

Transformer-XH: Multi-Evidence Reasoning with eXtra Hop Attention

Chen Zhao

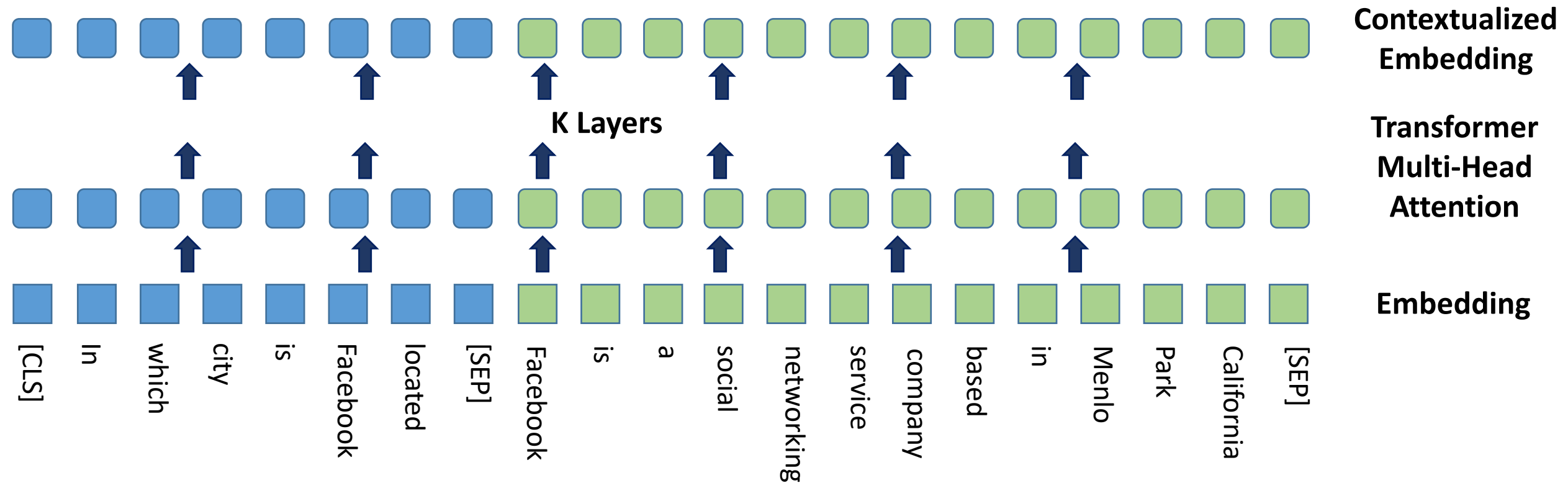
University of Maryland,
College Park

Chenyan Xiong, Corby Rosset,
Xia Song, Paul Bennett,
Saurabh Tiwary

Microsoft AI and Research

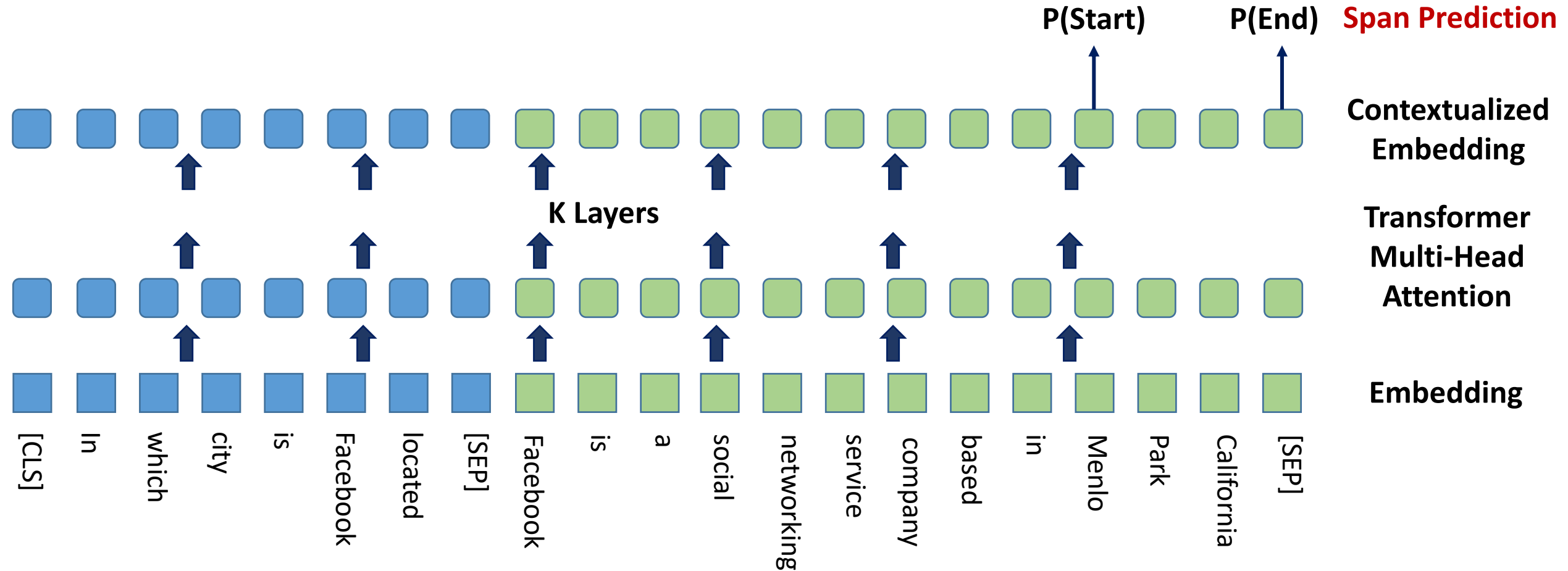
Transformers in Modeling Natural Language

