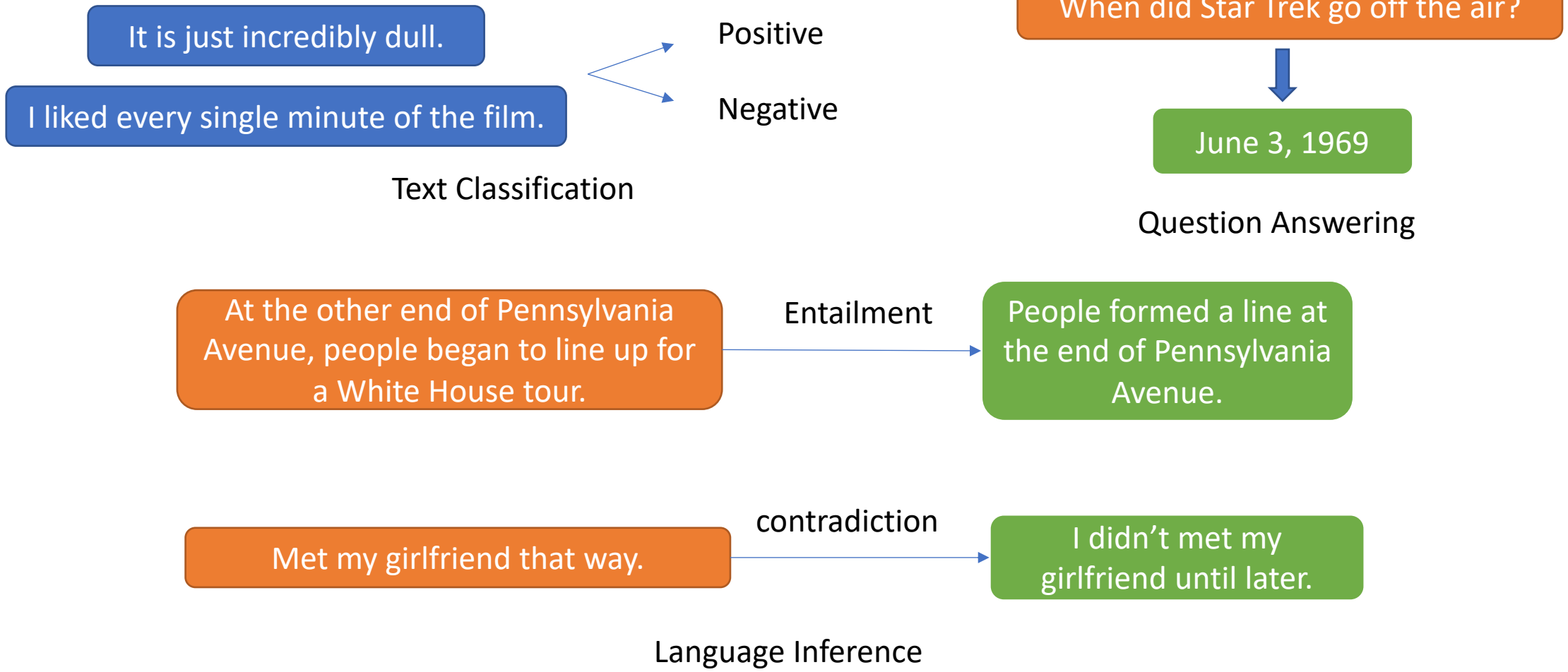
WSDM 2023 Tutorial

Knowledge-Augmented Methods for Natural Language **Understanding**

Yichong Xu
Cognitive Services Research, Microsoft

Natural Language Understanding



Natural Language Understanding with Knowledge

2016 Tour de France had a total of 198 musicians.

Shut Up is a song by Stormzy.

True

False

Fact Verification

When did Star Trek go off the air?

June 3, 1969

Question Answering

Kepulauan Sula — Country → Indonesia

Jerome Kersey — member of sports team → Portland Trail Blazers

Slot Filling/Knowledge Graph Completion

Cyclone Taylor , a professional ice hockey forward who led the [START_ENT] Vancouver Millionaires [END_ENT]

Entity Linking

Relevant Literature Reviews

- Yin, Da, et al. "A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models." *Spa-NLP workshop @ ACL 2022*
  - Overview of Knowledge-Intensive NLP methods
- Safavi, Tara, and Danai Koutra. "Relational World Knowledge Representation in Contextual Language Models: A Review." *EMNLP 2021.*
  - Ways to incorporate knowledge into LM pretraining

Background Knowledge for NLU

- Tasks and Benchmarks for Knowledge-rich NLU
- Knowledge Sources to Integrate in NLU
  - General-domain Knowledge
  - Domain-specific Knowledge
  - Commonsense Knowledge
- Integrating Knowledge into Language Models
  - Grounding Knowledge
  - Knowledge Representation
  - Fusing Knowledge Representations with LMs

Tasks and Benchmarks

- Question Answering
  - Open-Domain QA: Questions that requires external knowledge
    - NaturalQuestions (Kwiatkowski et al., 2019)
      What was the first capital city of Australia? -> Melbourne.
  - Commonsense QA: Questions that requires commonsense
    - CommonsenseQA (Talmor et al., 2018)
      What do all humans want to experience in their own home?
      👍 feel comfortable, 👎 work hard, 👎 fall in love, 👎 lay eggs, 👎 live forever

Tasks and Benchmarks

- Tasks based on Knowledge Graphs
  - Entity linking: Link mentions in sentences to entities in a KG
    - WNED-WIKI (Guo and Barbosa, 2018)
  - Slot filling/link prediction: Predict the tail entity given source entity and relation
    - Wikidata-based link prediction
    - LAMA probe: “The theory of relativity was developed by _____.”
  - Knowledge-base QA: Answer questions based on knowledge base
    - WebQuestions (Berant et al., 2013)
      What degrees did Barack Obama get?
      (bachelor_of_arts, juris_doctor)

Tasks and Benchmarks

- Fact verification: Verify a given claim from public knowledge
  - FEVER (Thorne et al., 2018)
    2016 Tour de France had a total of 198 musicians. -> REFUTE
- Dialogue Response Generation: generate response using given knowledge
  - Wizard of Wikipedia (Dinan et al., 2019)
- Commonsense Tasks
  - Coreference resolution: Winograd challenge (Levesque, 2011)
  - Next sentence prediction: Hellaswag (Zellers et al., 2019)

Knowledge Sources in NLU

- General open-domain Knowledge
- Domain-specific Knowledge
- Commonsense Knowledge

Wikipedia-based Knowledge

Unstructured Documents

Joseph Robinette Biden Jr. is an American politician who is the 46th and current president of the United States…

Structured Knowledge Graph

Joe Biden, occupation, politician

Joe Biden, position held, President of the United States

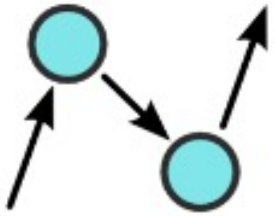
Semi-Structured Dictionary

United States of America: A country in North America, stretching from the Atlantic to the Pacific Ocean and ...

General Knowledge Sources

Domain-Specific Knowledge

Commonsense Knowledge



ConceptNet

- TransOMCS
- Commonsense Knowledge Graph(CSKG)

Name	Domain
ATOMIC	Human Interaction
ATOMIC^{20}_{20}	
ASER	Eventuality
ARC	Science

