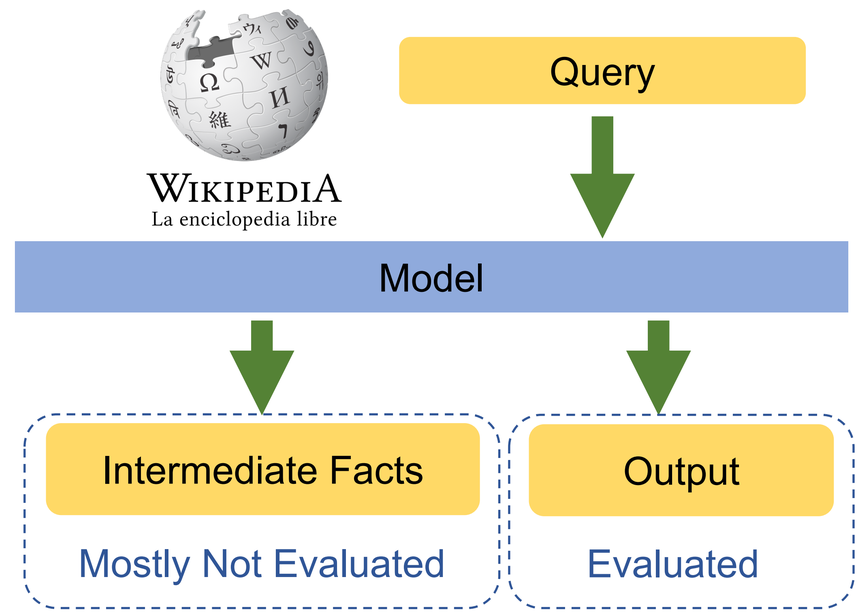
Revealing the Importance of Semantic Retrieval for Machine Reading at Scale

Yixin Nie yixin1@cs.unc.edu
Songhe Wang songhe17@cs.unc.edu
Mohit Bansal mbansal@cs.unc.edu

Background & Motivation

Machine Reading at Scale (**MRS**) (Chen et al. 2017)



Motivation:
- Effects of intermediate retrieval performance on downstream are not well-studied.
- Relationship between intermediate retrieved facts and the final output will give insights on model development