- Input sentence(s) in sequential form
- (New) standard solutions with pre-trained Transformers (e.g., BERT)



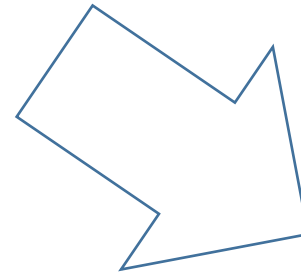
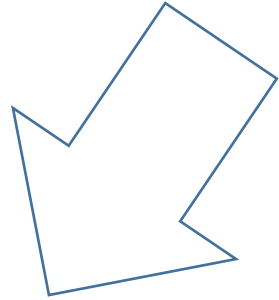
Transformer in Machine Reading

- (New) standard solutions with pre-trained Transformers (e.g., BERT)

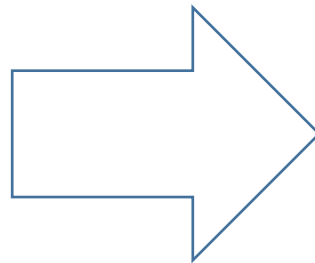


Structured Text Sequence

Facebook was founded by [Mark Zuckerberg](#), along with fellow [Harvard College](#) students and roommates.



Zuckerberg built a website called "Facemash" in 2003 while attending [Harvard University](#). The site

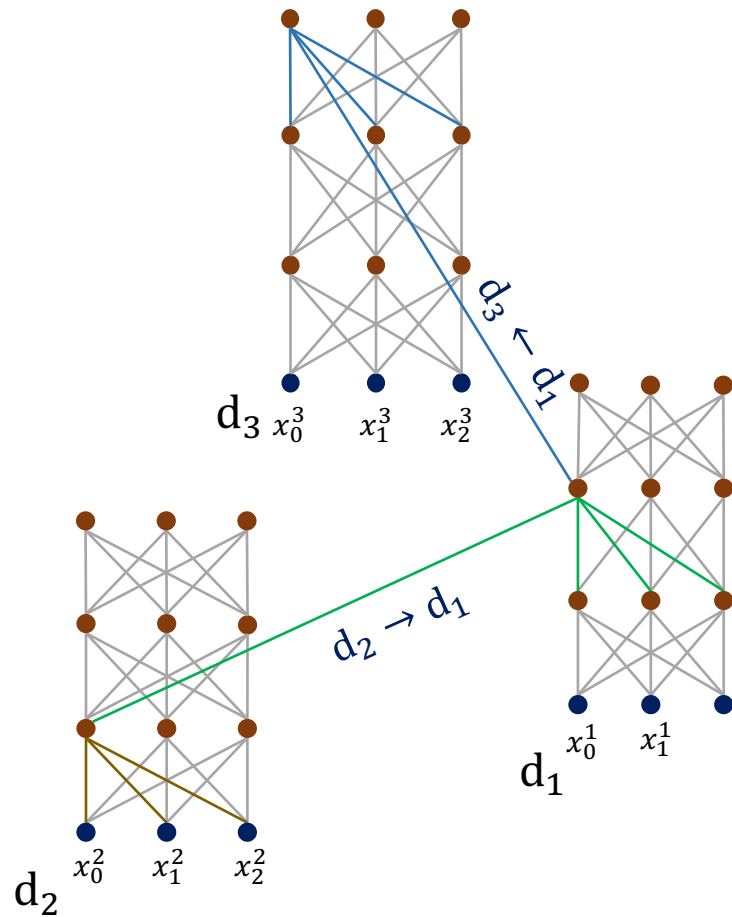


Harvard University is a private [Ivy League research university](#) in [Cambridge, Massachusetts](#), with about 6,800 undergraduate

