Stanford Question Answering Dataset (SQuAD)

(Rajpurkar et al., 2016)

Question: Which team won Super Bowl 50?

Passage

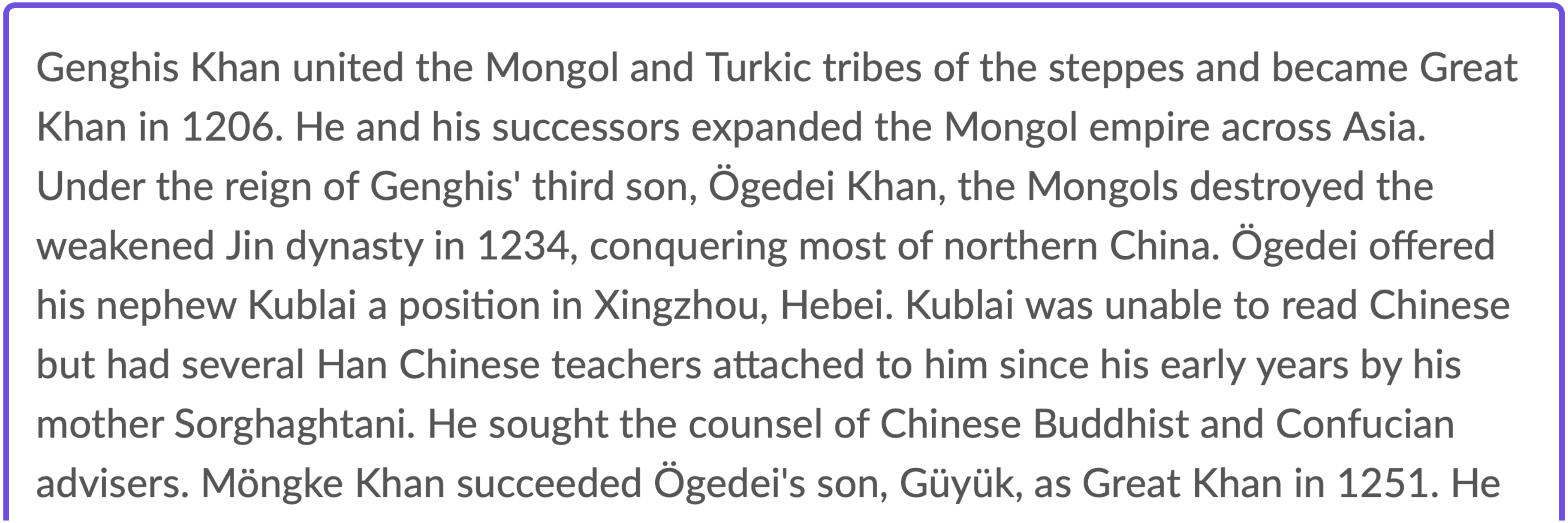
Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl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

100k examples

Answer must be a span in the passage

Extractive question answering/reading comprehension

SQuAD 2.0 No Answer Example



Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

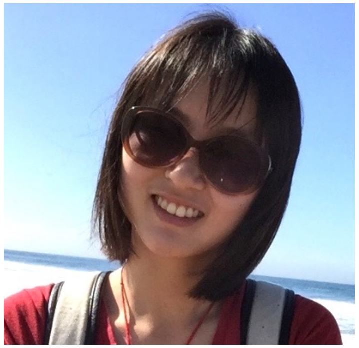
Prediction: 1234 [from Microsoft nlnet]

2. Stanford Attentive Reader

[Chen, Bolton, & Manning 2016]

