Globally Normalized Reader

Jonathan Raiman and John Miller
Baidu Silicon Valley Artificial Intelligence Lab
EMNLP

Motivation

Cast extractive QA as an iterative search problem can imporve the computation efficiency (without bi-attention and score all possible answer spans).

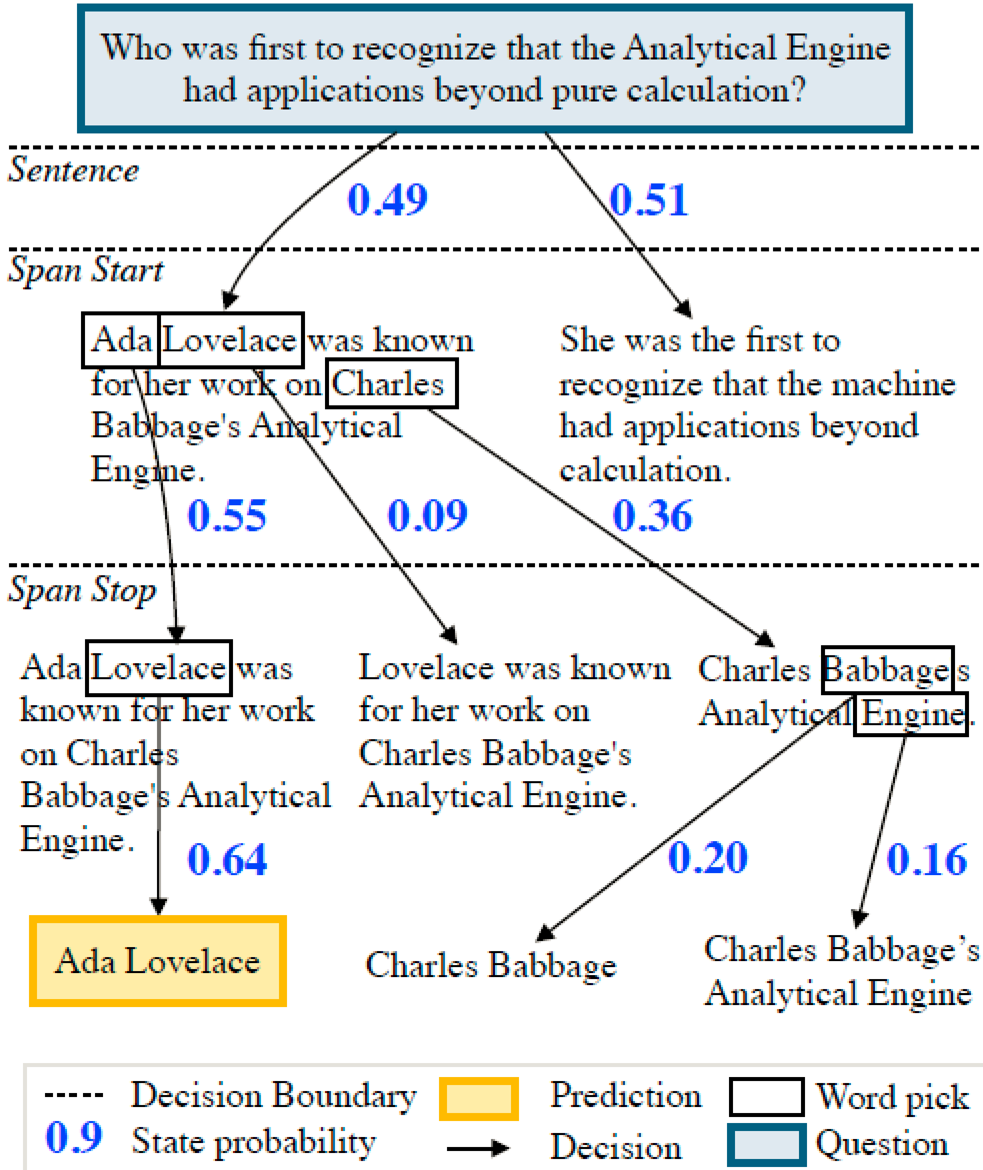
Task

input: <question, document>

output: <answer>

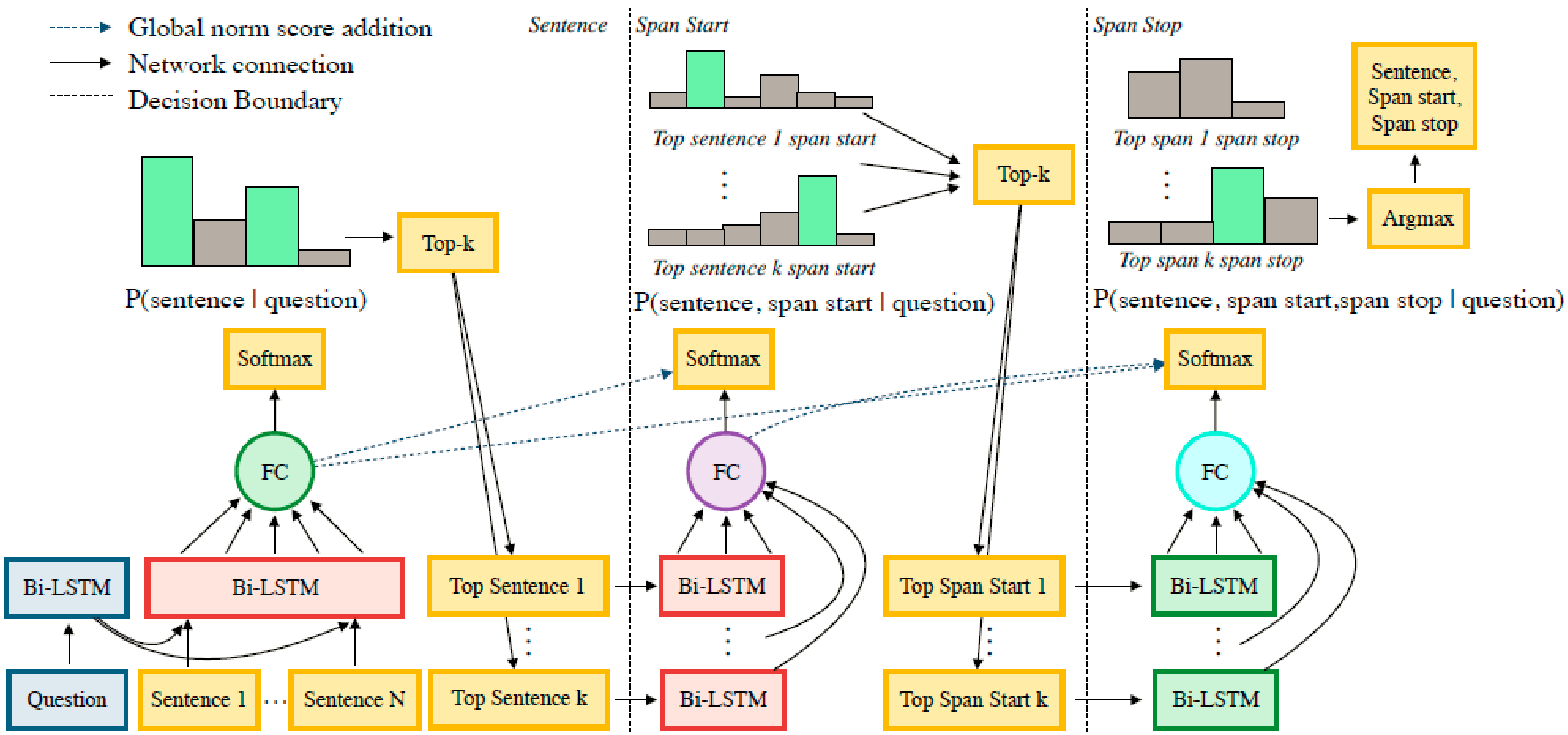
answer is a span of document.

Overview


Who was first to recognize that the Analytical Engine had applications beyond pure calculation?
Sentence
0.49
0.51
Span Start
Ada Lovelace was known for her work on Charles Babbage's Analytical Engine.
She was the first to recognize that the machine had applications beyond calculation.
0.55
0.09
0.36
Span Stop
