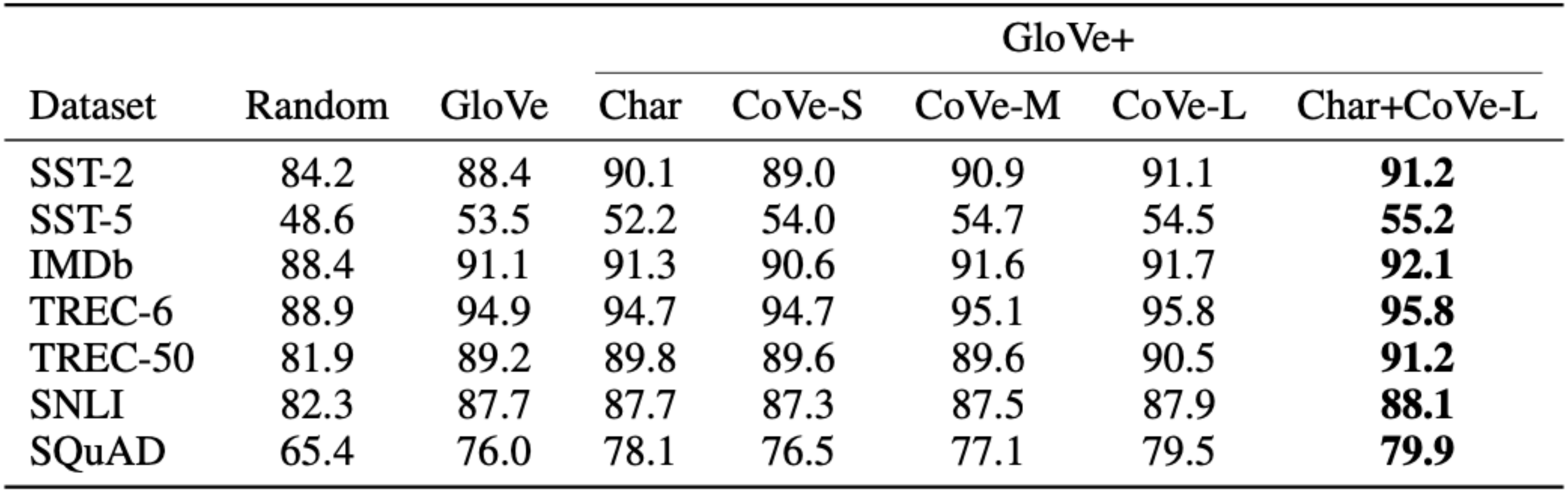
Question: What year was Super Bowl 50?

Answer: 2016

Evaluation

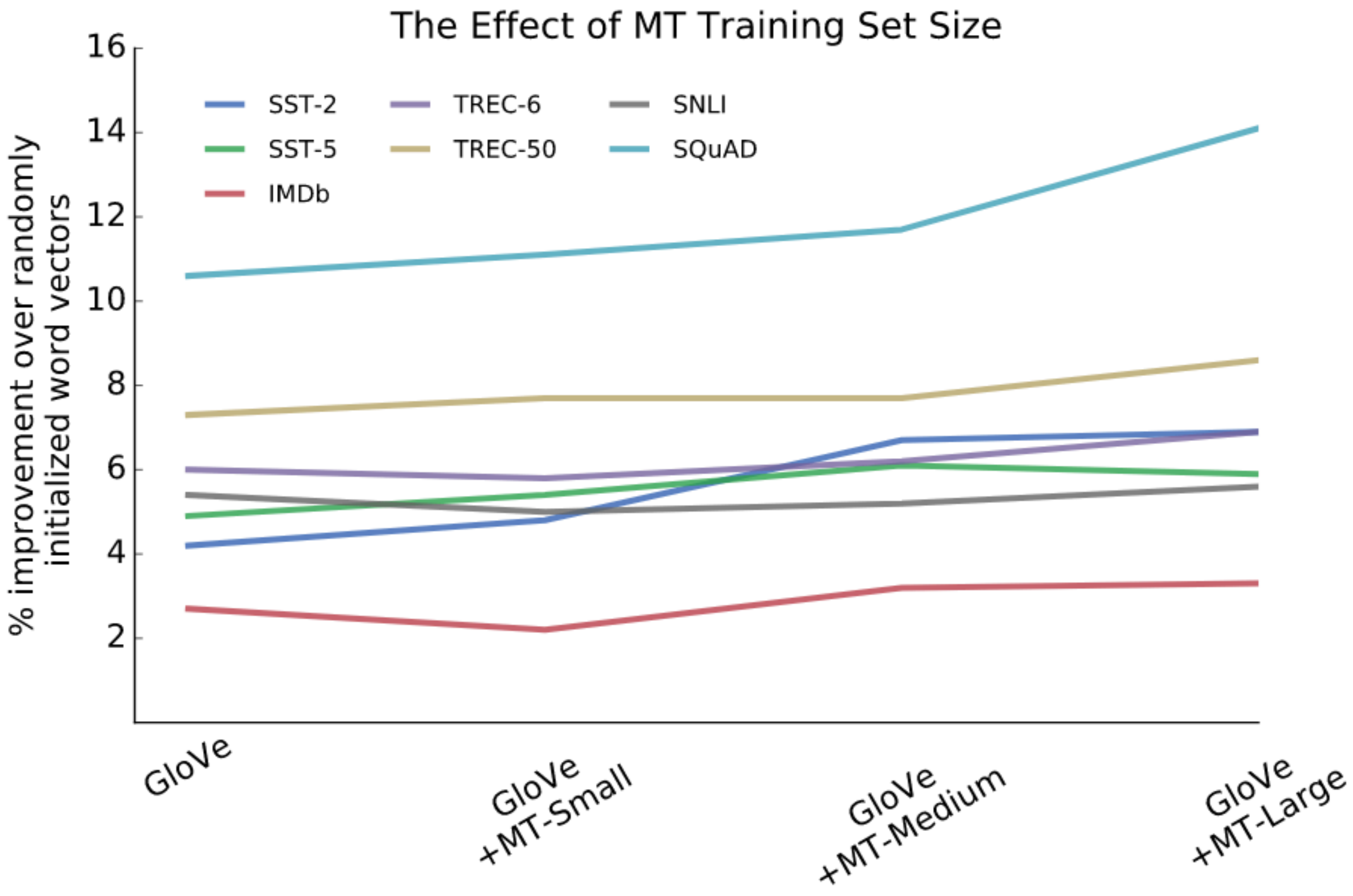

Comparing CoVe to GloVe
GloVe
GloVe+CoVe
% improvement over randomly initialized word vectors
18
16
14
12
10
8
6
4
2
SST-2
SST-5
IMDb
TREC-6
TREC-50
SNLI
SQuAD