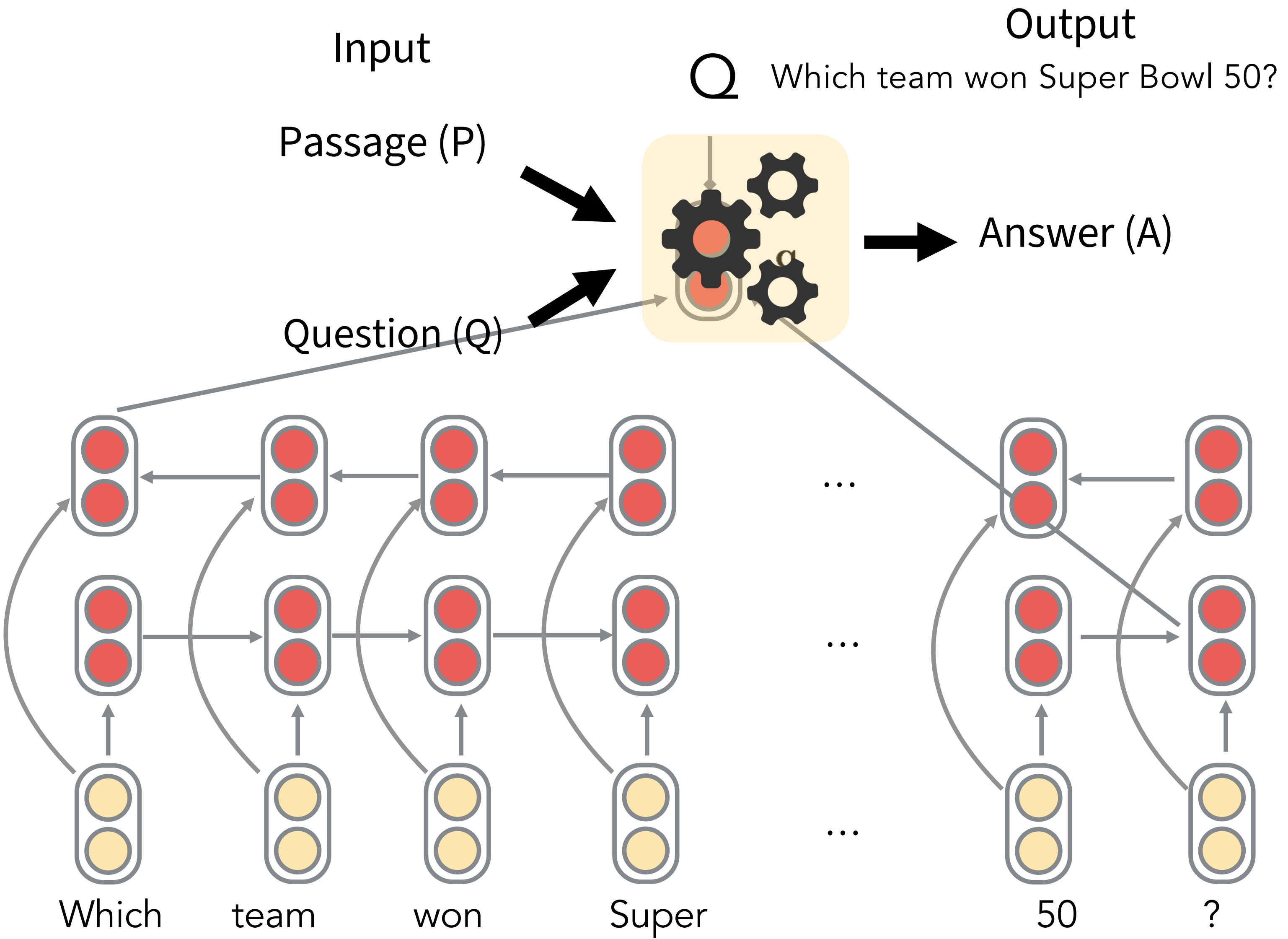
[Chen, Fisch, Weston & Bordes 2017] DrQA

[Chen 2018]



- Demonstrated a minimal, highly successful architecture for reading comprehension and question answering
- Became known as the Stanford Attentive Reader

The Stanford Attentive Reader



Input

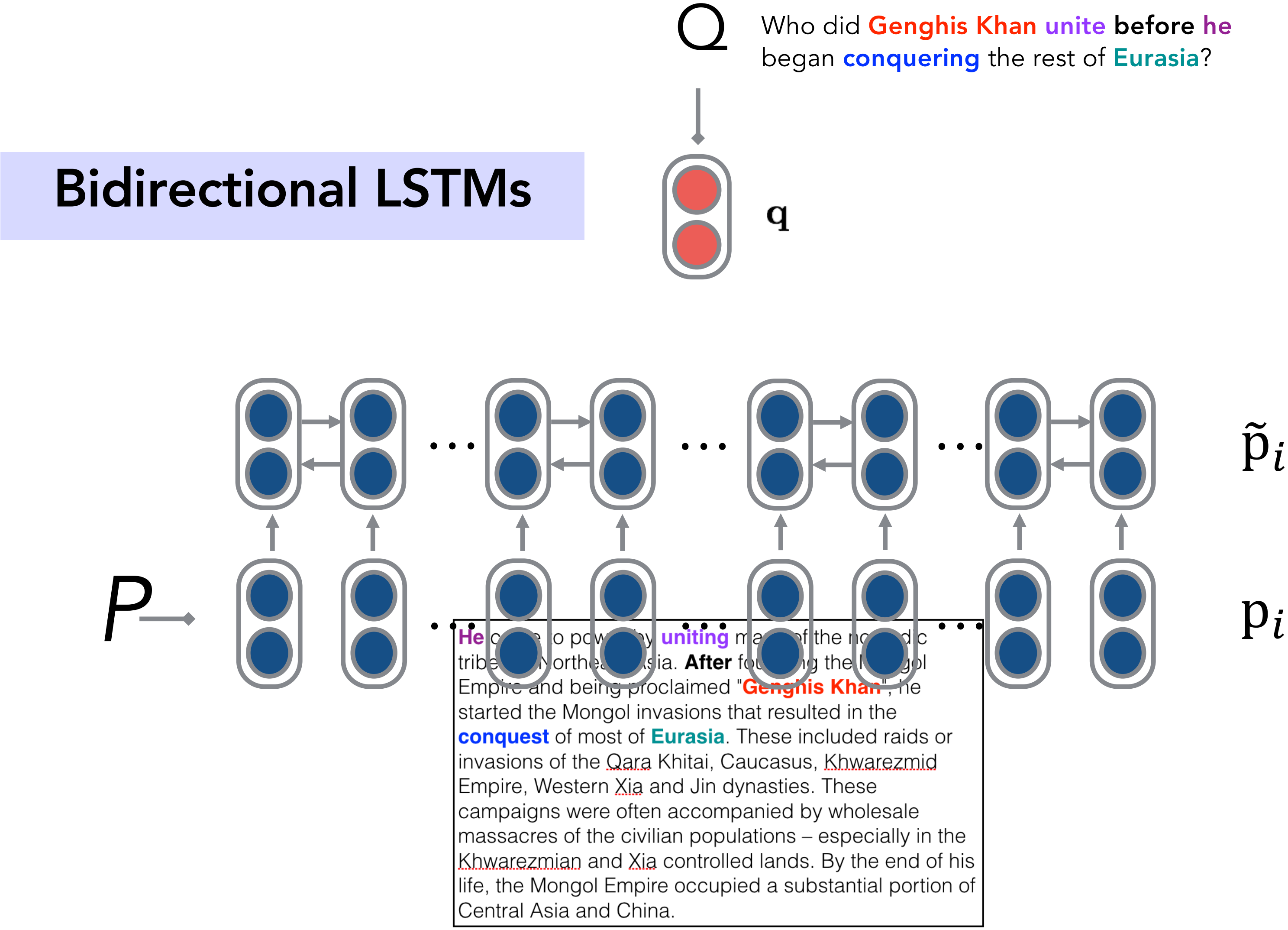
Passage (P)