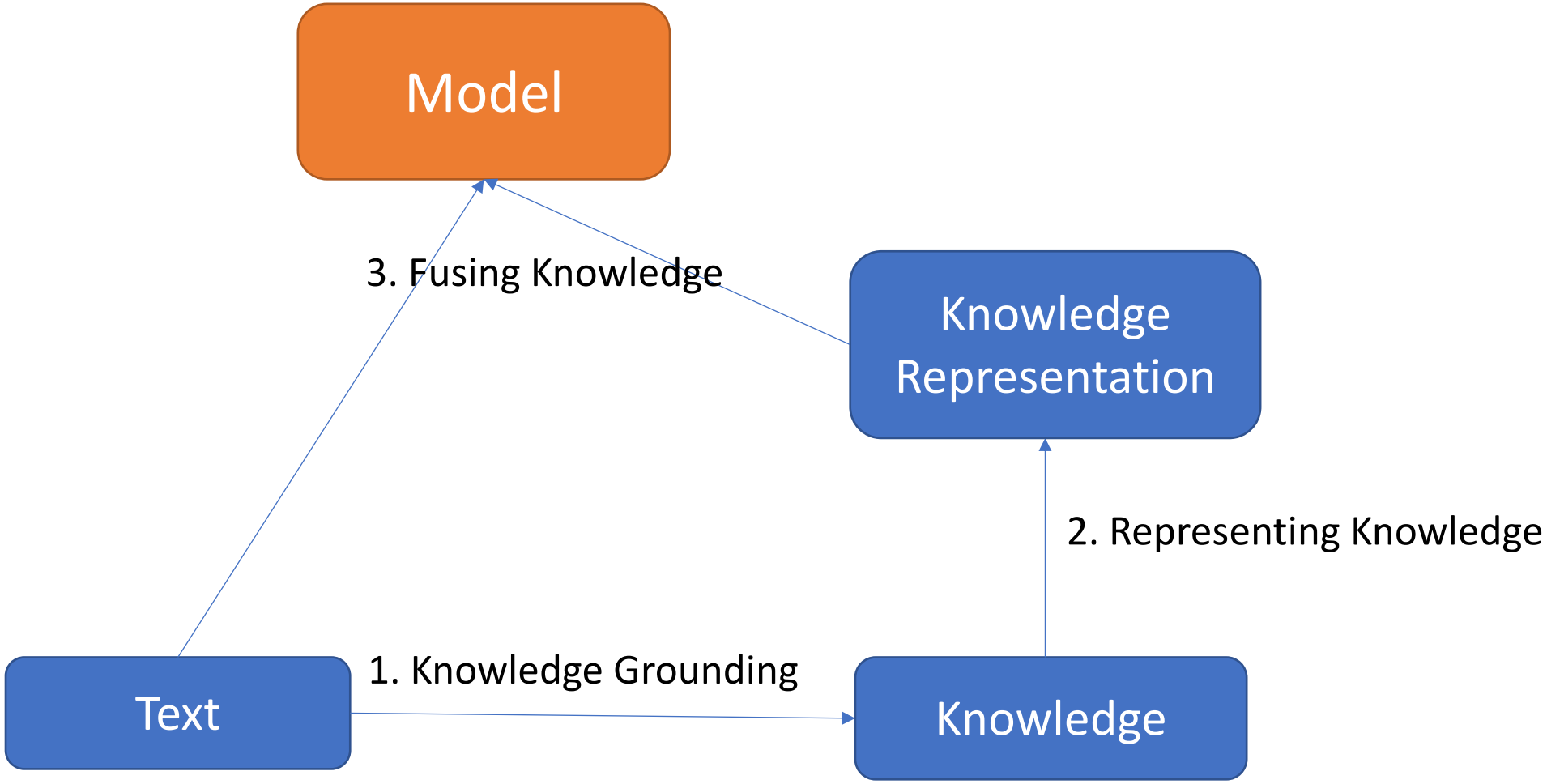
General Commonsense Knowledge

Specific Aspects of Commonsense

Commonsense Knowledge

Name	Domain	Examples
ATOMIC	Human Interaction	<personX adopts a pet, Effects, play with the pet>
ATOMIC_{20}^{20}		
ASER	Eventuality	I am hungry may result in I have lunch
ARC	Science	The earth's gravity acts on air molecules to create a force, that of the air pushing on the earth.

Knowledge Integration



Knowledge Grounding

- Entity linking: Link text to entities in Knowledge base
  - String match: JAKET (Yu et al., 2020)
  - Entity linking toolkits: ERNIE (Zhang et al., 2019)
  - Wikipedia hyperlinks for linking: Entity-as-Experts (Févry, et al., 2020)

Knowledge Grounding

- Retrieval-based methods
  - Sparse feature retrieval (e.g., tf-idf): DrQA (Chen et al., 2017)
  - Dense passage retrieval: DPR (Karpukhin et al., 2020)
  - Hybrid retrieval (sparse + dense): KFormer (Yao et al., 2022)

Knowledge Representation

- Entity Embeddings
  - Incorporate trained Entity Embeddings: ERNIE (TransE)
  - Learn Entity Embeddings: Entity-as-Experts
  - Contextual Embeddings
    - KEPLER (Wang et al., 2019): RoBERTa embedding of description <s> token
- Passage Embeddings
  - DPR: Use BERT to encode every passage as a vector
  - ColBERT: Encode every word as a contextual vector for retrieval

Knowledge Fusion

- Pre-Fusion Methods: Fuse in pretraining
  - Add knowledge-related loss in pretraining
    - KEPLER (Wang et al., 2019): Integrate entity embedding losses like TransE
    - ERICA (Qin et al., 2021): Contrastive learning on entities and relations
  - Pretrain MLM on corpus/KG
    - KELM(Agarwal et al., 2020): Linearize KG to natural sentences
    - Salient Span Masking: Mask entities and let the model predict them
    - ORQA (Xiong et al., 2020): Replace entities with negative samples and try to detect

From Yin et al., 2022

Knowledge Fusion

- Post-Fusion Methods: Fuse in finetuning
  - KEAR (Xu et al., 2021): Append knowledge as text
  - GENRE (De Cao et al., 2020): Train generation on entity linking
  - KFormer: Modify FFN matrix to incorporate knowledge

From Yin et al., 2022

Knowledge Fusion

- Hybrid methods: Fuse in both pretraining and finetuning
  - REALM: Retrieve passages to help predict entities in pretraining, and help QA in finetuning
  - ERNIE: Retrieve pretrained entity embeddings and incorporate into model
  - JAKET: Use GNN to process part of KG, and fuse representation into middle hidden layers

From Yin et al., 2022

Representative Papers

Representative Papers

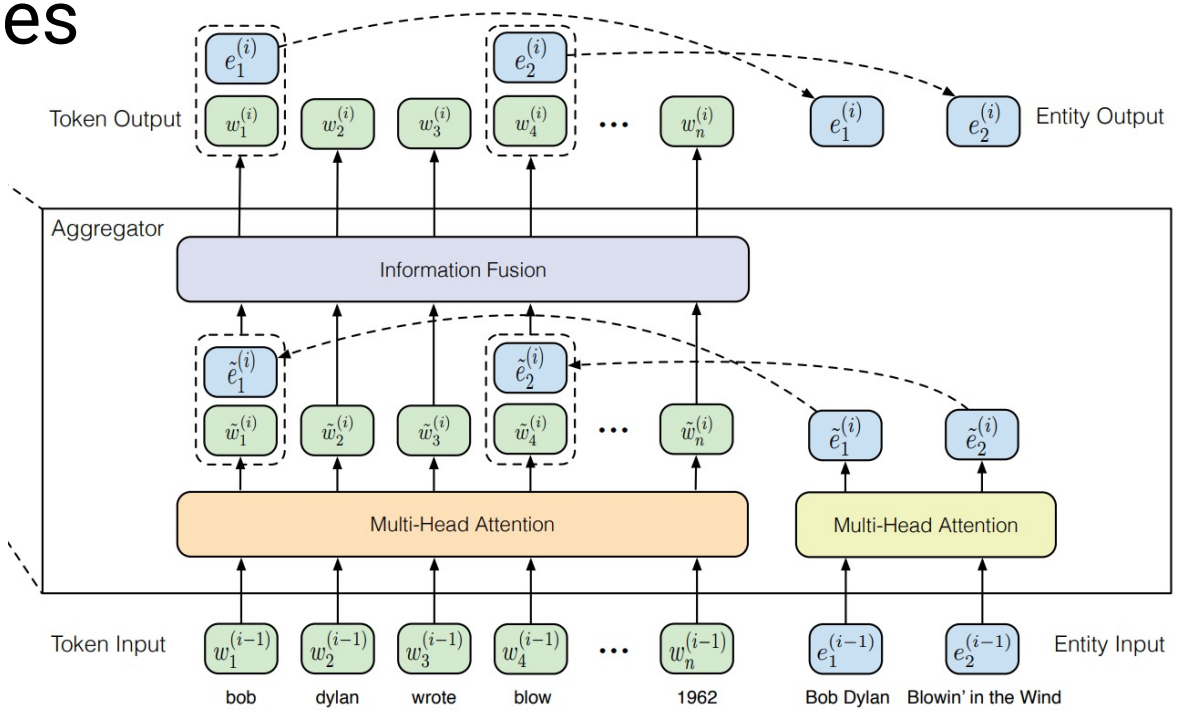
- Entity-Linking based methods
  - ERNIE
  - KEAR
  - Entity as Experts, FILM
  - K-BERT
- Retrieval based methods
  - DPR
  - REALM, RETRO, WebGPT
  - REINA
  - KFormer

Representative Papers

- **Entity-Linking based methods**
  - ERNIE
  - KEAR
  - Entity as Experts, FILM
  - K-BERT
- Retrieval based methods
  - DPR
  - REALM, RETRO, WebGPT
  - REINA
  - KFormer

ERNIE: Enhanced Language Representation with Informative Entities

- Compute TransE embeddings from Wikidata
- Use TAGME to link text to entities
- Add entity embeddings to the attention in Transformers



Bob Dylan wrote ***Blowin' in the Wind*** in 1962

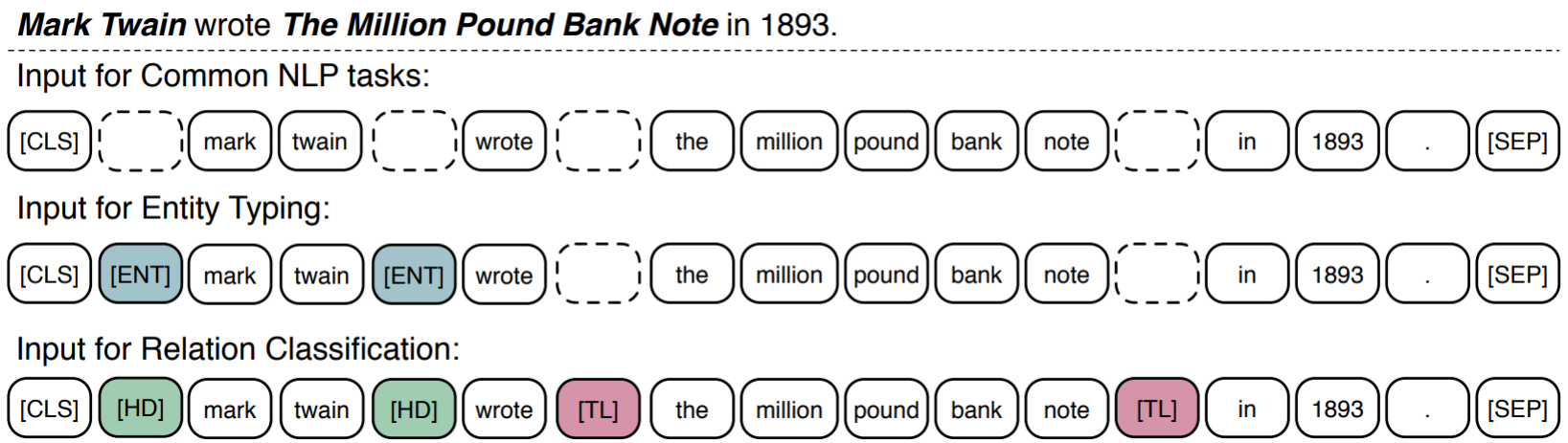
Zhang, Zhengyan, et al. "ERNIE: Enhanced Language Representation with Informative Entities." *ACL 2019.*

