Natural Language Processing with Deep Learning

CS224N/Ling284

Christopher Manning

Lecture 10: (Textual) Question Answering

Technical note: This is a “featured snippet” answer extracted from a web page, not a question answered using the (structured) Google Knowledge Graph (formerly known as Freebase).

2. Motivation: Question answering

- With massive collections of full-text documents, i.e., the web ☺, simply returning relevant documents is of limited use
- Rather, we often want **answers** to our **questions**
- Especially on mobile
- Or using a digital assistant device, like Alexa, Google Assistant, …

- We can factor this into two parts:
  1. Finding documents that (might) contain an answer
     - Which can be handled by traditional information retrieval/web search
     - (I teach cs276 next quarter which deals with this problem)
  2. Finding an answer in a paragraph or a document
     - This problem is often termed **Reading Comprehension**
     - It is what we will focus on today

A Brief History of Reading Comprehension

- Much early NLP work attempted reading comprehension
  - Schank, Abelson, Lehnert et al. c. 1977 – "Yale A.I. Project"
- Revived by Lynette Hirschman in 1999:
  - Could NLP systems answer human reading comprehension questions for 3rd to 6th graders? Simple methods attempted.
- Revived again by Chris Burges in 2013 with MCTest
  - Again answering questions over simple story texts
- Floodgates opened in 2015/16 with the production of large datasets which permit supervised neural systems to be built
  - Hermann et al. (NIPS 2015) DeepMind CNN/DM dataset
  - Rajpurkar et al. (EMNLP 2016) SQuAD
  - MS MARCO, TriviaQA, RACE, NewsQA, NarrativeQA, …

Machine Comprehension (Burges 2013)

- "A machine **comprehends** a passage of **text** if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question."

Towards the Machine Comprehension of Text: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

MCTest Reading Comprehension

Passage (*P*) + Question (*Q*) ⟶ Answer (*A*)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house…….

Q Why did Alyssa go to Miami?

A To visit some friends

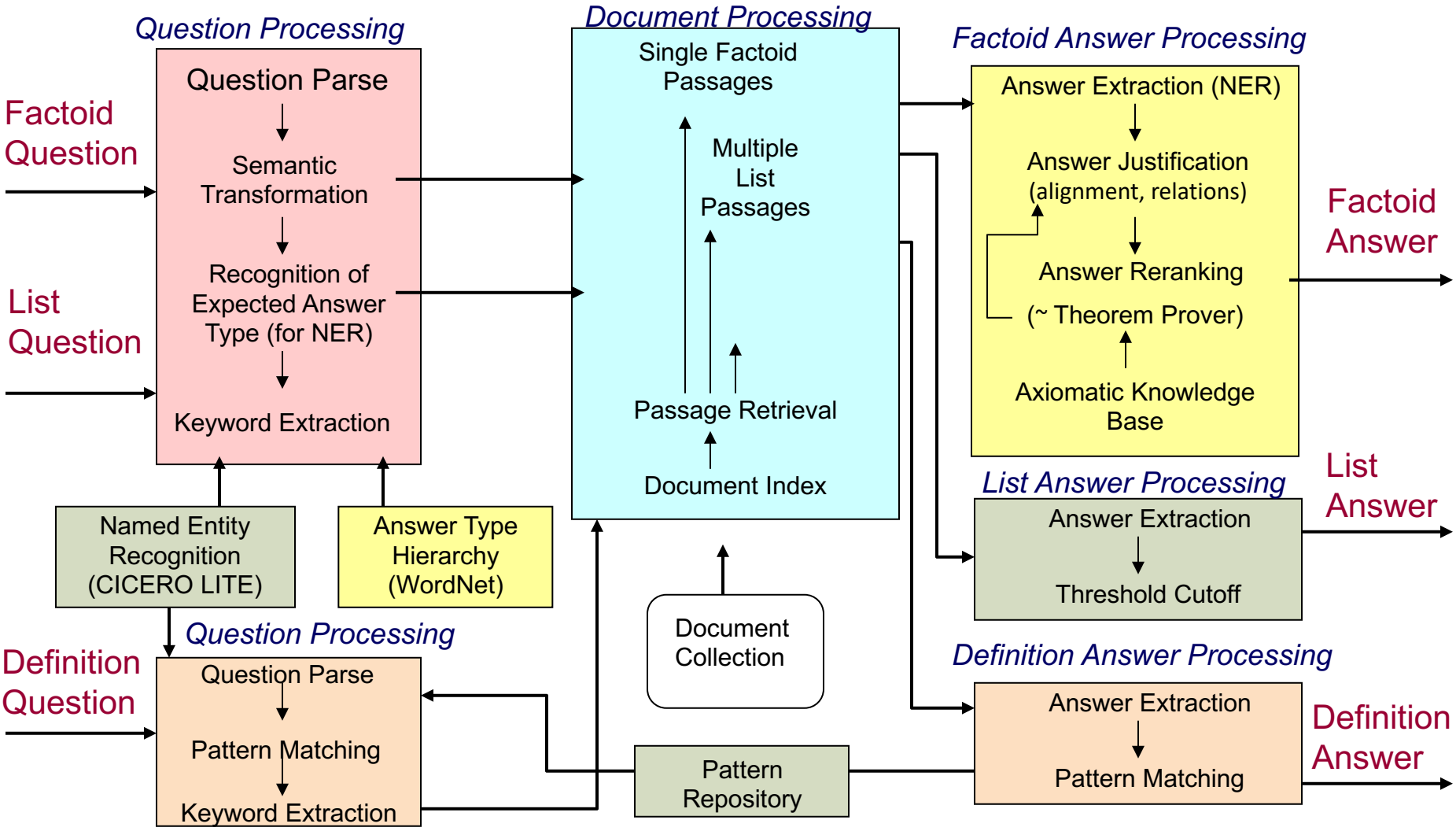
A Brief History of Open-domain Question Answering

- Simmons et al. (1964) did first exploration of answering questions from an expository text based on matching dependency parses of a question and answer
- Murax (Kupiec 1993) aimed to answer questions over an online encyclopedia using IR and shallow linguistic processing
- The NIST TREC QA track begun in 1999 first rigorously investigated answering fact questions over a large collection of documents
- IBM's Jeopardy! System (DeepQA, 2011) brought attention to a version of the problem; it used an ensemble of many methods
- DrQA (Chen et al. 2016) uses IR followed by neural reading comprehension to bring deep learning to Open-domain QA

Turn-of-the Millennium Full NLP QA:

[architecture of LCC (Harabagiu/Moldovan) QA system, circa 2003]

Complex systems but they did work fairly well on "factoid" questions



3. 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

A.k.a. extractive question answering

Stanford Question Answering Dataset (SQuAD)

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

Rather than taxation, what are private schools largely funded by?

Gold answers: ① tuition ② charging their students tuition ③ tuition

SQuAD evaluation, v1.1

- Authors collected 3 gold answers
- Systems are scored on two metrics:
  - Exact match: 1/0 accuracy on whether you match one of the 3 answers
  - F1: Take system and each gold answer as bag of words, evaluate
    Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$, harmonic mean F1 = $\frac{2PR}{P+R}$
    Score is (macro-)average of per-question F1 scores
- F1 measure is seen as more reliable and taken as primary
  - It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (**a**, **an**, **the** only)

SQuAD v1.1 leaderboard, end of 2016 (Dec 6)

		EM	F1
11	Fine-Grained Gating Carnegie Mellon University (Yang et al. '16)	62.5	73.3
12	Dynamic Chunk Reader IBM (Yu & Zhang et al. '16)	62.5	71.0
13	Match-LSTM with Ans-Ptr (Boundary) Singapore Management University (Wang & Jiang '16)	60.5	70.7
14	Match-LSTM with Ans-Ptr (Sequence) Singapore Management University (Wang & Jiang '16)	54.5	67.7
15	Logistic Regression Baseline Stanford University (Rajpurkar et al. '16)	40.4	51.0

Will your model outperform humans on the QA task?