Evaluation



			GloVe+				
Dataset	Random	GloVe	Char	CoVe-S	CoVe-M	CoVe-L	Char+CoVe-L
SST-2	84.2	88.4	90.1	89.0	90.9	91.1	**91.2**
SST-5	48.6	53.5	52.2	54.0	54.7	54.5	**55.2**
IMDb	88.4	91.1	91.3	90.6	91.6	91.7	**92.1**
TREC-6	88.9	94.9	94.7	94.7	95.1	95.8	**95.8**
TREC-50	81.9	89.2	89.8	89.6	89.6	90.5	**91.2**
SNLI	82.3	87.7	87.7	87.3	87.5	87.9	**88.1**
SQuAD	65.4	76.0	78.1	76.5	77.1	79.5	**79.9**

More MT —> Better CoVe



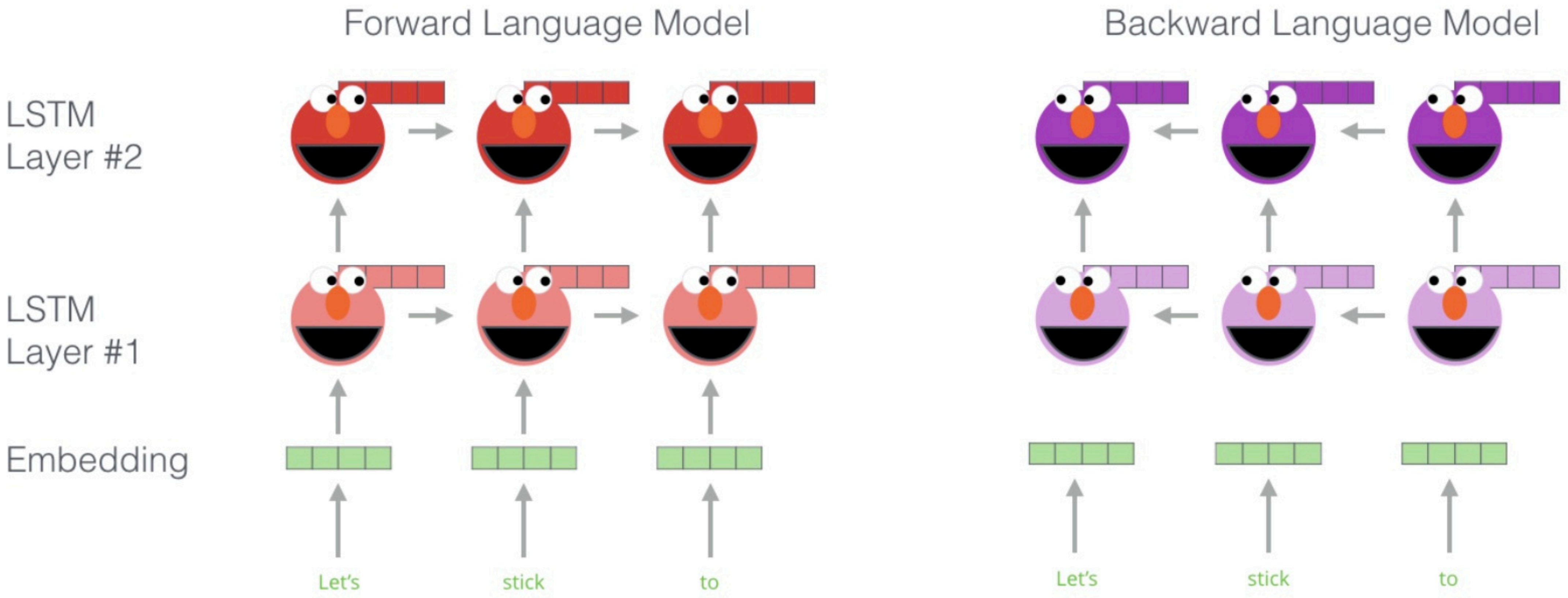
ELMo

Key idea:

- Train a forward LSTM-based LM and a backward LSTM-based LM on some large corpus

- Use the hidden states of the LSTMs for each token to compute a vector representation of each word