Question (Q)

Which team won Super Bowl 50?

Answer (A)

Stanford Attentive Reader



Bidirectional LSTMs

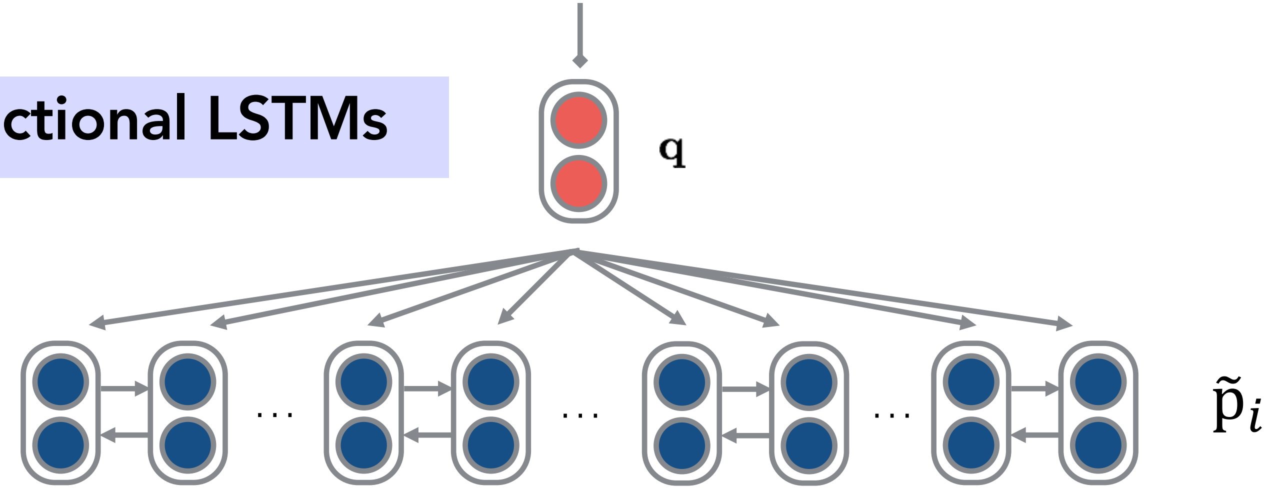
Q Who did **Genghis Khan** **unite** **before** **he** began **conquering** the rest of **Eurasia**?

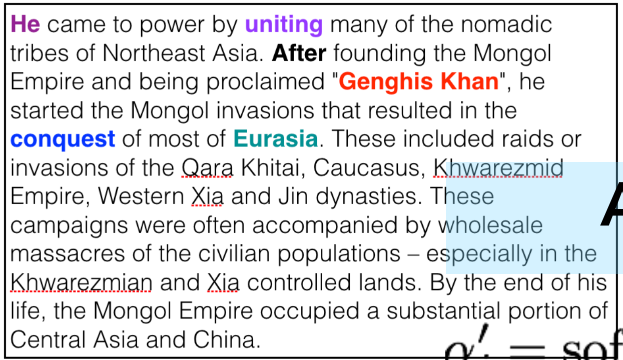
He came to power by **uniting** many of the nomadic tribes of Northeast Asia. **After** founding the Mongol Empire and being proclaimed "**Genghis Khan**", he started the Mongol invasions that resulted in the **conquest** of most of **Eurasia**. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmian and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.