Datasets: HotpotQA & FEVER

Overall Method

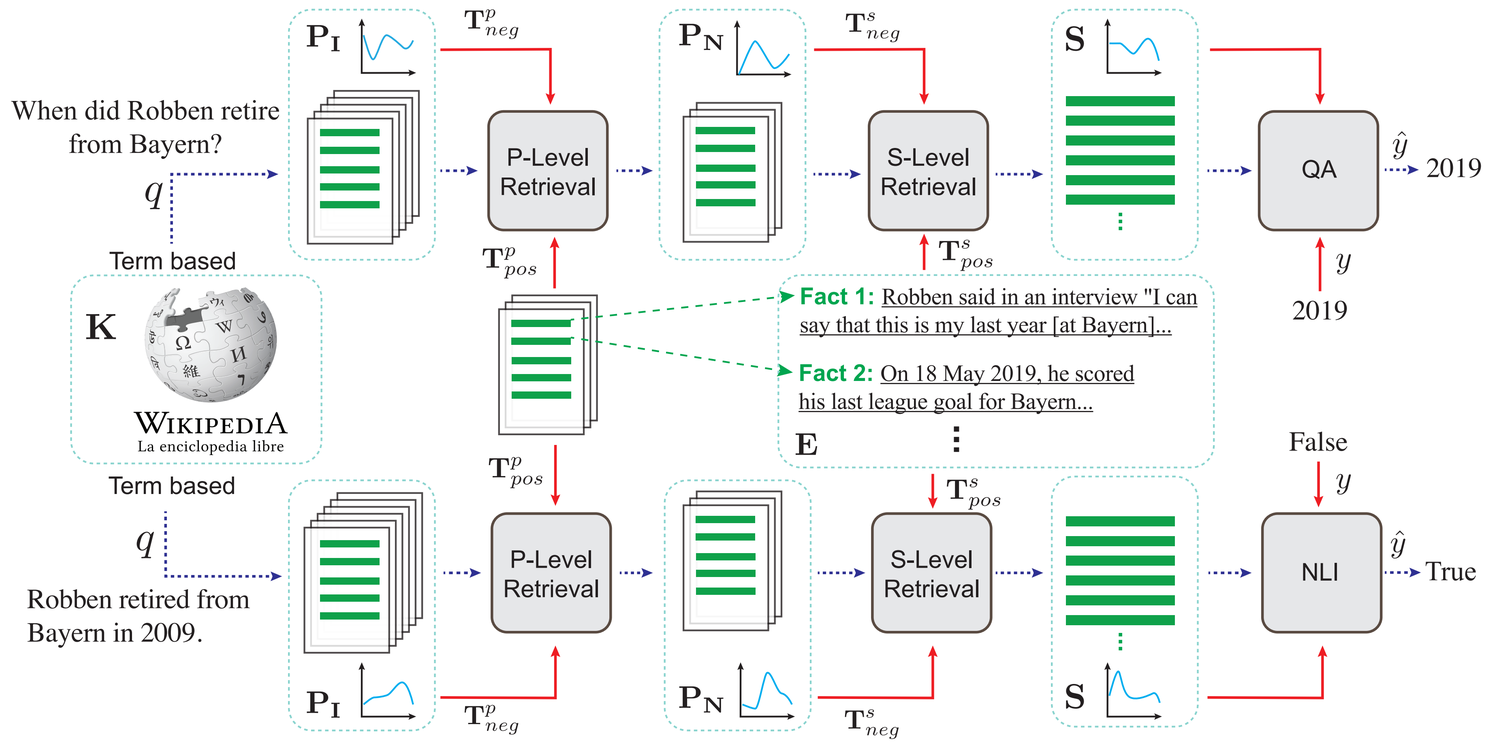


Figure 1: System Overview.

Input
- Retrieval $[CLS]\ Query\ [SEP]\ Context\ [SEP]$
- QA $[CLS]\ yes\ no\ Query\ [SEP]\ Context\ [SEP]$

Objective

$$\mathcal{J}_{retri} = -\sum_{i \in \mathbf{T}^{p/s}_{pos}} \log(\hat{p}_i) - \sum_{i \in \mathbf{T}^{p/s}_{neg}} \log(1-\hat{p}_i)$$

$$\mathcal{J}_{qa} = -\sum_{i} \left[\log(\hat{y}^s_i) + \log(\hat{y}^e_i)\right]$$

Benchmark Results

Method	Ans		Sup		Joint	
	EM	F1	EM	F1	EM	F1
Yang (2018)	24.7	34.4	5.3	41.0	2.5	17.7
Ding (2019)	37.6	49.4	23.1	58.5	12.2	35.3
whole pip.	**46.5**	**58.8**	**39.9**	**71.5**	**26.6**	**49.2**
Dev set						
Yang (2018)	24.0	32.9	3.9	37.7	1.9	16.2
MUPPET	30.6	40.3	16.7	47.3	10.9	27.0
Ding (2019)	37.1	48.9	22.8	57.7	12.4	34.9
whole pip.	**45.3**	**57.3**	**38.7**	**70.8**	**25.1**	**47.6**
Test set						

Table 1: HotpotQA test results.

Model	F1	LA	FS
Hanselowski (2018)	-	68.49	64.74
Yoneda (2018)	35.84	69.66	65.41
Nie (2019)	51.37	69.64	66.15
Full system (single)	**76.87**	**75.12**	**70.18**
Dev set			
Hanselowski (2018)	37.33	65.22	61.32
Yoneda (2018)	35.21	67.44	62.34
Nie (2019)	52.81	68.16	64.23
Full system (single)	**74.62**	**72.56**	**67.26**
Test set			

Table 2: FEVER test results.

