WSDM 2023 Tutorial

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

Meng Jiang and Wenhao Yu
University of Notre Dame

Related Materials

Tvswf z;!

A survey of knowledge-enhanced text generation

W Yu, C Zhu, Z Li, Z Hu, Q Wang, H Ji… ACM Computing Surveys (CSUR), 2022 - dl.acm.org

The goal of text generation is to make machines express in human language. It is one of most important yet challenging tasks in natural language processing (NLP). Since 2014,

☆ Save 99 Cite Cited by 38 Related articles All 6 versions »

- A survey of knowledge-enhanced text generation. In ACM Computing Survey
- DOI: https://dl.acm.org/doi/10.1145/3512467

Sf bejoh!Mjt uj!

Knowledge-enriched Text Generation Survey, Tutorial and Reading

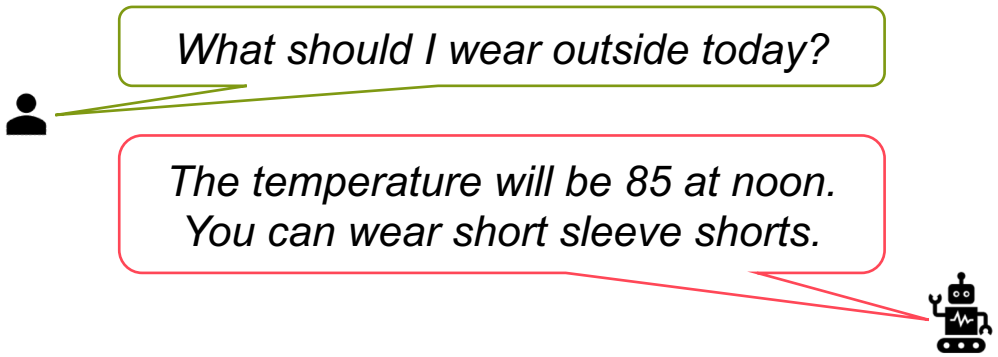
Status building PRs Welcome Unpin Unwatch 29 Fork 42 Starred 427

This repository contains a list of tutorials, papers, codes, datasets, leaderboards on the topic of **Knowledge-enhanced text generation**. If you found any error, please don't hesitate to open an issue or pull request.

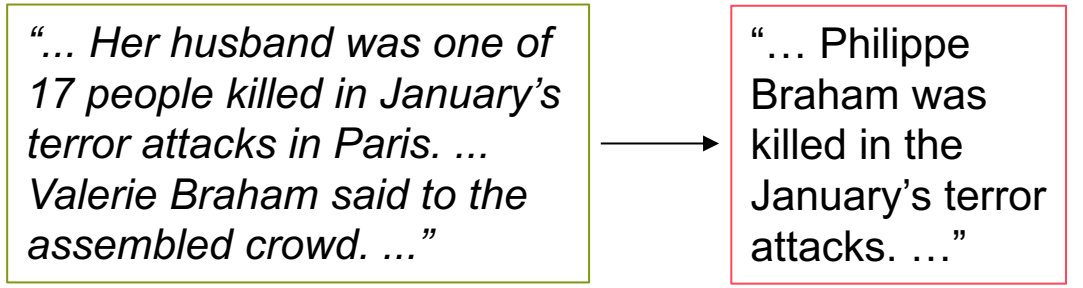
- Github: https://github.com/wyu97/KENLG-Reading

Text generation is everywhere in our life!

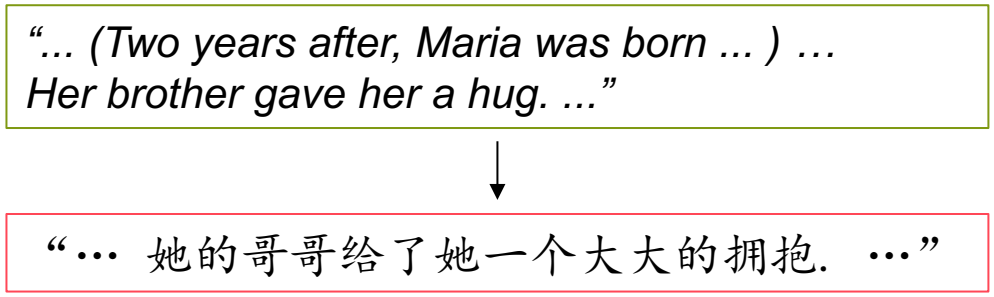
- Leveraging machine intelligence to make many things EASY and FAST



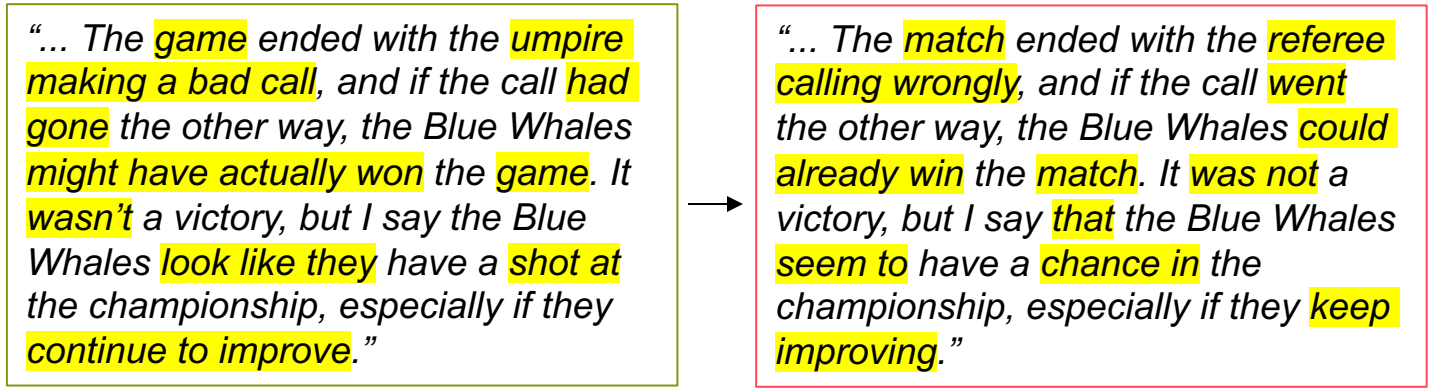
Dialog systems



Summarization



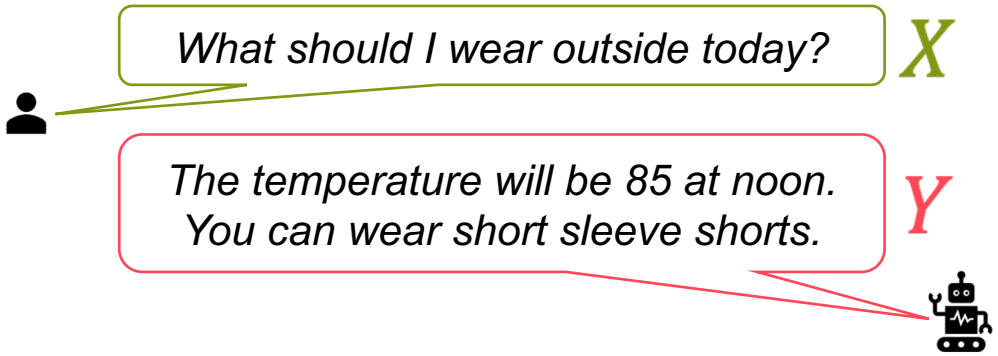
Machine Translation



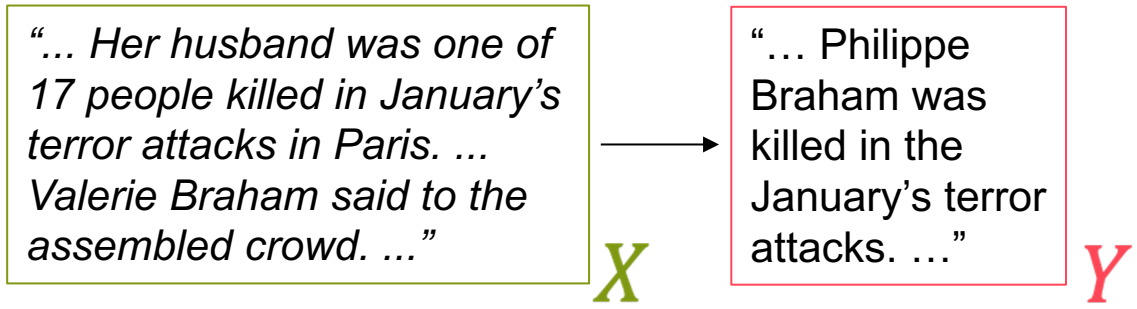
Paraphrasing

Text generation is everywhere in our life!