Stanford Attentive Reader

Q Who did **Genghis Khan** **unite** **before** **he** began **conquering** the rest of **Eurasia**?

Bidirectional LSTMs




He came to power by **uniting** many of the nomadic tribes of Northeast Asia. **After** founding the Mongol Empire and being proclaimed "**Genghis Khan**", he started the Mongol invasions that resulted in the **conquest** of most of **Eurasia**. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmian and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.

Attention

$\alpha_i = \boxed{\text{softmax}_i \left(\mathbf{q}^\top \mathbf{W}\right.}$

→ predict **start** token

Attention

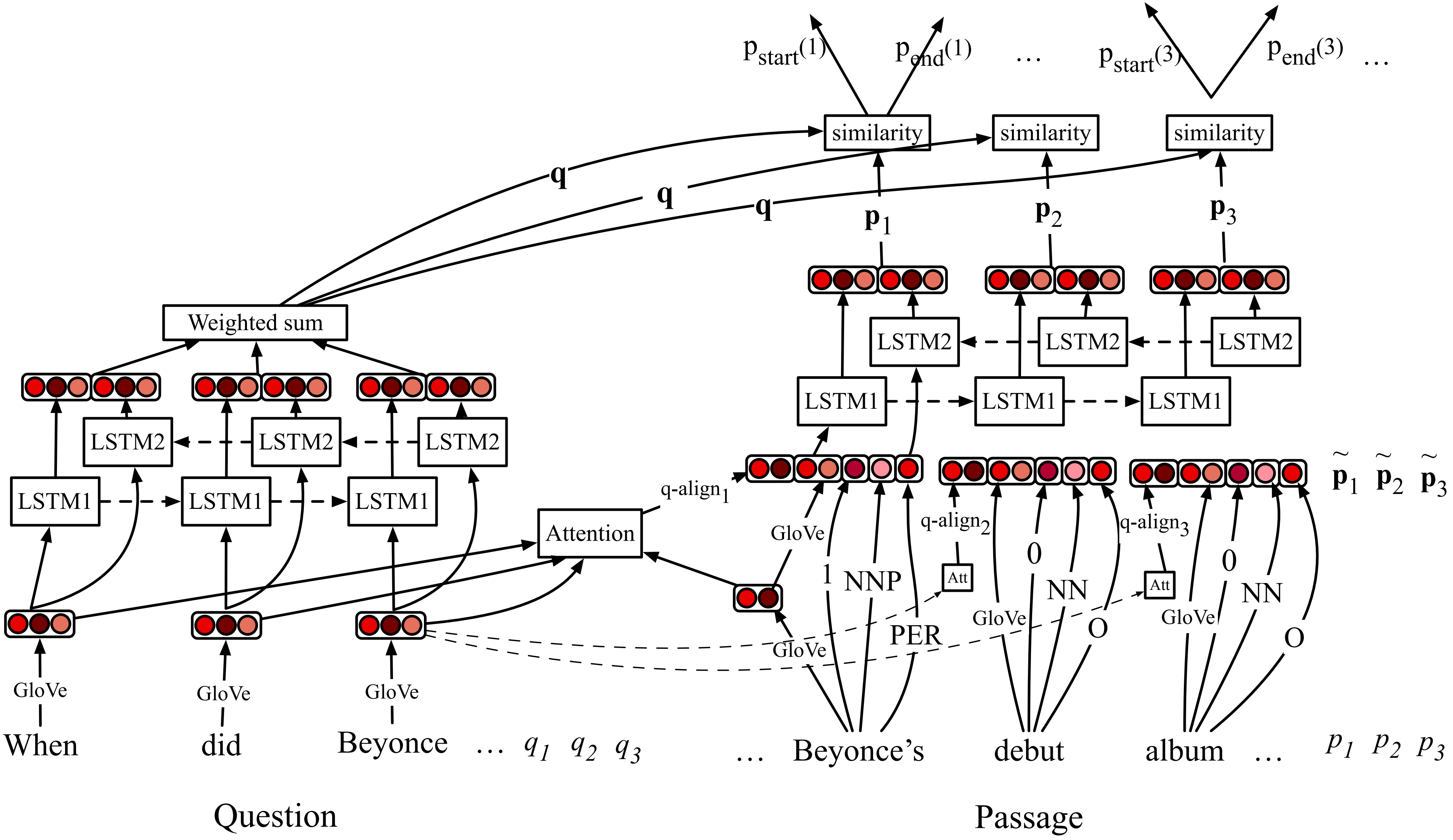
$\alpha'_i = \text{softmax}_i(\mathbf{q}^\top \mathbf{W}'_s \tilde{\mathbf{p}}_i)$