ELMo



# words in the sentence

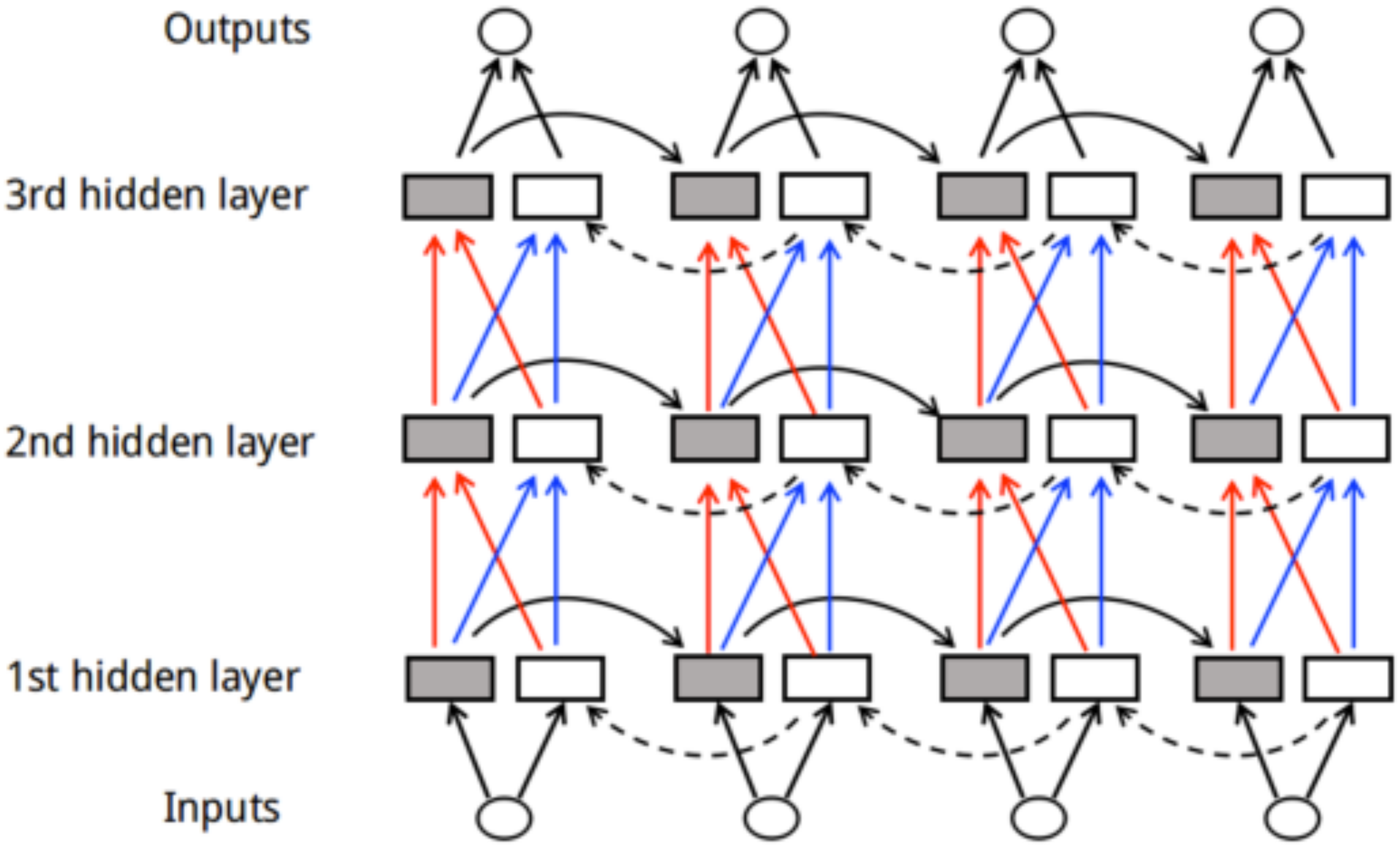
$$\sum_{k=1}^{N} \left(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \right)$$

input

softmax

ELMo != bidirectional LSTMs

- Essentially two LSTMs in different directions



How to use ELMo?

# of layers

$$\begin{aligned} R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\ &= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\}, \end{aligned}$$

$$\mathbf{h}_{k,0}^{LM} = \mathbf{x}_k^{LM}, \mathbf{h}_{k,j}^{LM} = [\overrightarrow{\mathbf{h}}_{k,j}^{LM}; \overleftarrow{\mathbf{h}}_{k,j}^{LM}]$$

$$\boxed{\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}}$$

- γ^{task}: allows the task model to scale the entire ELMo vector
- s_j^{task}: softmax-normalized weights across layers

How to use ELMo?

- Plug ELMo into any (neural) NLP model: freeze all the LMs weights and change the input representation to:

$$[\mathbf{x}_k; \mathbf{ELMo}_k^{task}]$$

- Could also insert ELMo into higher layers (will discuss soon)

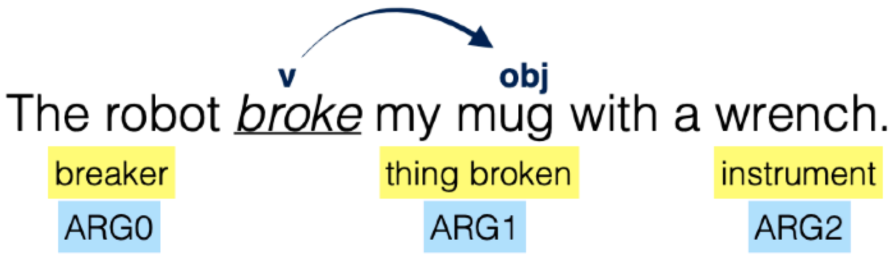
Training details

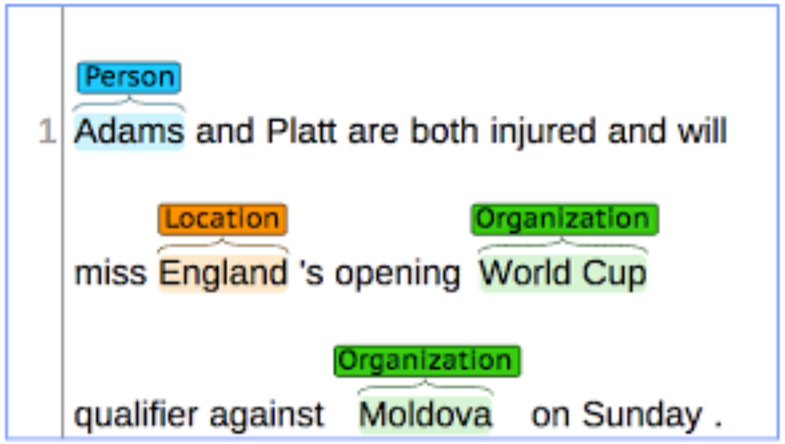
- Forward and backward LMs: 2 layers each
- Character embeddings to replace word embeddings
  - *Recommended reading: Character-Aware Neural Language Models*
  - 2048 char n-gram filters and 2 highway layers, 512 dim projection

  No-pretrained word embeddings!
- User 4096 dim hidden/cell LSTM states with 512 dim projections to next input
- A residual connection from the first to second layer
- Trained 10 epochs on 1B Word Benchmark

Evaluation

- SQuAD: question answering ✓
- SNLI: textual entailment ✓
- SRL: semantic role labeling
- Coref: coreference resolution
- NER: named entity recognition
- SST-5: sentiment analysis ✓