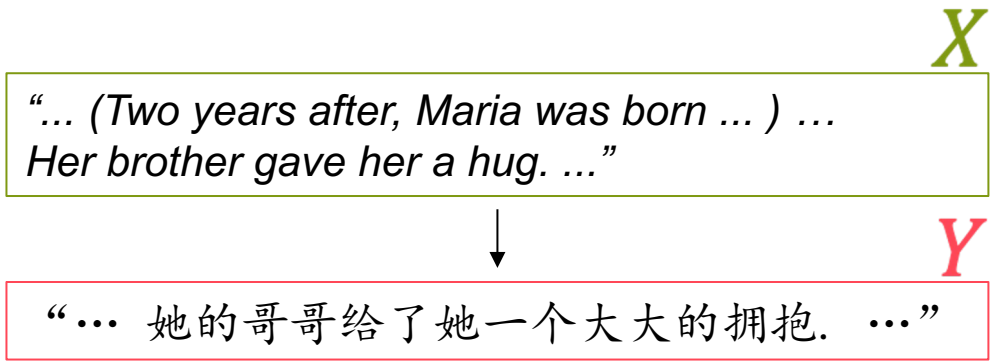
- $P(Y|X) = P(y_1, \cdots, y_m | x_1, \cdots, x_n) = \prod_{t=1}^{m} p(y_t | X, y_1, \cdots, y_{t-1})$, when Y is text and X is text.



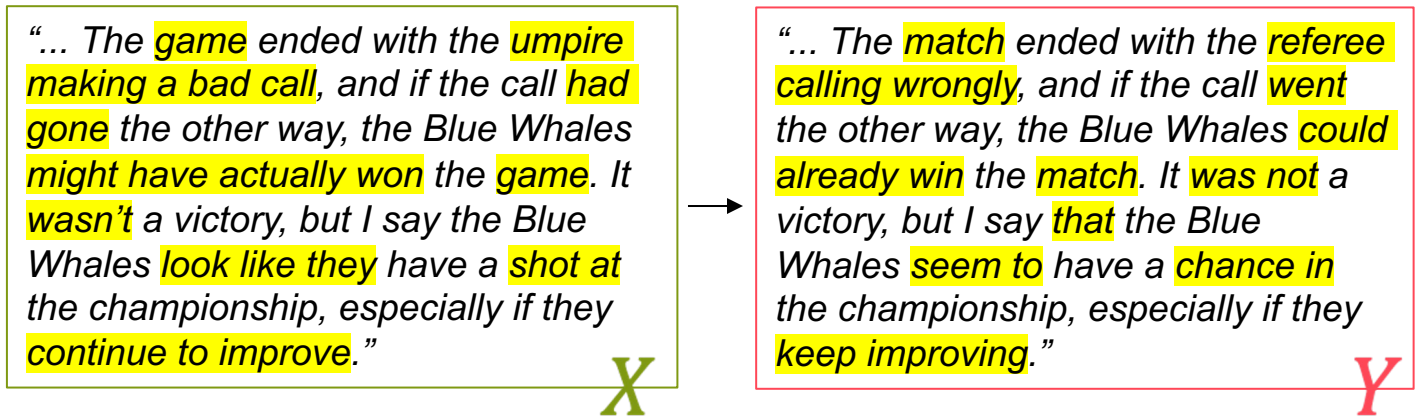
Dialog systems



Summarization

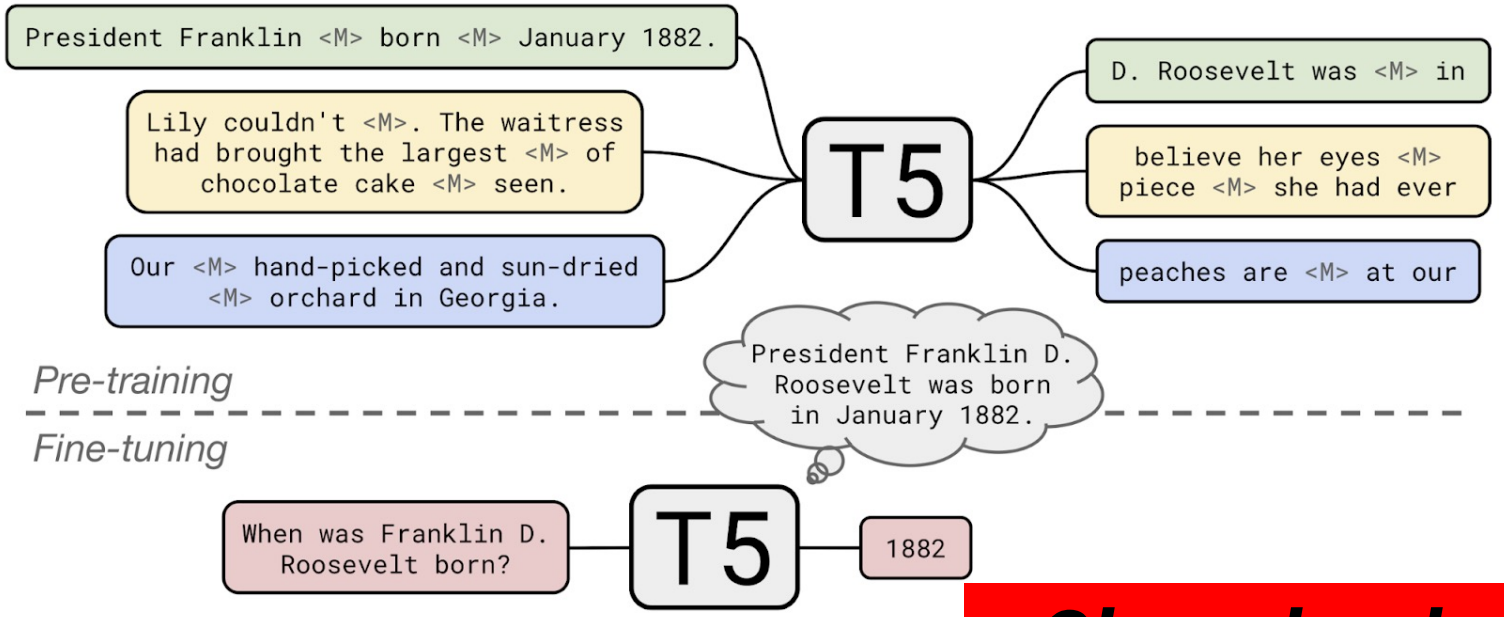


Machine Translation



Paraphrasing

Fine-tune PLMs on Downstream Tasks



Microsoft
UniLM, Turing-NLG

NVIDIA
Megatron LM

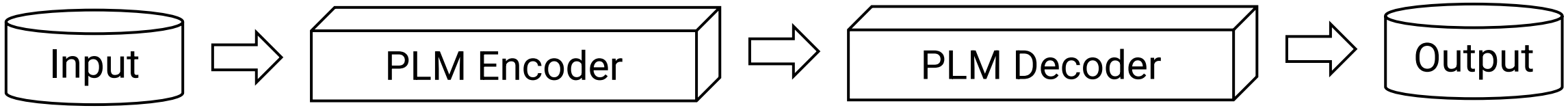
facebook
BART

OpenAI
GPT-2, GPT-3

Google
T5

Close-book

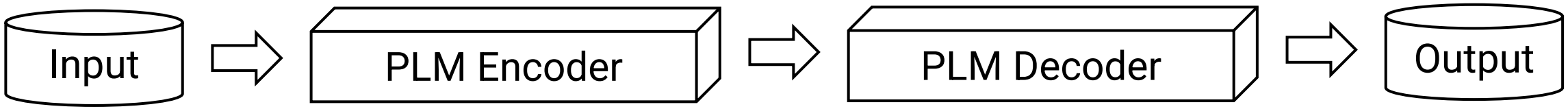
- Fine-tuning PLMs with *input-output* pairs of target data is the dominant paradigm in NLP research.



Input query: Miami Beach in Florida borders which ocean? Output Answer: Atlantic Ocean

Fine-tune PLMs on Downstream Tasks

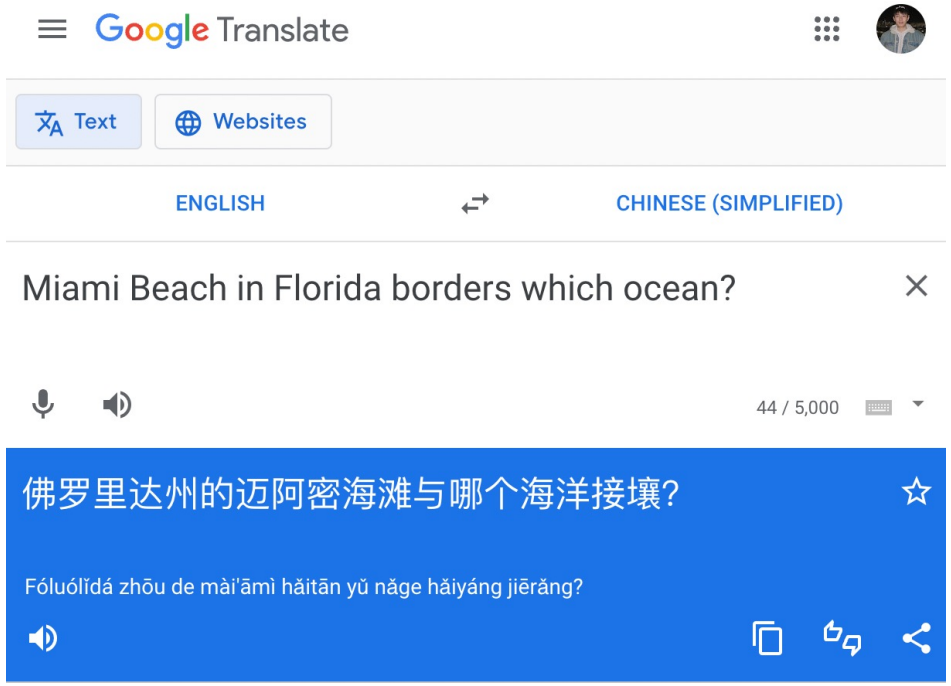
-- Fine-tuning PLMs with *input-output* of target data is the dominant paradigm.