	Human Performance Stanford University (Rajpurkar et al. '16)	82.3	91.2

SQuAD v1.1 leaderboard, end of 2016 (Dec 6)

Rank	Model	Test EM	Test F1
1	BiDAF (ensemble) Allen Institute for AI & University of Washington (Seo et al. '16)	73.3	81.1
2	Dynamic Coattention Networks (ensemble) Salesforce Research (Xiong & Zhong et al. '16)	71.6	80.4
2	r-net (ensemble) Microsoft Research Asia	72.1	79.7
	Best CS224N Default Final Project result in Winter 2017 class FNU Budianto (BiDAF variant, ensembled)	EM 68.5	F1 77.5
5	BiDAF (single model) Allen Institute for AI & University of Washington (Seo et al. '16)	68.0	77.3
5	Multi-Perspective Matching (ensemble) IBM Research	68.2	77.2

SQuAD v1.1 leaderboard, 2019-02-07 – it's solved!

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737
5 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133

SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one
- You don't have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
  - Have a threshold score for whether a span answers a question
- Or you could have a second component that confirms answering
  - Like Natural Language Inference (NLI) or "Answer validation"

SQuAD 2.0 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]

SQuAD 2.0 leaderboard, 2019-02-07

		EM	F1
36 Sep 13, 2018	BiDAF++ (single model) *UW and FAIR*	65.651	68.866
37 Jun 27, 2018	BSAE AddText (single model) *reciTAL.ai*	63.338	67.422
38 Aug 14, 2018	eeAttNet (single model) *BBD NLP Team* https://www.bbdservice.com	63.327	66.633
38 May 30, 2018	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	63.372	66.251
39 Nov 27, 2018	Tree-LSTM + BiDAF + ELMo (single model) *Carnegie Mellon University*	57.707	62.341
39 May 30, 2018	BiDAF + Self Attention (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	59.332	62.305
40 May 30, 2018	BiDAF-No-Answer (single model) *University of Washington [modified by Stanford]*	59.174	62.093

SQuAD 2.0 leaderboard, 2019-02-07

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035

Good systems are great, but still basic NLU errors

The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

What dynasty came before the Yuan?

Gold Answers: ① Song dynasty ② Mongol Empire
③ the Song dynasty

Prediction: Ming dynasty [BERT (single model) (Google AI)]

SQuAD limitations

- SQuAD has a number of other key limitations too:
  - Only span-based answers (no yes/no, counting, implicit why)
  - Questions were constructed looking at the passages
    - Not genuine information needs
    - Generally greater lexical and syntactic matching between questions and answer span than you get IRL
  - Barely any multi-fact/sentence inference beyond coreference

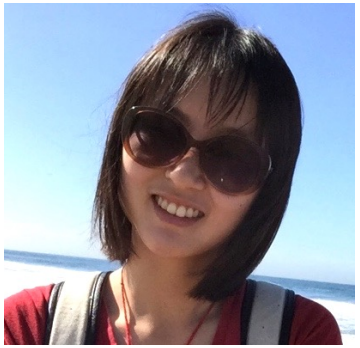
- Nevertheless, it is a well-targeted, well-structured, clean dataset
  - It has been the most used and competed on QA dataset
  - It has also been a useful starting point for building systems in industry (though in-domain data always really helps!)
  - And we're using it (SQuAD 2.0)

4. Stanford Attentive Reader

[Chen, Bolton, & Manning 2016]

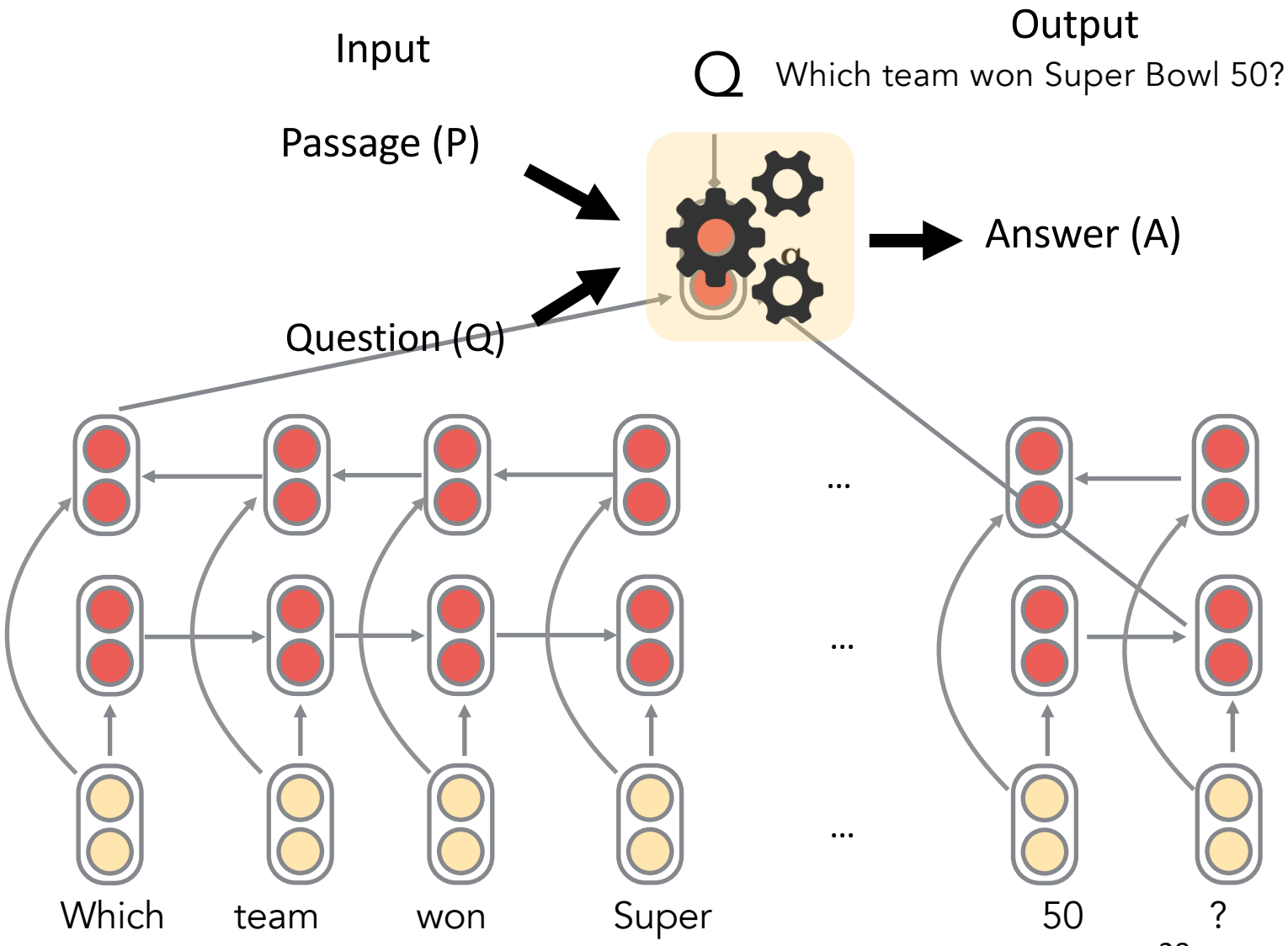
[Chen, Fisch, Weston & Bordes 2017] DrQA