This Work

- Transformer-XH for **structured** text:
 - Transformer with e**X**tra **H**op attentions
 - Global representations of **multiple connected** text pieces
- Strong performance on different multi-evidence reasoning tasks
 - Multi-hop QA (Hotpot QA)
 - Multi-evidence Fact Verification (FEVER)

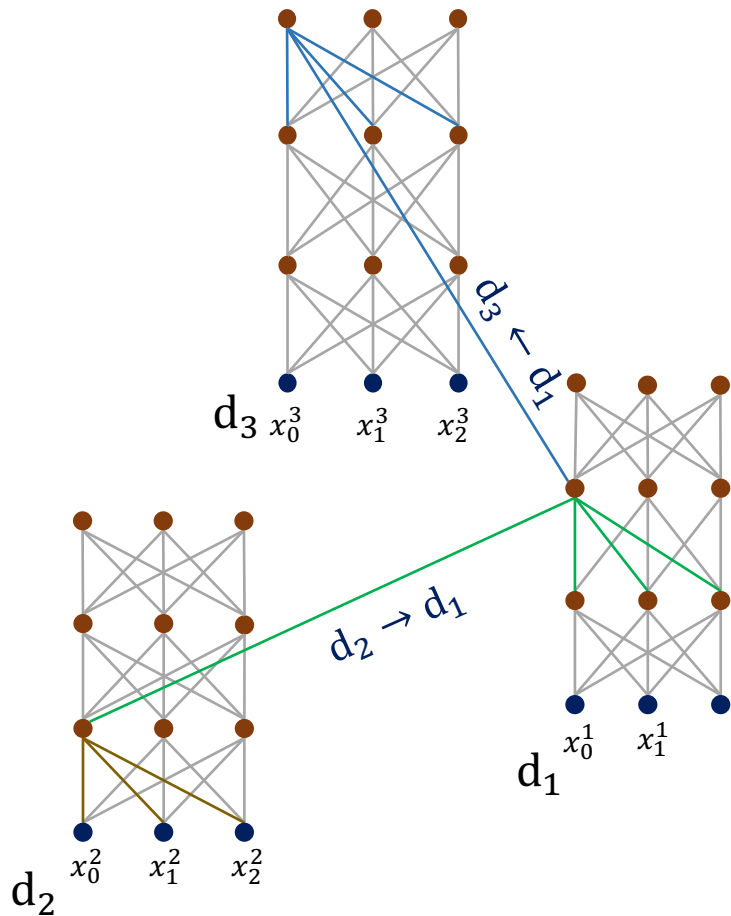
Transformer-XH: In Sequence Attention



In sequence (τ) attention in layer l , token i :

$$h_{\tau,i}^l = \sum_j \text{Softmax}_j \left(\frac{q_{\tau,i}^T \cdot k_{\tau,j}}{\sqrt{d_k}} \right) \cdot v_{\tau,j}$$

Transformer-XH: Extra Hop Attention



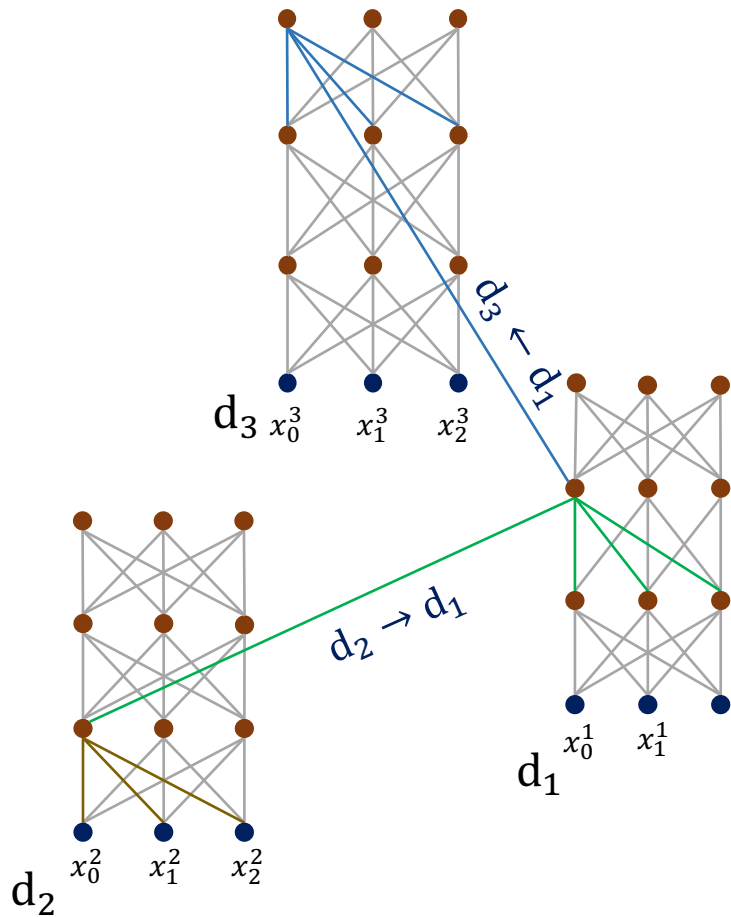
In sequence (τ) attention in layer l , token i :