- **Machine Translation**

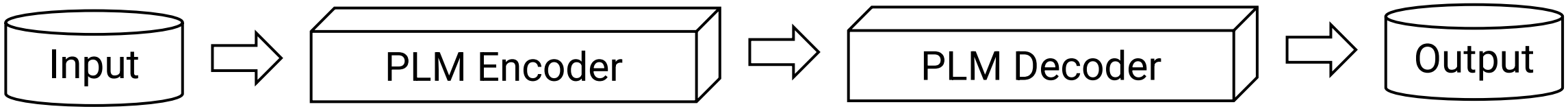
Input: Miami Beach in Florida borders which ocean?

Output: 佛罗里达州的迈阿密海滩与哪个海洋接壤？ (from Google Translate)



Fine-tune PLMs on Downstream Tasks

-- Fine-tuning PLMs with *input-output* of target data is the dominant paradigm.



- **Summarization**

Input: BERT paper full text

Output: TLDR (from semantic scholar)

(A new language representation model, BERT, designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning both left and right context in all layers)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova • Computer Science • NAACL • 2019

TLDR A new language representation model, **BERT**, designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers, which can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks.

Abstract

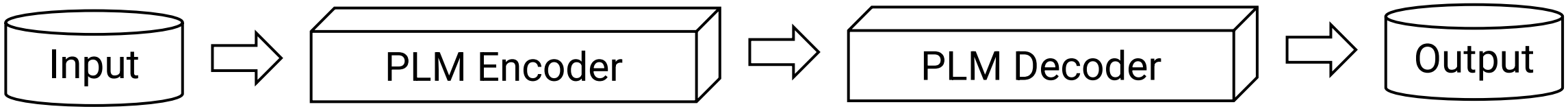
We introduce a new language representation model called **BERT**, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), **BERT** is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained **BERT** model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. **BERT** is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5 (7.7 point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement). Collapse

31,856 PDF • View on ACL Save Alert Cite

SEMANTIC SCHOLAR

Fine-tune PLMs on Downstream Tasks

-- Fine-tuning PLMs with *input-output* of target data is the dominant paradigm.



- Summarization

Though the input and output are different (e.g., language, length), the contents are very similar, and globally, under the same topic.

LITTLE SEMANTIC GAP

sentations from unlabeled text by jointly cond-
itioning both left and right context in all layers)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement). Collapse

31,856 PDF View on ACL Save Alert Cite SEMANTIC SCHOLAR

Why knowledge is needed in NLG?

Question/Answer Generation (e.g., open-domain QA, question generation)

- Query: Who did Hawaii belong to before 1895?

Hawaiian Kingdom

Question-answer gap

Hawaii is the most recent state … **On Jan 17, 1895**. The United States Minister to the **Hawaiian Kingdom** conspired with U.S. citizens to overthrow the monarchy.

In 1895, United States Public Law acknowledged that "the overthrow of **Hawaiian Kingdom** occurred with the active participation of agents …

The **Hawaiian Kingdom**, or Kingdom of Hawaii, was a sovereign state located in the Hawaiian Islands formed in 1795.

Subject: **Hawaiian Kingdom**
Relation: end time
Object: **1895**

Why knowledge is needed in NLG?

Creative Generation (e.g., story generation, abductive generation)

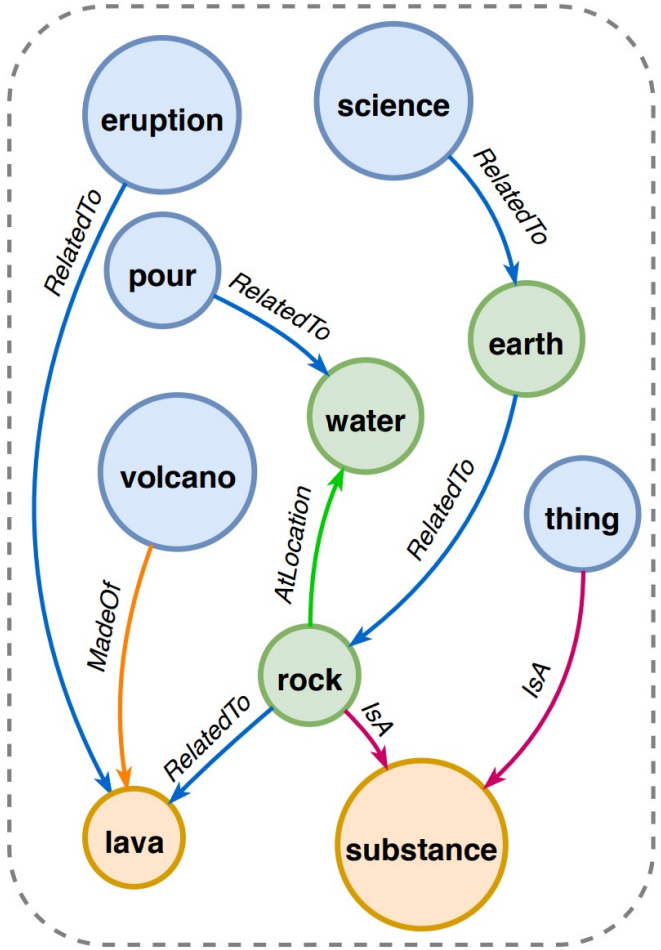
Storyline content semantic gap

- (ROCStories) Input:

(S1) Mr. Egg was presenting a volcanic *eruption* to the *science* class.
(S2) He has a diagram of a *volcano* that looked like it was made of tinfoil
(S3) He then took out a huge *thing* of vinegar and started to *pour* it in!
(S4) The class had no clue what was going on and looked on in astonishment.

- (ROCStories) Output:

(S5) The volcano then exploded with *substance* that looked like *lava*!



Why knowledge is needed in NLG?

Explanation Generation (e.g., choice explanation, counterfactual explanation)

Input: Piano is a kind of sport .

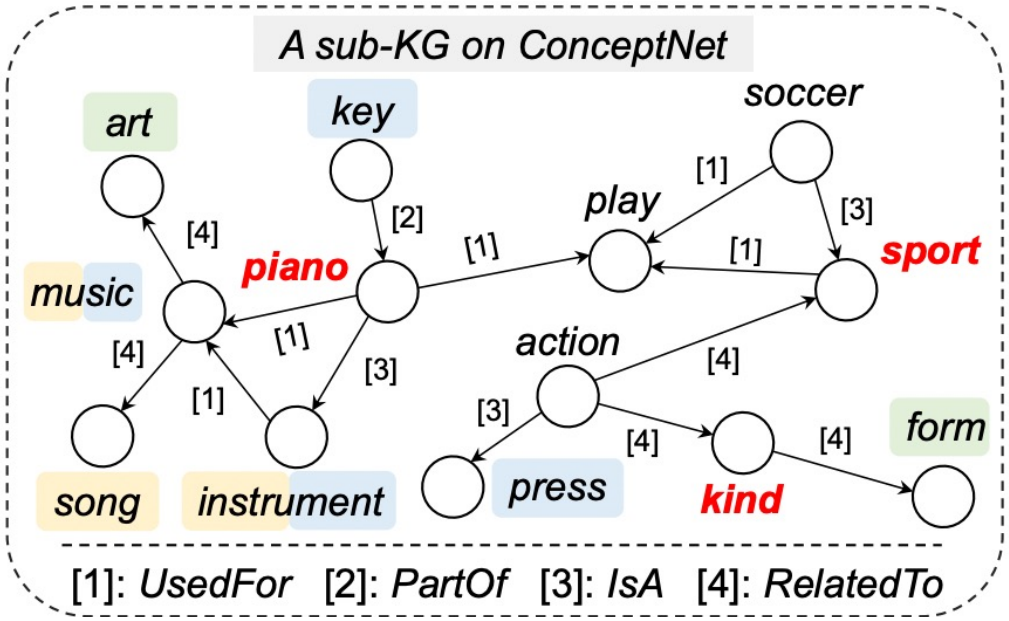
Statement-explanation gap

Outputs: *3* different explanations

(1) You can produce music when pressing keys on the piano, so it is an instrument .