ERNIE: Enhanced Language Representation with Informative Entities

- Pretraining loss
  - Entity denoising: Let the model predict a masked or replaced entity
    - 5% of entities are replaced and 15% are masked

$$p(e_j|w_i) = \frac{\exp(\texttt{linear}(\boldsymbol{w}_i^o) \cdot \boldsymbol{e}_j)}{\sum_{k=1}^{m} \exp(\texttt{linear}(\boldsymbol{w}_i^o) \cdot \boldsymbol{e}_k)}$$

- Finetuning
  - Add special tokens around entities to emphasize



ERNIE: Enhanced Language Representation with Informative Entities

- Initialize from BERT and continue pretrain on Wikipedia
- Improved performance compared with BERT

Model	Acc.	Macro	Micro
NFGEC (Attentive)	54.53	74.76	71.58
NFGEC (LSTM)	55.60	75.15	71.73
BERT	52.04	75.16	71.63
ERNIE	**57.19**	**76.51**	**73.39**

Results on FIGER (Entity Typing)

Model	FewRel			TACRED		
	P	R	F1	P	R	F1
CNN	69.51	69.64	69.35	70.30	54.20	61.20
PA-LSTM	-	-	-	65.70	64.50	65.10
C-GCN	-	-	-	69.90	63.30	66.40
BERT	85.05	85.11	84.89	67.23	64.81	66.00
ERNIE	88.49	88.44	**88.32**	69.97	66.08	**67.97**

Relation Classification Results

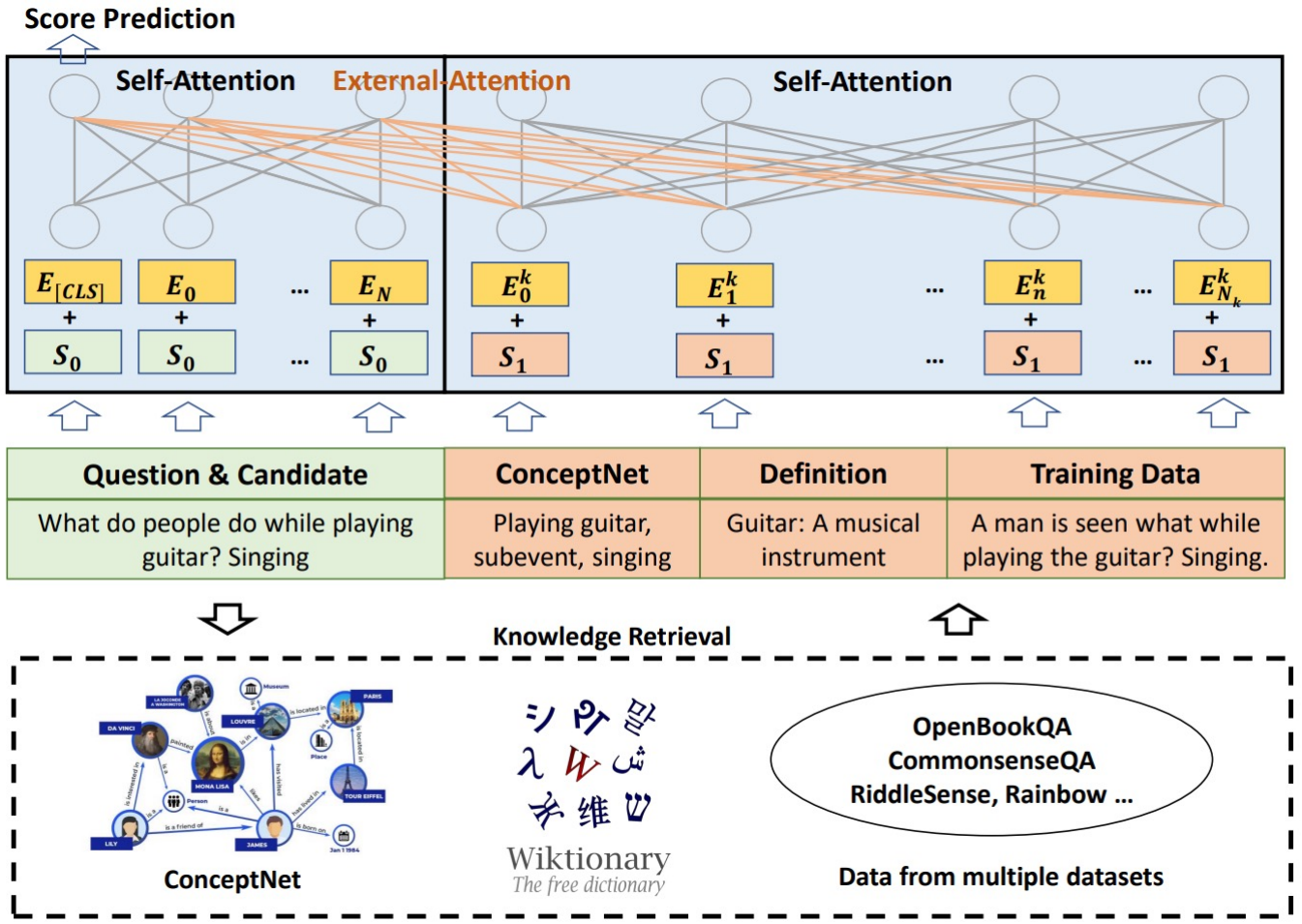
Model	MNLI-(m/mm) 392k	QQP 363k	QNLI 104k	SST-2 67k
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	-	93.5
ERNIE	84.0/83.2	71.2	91.3	93.5

Model	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k
$\text{BERT}_{\text{BASE}}$	52.1	85.8	88.9	66.4
ERNIE	52.3	83.2	88.2	68.8

GLUE (General Domain NLU) Results

KEAR: Augmenting Self-Attention with External Attention

- Knowledge for commonsense question answering
- Use string match to find entities
- Retrieve knowledge from ConceptNet, Wiktionary, and training data



Xu, Yichong, et al. "Human Parity on CommonsenseQA: Augmenting Self-Attention with External Attention." *IJCAI 2022.*

KEAR: Augmenting Self-Attention with External Attention

- Human parity result on CommonsenseQA benchmark

Encoder	CSQA	MNLI	#Para
Fine-tuned GPT-3	73.0	82.1	175B
RoBERTa-large	76.7	90.2	355M
ALBERT-xxlarge	81.2	90.6	235M
ELECTRA-base	75.0	88.8	110M
ELECTRA-large	81.3	90.9	335M
DeBERTa-xlarge	82.9	91.7	900M
DeBERTa-xxlarge	83.8	91.7	1.5B
DeBERTaV3-large	84.6	91.8	418M
T5-11B	83.5[1]	91.3	11B

Method	Dev Acc(%)
ELECTRA-large + VAT + KEAR	88.7
DeBERTa-xxlarge + KEAR	90.8
DeBERTaV3-large + KEAR	**91.2**
Ensemble (39 models w/ KEAR)	**93.4**

Dev Set Performance

Method	Single	Ensemble
BERT+OMCS	62.5	-
RoBERTa	72.1	72.5
RoBERTa+KEDGN	-	74.4
ALBERT	-	76.5
RoBERTa+MHGRN	75.4	76.5
ALBERT + HGN	77.3	80.0
T5	78.1	-
UnifiedQA	79.1	-
ALBERT+KCR	79.5	-
ALBERT + KD	80.3	80.9
ALBERT + SFR	-	81.8
DEKCOR	80.7	83.3
Human	-	88.9
KEAR (ours)	**86.1**	**89.4**

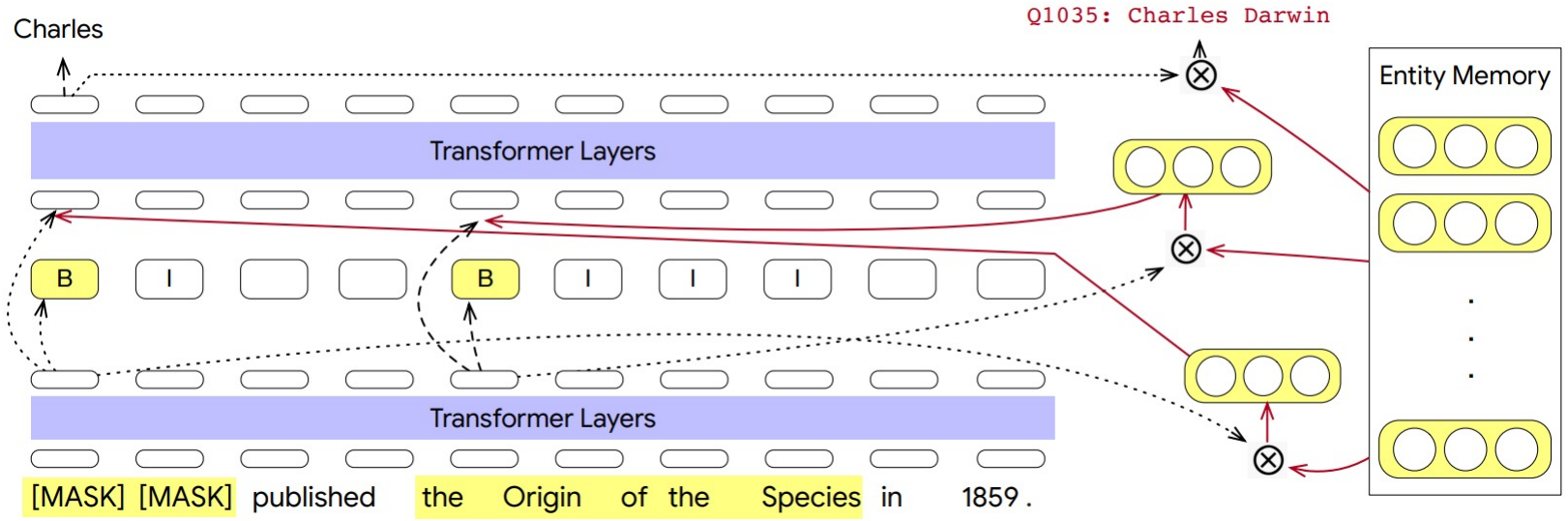
Test Set Performance

Entities as Experts: Sparse Memory Access with Entity Supervision

- Use a memory to store all (learned) entity embeddings
- For each entity mention, compute a pseudo embedding from first and last token: $h_{m_i} = \mathbf{W_f}[x^l_{s_{m_i}} || x^l_{t_{m_i}}]$
- Find k nearest neighbors of pseudo embedding and weighted sum:

$$E_{m_i} = \sum_{e_j \in \texttt{topK}(\mathcal{E}, h_{m_i}, k)} \alpha_j \cdot (\texttt{EntEmbed}(e_j))$$

$$\alpha_j = \frac{\exp(\texttt{EntEmbed}(e_j) \cdot h_{m_i})}{\sum_{e \in \texttt{topK}(\mathcal{E}, h_{m_i}, k)} \exp(\texttt{EntEmbed}(e) \cdot h_{m_i})}$$