Ablation

Method	P-Level Retrieval			S-Level Retrieval				Answer		Joint	
	Prec.	Rec.	F1	EM	Prec.	Rec.	F1	EM	F1	EM	F1
Whole Pip.	35.17	87.93	50.25	**39.86**	**75.60**	**71.15**	**71.54**	**46.50**	**58.81**	**26.60**	**49.16**
Pip. w/o p-level	6.02	89.53	11.19	0.58	29.57	60.71	38.84	31.23	41.30	0.34	19.71
Pip. w/o s-level	35.17	87.92	50.25	-	-	-	-	44.77	56.71	-	-

Table 3: Ablation over the paragraph-level and sentence-level neural retrieval sub-modules on HotpotQA.

Method	P-Level Retrieval				S-Level Retrieval				Verification		
	Orcl.	Prec.	Rec.	F1	Orcl.	Prec.	Rec.	F1	LA	FS	L-F1 (S/R/N)
Whole Pip.	94.15	48.84	91.23	63.62	88.92	**71.29**	83.38	76.87	**70.18**	**75.01**	81.7/75.7/**67.1**
Pip. w/o p-level	**94.69**	18.11	**92.03**	30.27	**91.07**	44.47	**86.60**	58.77	61.55	67.01	76.5/72.7/40.8
Pip. w/o s-level	94.15	48.84	91.23	63.62	-	-	-	-	55.92	61.04	72.1/67.6/27.7

Table 4: Ablation over the paragraph-level and sentence-level neural retrieval sub-modules on FEVER.

Analysis

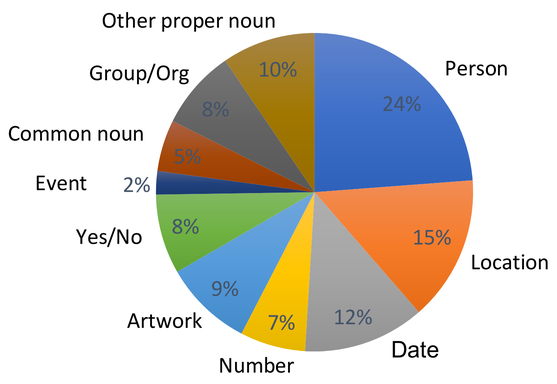


Figure 2: Proportion of answer types.

Answer Type	Total	Correct	Acc. (%)
Person	50	28	56.0
Location	31	14	45.2
Date	26	13	50.0
Number	14	4	28.6
Artwork	19	7	36.8
Yes/No	17	12	**70.6**
Event	5	2	40.0
Common noun	11	3	27.3
Group/Org	17	6	35.3
Other PN	20	9	45.0
Total	200	98	49.0

Table 5: System performance on different answer types. “PN”= Proper Noun.

Question: Wojtek Wolski played for what team based in the Miami metropolitan area?
GT Answer: Florida Panthers
GT Facts:
[**Florida Panthers,0**]: The Florida Panthers are a professional ice hockey team based in the Miami metropolitan area. (*P-Score* : 0.99; *S-Score* : 0.98)
[**Wojtek Wolski,1**]: In the NHL, he has played for the Colorado Avalanche, Phoenix Coyotes, New York Rangers, Florida Panthers, and the Washington Capitals. (*P-Score* : 0.98; *S-Score* : 0.95)
Distracting Fact:
[**History of the Miami Dolphins,0**]: The Miami Dolphins are a professional American football franchise based in the Miami metropolitan area. (*P-Score* : 0.56; *S-Score* : 0.97)
Wrong Answer : The Miami Dolphins

Figure 3: P-Score and S-Score are the retrieval score at paragraph and sentence level, respectively.