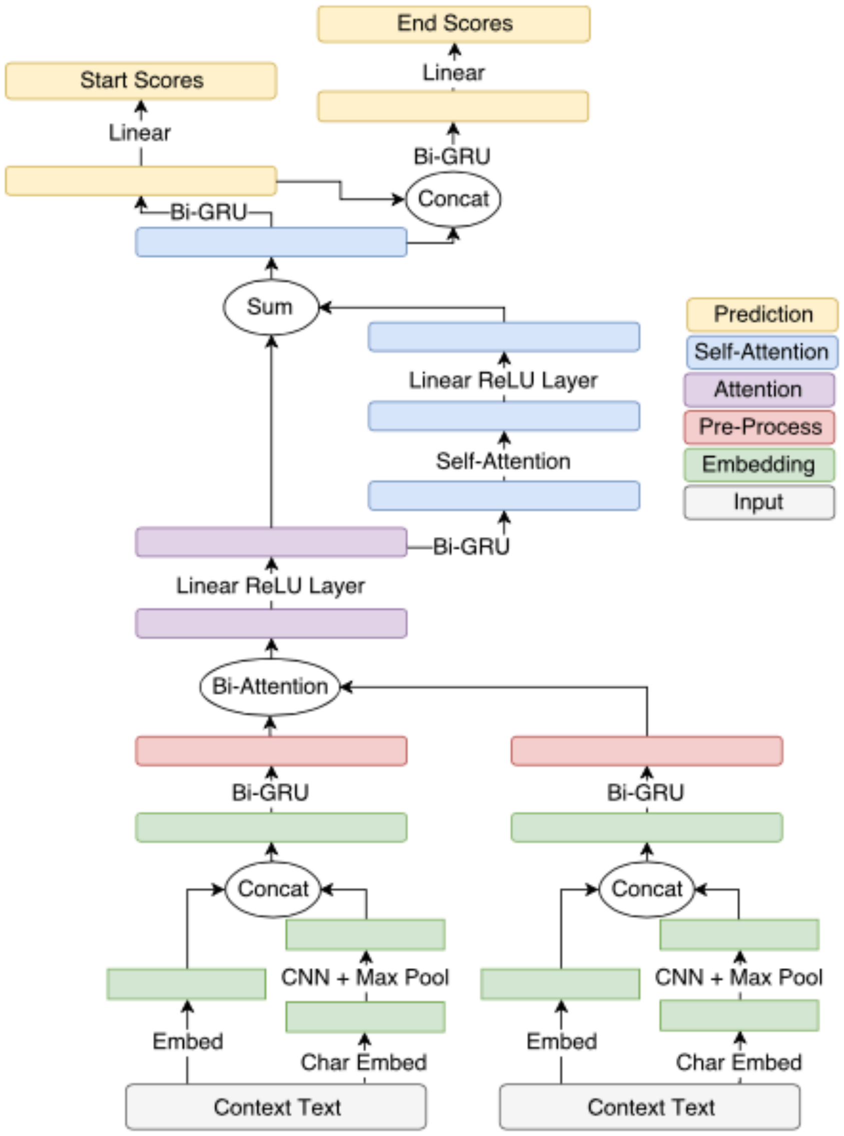
1 Adams and Platt are both injured and will miss England 's opening World Cup qualifier against Moldova on Sunday .

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

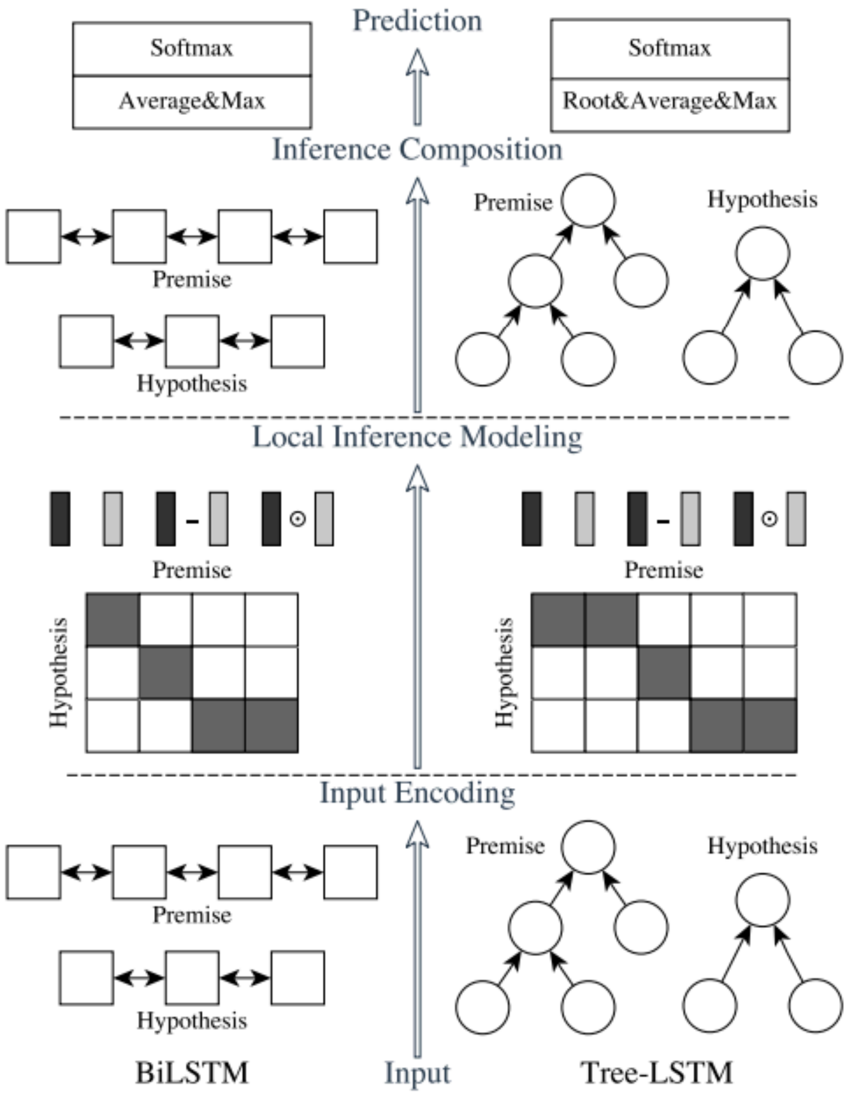
Baseline models

- SQuAD: (Clark and Gardner, 2017)
- Textual entailment: (Chen et al, 2017)
- Semantic role labeling: (He et al, 2017)
- Coreference resolution: (Lee et al, 2017)
- Named entity recognition: pre-trained word embeddings, a character-based CNN representation, two biLSTM layers and a conditional random field (CRF) loss
- Sentiment analysis: (McCann et al, 2017)