$$h_{\tau,i}^l = \sum_j \text{Softmax}_j \left(\frac{q_{\tau,i}^T \cdot k_{\tau,j}}{\sqrt{d_k}} \right) \cdot v_{\tau,j}$$

Extra hop attention between sequences ($\tau \rightarrow \eta$):

$$\hat{h}_{\tau,0}^l = \sum_{\eta; e_{\tau\eta}=1} \text{Softmax}_{\eta} \left(\frac{\hat{q}_{\tau,0}^T \cdot \hat{k}_{\tau,0}}{\sqrt{d_k}} \right) \cdot \hat{v}_{\eta,0}$$

Transformer-XH: Layer Representation



In sequence (τ) attention in layer l , token i :

$$h_{\tau,i}^l = \sum_j \text{Softmax}_j \left(\frac{q_{\tau,i}^T \cdot k_{\tau,j}}{\sqrt{d_k}} \right) \cdot v_{\tau,j}$$

Extra hop attention between sequences ($\tau \rightarrow \eta$):

$$\hat{h}_{\tau,0}^l = \sum_{\eta; e_{\tau\eta}=1} \text{Softmax}_{\eta} \left(\frac{\hat{q}_{\tau,0}^T \cdot \hat{k}_{\tau,0}}{\sqrt{d_k}} \right) \cdot \hat{v}_{\eta,0}$$

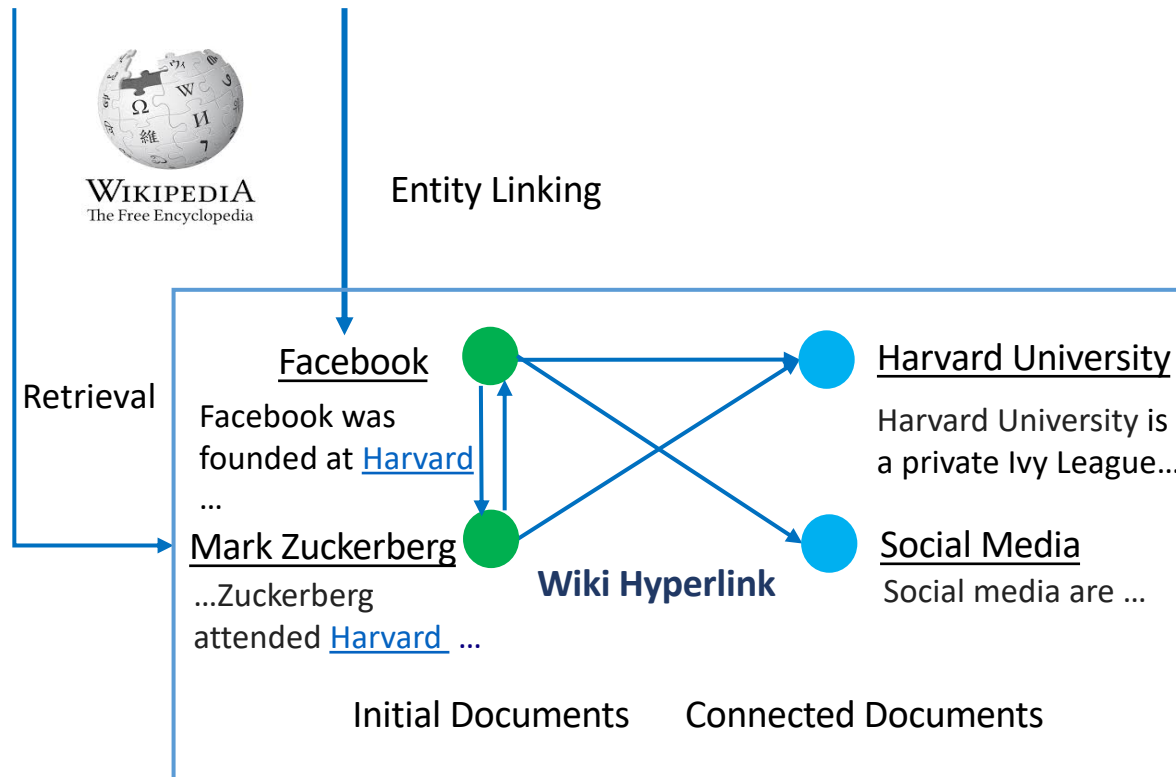
Combine two attentions:

$$\begin{aligned} \tilde{h}_{\tau,0}^l &= \text{Linear}([h_{\tau,0}^l \circ \hat{h}_{\tau,0}^l]) \\ \tilde{h}_{\tau,l}^l &= h_{\tau,i}^l; \forall i \neq 0 \end{aligned}$$

Transformer-XH in Multi-hop QA

Input Question:

In which city was Facebook launched?

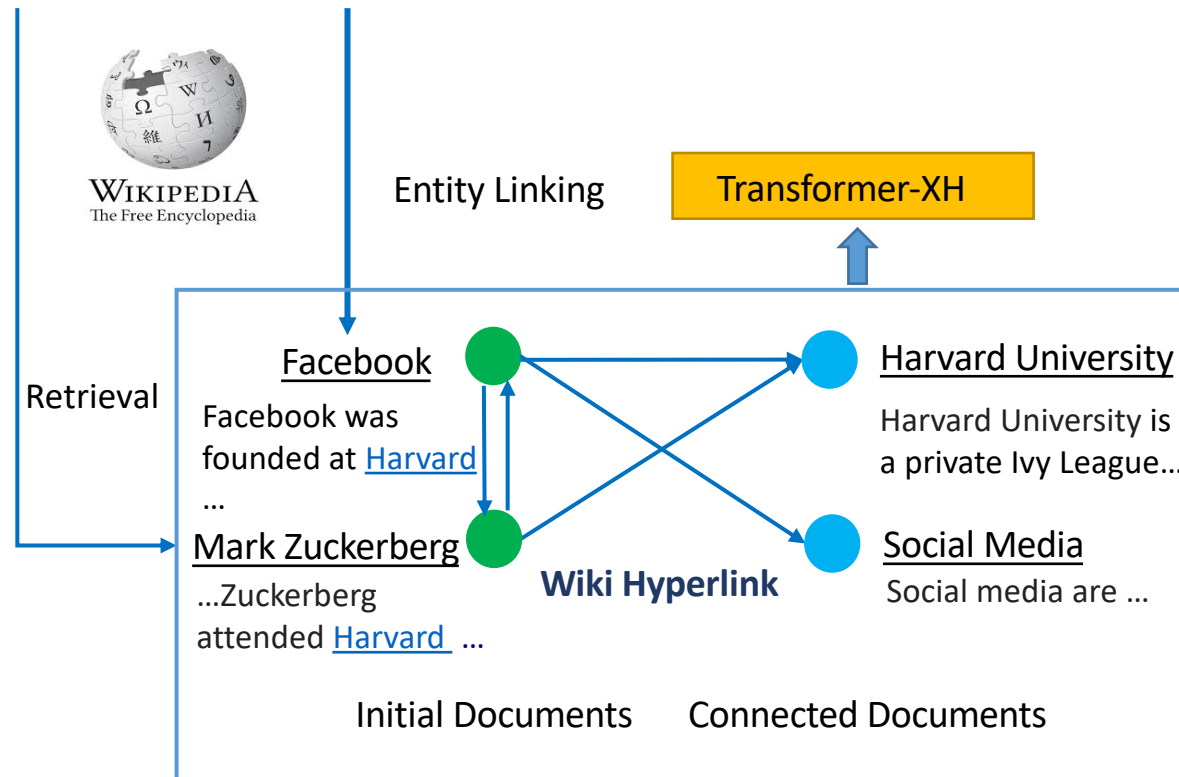


- Evidence graph construction

Transformer-XH in Multi-hop QA

Input Question:

In which city was Facebook launched?



- Evidence graph construction
- Transformer-XH

Transformer-XH in Multi-hop QA

Input Question:

In which city was Facebook launched?



WIKIPEDIA
The Free Encyclopedia

Answer:

Harvard University is a private Ivy League research university in Cambridge, Massachusetts...