Févry, Thibault, et al. "Entities as Experts: Sparse Memory Access with Entity Supervision." *EMNLP 2020.*

Entities as Experts: Sparse Memory Access with Entity Supervision

- Train on Wikipedia with MLM and entity linking
  - Hyperlinks in Wiki provides entity linking
  - Pseudo embedding should be close to the true embedding

$$\texttt{ELLoss} = \sum_{m_i} \alpha_i \cdot \mathbb{1}_{e_{m_i} \neq e_\varnothing}$$

$$\alpha_i = \frac{\exp(\texttt{EntEmbed}(e_{m_i}) \cdot h_{m_i})}{\sum_{e \in \mathcal{E}} \exp(\texttt{EntEmbed}(e) \cdot h_{m_i})}$$

- In prediction, predict an entity by generating a pseudo embedding from last layer, and find closest in memory

Entities as Experts: Sparse Memory Access with Entity Supervision

- Results on LAMA probe and Open-domain QA

Model	Concept Net	RE	SQuAD	T -REx	Avg.
BERT-base	15.6	9.8	14.1	31.1	17.7
BERT-large	**19.2**	**10.5**	17.4	32.3	19.9
MM-base	10.4	9.2	16.0	29.7	16.3
MM-large	12.4	6.5	**24.4**	31.4	18.7
EAE-unsup	10.6	8.4	23.1	30.0	18.0
No EAE	10.3	9.2	18.5	31.8	17.4
EAE	10.7	9.4	22.4	**37.4**	**20.0**

LAMA probe

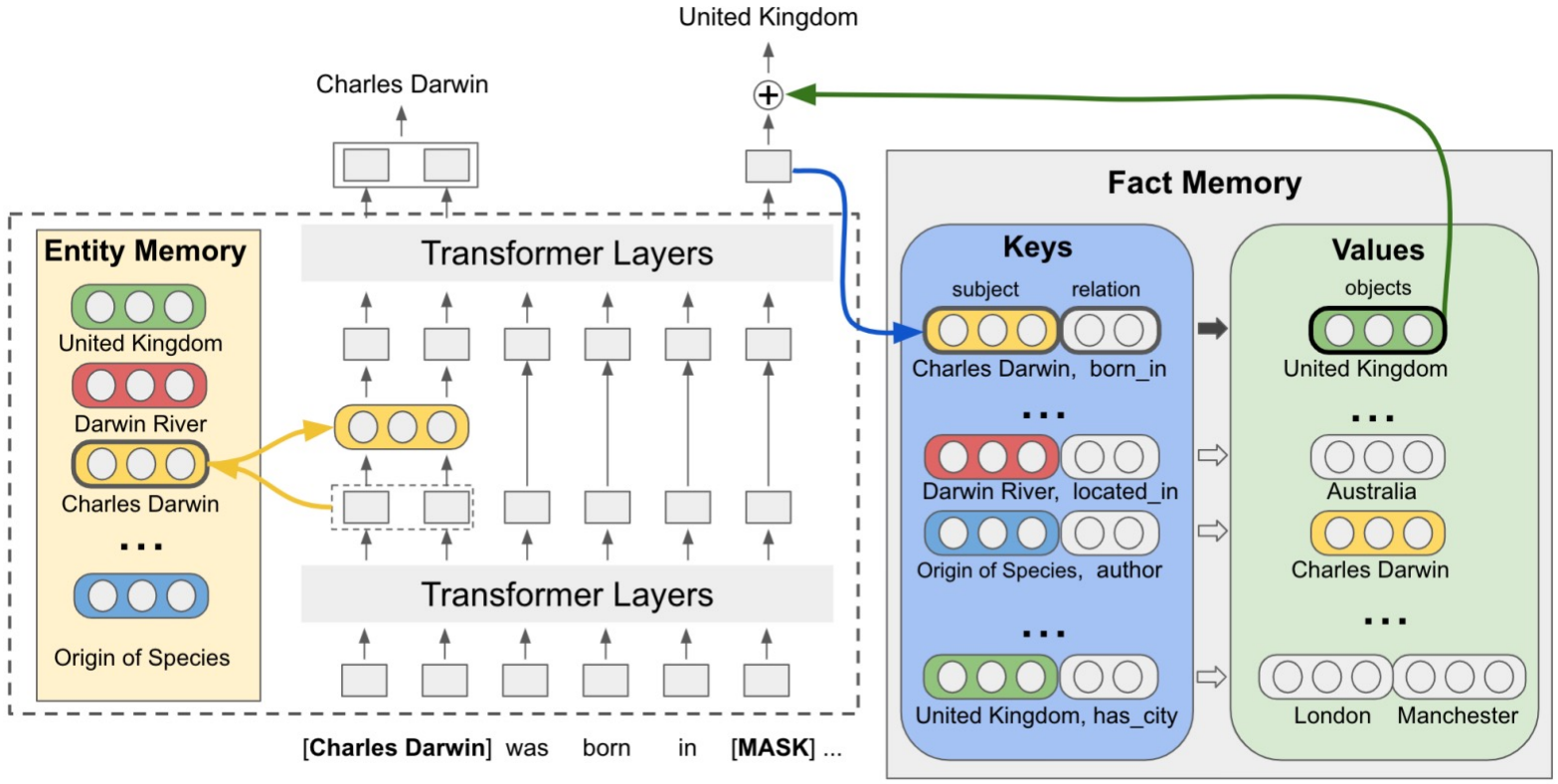
	# Params	TQA Dev	TQA Wiki Test	Web Q
Open-Book: Span Selection - Oracle 100%				
BM25+BERT	110m	47.1	-	17.7
ORQA	330m	45.0	-	36.4
GR	110m	55.4	-	31.6
Closed-Book: Nearest Neighbor				
ORACLE	-	63.6	-	-
TFIDF	-	23.5	-	-
BERT-base	110m	28.9	-	-
Closed-Book: Generation - Oracle 100%				
T5-Base	220m	-	29.1	29.1
T5-Large	770m	-	35.9	32.2
T5-3B	3B	-	43.4	34.4
T5-11B	11B	42.3	50.1	37.4
T5-11B+SSM	11B	53.3	61.6	43.5
Closed-Book: Entity Prediction				
ORACLE	-	85.0	-	91.0
RELIC	3B	35.7	-	-
No EAE	366m	37.7	-	33.4
EAE	367m	43.2	53.4	39.0
EAE, emb 512	623m	45.7	-	38.7

Open-domain QA

FILM: Adaptable and Interpretable Neural Memory Over Symbolic Knowledge

- Extends EaE by further adding another facts memory with relations
- In addition to entity embedding, build a query to retrieve the (head, rel) pair in facts memory

$$\mathbf{v}_{m_{\text{ans}}} = \mathbf{W}_f^T \left[\mathbf{h}_{s_{\text{ans}}}^{(T)}; \mathbf{h}_{t_{\text{ans}}}^{(T)}\right]$$



Verga, Pat, et al. "Adaptable and interpretable neural memoryover symbolic knowledge." *ACL 2021.*

FILM: Adaptable and Interpretable Neural Memory Over Symbolic Knowledge

- Improves over EaE on LAMA and ODQA

Model	P@1
K-Adapter †	29.1
BERT-Large †	33.9
BERT-KNN ‡	38.7
EaE	38.6
FILM	**44.2**

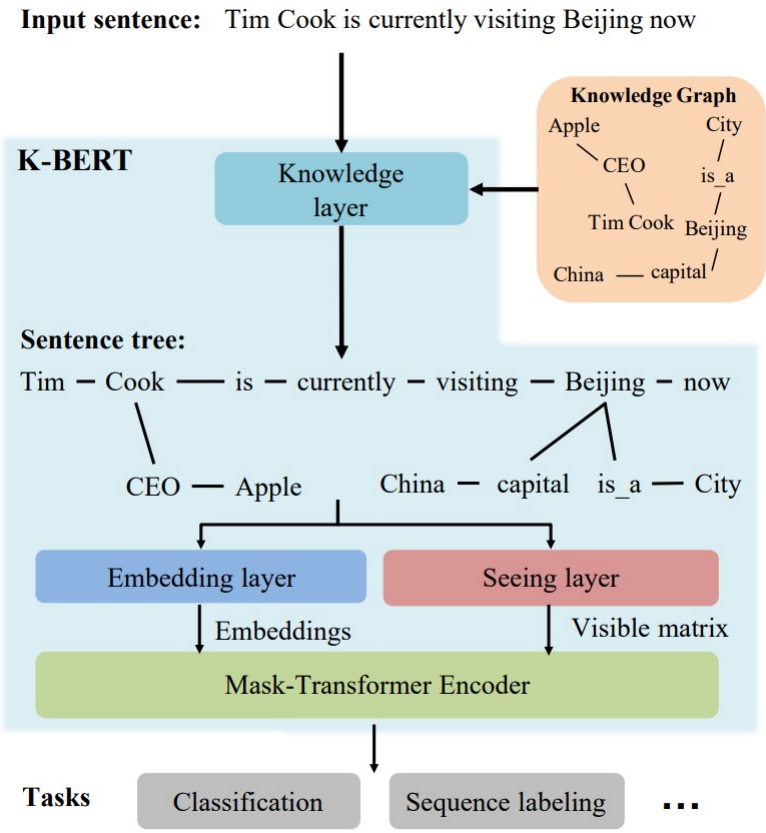
LAMA probe

	Model	WebQuestionsSP				TriviaQA			
		Full Dataset		Wikidata Answer		Full Dataset		Wikidata Answer	
		Total	No Overlap	Total	No Overlap	Total	No Overlap	Total	No Overlap
Closed-book	**FILM**	**54.7**	**36.4**	**78.1**	**72.2**	**29.1**	**15.6**	**37.3**	**28.4**
	EaE	47.4	25.1	62.4	42.9	19.0	9.1	24.4	17.1
	T5-11B	49.7	31.8	61.0	48.5	–	–	–	–
	BART-Large	30.4	5.6	36.7	8.3	26.7	0.8	30.6	1.0
Open-Book	RAG	**50.1**	**30.7**	**62.5**	**45.1**	56.8	29.2	64.9	45.2
	DPR	48.6	34.1	56.9	45.1	57.9	31.6	66.3	48.8
	FID	–	–	–	–	**67.6**	**42.8**	**76.5**	**64.5**
	EmQL†	75.5	-	74.6	-	-	-	-	-