→ predict **end** token

SQuAD 1.1 Results (single model, c. Feb 2017)

	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]	79.7
Human performance	91.2

Stanford Attentive Reader++



Training objective: $\mathcal{L} = -\sum \log P^{(\text{start})}(a_{\text{start}}) - \sum \log P^{(\text{end})}(a_{\text{end}})$

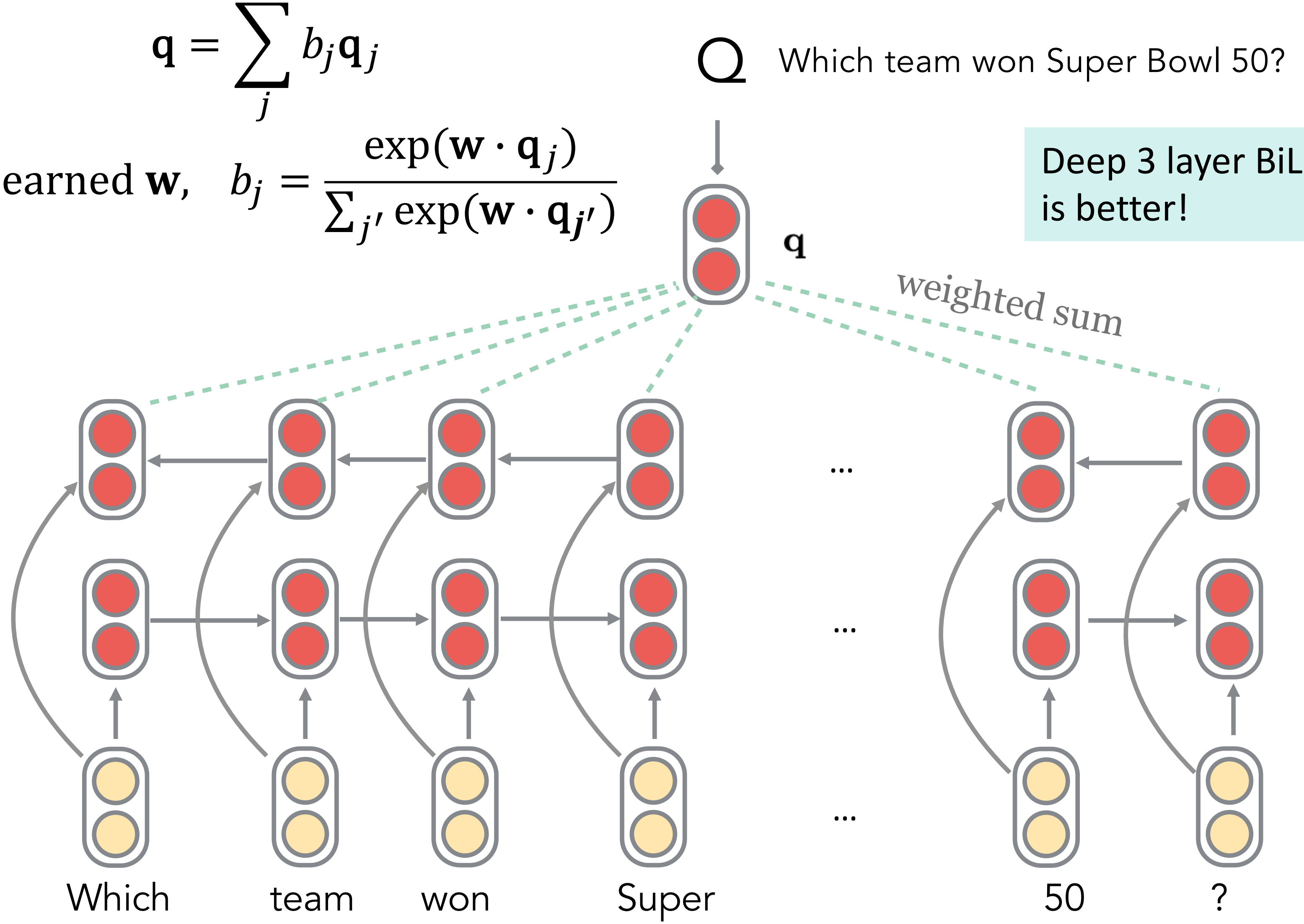
(Chen et al., 2018)

Stanford Attentive Reader++

$$\mathbf{q} = \sum_j b_j \mathbf{q}_j$$

For learned $\mathbf{w}$, $\quad b_j = \dfrac{\exp(\mathbf{w} \cdot \mathbf{q}_j)}{\sum_{j'} \exp(\mathbf{w} \cdot \mathbf{q}_{j'})}$

Deep 3 layer BiLSTM is better!