Entity Linking

● Harvard University

Node

Span

Transformer-XH

Retrieval

Facebook

Facebook was founded at Harvard

...

Mark Zuckerberg

...Zuckerberg attended Harvard ...

Wiki Hyperlink

● Harvard University

Harvard University is a private Ivy League...

● Social Media

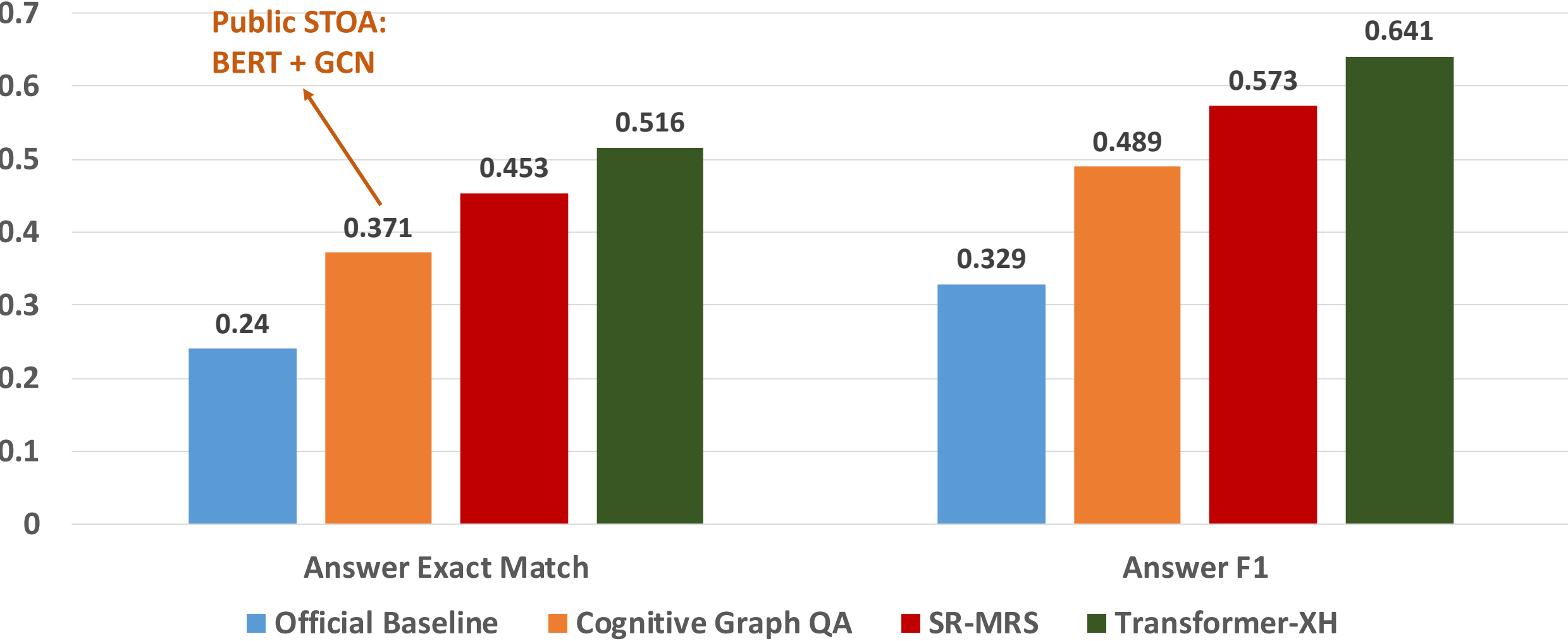
Social media are ...