We change the intermediate facts by modifying the threshold at both paragraph and sentence level and plot its effects.
Paragraph-level retrieval matters most for QA.

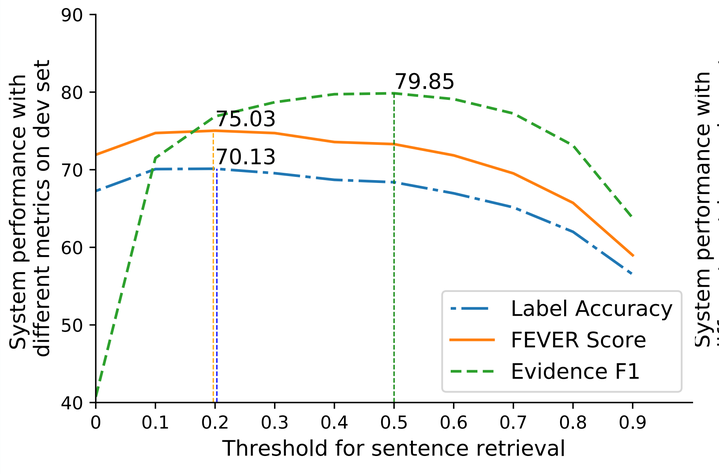


Figure 4: Effect of Sentence Retrieval for Fact Verification.

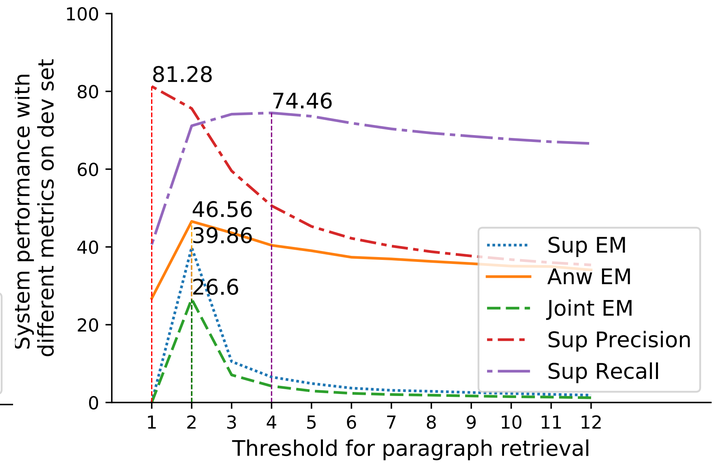


Figure 5: Effect of Paragraph Retrieval for QA.

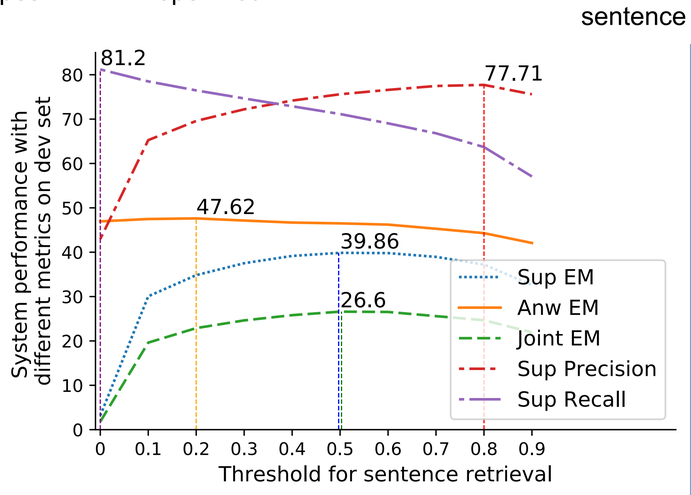


Figure 6: Effect of Sentence Retrieval for QA.

Acknowledgement:
This work was supported by awards from Verisk, Google, Facebook, Salesforce, and Adobe.

Reference:
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. *(ACL 2017)*
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *(EMNLP 2018)*