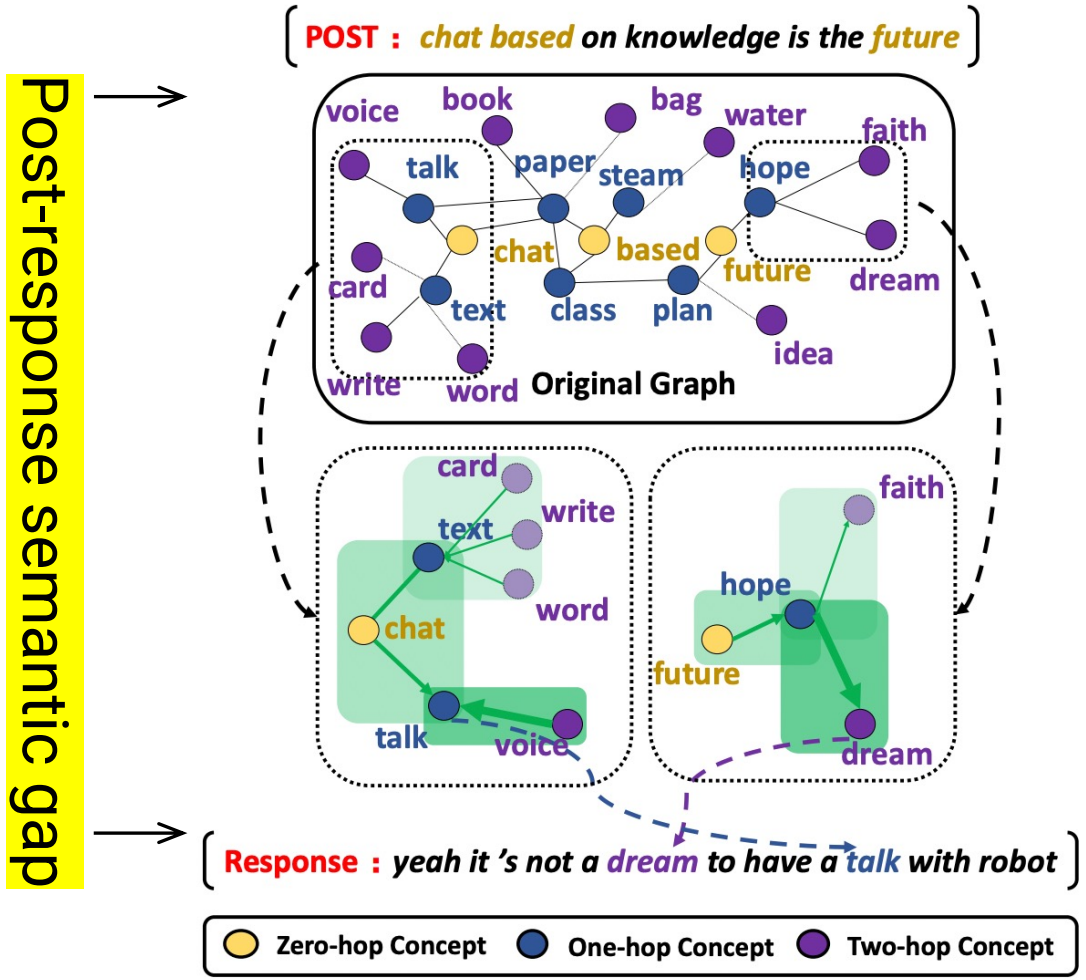
(2) Piano is a musical instrument used in songs to produce different musical tones .

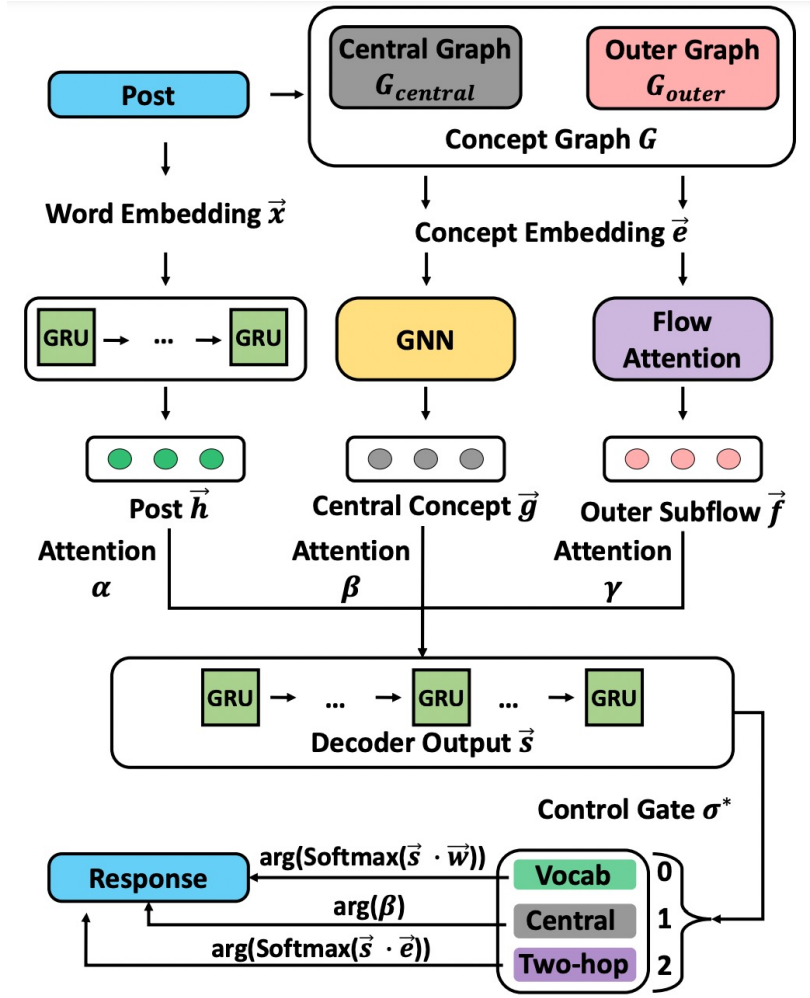
(3) Piano is a kind of art form .



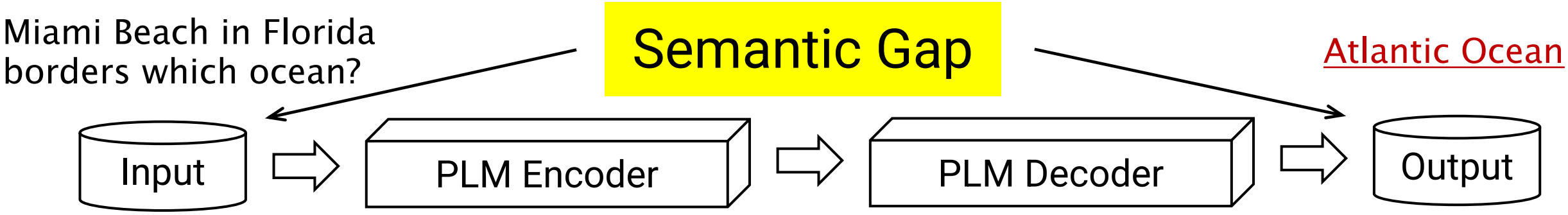
Why knowledge is needed in NLG?

Response Generation (e.g., chit-chat conversation, open-domain dialogue)





Why Close-book does not work?

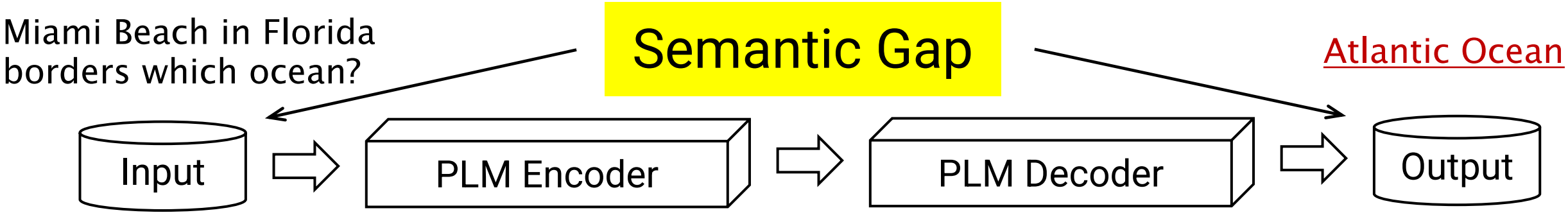


- When fine-tuned on the downstream tasks, the LMs might forget previously learned knowledge during pre-training, leading to catastrophic forgetting.