Stanford Attentive Reader++

- $\mathbf{p}_i$: Vector representation of each token in passage

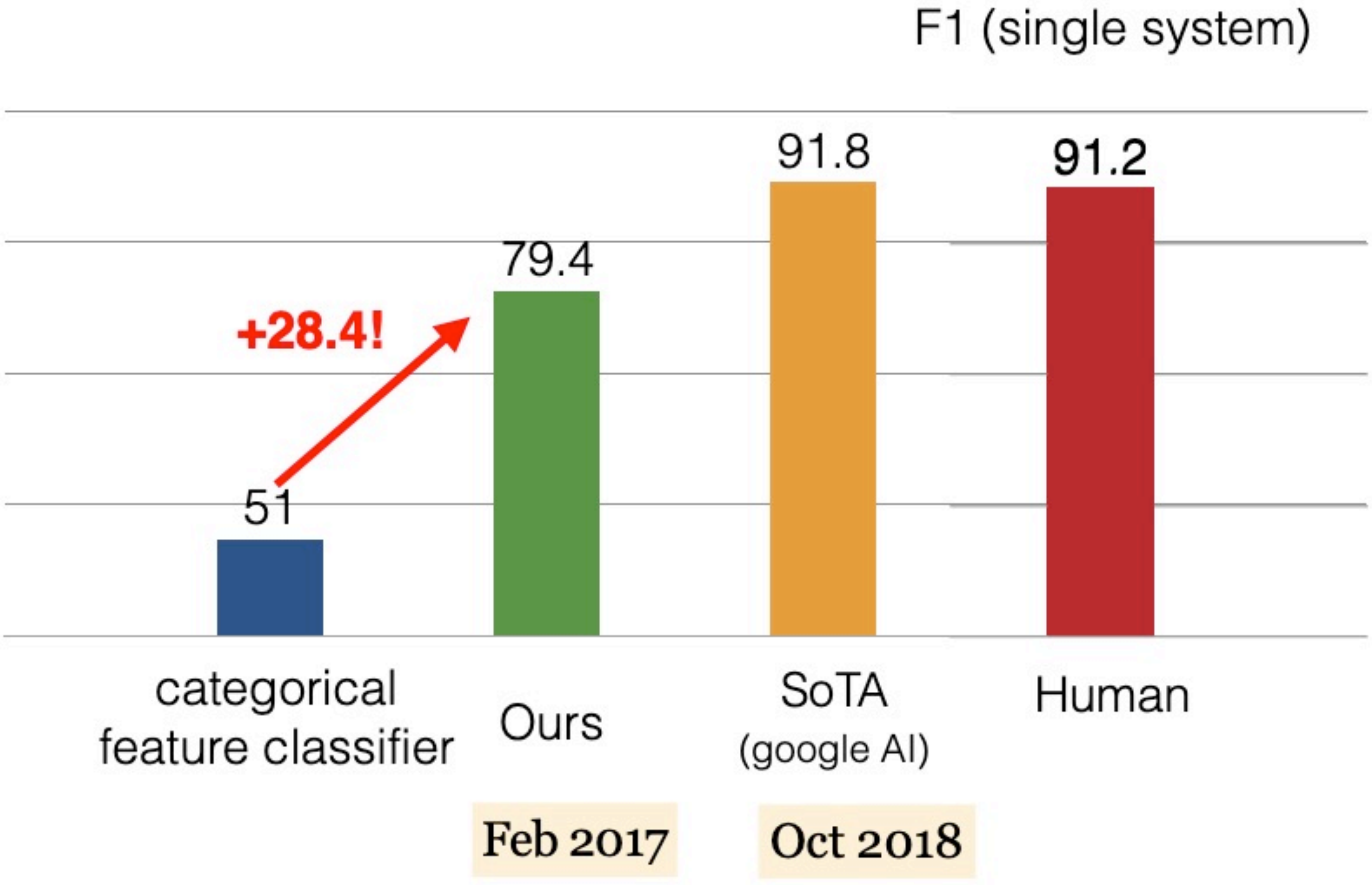
Made from concatenation of

- Word embedding (GloVe 300d)
- Linguistic features: POS & NER tags, one-hot encoded
- Term frequency (unigram probability)
- Exact match: whether the word appears in the question
  - 3 binary features: exact, uncased, lemma
- Aligned question embedding ("car" vs "vehicle")

$$f_{align}(p_i) = \sum_j a_{i,j} \mathbf{E}(q_j) \qquad q_{i,j} = \frac{\exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_j)))}{\sum_{j'} \exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q'_j)))}$$