Baseline models

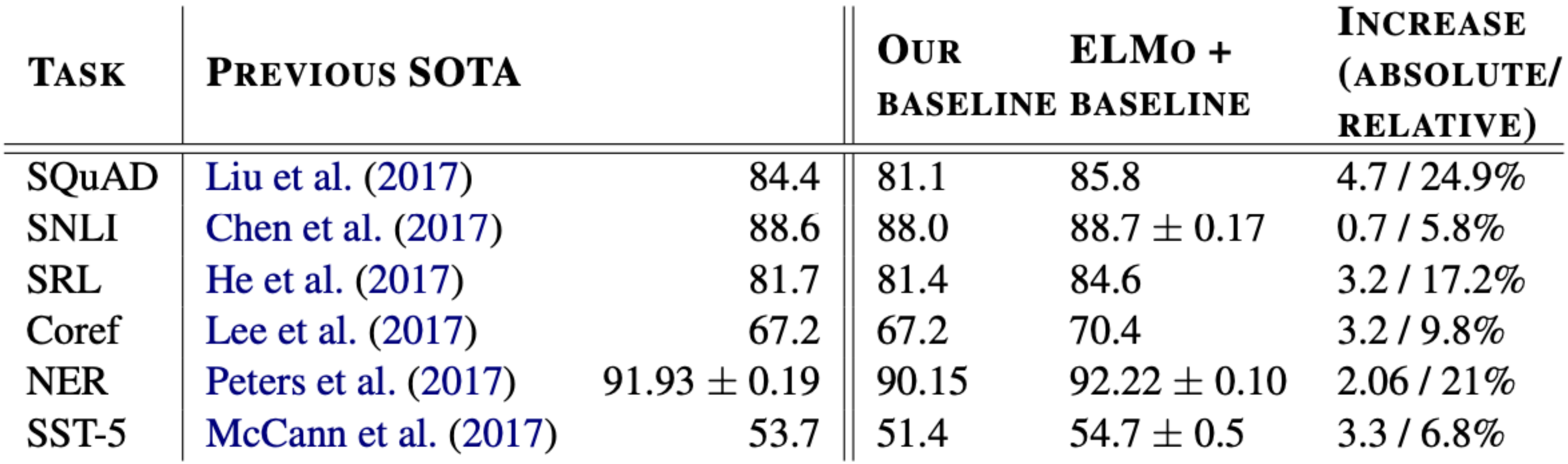


(Clark and Gardner, 2017)



(Chen et al, 2017)