[Chen 2018]



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

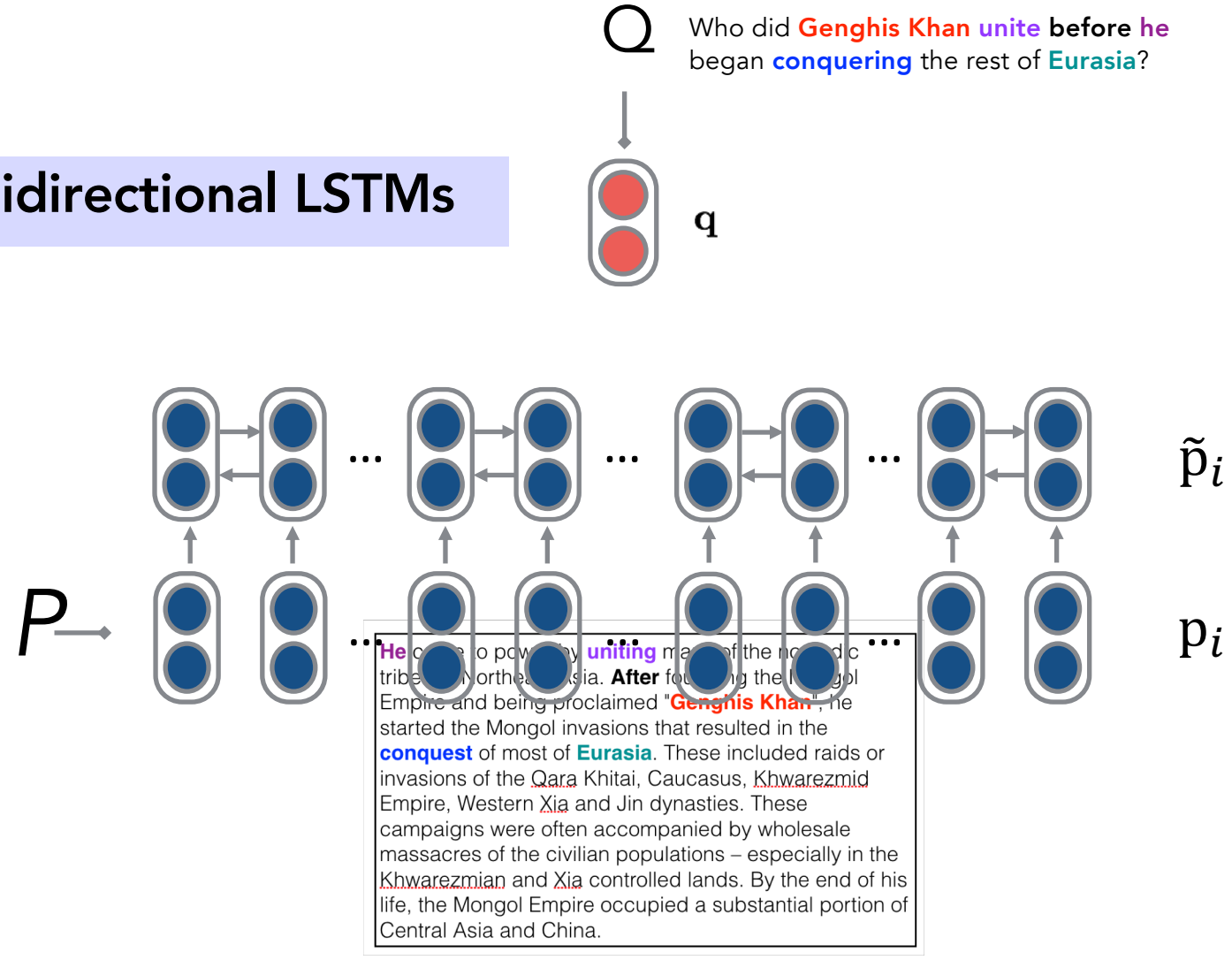
The Stanford Attentive Reader



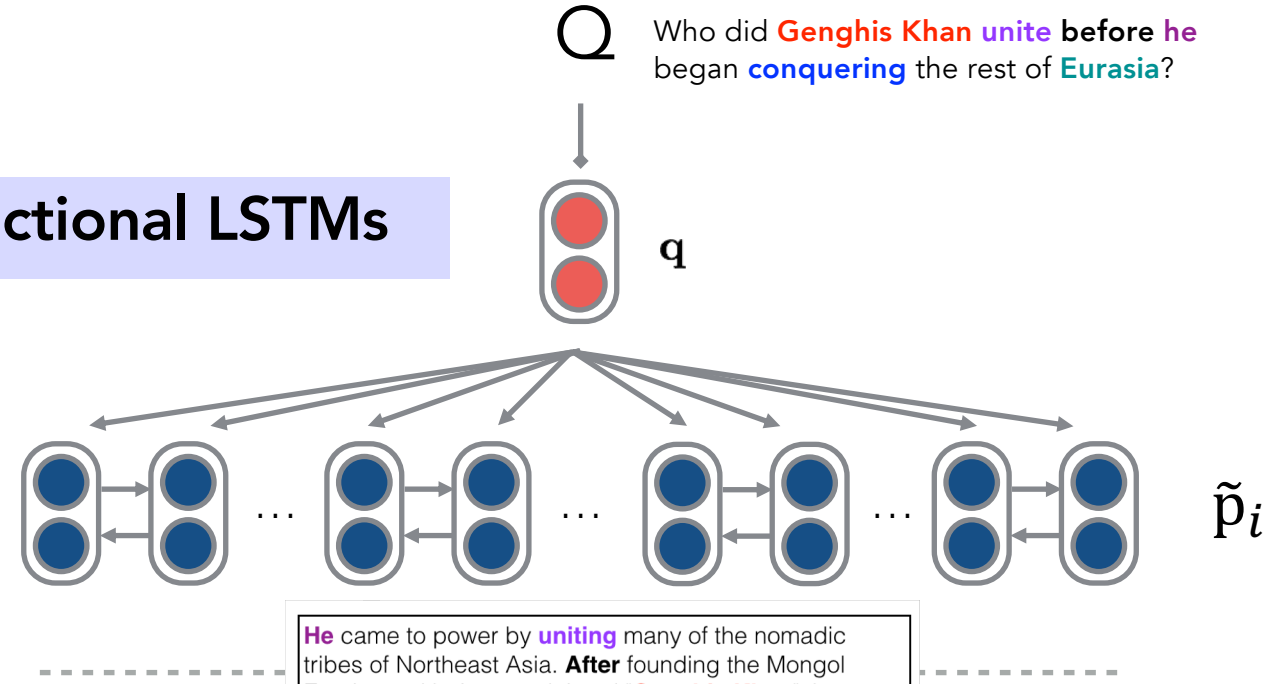
Stanford Attentive Reader

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

Bidirectional LSTMs



Stanford Attentive Reader