Open-domain QA

K-BERT: Enabling Language Representation with Knowledge Graph

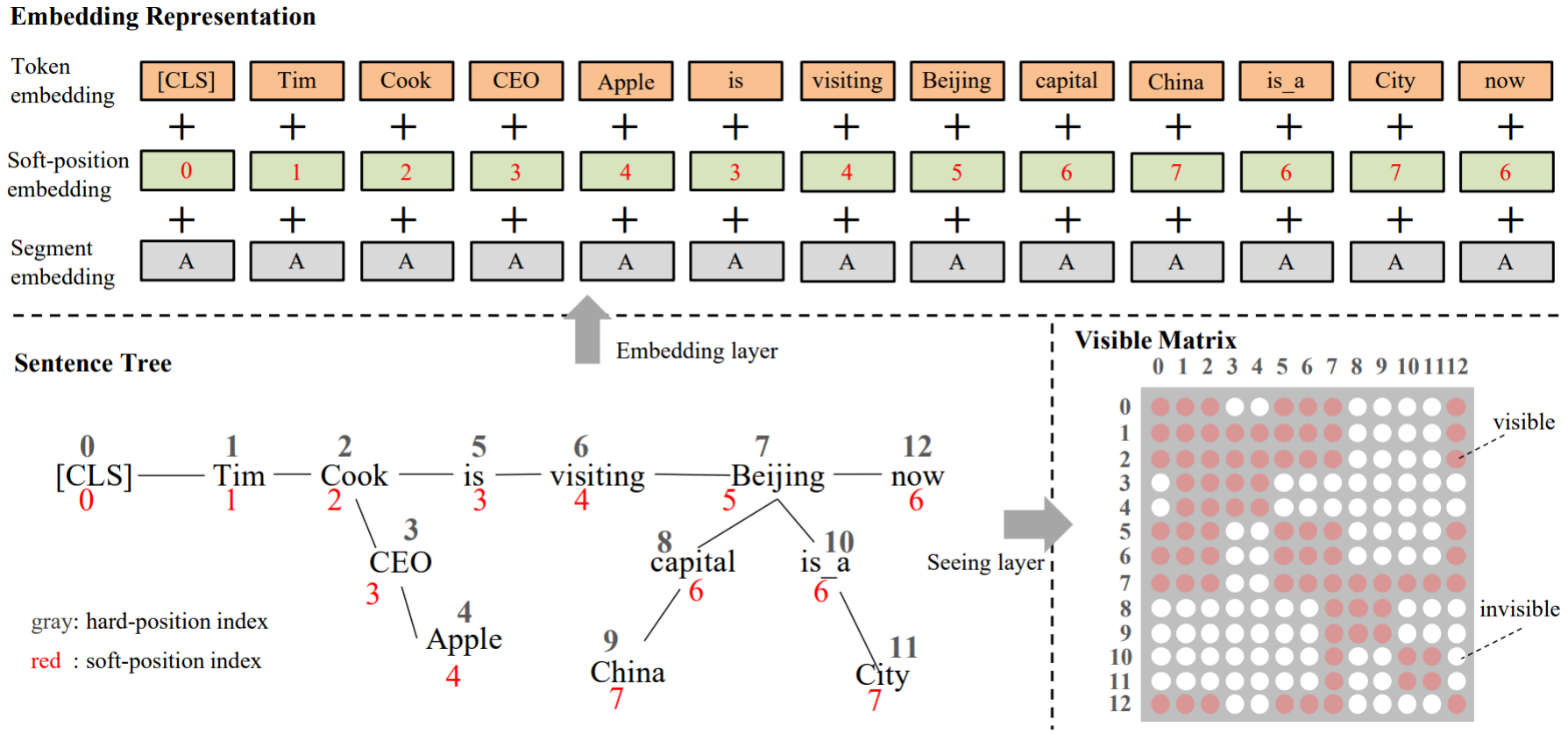
- Link entities by string matching
- Retrieve a subgraph of KG and form a sentence tree
- Manipulate attention matrix such that different branches cannot see each other



Liu, Weijie, et al. "K-bert: Enabling language representation with knowledge graph." *AAAI 2020.*

K-BERT: Enabling Language Representation with Knowledge Graph

- Also change the position embeddings so that the tree is “flattened” in every branch



K-BERT: Enabling Language Representation with Knowledge Graph

Table 1: Results of various models on sentence classification tasks on open-domain tasks (*Acc. %*)

Models\Datasets	Book_review		Chnsenticorp		Shopping		Weibo		XNLI		LCQMC	
	Dev	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*
Pre-trainied on WikiZh by Google.												
Google BERT	88.3	**87.5**	93.3	94.3	96.7	96.3	98.2	98.3	76.0	75.4	88.4	86.2
K-BERT (HowNet)	88.6	87.2	**94.6**	**95.6**	**97.1**	**97.0**	98.3	98.3	**76.8**	**76.1**	**88.9**	86.9
K-BERT (CN-DBpedia)	**88.6**	87.3	93.9	95.3	96.6	96.5	98.3	98.3	76.5	76.0	88.6	**87.0**
Pre-trained on WikiZh and WebtextZh by us.												
Our BERT	**88.6**	87.9	94.8	95.7	96.9	**97.1**	98.2	98.2	77.0	76.3	89.0	86.7
K-BERT (HowNet)	88.5	87.4	**95.4**	95.6	96.9	96.9	98.3	**98.4**	**77.2**	**77.0**	**89.2**	**87.1**
K-BERT (CN-DBpedia)	88.8	87.9	95.0	**95.8**	**97.1**	97.0	98.3	98.3	76.2	75.9	89.0	86.9

Results in general open-domain NLU

Results in specific domains (Finance, Law, Medicine)

Table 3: Results of various models on specific-domain tasks (*%*).

Models\Datasets	Finance_Q&A			Law_Q&A			Finance_NER			Medicine_NER		
	P.	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*
Pre-trained on WikiZh by Google.												
Google BERT	81.9	86.0	83.9	83.1	90.1	86.4	84.8	87.4	86.1	91.9	93.1	92.5
K-BERT (HowNet)	83.3	84.4	83.9	83.7	91.2	87.3	86.3	89.0	**87.6**	93.2	93.3	93.3
K-BERT (CN-DBpedia)	81.5	88.6	**84.9**	82.1	93.8	**87.5**	86.1	88.7	87.4	93.9	93.8	93.8
K-BERT (MedicalKG)	-	-	-	-	-	-	-	-	-	94.0	94.4	**94.2**
Pre-trained on WikiZh and WebtextZh by us.												
Our BERT	82.1	86.5	84.2	83.2	91.7	87.2	84.9	87.4	86.1	91.8	93.5	92.7
K-BERT (HowNet)	82.8	85.8	84.3	83.0	92.4	87.5	86.3	88.5	87.3	93.5	93.8	93.7
K-BERT (CN-DBpedia)	81.9	87.1	**84.4**	83.1	92.6	**87.6**	86.3	88.6	**87.4**	93.9	94.3	94.1
K-BERT (MedicalKG)	-	-	-	-	-	-	-	-	-	94.1	94.3	**94.2**

Representative Papers

- Entity-Linking based methods
  - ERNIE
  - KEAR
  - Entity as Experts, FILM
  - K-BERT
- **Sf usjf wbnhcbt f e!n f ui pet !**
  - DPR
  - REALM, RETRO, WebGPT
  - REINA
  - KFormer

DPR: Dense Passage Retrieval for Open-Domain Question Answering

- Given a question, we want to retrieve relevant passages to answer the question
- Use a BERT model to encode both the question and passages
- Train the BERT to get similar representation for Q and positive P:

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}}.$$

- For a new question, use fast nearest neighbor search to find closest passages

Karpukhin, Vladimir, et al. "Dense Passage Retrieval for Open-Domain Question Answering." *EMNLP 2020.*

DPR: Dense Passage Retrieval for Open-Domain Question Answering

Training	Retriever	Top-20					Top-100				
		NQ	TriviaQA	WQ	TREC	SQuAD	NQ	TriviaQA	WQ	TREC	SQuAD
None	BM25	59.1	66.9	55.0	70.9	68.8	73.7	76.7	71.1	84.1	80.0
Single	DPR	78.4	79.4	73.2	79.8	63.2	85.4	**85.0**	81.4	89.1	77.2
	BM25 + DPR	76.6	79.8	71.0	85.2	**71.5**	83.8	84.5	80.5	92.7	**81.3**
Multi	DPR	**79.4**	78.8	**75.0**	**89.1**	51.6	**86.0**	84.7	**82.9**	93.9	67.6
	BM25 + DPR	78.0	**79.9**	74.7	88.5	66.2	83.9	84.4	82.3	**94.1**	78.6