Initial Documents

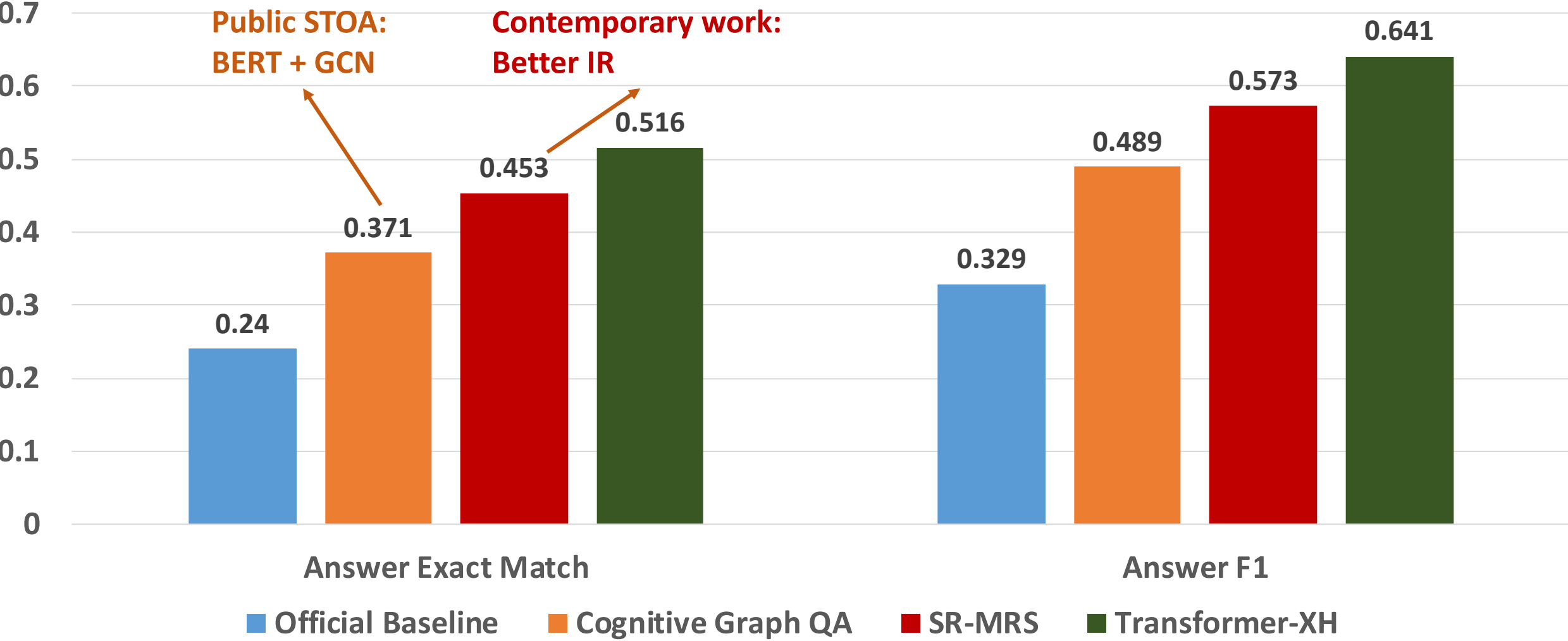
Connected Documents

- Evidence graph construction
- Transformer-XH
- Task specific Layers

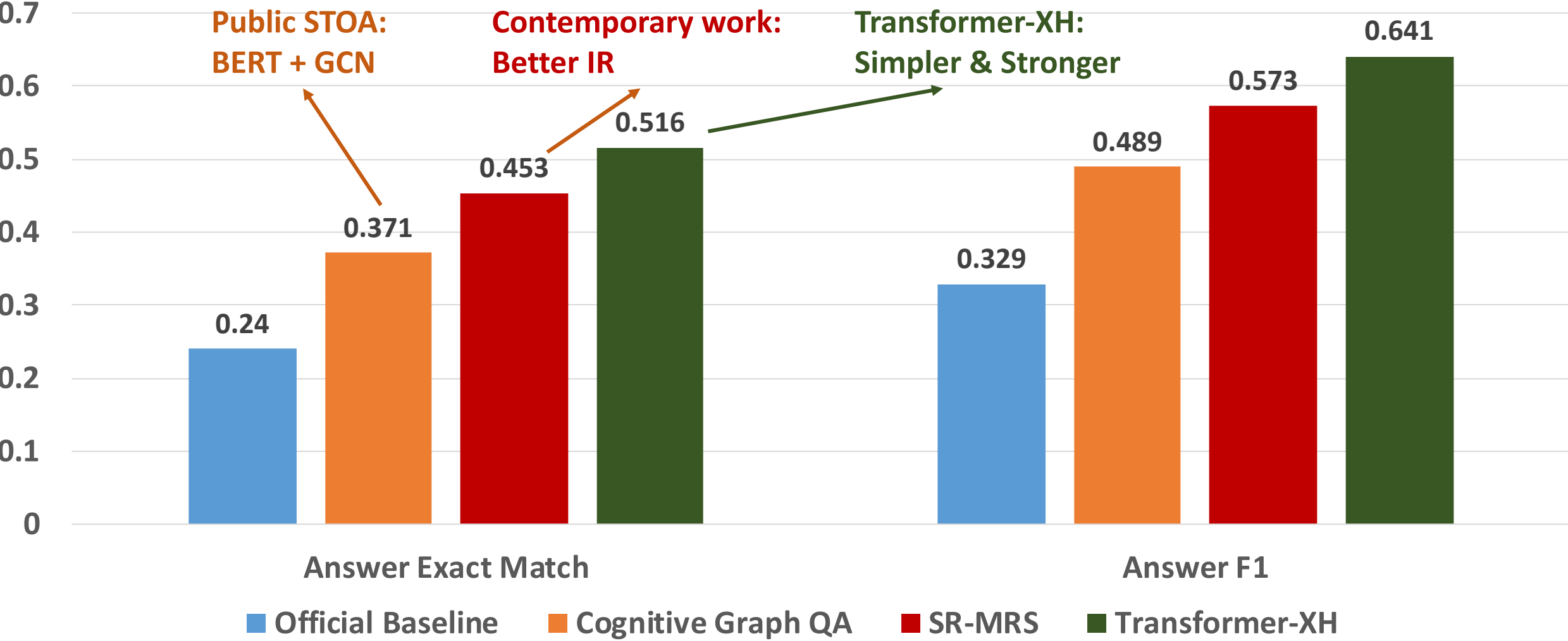
Results on Hotpot QA Full-Wiki Test Set



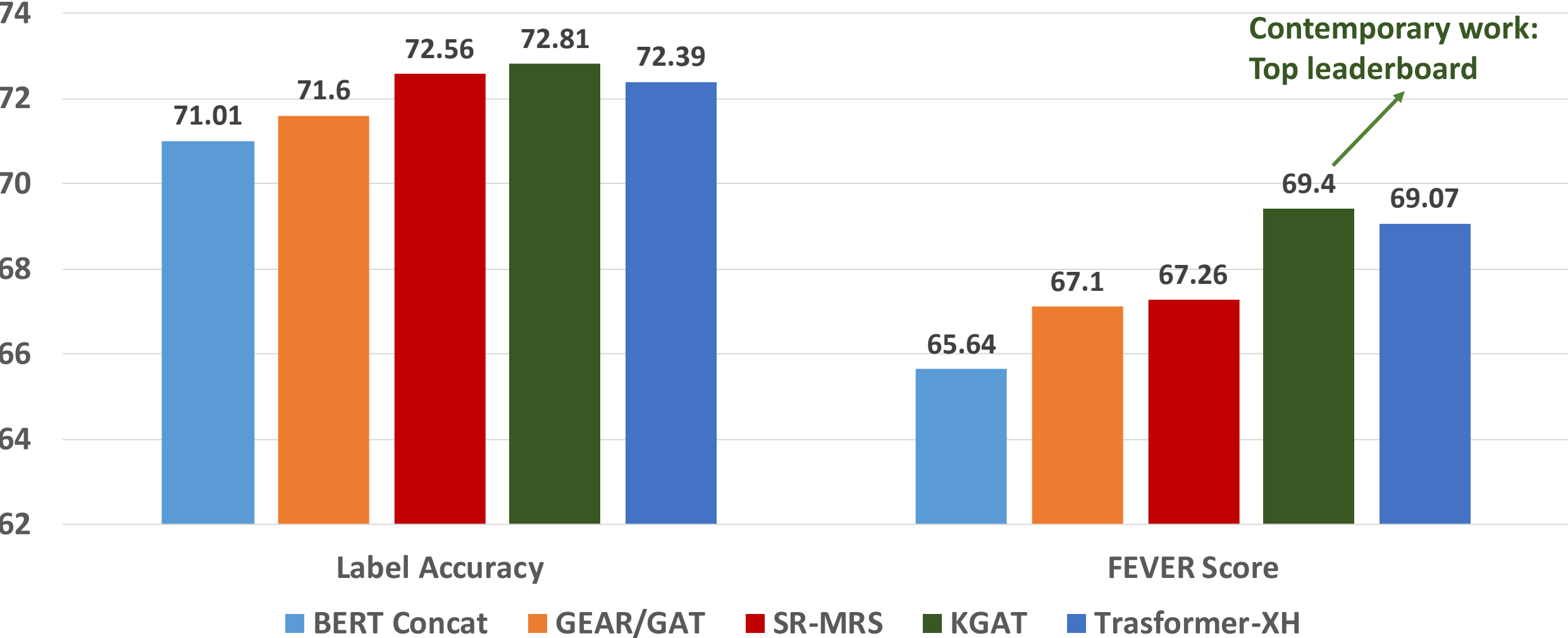
Results on Hotpot QA Full-Wiki Test Set



Results on Hotpot QA Full-Wiki Test Set



Results on FEVER 1.0



Summary

- Transformer-XH: Simple but powerful adaption of Transformer for structured text
- It learns better representations via extra hop attention
- Strong performance on multiple tasks

Thanks!