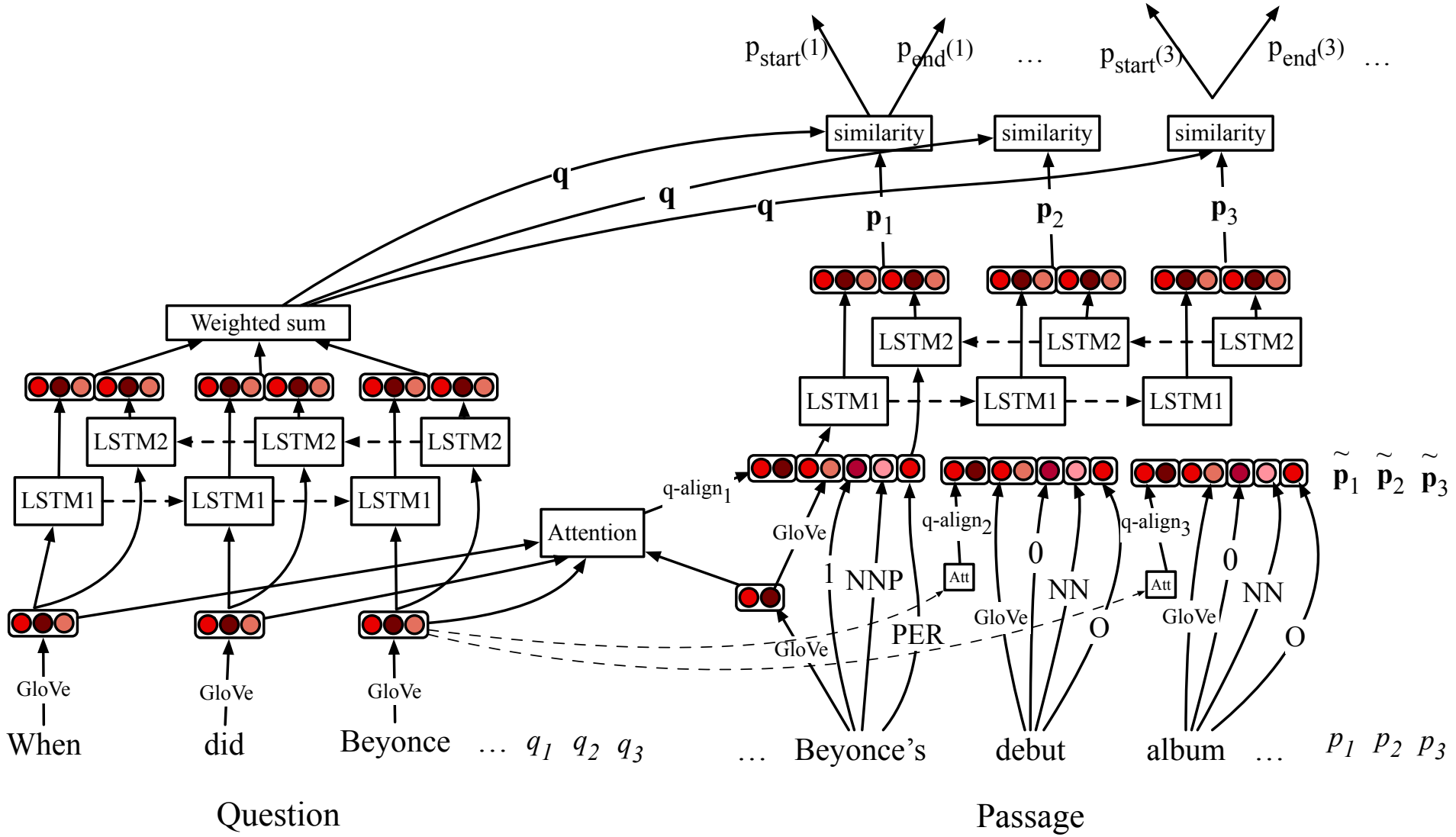
→ predict **start** token

→ 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:

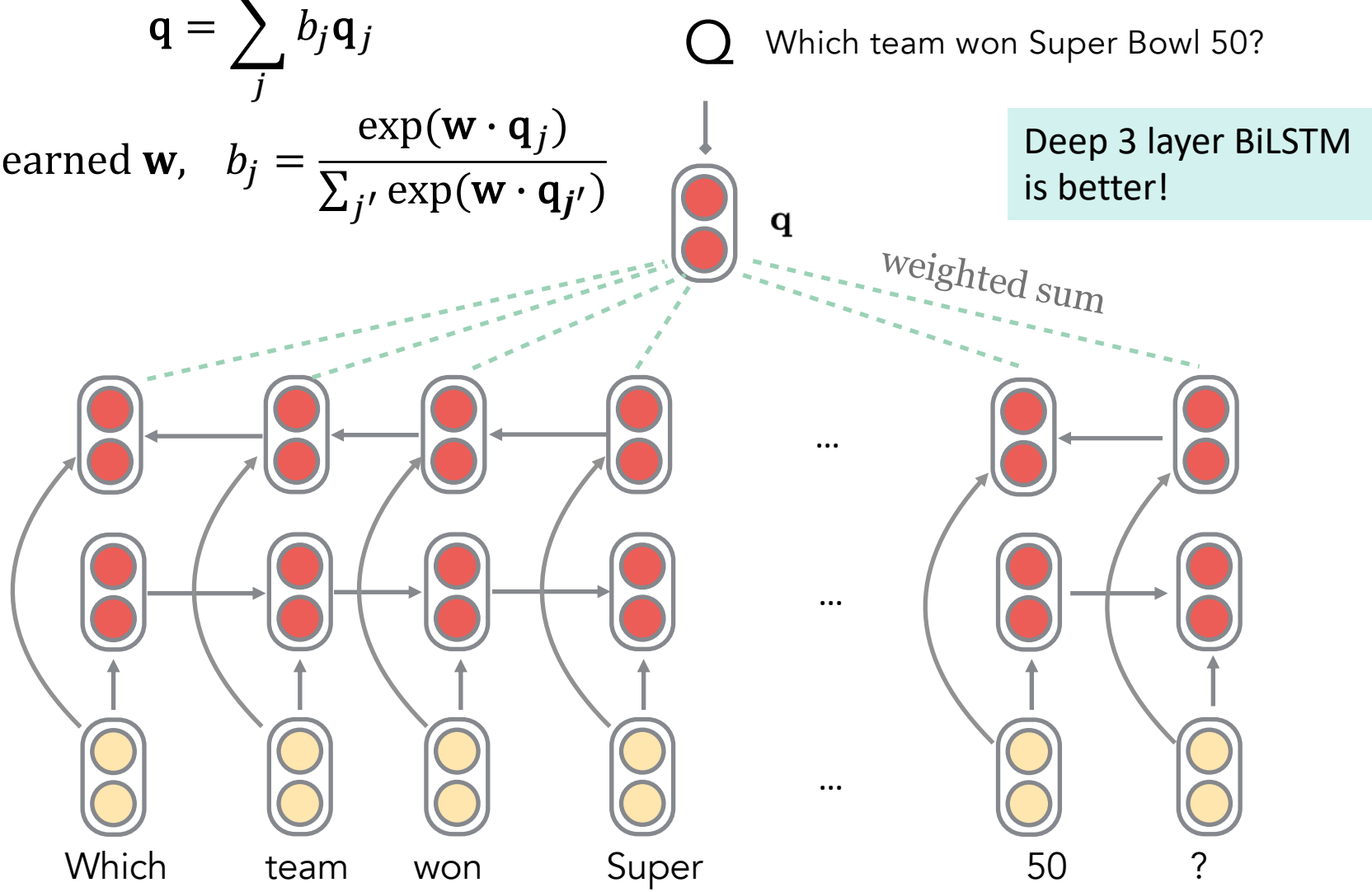
$$\mathcal{L} = -\sum \log P^{(\text{start})}(a_{\text{start}}) - \sum \log P^{(\text{end})}(a_{\text{end}})$$

(Chen et al., 2016; Chen et al., 2017)