Where α is a simple one layer FFNN

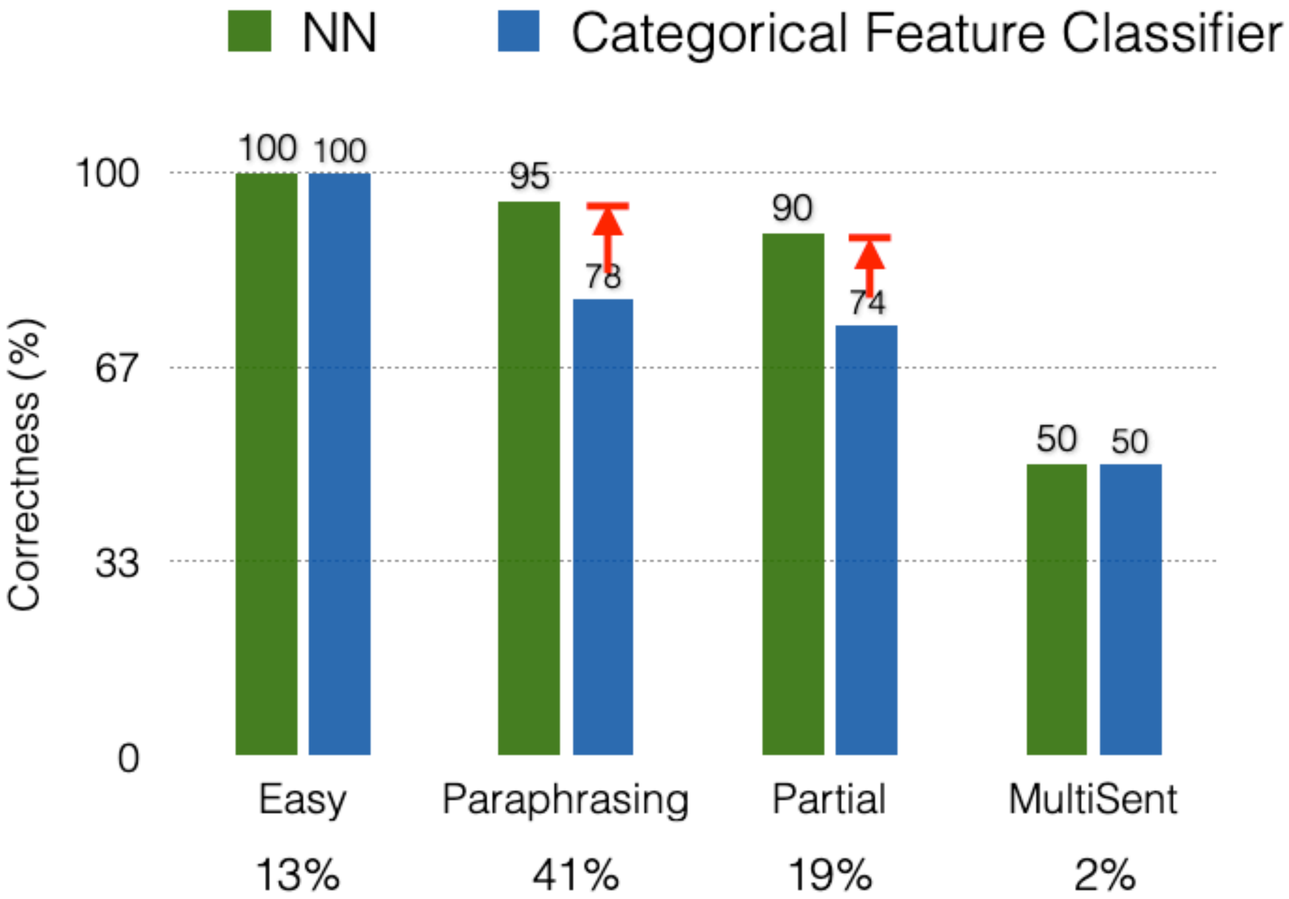
A big win for neural models



x

(Chen, Bolton, Manning, 2016)