Ada Lovelace was known for her work on Charles Babbage's Analytical Engine.
Lovelace was known for her work on Charles Babbage's Analytical Engine.
Charles Babbage's Analytical Engine.
0.64
0.20
0.16
Ada Lovelace
Charles Babbage
Charles Babbage's Analytical Engine
Decision Boundary
Prediction
Word pick
0.9
State probability
Decision
Question

Model

Architecture



Question Encoding

BiLSTM(q)= $[(h_1^{fwd}, h_1^{bwd}), (h_2^{fwd}, h_2^{bwd}), ..., (h_l^{fwd}, h_l^{bwd})]$

Aligned Question Embedding

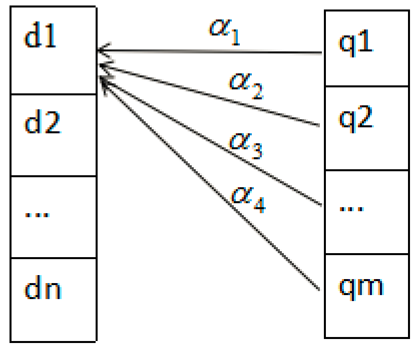
$$s_j = w_q^\top \mathrm{MLP}([h_j^{\mathrm{bwd}}; h_j^{\mathrm{fwd}}])$$

$$\alpha_j = \frac{\exp(s_j)}{\sum_{j'=1}^{\ell} \exp(s_{j'})}$$

$$q^{\mathrm{indep}} = \sum_{j=1}^{\ell} \alpha_j [h_j^{\mathrm{bwd}}; h_j^{\mathrm{fwd}}],$$