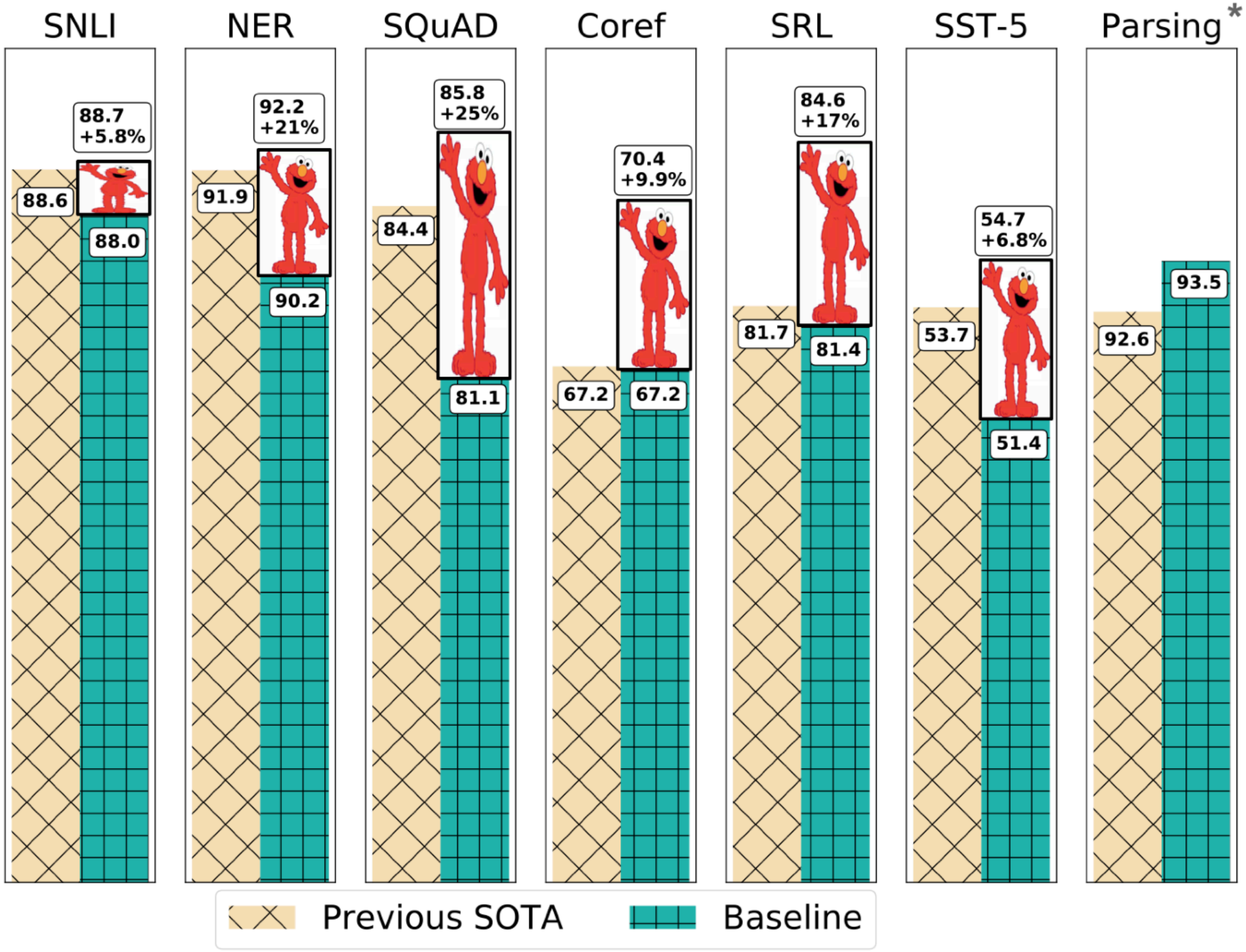
Experimental results



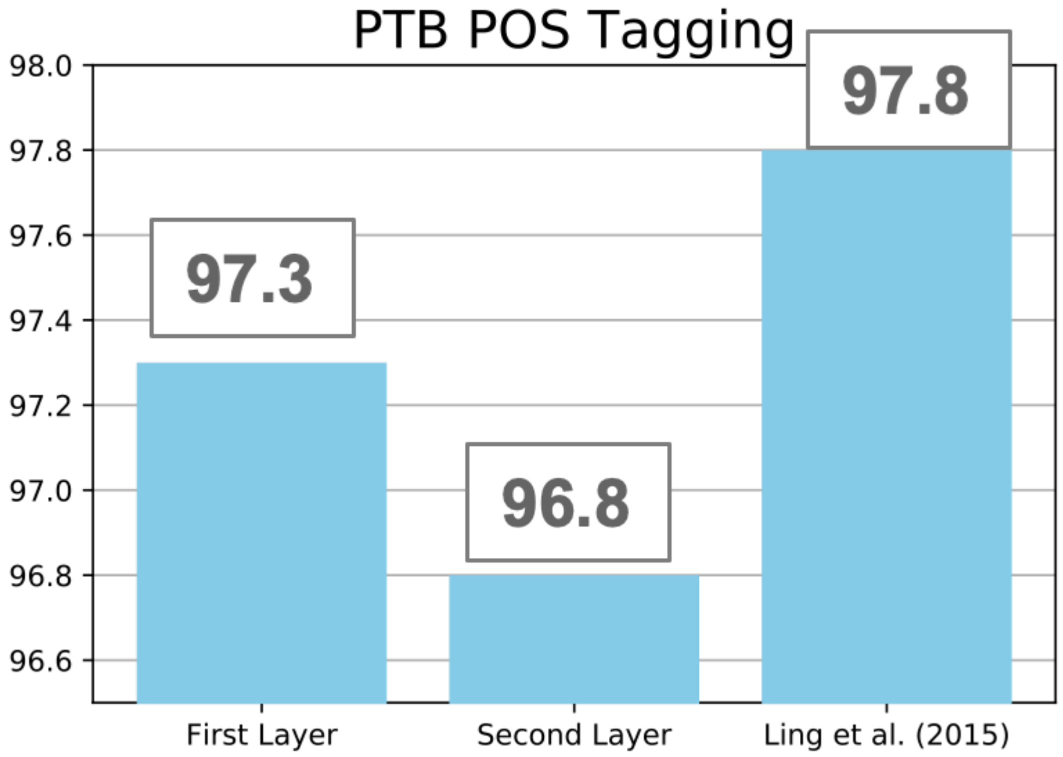
TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

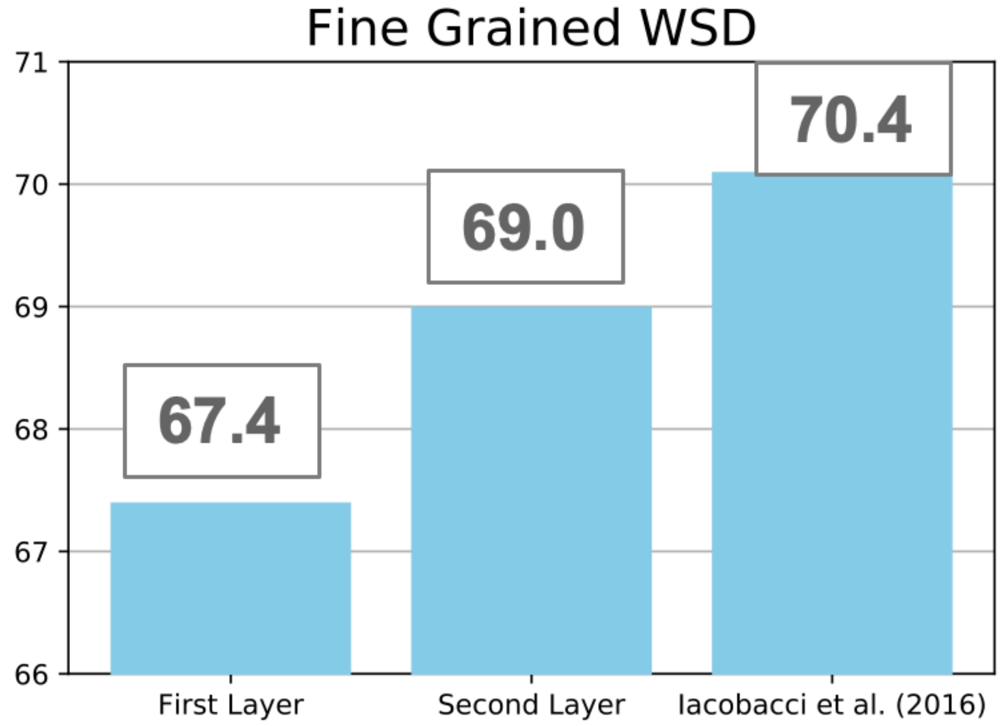
- SQuAD: “The increase of 4.7% with ELMo is also significantly larger then the 1.8% improvement from adding CoVe to a baseline model”
- SST-5: “Replacing CoVe with ELMo results in a 1.0% absolute accuracy”

Experimental results



What information is captured by ELMo representations?