Stanford Attentive Reader++

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

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



Stanford Attentive Reader++

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

Made from concatenation of

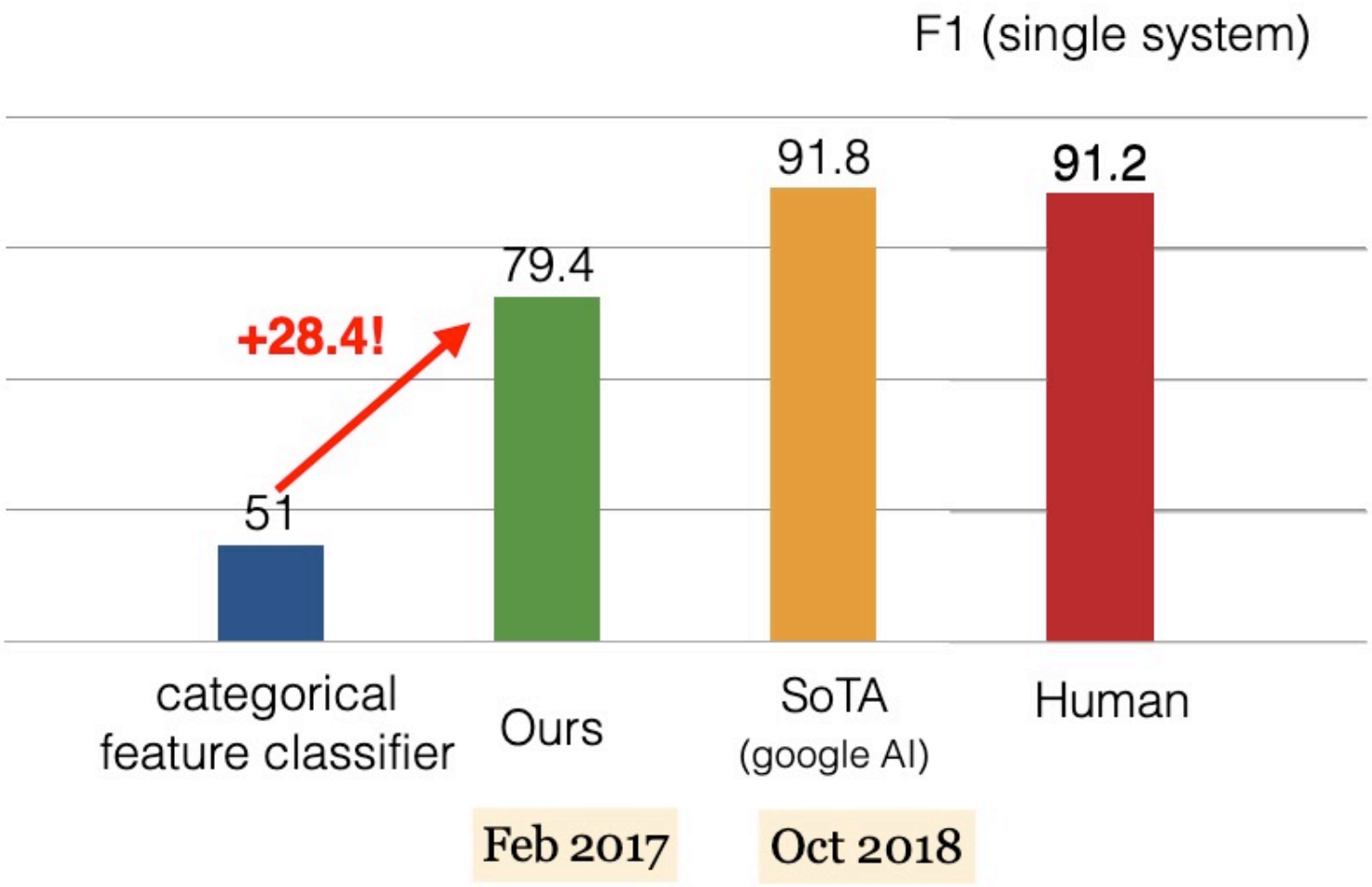
- Word embedding (GloVe 300d)
- Linguistic features: POS & NER tags, one-hot encoded
- Term frequency (unigram probability)
- Exact match: whether the word appears in the question
  - 3 binary features: exact, uncased, lemma
- Aligned question embedding ("car" vs "vehicle")