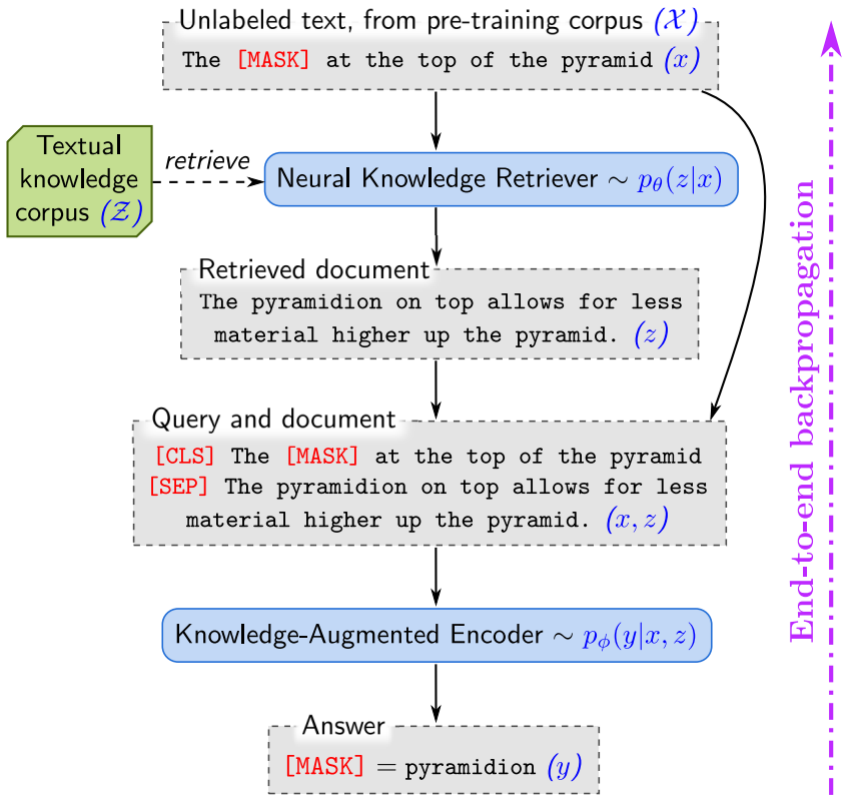
Retrieval Accuracy

Training	Model	NQ	TriviaQA	WQ	TREC	SQuAD
Single	BM25+BERT (Lee et al., 2019)	26.5	47.1	17.7	21.3	33.2
Single	ORQA (Lee et al., 2019)	33.3	45.0	36.4	30.1	20.2
Single	HardEM (Min et al., 2019a)	28.1	50.9	-	-	-
Single	GraphRetriever (Min et al., 2019b)	34.5	56.0	36.4	-	-
Single	PathRetriever (Asai et al., 2020)	32.6	-	-	-	**56.5**
Single	$\text{REALM}_{\text{Wiki}}$ (Guu et al., 2020)	39.2	-	40.2	46.8	-
Single	$\text{REALM}_{\text{News}}$ (Guu et al., 2020)	40.4	-	40.7	42.9	-
Single	BM25	32.6	52.4	29.9	24.9	38.1
	DPR	**41.5**	56.8	34.6	25.9	29.8
	BM25+DPR	39.0	57.0	35.2	28.0	36.7
Multi	DPR	**41.5**	56.8	**42.4**	49.4	24.1
	BM25+DPR	38.8	**57.9**	41.1	**50.6**	35.8

End-to-End Accuracy

REALM: Retrieval-Augmented Language Model Pre-Training

- Apply dense retrieval for LM pretraining
- Retrieve relevant passages to help MLM in pretraining, and downstream tasks in finetuning
- Jointly train the retriever and language model



Guu, Kelvin, et al. "Realm: Retrieval-augmented language model pre-training." *ICML 2020.*

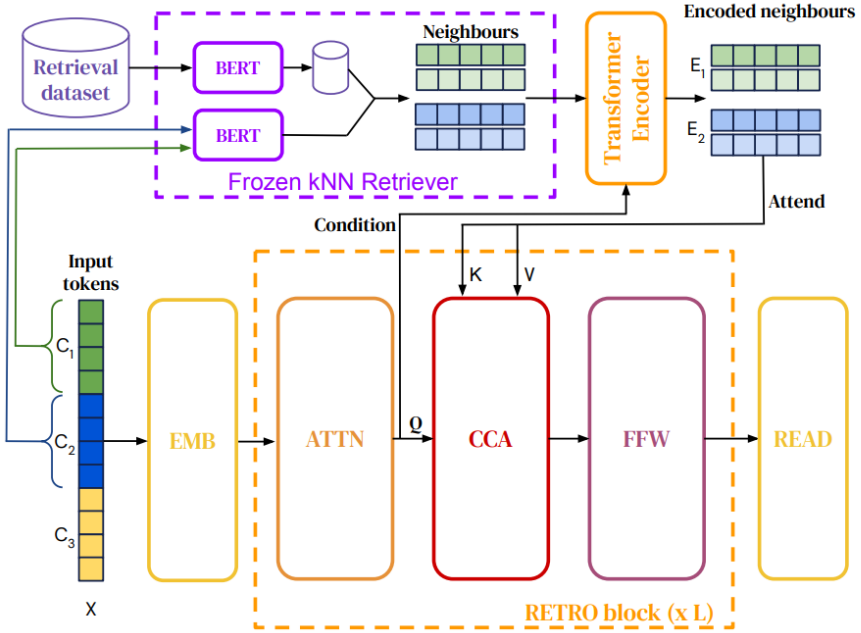
REALM: Retrieval-Augmented Language Model Pre-Training

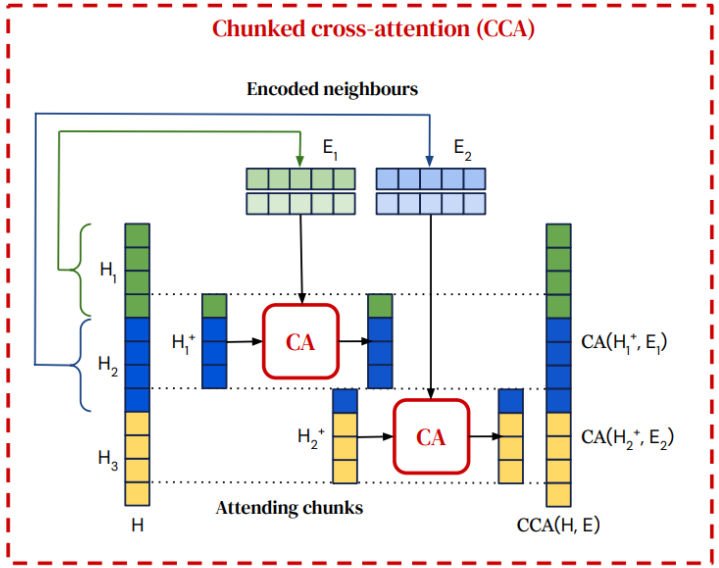
Name	Architectures	Pre-training	NQ (79k/4k)	WQ (3k/2k)	CT (1k /1k)	# params
BERT-Baseline (Lee et al., 2019)	Sparse Retr.+Transformer	BERT	26.5	17.7	21.3	110m
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1	-	223m
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2	-	738m
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4	-	11318m
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7	25.7	34m
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-	110m
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-	110m
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-	110m
ORQA (Lee et al., 2019)	Dense Retr.+Transformer	ICT+BERT	33.3	36.4	30.1	330m
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2	**46.8**	330m
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**	42.9	330m

RETRO: Improving language models by retrieving from trillions of tokens

- Similar to REALM, but retrieve for text chunks (64 tokens)
- A fixed BERT encoder gives representations of chunks
- Every chunk attends to retrieved chunks of itself and neighboring chunks

Model	Test Accuracy
REALM (Guu et al., 2020)	40.4
DPR (Karpukhin et al., 2020)	41.5
RAG (Lewis et al., 2020)	44.5
EMDR^2 (Sachan et al., 2021)	52.5
FID (Izacard and Grave, 2021)	51.4
FID + Distill. (Izacard et al., 2020)	**54.7**
Baseline 7B (closed book)	30.4
RETRO 7.5B (DPR retrieval)	45.5

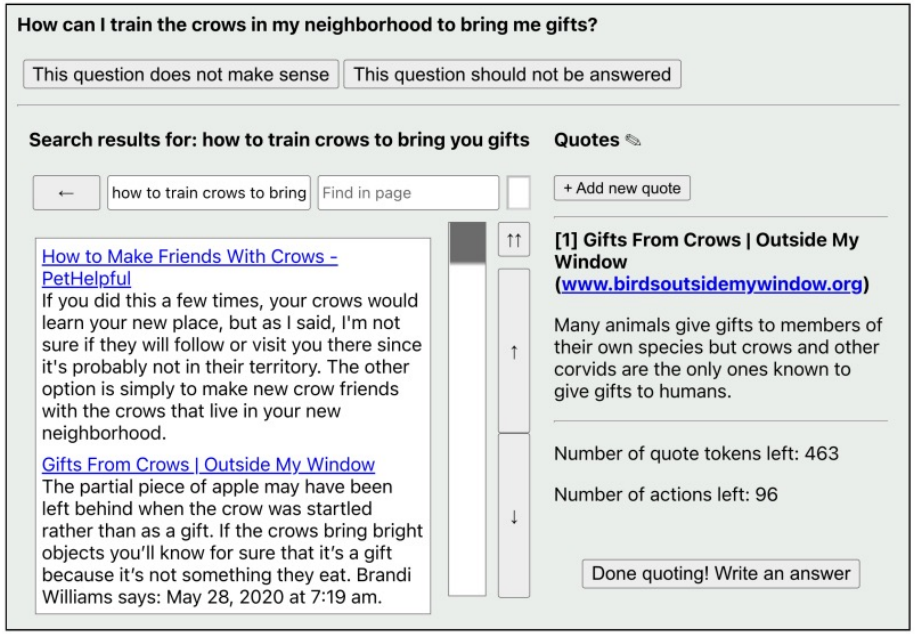




Borgeaud, Sebastian, et al. "Improving language models by retrieving from trillions of tokens." *arXiv* 2021.