Model		Open Natural Questions				TriviaQA				WebQuestions			
		Total	Question Overlap	Answer Overlap Only	No Overlap	Total	Question Overlap	Answer Overlap Only	No Overlap	Total	Question Overlap	Answer Overlap Only	No Overlap
Closed book	T5-11B+SSM	36.6	77.2	22.2	9.4	-	-	-	-	44.7	82.1	44.5	22.0
	BART	26.5	67.6	10.2	0.8	26.7	67.3	16.3	0.8	27.4	71.5	20.7	1.6
Nearest Neighbor	Dense	26.7	69.4	7.0	0.0	28.9	81.5	11.2	0.0	26.4	78.8	17.1	0.0
	TF-IDF	22.2	56.8	4.1	0.0	23.5	68.8	5.1	0.0	19.4	63.9	8.7	0.0

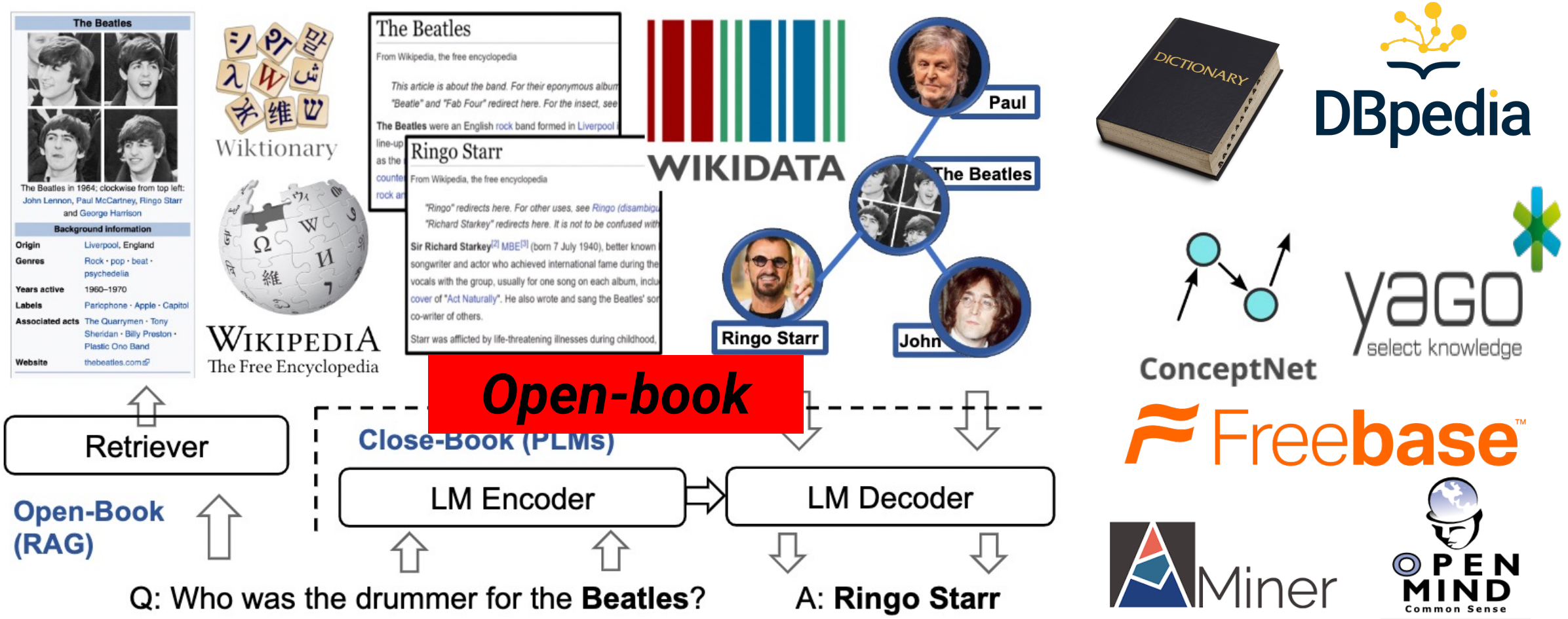
[1] Question and Answer Test-Train Overlap in Open-Domain QA Datasets. EACL 2021.

Why Close-book does not work?



- When fine-tuned on the downstream tasks, the LMs might forget previously learned knowledge during pre-training, leading to the catastrophic forgetting.
- The LMs make predictions by only “looking up information” stored in its parameters, leading to inferior performance and interpretability.
- They are usually trained offline, rendering the model agnostic to the latest information, e.g., asking BERT (released at 2018) about COVID-19.
- They are often expensive to train (e.g., GPT-3, T5-11B, GLaM-1.2T).

Open-book: Bridge the Semantic Gap



-- **Knowledge-enhanced methods**: Knowledge is **retrieved** based on the input text. The Language model make predictions by *reading and reasoning* over retrieved information.

Open-book: Bridge the Semantic Gap

Structure Knowledge (i.e., knowledge graph)

Unstructured Knowledge (i.e., grounded document)

A Survey of Knowledge-Enhanced Text Generation (Yu et al., ACM Computing Survey 2022)

Encyclopedic Knowledge (i.e., Wikipedia, Aminer)

Commonsense Knowledge (i.e., OMCS, ConceptNet)

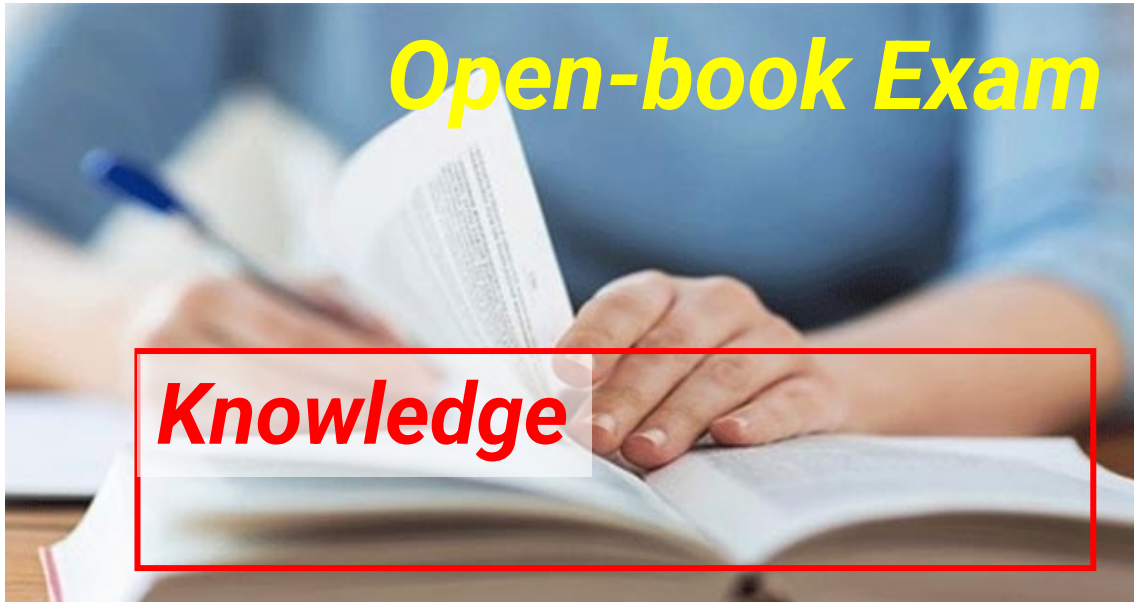
A Survey of Knowledge-Intensive Natural Language Processing (Yin et al., on arXiv 2022)

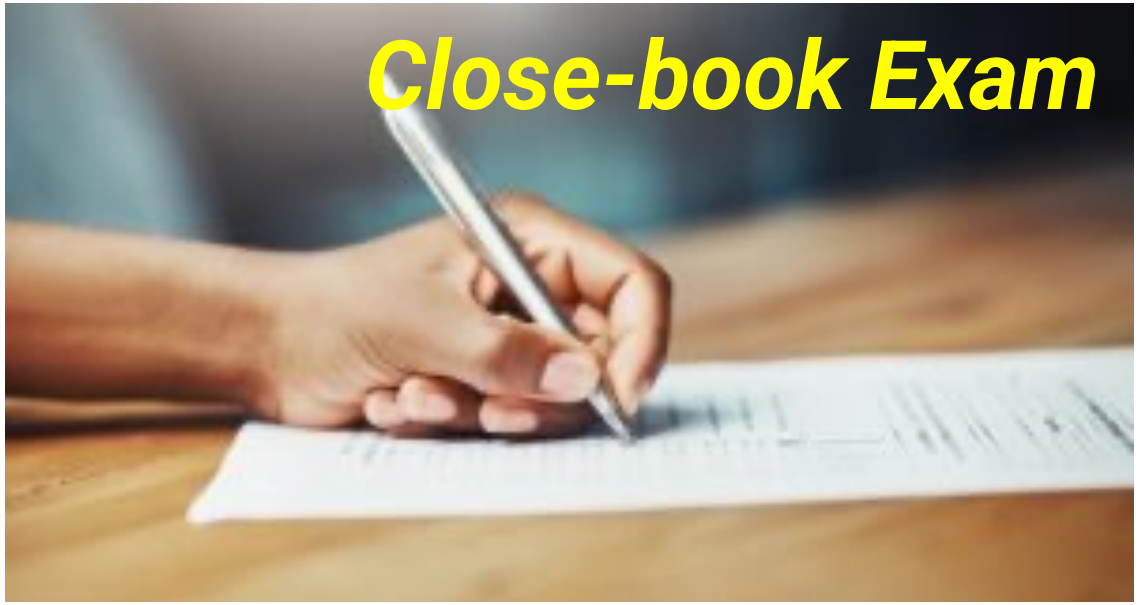
Summary: Close-book & Open-book

Dnpt f .cppl N pef rn; **Knowledge** is learnt into a language model (LM) **parameters**. During fine-tuning, only feed *input text* into LMs and make predictions. Examples include BART, T5, GPT-2/GPT-3, UniLM.

Pqf o.cppl N pef rn; We also refer it as knowledge-enhanced methods. **Knowledge** is **retrieved** based on the input text. The Language model make predictions by *reading and reasoning* over the retrieved texts. Examples include JointGT, REALM, KG-BART, RAG (*will be covered in this tutorial*).