$$q = [h_1^{bwd}; h_l^{fwd}; q^{indep}]$$

Question-Aware Document Encoding



$$\bar{d}_1 = \sum_{i=1}^{m} \alpha_i q_i$$

$$s_{i,j,k} = \mathrm{MLP}(d_{i,j})^\top \mathrm{MLP}(q_k)$$

$$\alpha_{i,j,k} = \frac{\exp(s_{i,j,k})}{\sum_{k'=1}^{\ell} \exp(s_{i,j,k'})}$$

$$q_{i,j}^{\text{align}} = \sum_{k=1}^{\ell} \alpha_{i,j,k} q_k.$$

Document Encoding

$d_{i,j} = [e_{i,j}; q; f_{i,j}; q_{i,j}^{align}]$

$f_{i,j}$ = 1 if $D_{i,j}$ in Q else 0

BiLSTM(document) = $[(h_{1,1}^{fwd}, h_{1,1}^{bwd}), ..., (h_{n,m_n}^{fwd}, h_{n,m_n}^{bwd})]$

Answer Selection

Local Normalization

$$\begin{aligned}\mathbb{P}(a|d,q) =&\mathbb{P}_{\text{sent}}(i|d,q) \cdot \mathbb{P}_{\text{sw}}(j|i,d,q)\cdot \\ &\mathbb{P}_{\text{ew}}(k|j,i,d,q).\end{aligned}$$

$$\mathbb{P}_{\text{sent}}(i|d,q) = \frac{\exp(\phi_{\text{sent}}(d_i))}{\sum_{x=1}^{n}\exp(\phi_{\text{sent}}(d_x))},$$

$$\mathbb{P}_{\text{sw}}(j|i,d,q) = \frac{\exp(\phi_{\text{sw}}(d_{i,j}))}{\sum_{x=1}^{m_i}\exp(\phi_{\text{sw}}(d_{i,x}))},$$

$$\mathbb{P}_{\text{ew}}(k|j,i,d,q) = \frac{\exp(\phi_{\text{ew}}(d_{i,j:k}))}{\sum_{x=j}^{m_i}\exp(\phi_{\text{ew}}(d_{i,j:x}))}.$$

Global Normalization

$$\text{score}(a, d, q) = \phi_{\text{sent}}(d_i) + \phi_{\text{sw}}(d_{i,j}) + \phi_{\text{ew}}(d_{i,j:k}).$$

$$\mathbb{P}(a \mid d, q) = \frac{\exp(\text{score}(a, d, q))}{Z},$$

$$Z = \sum_{a' \in \mathcal{A}(d)} \exp(\text{score}(a', d, q)).$$

Beam Search

$$Z \approx \sum_{a' \in \mathcal{B}} \exp(\text{score}(a', d, q)).$$

Data Augmentation

1.Locate named entities in document and question.
2.Collect surface variation for each entity type:

$$human -> \{AdaLovelace, DanielKahnemann, ...\}$$

$$country -> \{USA, France, ...\}$$

3.Generage new document-question-answer examples by swapping each named entity in an original triplet with a surface variant that shares the type.

Experiments

DataSet: SQuAD

Model	EM	F1
Human (Rajpurkar et al., 2016)	80.3	90.5
Single model		
Sliding Window (Rajpurkar et al., 2016)	13.3	20.2
Match-LSTM (Wang and Jiang, 2016)	64.1	73.9
DCN (Xiong et al., 2016)	65.4	75.6
Rasor (Lee et al., 2016)	66.4	74.9
Bi-Attention Flow (Seo et al., 2016)	67.7	77.3
R-Net(Wang et al., 2017)	**72.3**	**80.6**
Globally Normalized Reader w/o Type Swaps (Ours)	66.6	75.0
Globally Normalized Reader (Ours)	68.4	76.21

Experiments

Model	B	EM	F1	Sentence
Local, $T = 10^4$	1	65.7	74.8	**89.0**
	2	66.6	75.0	88.3
	10	66.7	75.0	88.6
	32	66.3	74.6	88.0
	64	66.6	75.0	88.8
Global, $T = 10^4$	1	58.8	68.4	84.5
	2	64.3	73.0	86.8
	10	66.6	75.2	88.1
	32	**68.4**	**76.21**	88.4
	64	67.0	75.6	88.4

Experiments

Model	T	EM	F1	Sentence
Local	0	65.8	74.0	88.0
Local	10^3	66.3	74.6	88.9
Local	10^4	66.7	74.9	**89.0**
Local	$5 \cdot 10^4$	66.7	75.0	**89.0**
Local	10^5	66.2	74.5	88.6
Global	0	66.6	75.0	88.2
Global	10^3	66.9	75.0	88.1
Global	10^4	**68.4**	**76.21**	88.4
Global	$5 \cdot 10^4$	66.8	75.3	88.3
Global	10^5	66.1	74.3	86.9