WebGPT: Browser-assisted question-answering with human feedback

- Teach a GPT-3 model to use a web browser
- The model issues commands to browser to retrieve knowledge



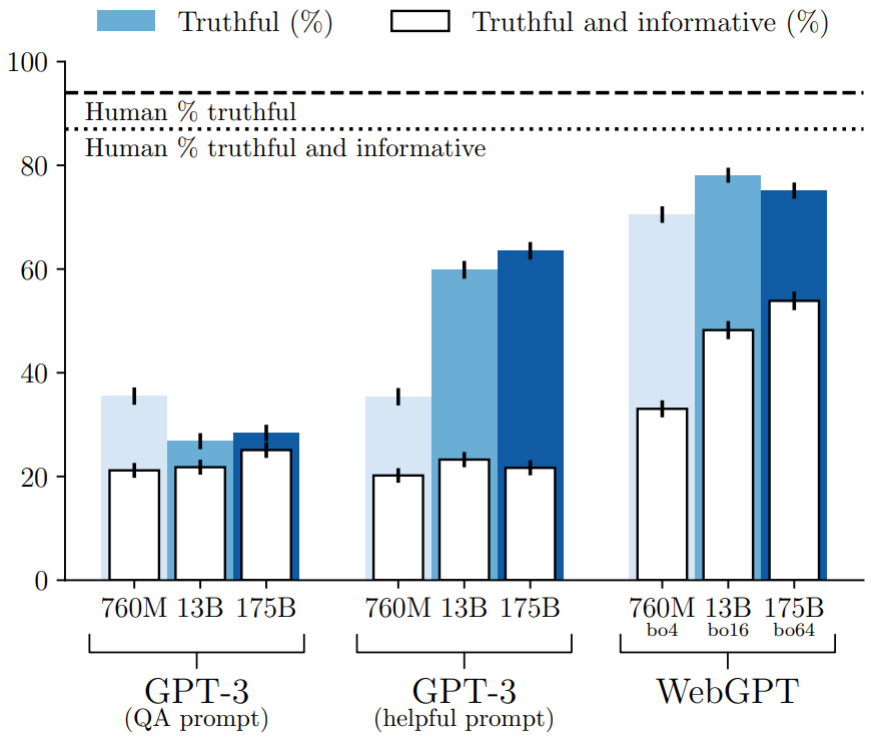
(a) Screenshot from the demonstration interface.

```
♦Question
How can I train the crows in my neighborhood to bring me gifts?

♦Quotes
From Gifts From Crows | Outside My Window (www.birdsoutsidemywindow.org)
> Many animals give gifts to members of their own species but crows and
other corvids are the only ones known to give gifts to humans.

♦Past actions
Search how to train crows to bring you gifts
Click Gifts From Crows | Outside My Window www.birdsoutsidemywindow.org
Quote
Back

♦Title
Search results for: how to train crows to bring you gifts

♦Scrollbar: 0 - 11
♦Text
【0†How to Make Friends With Crows - PetHelpful†pethelpful.com】
If you did this a few times, your crows would learn your new place, but
as I said, I'm not sure if they will follow or visit you there since it's
probably not in their territory. The other option is simply to make new
crow friends with the crows that live in your new neighborhood.

【1†Gifts From Crows | Outside My Window†www.birdsoutsidemywindow.org】
The partial piece of apple may have been left behind when the crow was
startled rather than as a gift. If the crows bring bright objects you'll
know for sure that it's a gift because it's not something they eat.
Brandi Williams says: May 28, 2020 at 7:19 am.

♦Actions left: 96
♦Next action
```

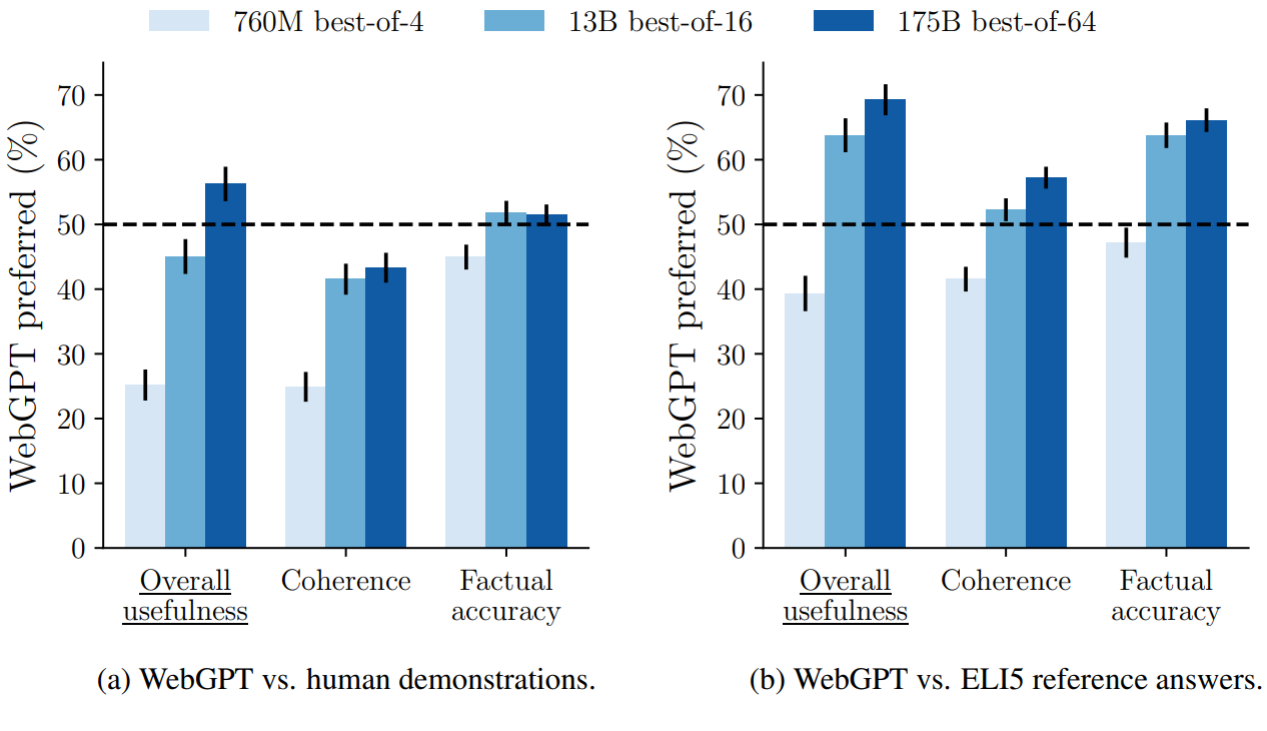
(b) Corresponding text given to the model.

Nakano, Reiichiro, et al. "WebGPT: Browser-assisted question-answering with human feedback." *arXiv 2021.*