Open-book **(pros)** v.s. Close-book

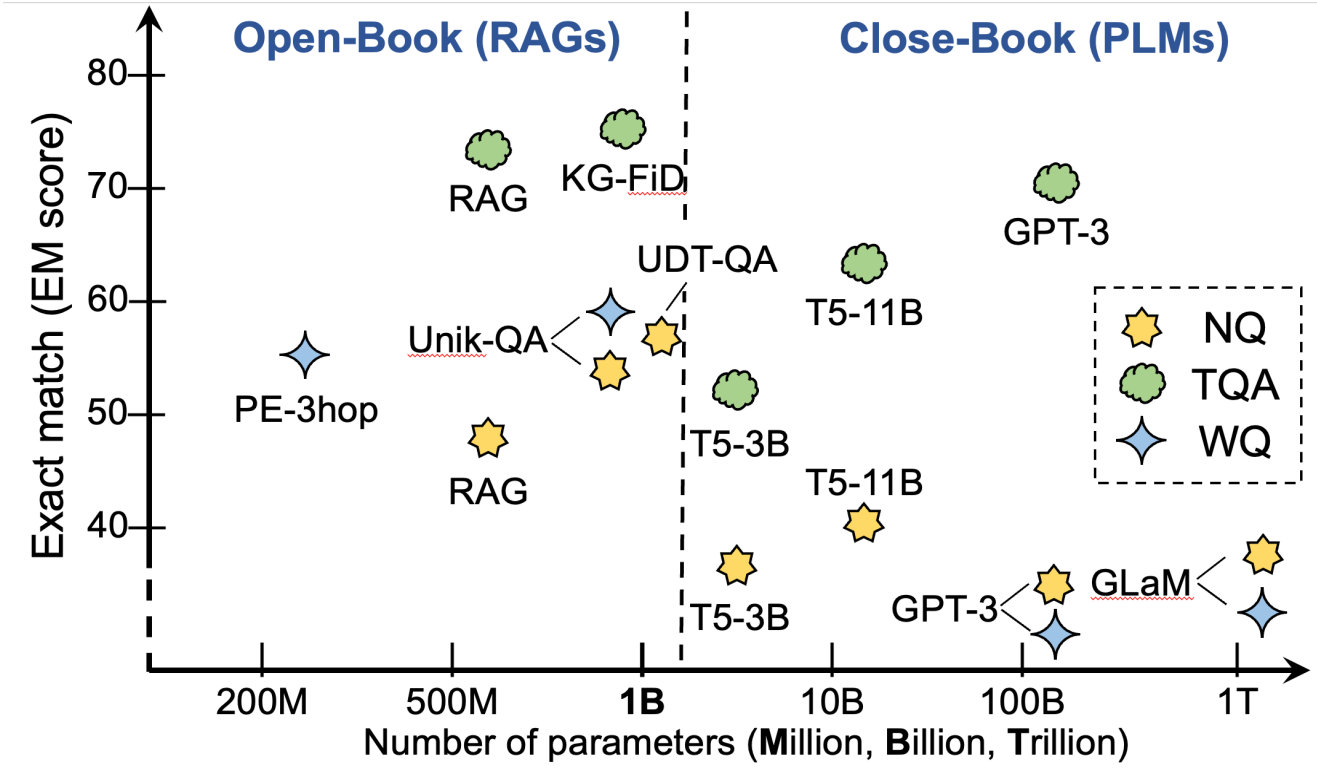




- The knowledge is not implicitly stored in model parameters, but is explicitly acquired in a plug-and-play manner, leading to great scalability and interpretability.
- Instead of generating from scratch, the paradigm generating text from some retrieved references, which potentially alleviates the difficulty of text generation.

Open-book **(SoTA)** v.s. Close-book **(SoTA)**

- From the existing literature, the performance of the open-book method is usually significantly better than that of the close-book method. At the same time, open-book methods are trained with less parameters.



	Close-book SoTA	Open-book SoTA
Model	**GLaM-1.2T**	**UnikQA-990M**
WQ	**25.3**	**57.8**
Model	GPT-175B	KGFiD-994M
TQA	68.0	72.5
Model	T5-11B	UDTQA-990M
NQ	42.3	55.2

Outline of Knowledge-enhanced NLG

	Knowledge-enhanced pre-training (self-supervised training)	**Knowledge-enhanced fine-tuning (end-to-end training)**
Structured knowledge	KG-BART [AAAI 2021] JointGT [ACL 2021]	GRF [EMNLP 2020] KGMoE [ACL 2022]
Unstructured knowledge	REALM [ICML 2020] CALM [ICLR 2021]	RAG [Neurips 2020], FiD [EACL 2021], Re-T5 [ACL 2021]

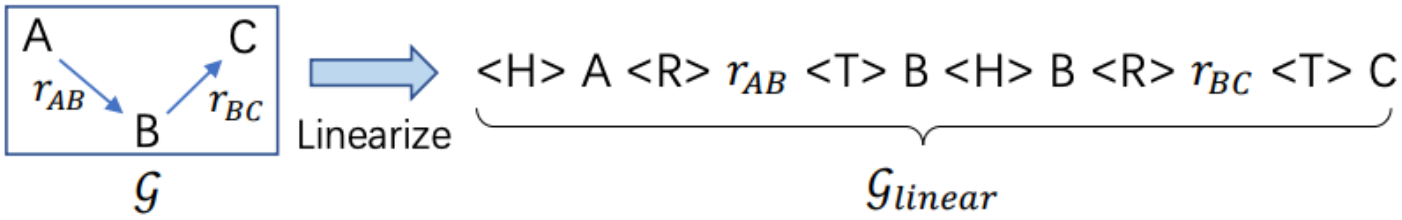
- Covered NLP applications: Open-domain question answering, commonsense reasoning, dialogue system, question generation, data-to-text generation …

[KG in PT] JointGT in ACL 2021

JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs. ACL'21

- First work of jointly training knowledge graph and text representation for NLG
- Pre-training dataset: 7M KG-text pairs extracted from Wikipedia/Wikidata

- **Feed KG triples into Transformer:**



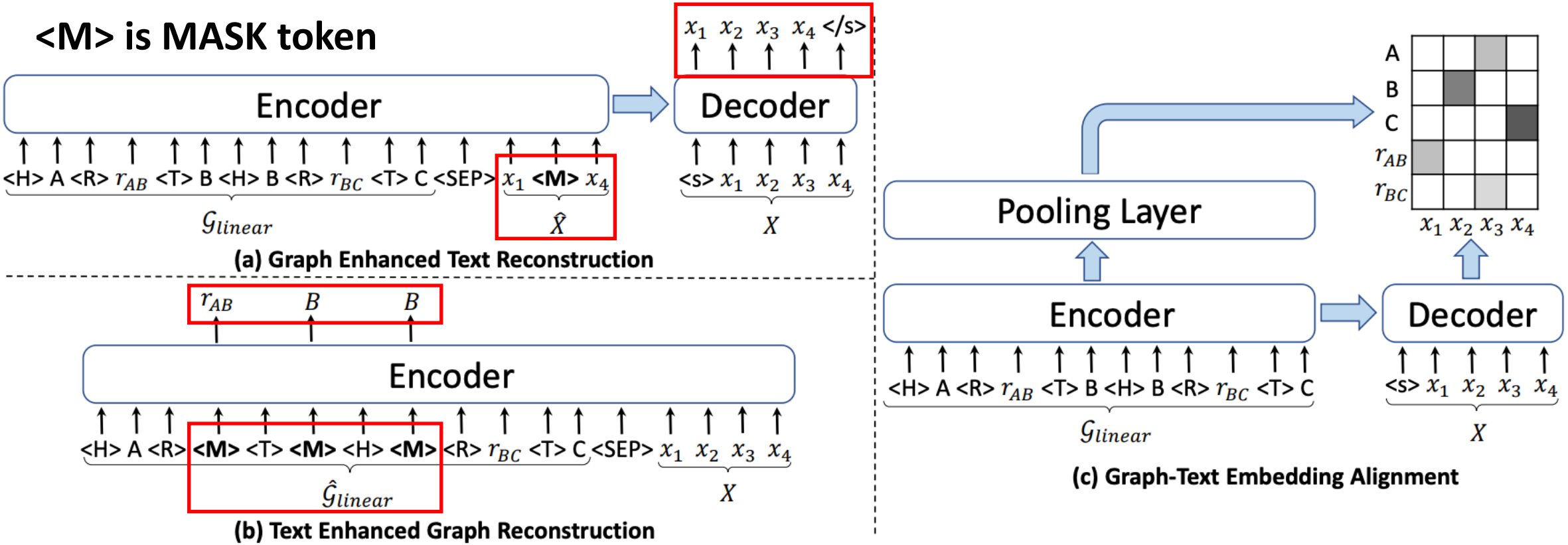
SubKG -> Sequence <H>: head entity, <R>: relation, <T>: tail entity

- **Three novel pre-training tasks:** Graph Enhanced Text Reconstruction, Text Enhanced Graph Reconstruction, Graph-Text Embedding Alignment

[KG in PT] JointGT in ACL 2021 (cont.)

JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs. ACL'21

- Two novel generation pre-training tasks; one graph-text alignment objective



[KG in PT] JointGT in ACL 2021 (cont.)

JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs. ACL'21

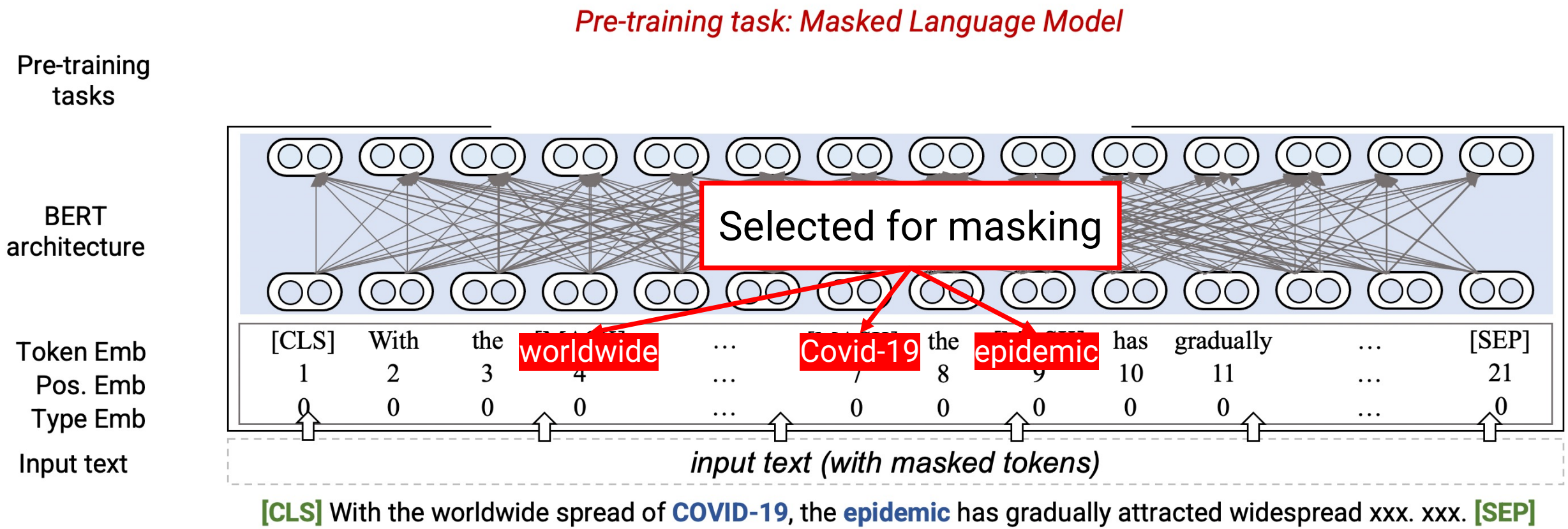
- Downstream tasks: data-to-text generation, knowledge base question generation

Dataset	#Param	WebNLG(U)			WebQuestions			PathQuestions		
Model		BLEU	METEOR	ROUGE	BLEU	METEOR	ROUGE	BLEU	METEOR	ROUGE
SOTA-NPT	-	61.00†	42.00†	71.00†	29.45‡	30.96‡	55.45‡	61.48‡	44.57‡	77.72‡
KGPT	177M	64.11♯	46.30♯	74.57♯	-	-	-	-	-	-
BART	140M	64.55	46.51	75.13	29.61	31.48	55.42	63.74	47.23	77.76
T5	220M	64.42	46.58	74.77	28.78	30.55	55.12	58.95	44.72	76.58
JointGT (BART)	160M	65.92	47.15	**76.10****	**30.02***	**32.05****	**55.60**	**65.89****	**48.25****	**78.87****
JointGT (T5)	265M	**66.14****	**47.25****	75.91	28.95	31.29	54.47	60.45	45.38	77.59

Table: Results on three text generation tasks, including WebNLG, WebQuestions and PathQuestions.

[Text in PT] REALM in ICML 2020

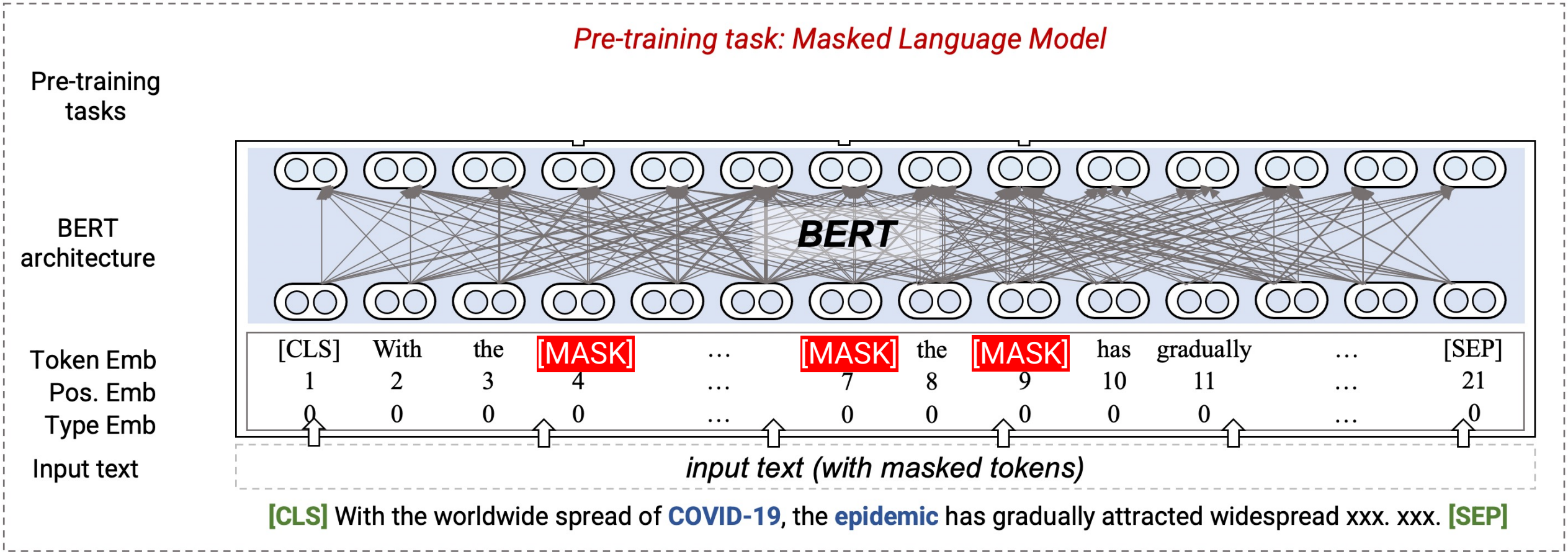
REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20



Input text: With the worldwide spread of COVID-19, the epidemic has gradually attracted widespread attention and discussion on social media platforms.

[Text in PT] REALM in ICML 2020

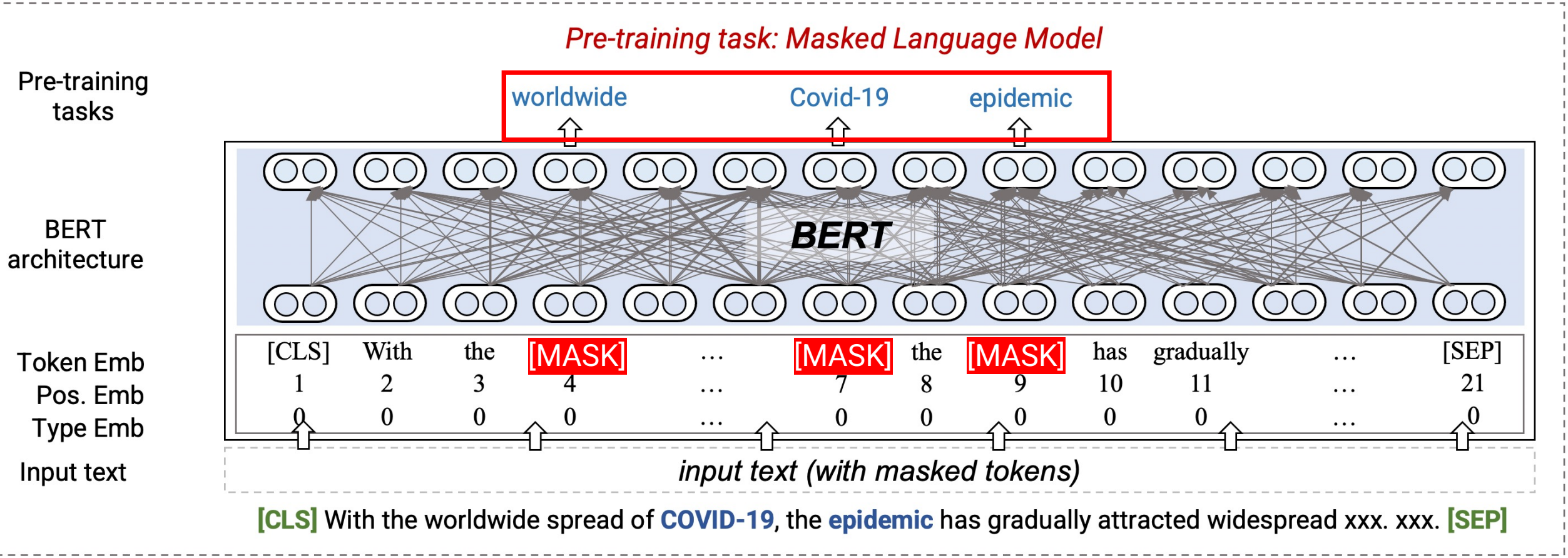
REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20



Input text: With the worldwide spread of COVID-19, the epidemic has gradually attracted widespread attention and discussion on social media platforms.