$$f_{align}(p_i) = \sum_j a_{i,j}\mathbf{E}(q_j)$$

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

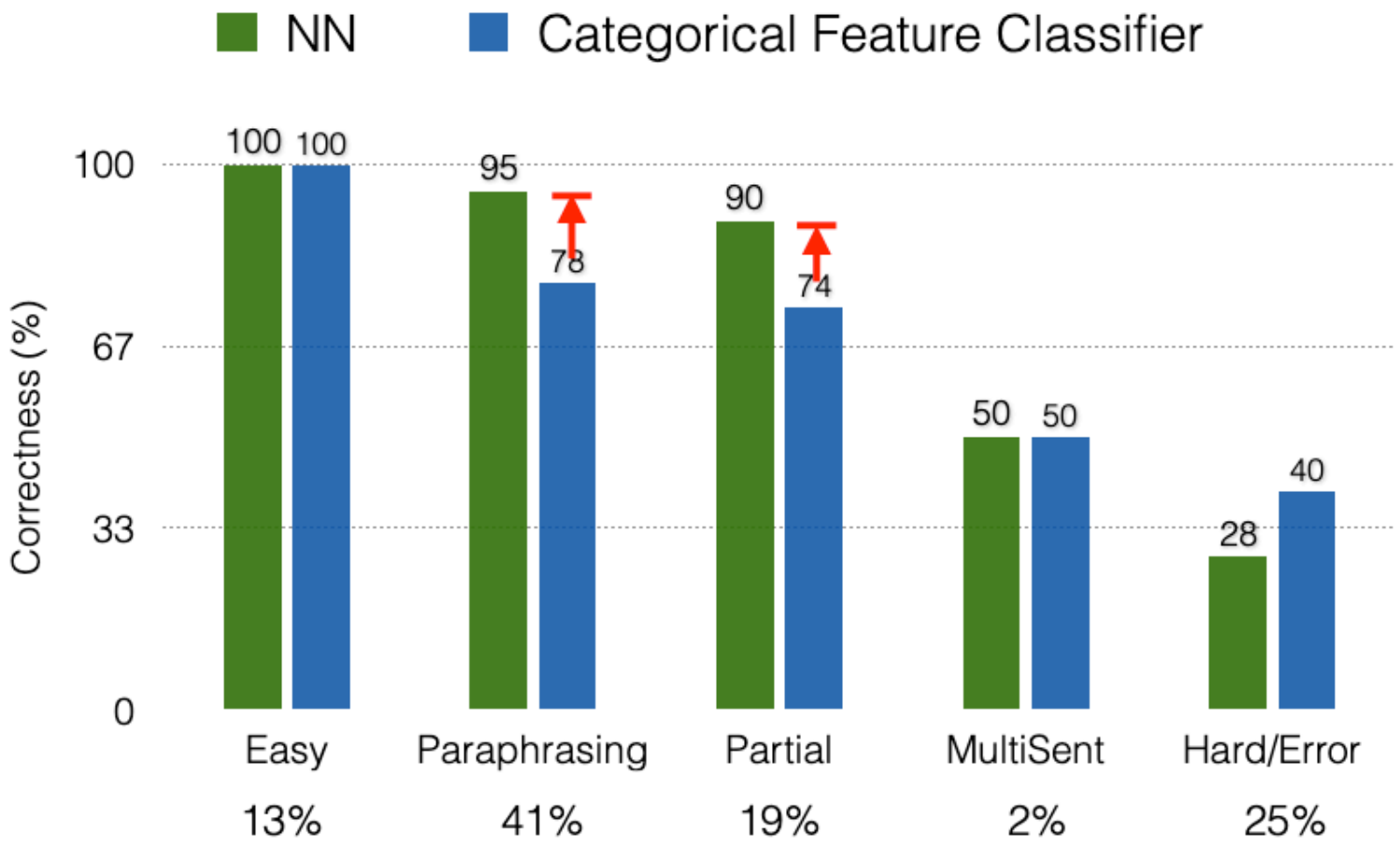
Where α is a simple one layer FFNN

A big win for neural models



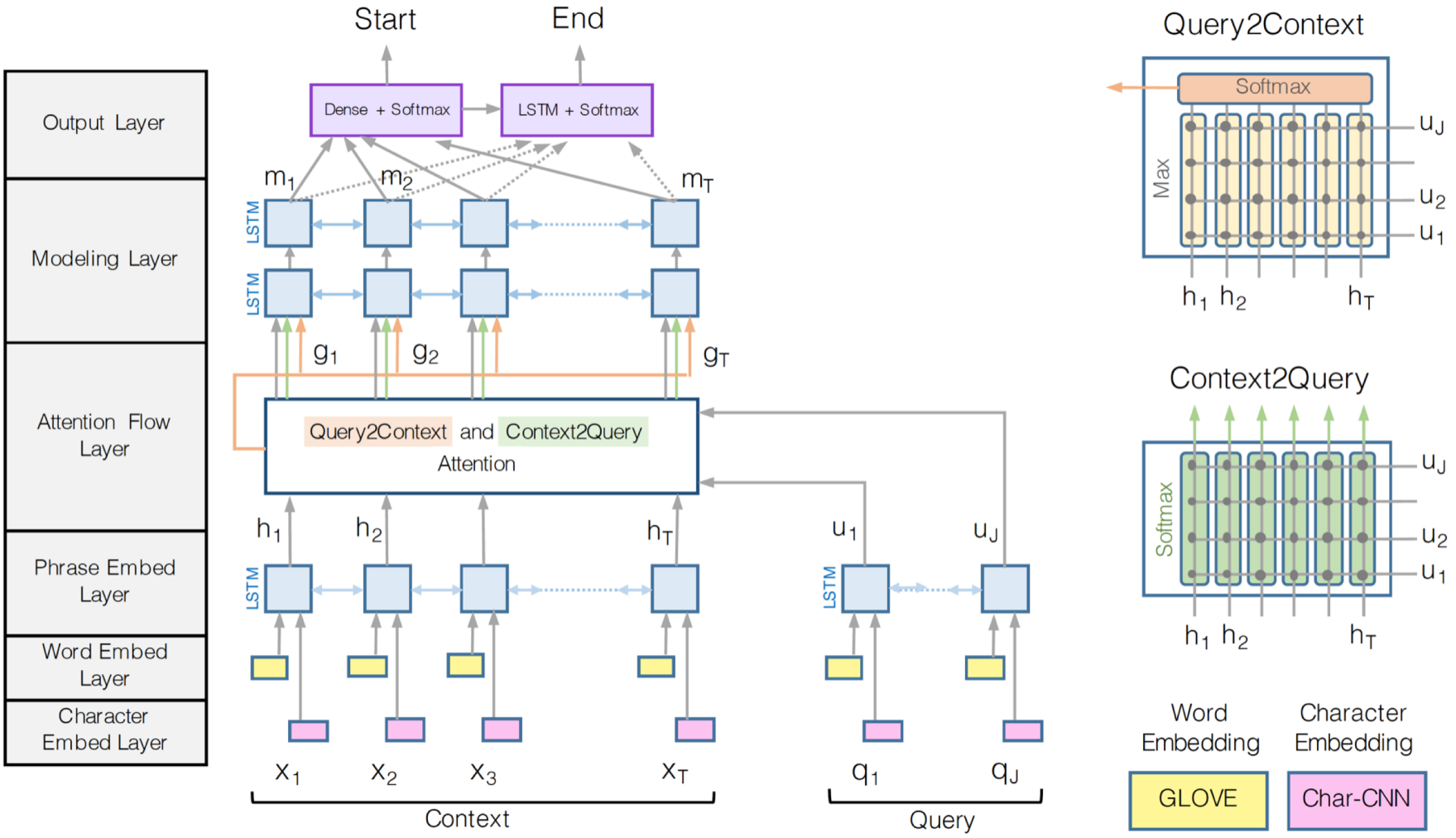
(Chen, Bolton, Manning, 2016)

What do these neural models do?



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

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



BiDAF

- 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)

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

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

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

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

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

BiDAF

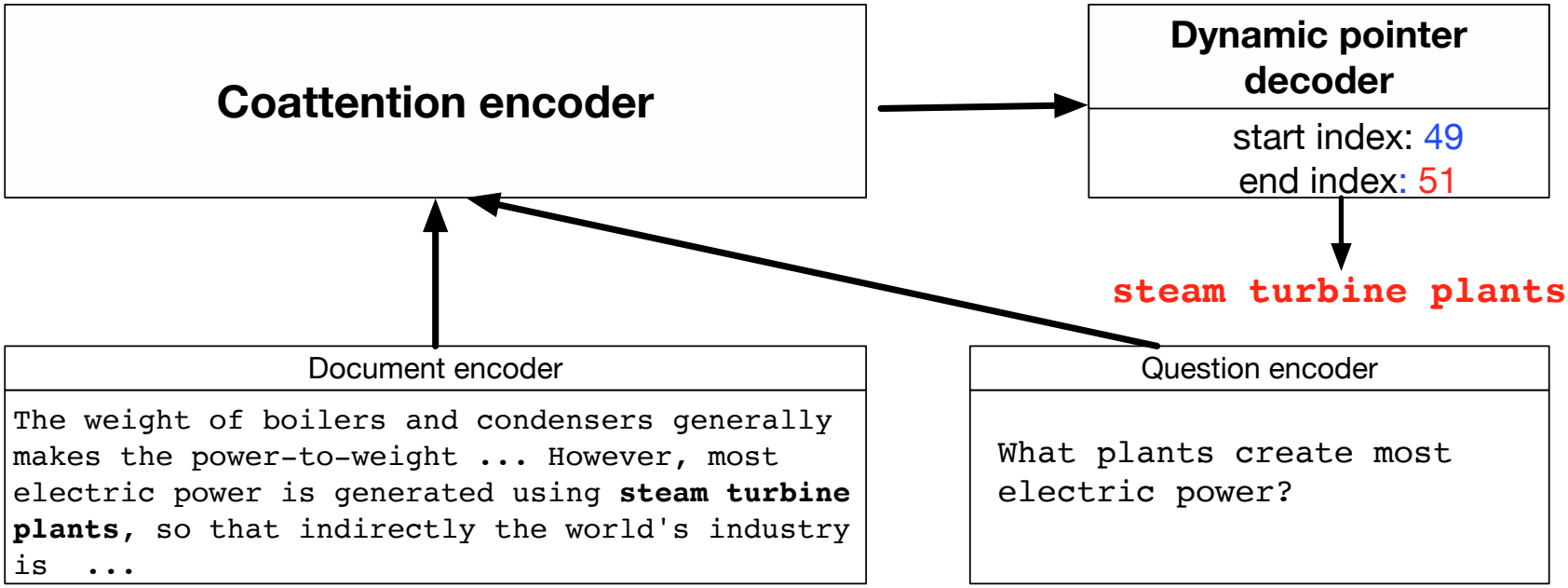
- There is then a “modelling” layer:
  - Another deep (2-layer) BiLSTM over the passage
- And answer span selection is more complex:
  - Start: Pass output of BiDAF and modelling layer concatenated to a dense FF layer and then a softmax
  - End: Put output of modelling layer M through another BiLSTM to give M_2 and then concatenate with BiDAF layer and again put through dense FF layer and a softmax