WebGPT: Browser-assisted question-answering with human feedback



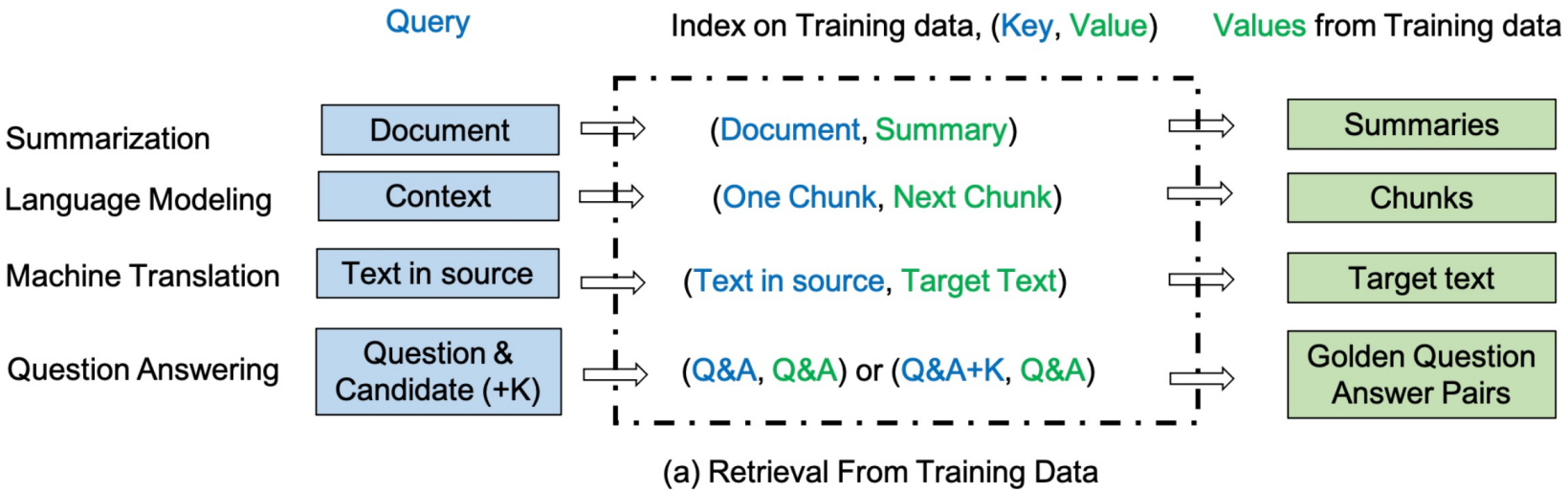
Results on TruthfulQA



Results on ELI5

REINA: Training Data is More Valuable than You Think

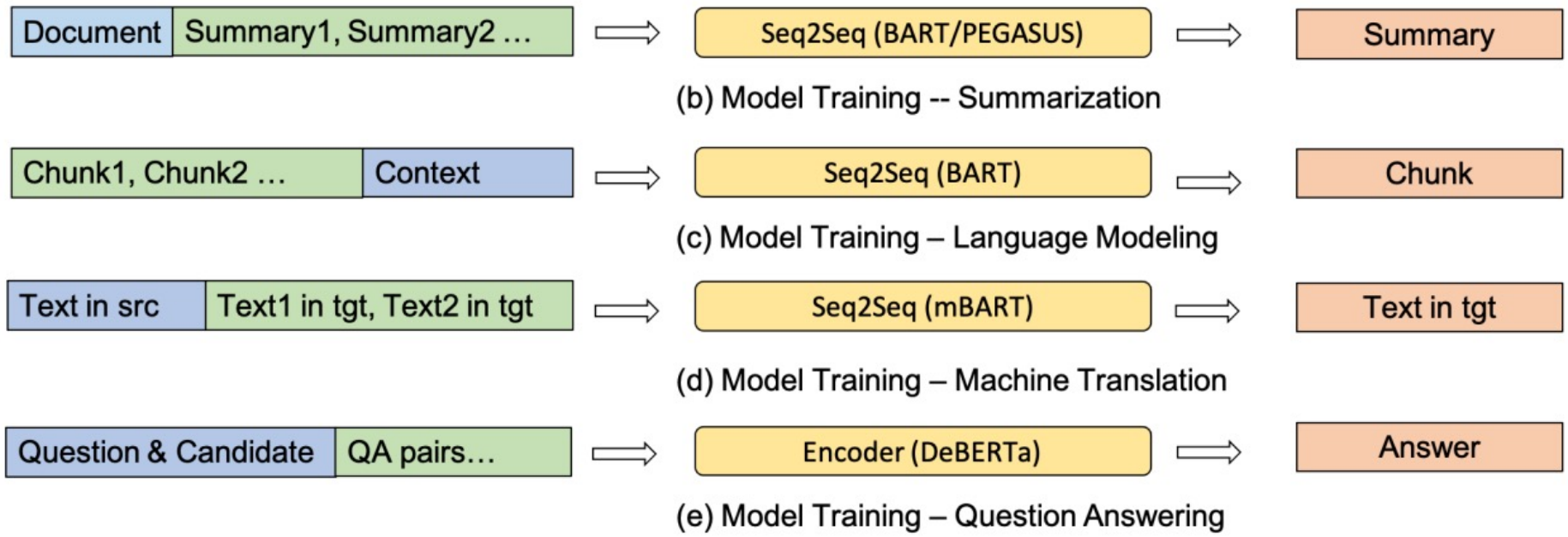
- Retrieve similar samples from training data
- Build training data as a (input, output) key-value index
- Given new input, retrieve similar inputs and use the output as knowledge



(a) Retrieval From Training Data

REINA: Training Data is More Valuable than You Think

- Append the knowledge (output of similar inputs) to the input



Wang, Shuohang, et al. "Training Data is More Valuable than You Think: A Simple and Effective Method by Retrieving from Training Data." *ACL 2022.*

REINA: Training Data is More Valuable than You Think

	BigPatent			XSum			WikiHow			Multi-News			NEWSROOM		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Earlier SOTA	37.5	10.6	22.7	45.1	22.2	37.2	28.5	9.2	26.5	43.4	14.8	17.4	39.9	28.3	36.8
PEGASUS	53.6	33.2	42.3	47.2	24.6	39.3	43.1	19.7	34.8	**47.5**	**18.7**	**24.9**	**45.2**	**33.5**	**41.3**
PEGASUS	38.4	13.5	26.3	46.6	23.9	38.6	35.9	15.3	30.3	43.1	15.4	22.6	41.7	30.7	37.8
REINA (PG)	44.6	21.5	33.0	**48.2**	**26.0**	**40.2**	36.8	16.7	31.0	45.0	17.1	23.8	41.4	30.5	37.5
BART-base	44.2	16.9	28.4	41.0	18.2	33.3	43.3	18.1	33.9	44.8	16.4	23.3	41.3	29.1	37.5
REINA (B)	59.5	42.6	50.6	43.2	21.0	35.5	44.2	19.4	34.9	45.1	16.9	23.6	41.2	29.0	37.5
BART-large	44.9	17.5	28.9	44.7	21.6	36.5	43.4	19.0	34.9	44.1	16.6	22.7	41.6	29.4	38.0
REINA (L)	**60.7**	**43.3**	**51.3**	46.5	24.1	38.6	**44.2**	**20.4**	**35.8**	46.9	17.7	24.0	42.5	30.2	38.7

(1) Summarization

	CSQA	aNLI	PIQA
Dev Set results			
DeBERTa	84.0	88.8	85.6
REINA (w/o K)	**88.8**	88.6	85.5
REINA (w/ K)	86.8	**89.6**	**86.9**
Test Set results			
CALM	71.8	82.4	76.9
UNICORN	79.3	87.3	90.1
DEKCOR	83.3	-	-
DeBERTa	-	86.8	85.1
REINA	**84.6**	**88.0**	**85.4**

(2) Question Answering

	WikiText103	WikiText2
Transformer-XL	18.30	-
kNN-LM	15.79	-
GPT-2	17.48	18.34
BART-Base	15.88	20.41
REINA (B)	14.76	20.78
BART-Large	12.10	**15.11**
REINA (L)	**11.36**	15.62

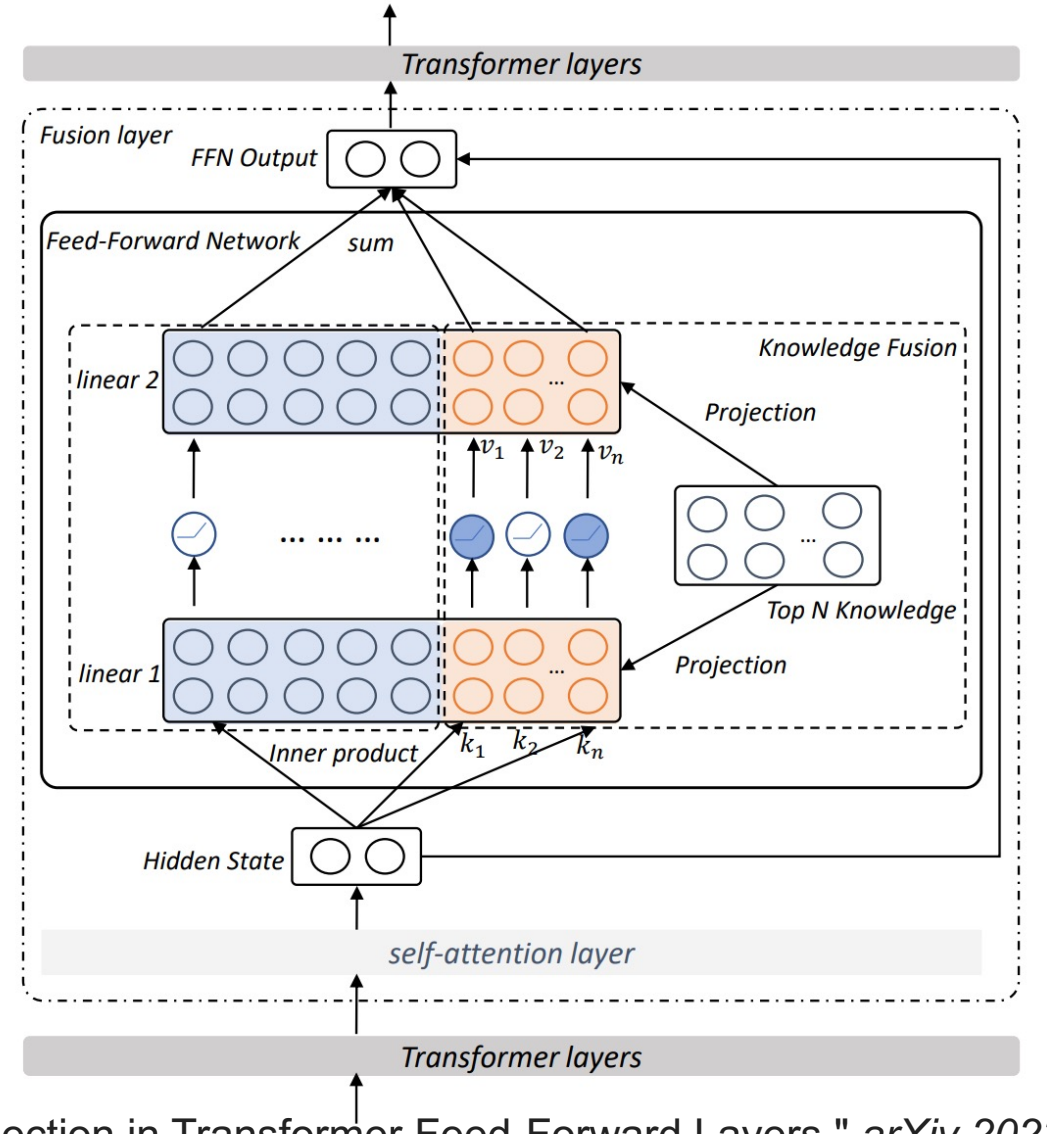
(4) Language Modeling

	WMT16			
	en2tr	tr2en	en2de	de2en
XLM	-	-	26.4	34.3
mBART	18.4	23.1	32.6	37.0
REINA	**18.8**	**23.6**	**32.9**	37.0

(3) Machine Translation

Kformer: Knowledge Injection in Transformer Feed-Forward Layers

- For any sentence, retrieve top facts using ElasticSearch
- Embed knowledge sentence as the average token embedding
- Append knowledge embedding into the FFN matrix in finetuning
- Improvements on SocialIQA, MedQA and LAMA probe



Yao, Yunzhi, et al. "Kformer: Knowledge Injection in Transformer Feed-Forward Layers." *arXiv 2022.*

Thank you!