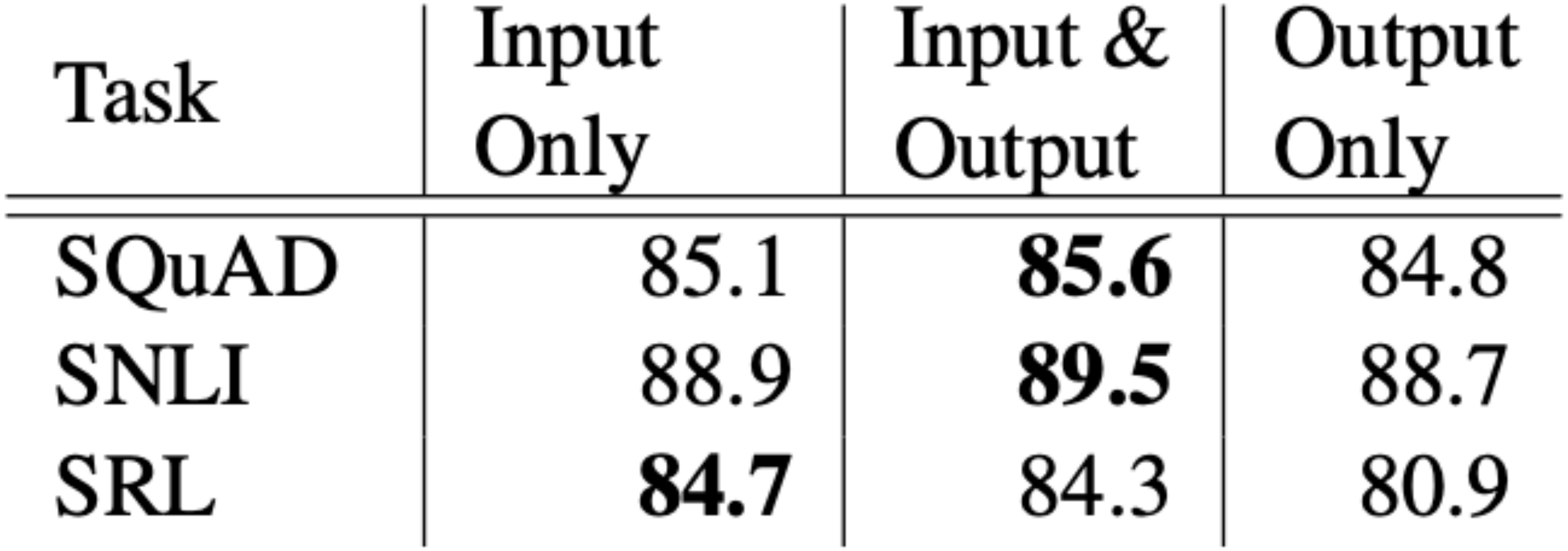


First Layer > Second Layer

Second Layer > First Layer

syntactic information is better represented at lower layers
while semantic information is captured a higher layers

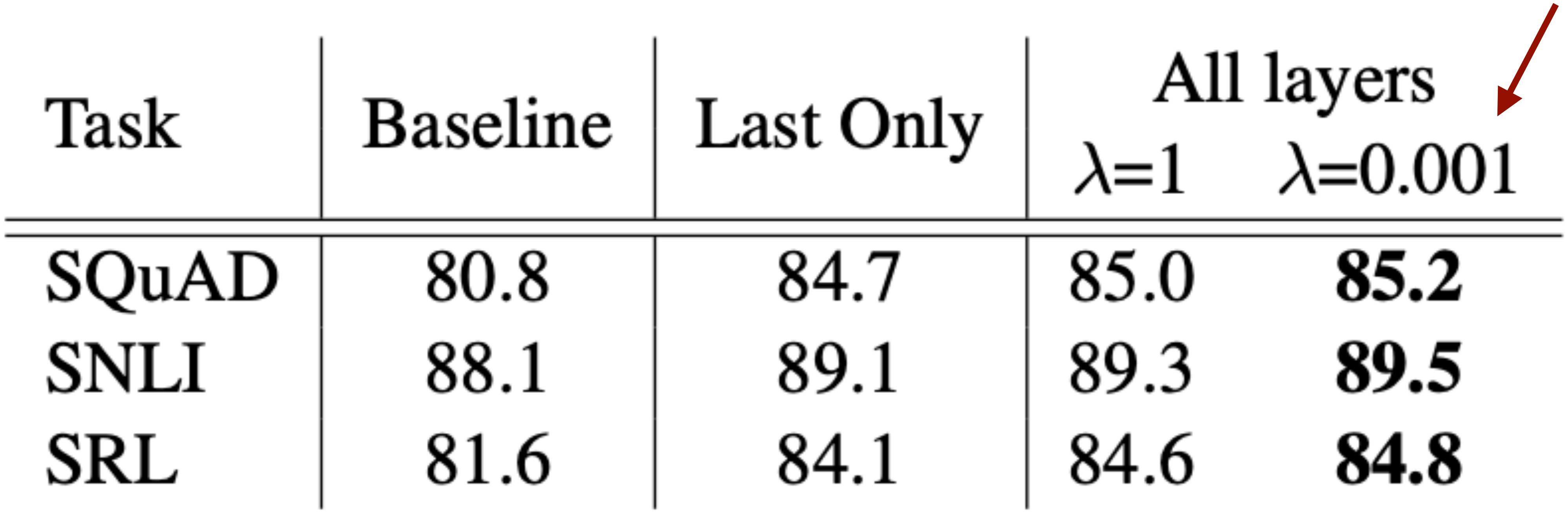
Where to include ELMo?



Task	Input Only	Input & Output	Output Only
SQuAD	85.1	**85.6**	84.8
SNLI	88.9	**89.5**	88.7
SRL	**84.7**	84.3	80.9