6. Recent, more advanced architectures

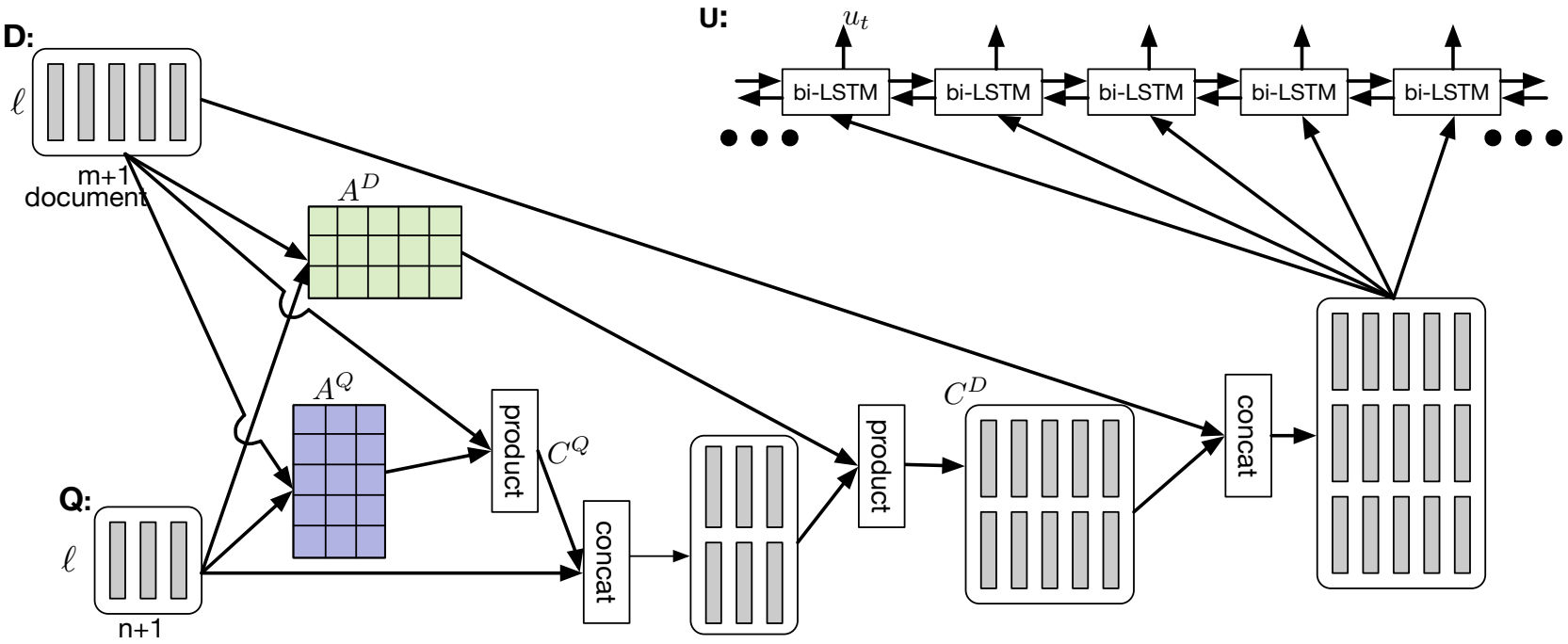
- Most of the work in 2016, 2017, and 2018 employed progressively more complex architectures with a multitude of variants of attention – often yielding good task gains

Dynamic Coattention Networks for Question Answering (Caiming Xiong, Victor Zhong, Richard Socher ICLR 2017)

- Flaw: Questions have input-independent representations
- Interdependence needed for a comprehensive QA model



Coattention Encoder



Coattention layer

- Coattention layer again provides a two-way attention between the context and the question
- However, coattention involves a second-level attention computation:
  - attending over representations that are themselves attention outputs
- We use the C2Q attention distributions α_i to take weighted sums of the Q2C attention outputs $\mathbf{b}_j$. This gives us second-level attention outputs $\mathbf{s}_i$:

$$\boldsymbol{s}_i = \sum_{j=1}^{M+1} \alpha_j^i \boldsymbol{b}_j \in \mathbb{R}^l \quad \forall i \in \{1, \ldots, N\}$$

Co-attention: Results on SQUAD Competition

Model	Dev EM	Dev F1	Test EM	Test F1
Ensemble				
DCN (Ours)	**70.3**	**79.4**	**71.2**	**80.4**
Microsoft Research Asia *	–	–	69.4	78.3
Allen Institute *	69.2	77.8	69.9	78.1
Singapore Management University *	67.6	76.8	67.9	77.0
Google NYC *	68.2	76.7	–	–
Single model				
DCN (Ours)	65.4	**75.6**	**66.2**	**75.9**
Microsoft Research Asia *	65.9	75.2	65.5	75.0
Google NYC *	**66.4**	74.9	–	–
Singapore Management University *	–	–	64.7	73.7
Carnegie Mellon University *	–	–	62.5	73.3
Dynamic Chunk Reader (Yu et al., 2016)	62.5	71.2	62.5	71.0
Match-LSTM (Wang & Jiang, 2016)	59.1	70.0	59.5	70.3
Baseline (Rajpurkar et al., 2016)	40.0	51.0	40.4	51.0
Human (Rajpurkar et al., 2016)	81.4	91.0	82.3	91.2

Results are at time of ICLR submission

See https://rajpurkar.github.io/SQuAD-explorer/ for latest results

FusionNet (Huang, Zhu, Shen, Chen 2017)

Attention functions

MLP (Additive) form:

$$S_{ij} = s^T \tanh(W_1 c_i + W_2 q_j)$$

Space: O(mnk), W is kxd

Bilinear (Product) form:

$$S_{ij} = c_i^T W q_j$$

$$S_{ij} = c_i^T U^T V q_j$$

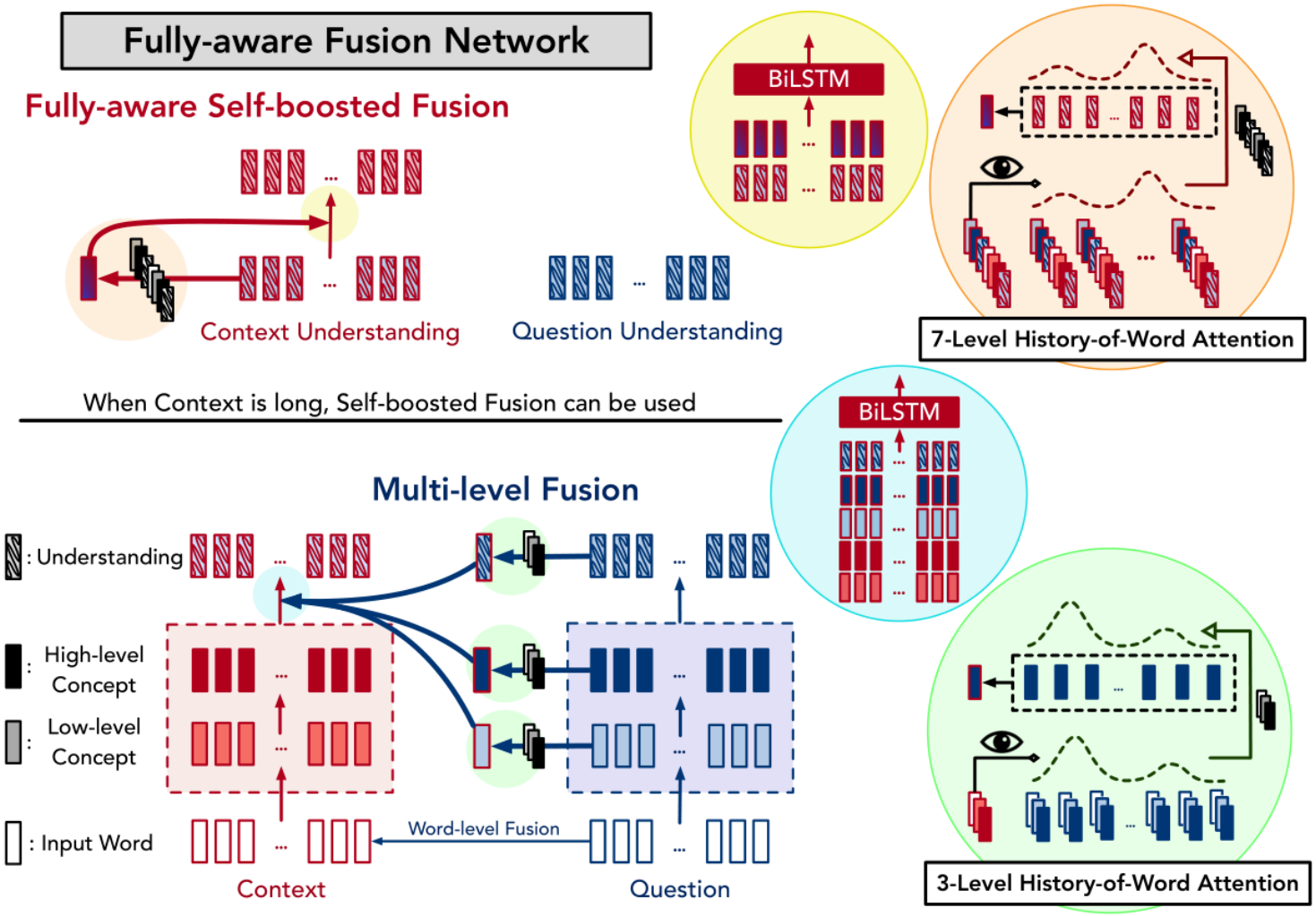
Space: O((m+n)k)