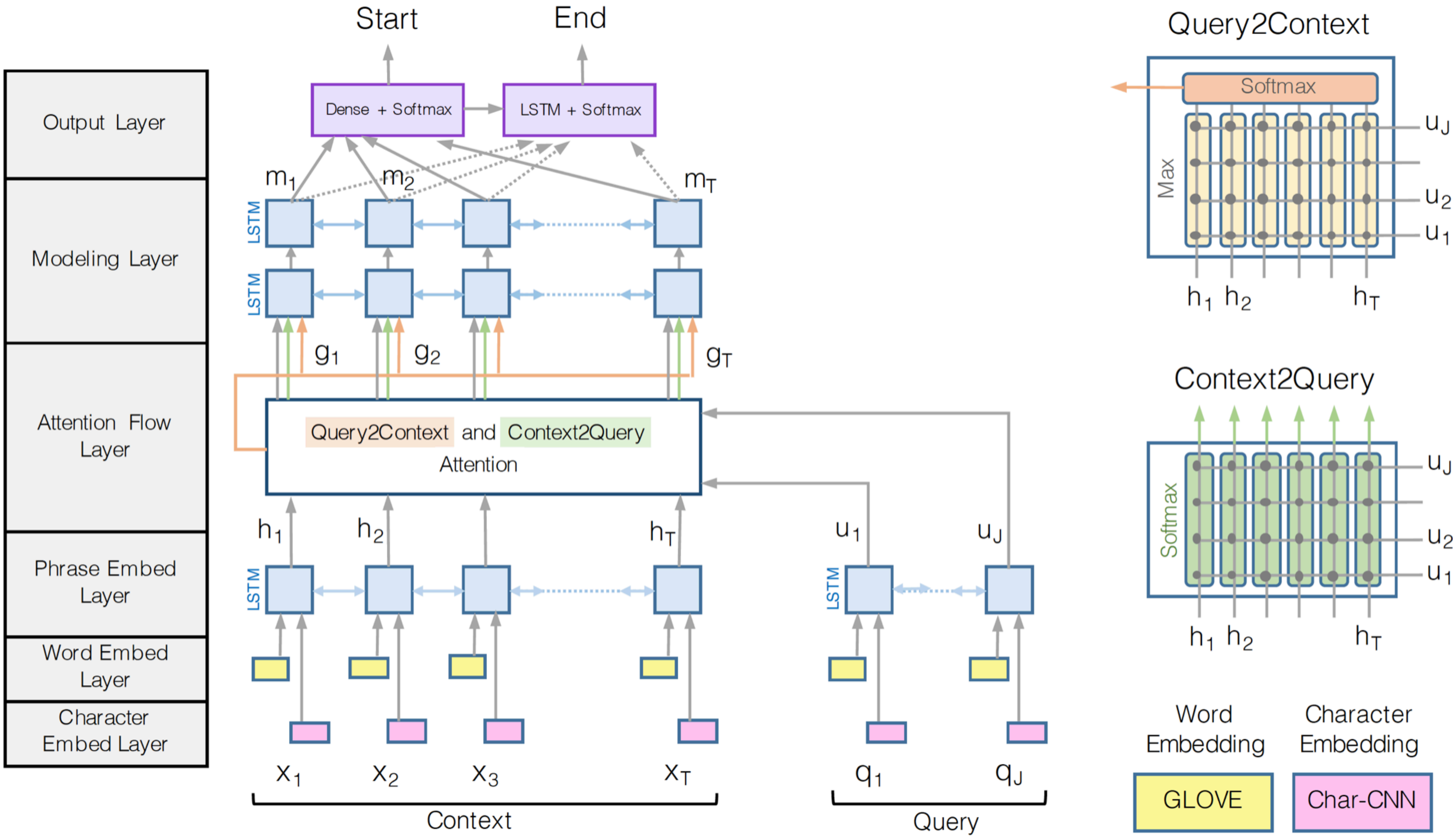
What do these neural models do?



X

3. BiDAF: Bi-Directional Attention Flow for Machine Comprehension

(Seo, Kembhavi, Farhadi, Hajishirzi, ICLR 2017)



BiDAF – Roughly the CS224N DFP baseline

- There are variants of and improvements to the BiDAF architecture over the years, but the central idea is **the Attention Flow layer**
- **Idea:** attention should flow both ways – from the context to the question and from the question to the context
- Make similarity matrix (with **w** of dimension 6*d*):

$$\boldsymbol{S}_{ij} = \boldsymbol{w}_{\text{sim}}^{T}[\boldsymbol{c}_i; \boldsymbol{q}_j; \boldsymbol{c}_i \circ \boldsymbol{q}_j] \in \mathbb{R}$$

- Context-to-Question (C2Q) attention:
  (which query words are most relevant to each context word)

$$\alpha^i = \text{softmax}(\boldsymbol{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \ldots, N\}$$

$$\boldsymbol{a}_i = \sum_{j=1}^{M} \alpha_j^i \boldsymbol{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \ldots, N\}$$

BiDAF

- **Attention Flow Idea:** attention should flow both ways – from the context to the question and from the question to the context
- Question-to-Context (Q2C) attention:
  (the weighted sum of the most important words in the context with respect to the query – slight asymmetry through max)

$$\boldsymbol{m}_i = \max_j \boldsymbol{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \dots, N\}$$

$$\beta = \mathrm{softmax}(\boldsymbol{m}) \in \mathbb{R}^N$$

$$\boldsymbol{c}' = \sum_{i=1}^{N} \beta_i \boldsymbol{c}_i \in \mathbb{R}^{2h}$$

- For each passage position, output of BiDAF layer is:

$$\boldsymbol{b}_i = [\boldsymbol{c}_i; \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{c}'] \in \mathbb{R}^{8h} \quad \forall i \in \{1, \dots, N\}$$

BiDAF

- There is then a “modelling” layer:
  - Another deep (2-layer) BiLSTM over the passage
- And answer span selection is more complex:
  - Start: Pass output of BiDAF and modelling layer concatenated to a dense FF layer and then a softmax
  - End: Put output of modelling layer M through another BiLSTM to give M_2 and then concatenate with BiDAF layer and again put through dense FF layer and a softmax
    - Editorial: Seems very complex, but it does seem like you should do a bit more than Stanford Attentive Reader, e.g., conditioning end also on start