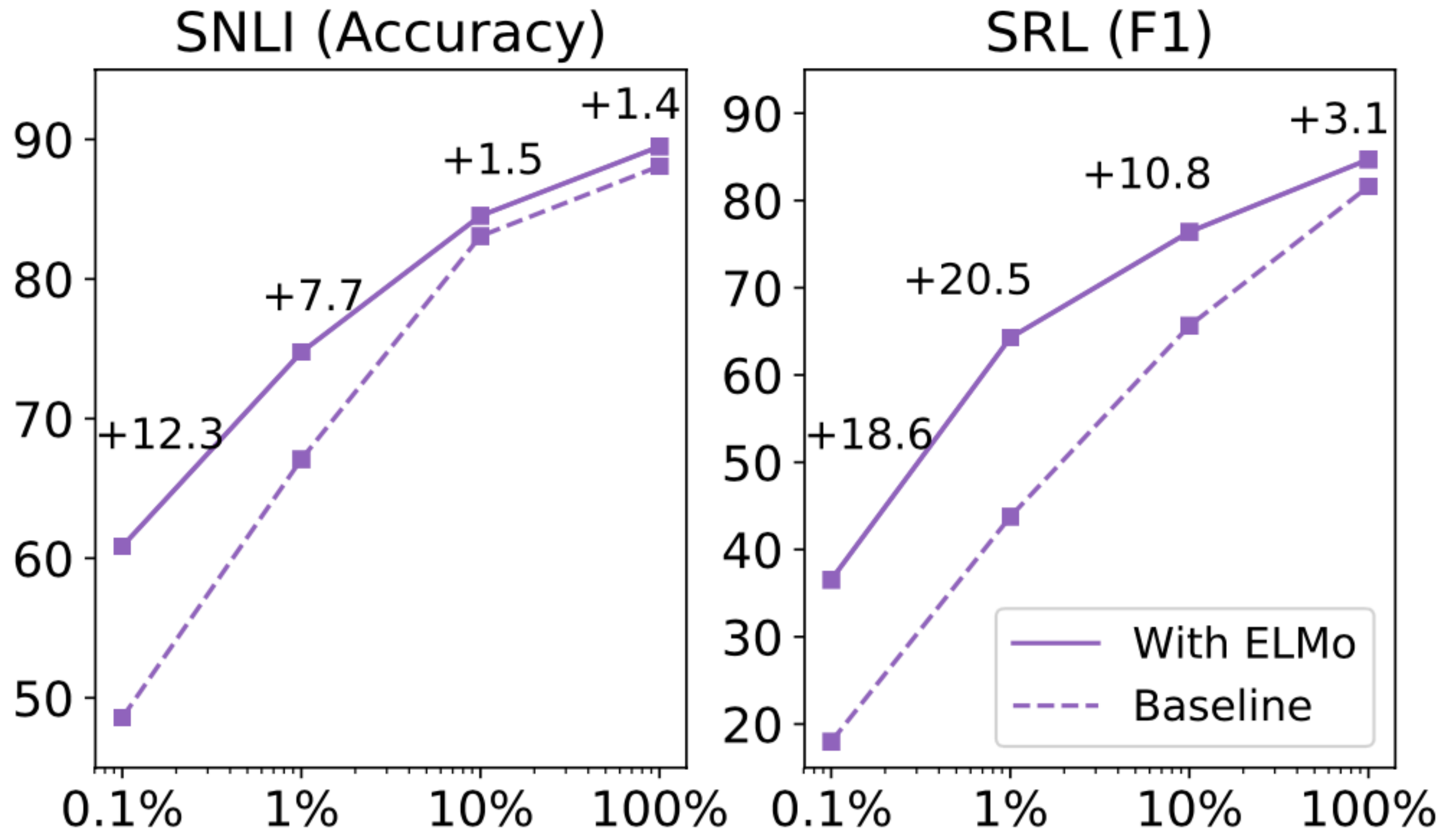
Using all layers > last layer

Reg. Parameter

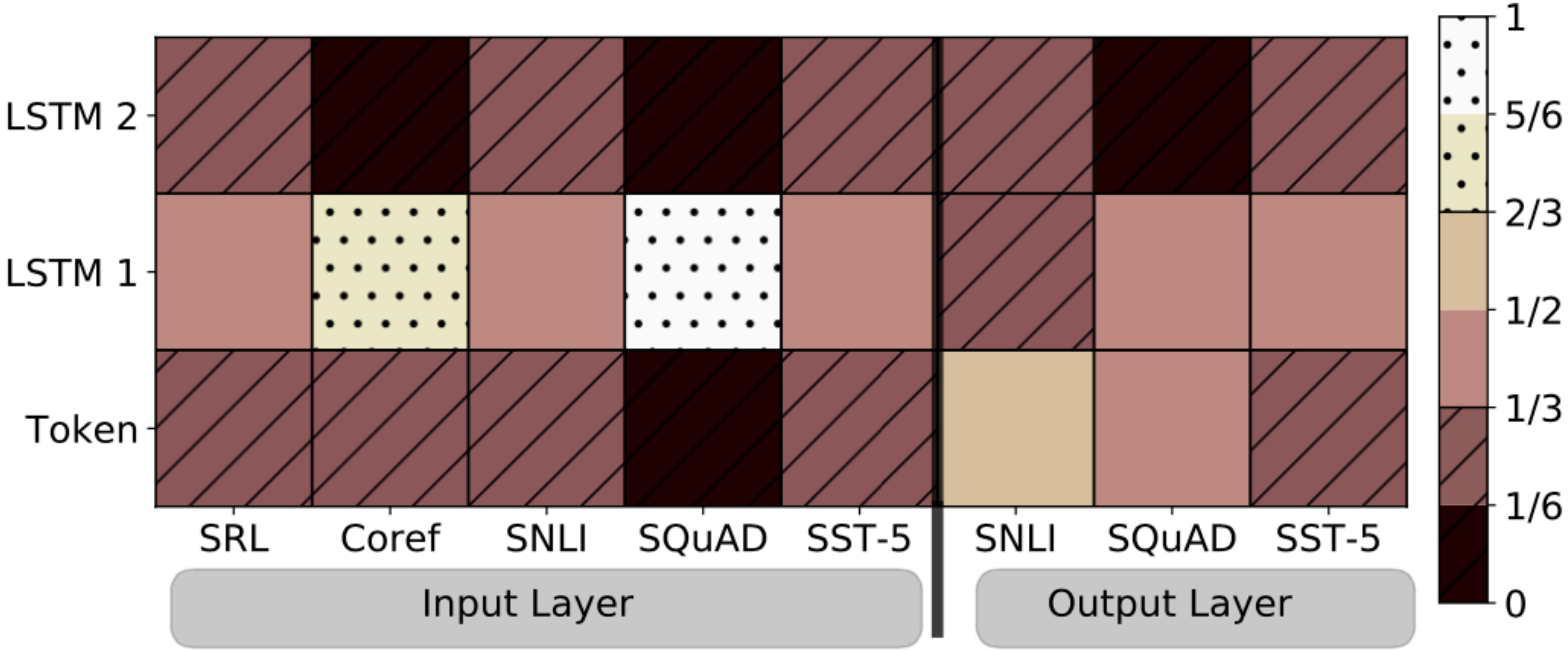


Task	Baseline	Last Only	All layers	
			λ=1	λ=0.001
SQuAD	80.8	84.7	85.0	**85.2**
SNLI	88.1	89.1	89.3	**89.5**
SRL	81.6	84.1	84.6	**84.8**

Sample Efficiency



Visualization of softmax normalized weights



Discussion: CoVe vs ELMo

- Data dependence: monolingual data vs parallel data
- Training efficiency: ELMo slightly better than CoVe
- Objectives: predicting each word in the sentence vs “translating source sentence as a whole”
- Directionality: bidirectional encoder vs forward + backward LMs

Limitations of ELMo/CoVe

- Task-specific architectures: Contextualized word embeddings are used as an augmentation to static word embeddings
- Trained on single sentences
- Training corpus is much smaller than those used for training word2vec/GloVe vectors