$$S_{ij} = c_i^T W^T D W q_j$$

1. Smaller space
2. Non-linearity

$$S_{ij} = Relu(c_i^T W^T) D Relu(W q_j)$$

FusionNet tries to combine many forms of attention



Multi-level inter-attention

$$\{\boldsymbol{m}_i^{(k),C}\}_{i=1}^m = \text{Attn}(\{\text{HoW}_i^C\}_{i=1}^m, \{\text{HoW}_i^Q\}_{i=1}^n, \{\boldsymbol{h}_i^{Q,k}\}_{i=1}^n), 1 \leq k \leq K+1$$

$$\text{HoW}_i^C = [\text{GloVe}(w_i^C); \text{BERT}_{w_i^C}; \boldsymbol{h}_i^{C,1}; ..., \boldsymbol{h}_i^{C,k}],$$

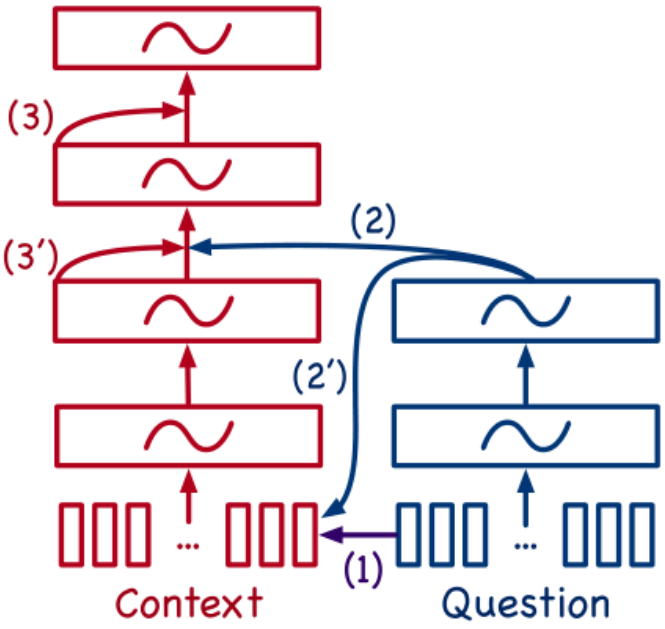
$$\text{HoW}_i^Q = [\text{GloVe}(w_i^Q); \text{BERT}_{w_i^Q}; \boldsymbol{h}_i^{Q,1}; ..., \boldsymbol{h}_i^{Q,k}].$$

After multi-level inter-attention, use RNN, self-attention and another RNN to obtain the final representation of context: $\{\boldsymbol{u}_i^C\}$

Recent, more advanced architectures

- Most of the work in 2016, 2017, and 2018 employed progressively more complex architectures with a multitude of variants of attention – often yielding good task gains

Architectures	(1)	(2)	(2')	(3)	(3')
Match-LSTM (Wang and Jiang, 2016)		✓			
DCN (Xiong et al., 2017)		✓			✓
FastQA (Weissenborn et al., 2017)	✓				
FastQAExt (Weissenborn et al., 2017)	✓	✓		✓	
BiDAF (Seo et al., 2017)		✓			✓
RaSoR (Lee et al., 2016)	✓		✓		
DrQA (Chen et al., 2017)	✓				
MPCM (Wang et al., 2016)	✓	✓			
Mnemonic Reader (Hu et al., 2017)	✓	✓		✓	
R-net (Wang et al., 2017b)		✓		✓	



7. ELMo and BERT preview

Contextual word representations
Using language model-like objectives

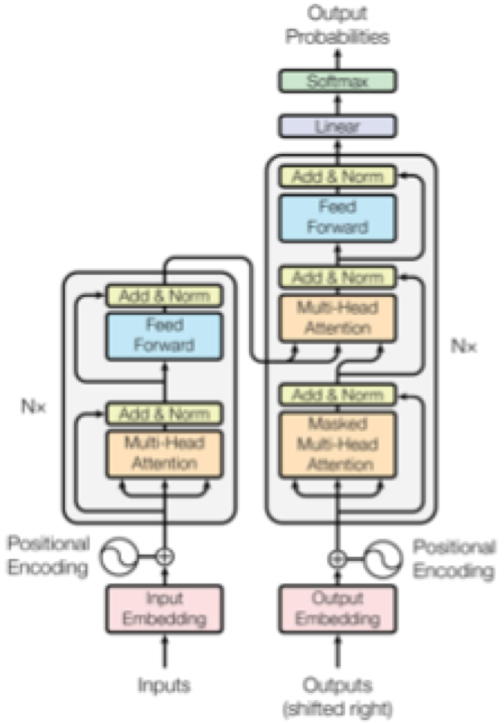
Elmo

(Peters et al, 2018)

Bert

(Devlin et al, 2018)

The transformer architecture used in BERT is sort of attention on steroids. More later!



(Vaswani et al, 2017)

Look at SDNet as an example of how to use BERT as submodule: https://arxiv.org/abs/1812.03593

SQuAD 2.0 leaderboard, 2019-02-07

		EM	F1
36 Sep 13, 2018	BiDAF++ (single model) *UW and FAIR*	65.651	68.866
37 Jun 27, 2018	BSAE AddText (single model) *reciTAL.ai*	63.338	67.422
38 Aug 14, 2018	eeAttNet (single model) *BBD NLP Team* https://www.bbdservice.com	63.327	66.633
38 May 30, 2018	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	63.372	66.251
39 Nov 27, 2018	Tree-LSTM + BiDAF + ELMo (single model) *Carnegie Mellon University*	57.707	62.341
39 May 30, 2018	BiDAF + Self Attention (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	59.332	62.305
40 May 30, 2018	BiDAF-No-Answer (single model) *University of Washington [modified by Stanford]*	59.174	62.093