[Text in PT] REALM in ICML 2020

REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20



Input text: With the worldwide spread of COVID-19, the epidemic has gradually attracted widespread attention and discussion on social media platforms.

[Text in PT] REALM in ICML 2020 (cont.)

REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20

- It is the first retrieval-augmented pre-training model. The task is masked language model.

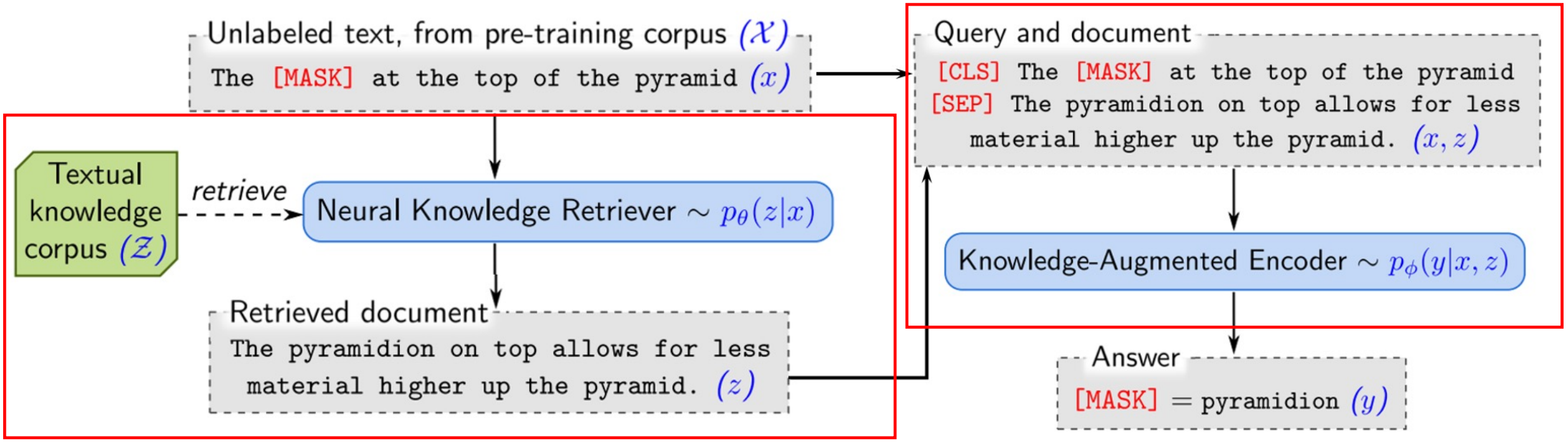
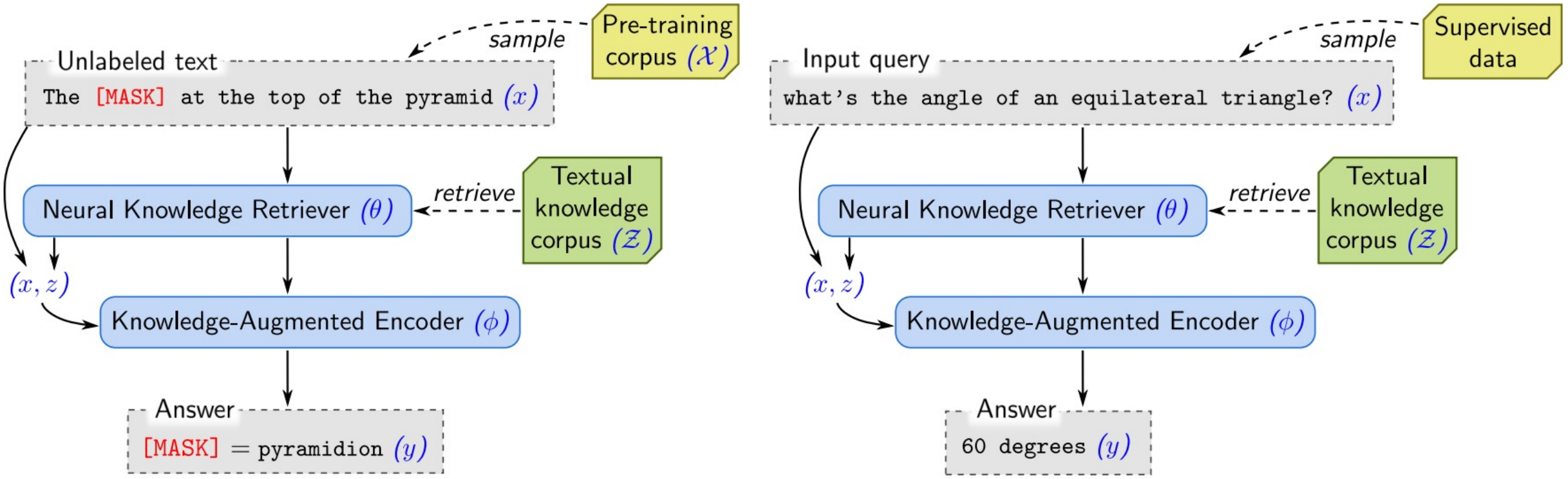


Figure: REALM augments language model pre-training with a neural knowledge retriever that retrieves knowledge from a textual knowledge corpus, Z (e.g., all of Wikipedia).

[Text in PT] REALM in ICML 2020 (cont.)



Left: Unsupervised pre-training. The knowledge retriever and knowledge-augmented encoder are jointly pre-trained on the MLM task. ***Right:*** Supervised fine-tuning. After the parameters are pre-trained, the model is then fine-tuned on a task of primary interest, using supervised examples.

[Text in PT] REALM in ICML 2020 (cont.)

REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20

Name	Architectures	Pre-training	NQ (79k/4k)	WQ (3k/2k)	CT (1k /1k)
Close-book Models	Sparse Retr.+Transformer	BERT	26.5	17.7	21.3
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1	-
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2	-
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4	-
Open-book Models	Sparse Retr.+DocReader		-	20.7	25.7
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2	**46.8**
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**	42.9

Sparse v.s. Dense

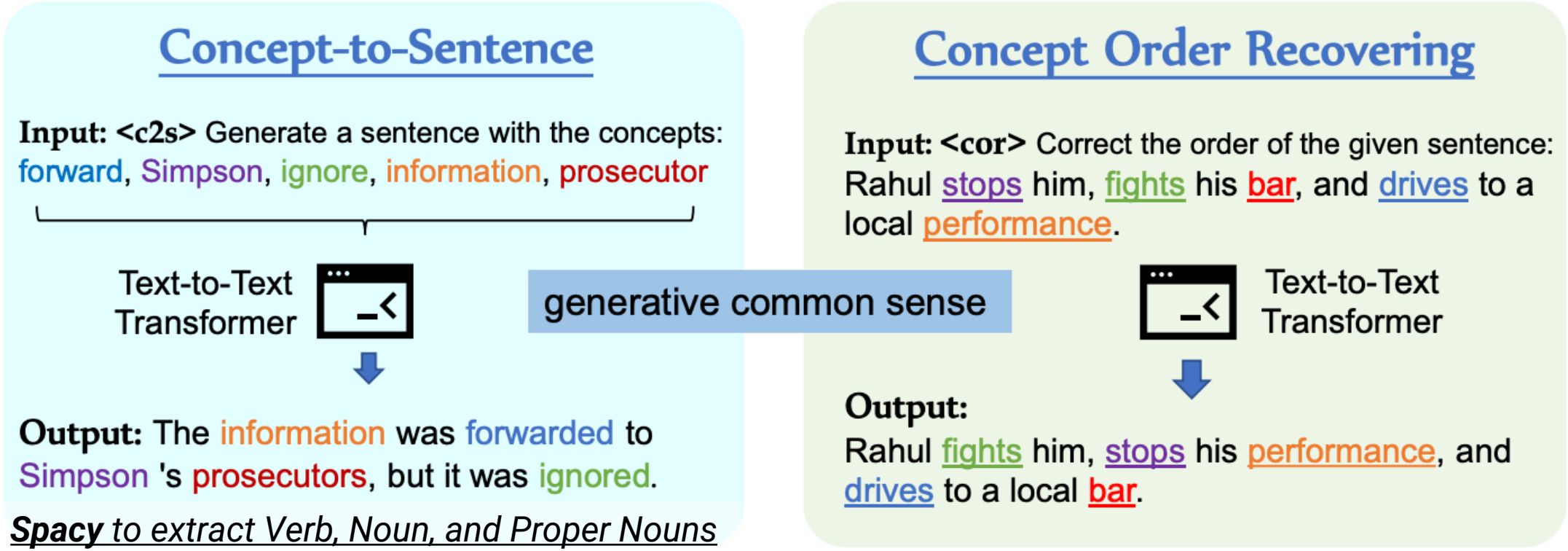
Table: Test results on Open-QA benchmarks. REALM performs the best among all baseline methods.

Open-book (REALM) > Close-book (11B) Dense (REALM) > Sparse (BM25+Transformer)

[Text in PT] CALM in ICLR 2021

Pre-training Text-to-Text Transformers for Concept-centric Common Sense. ICLR'21

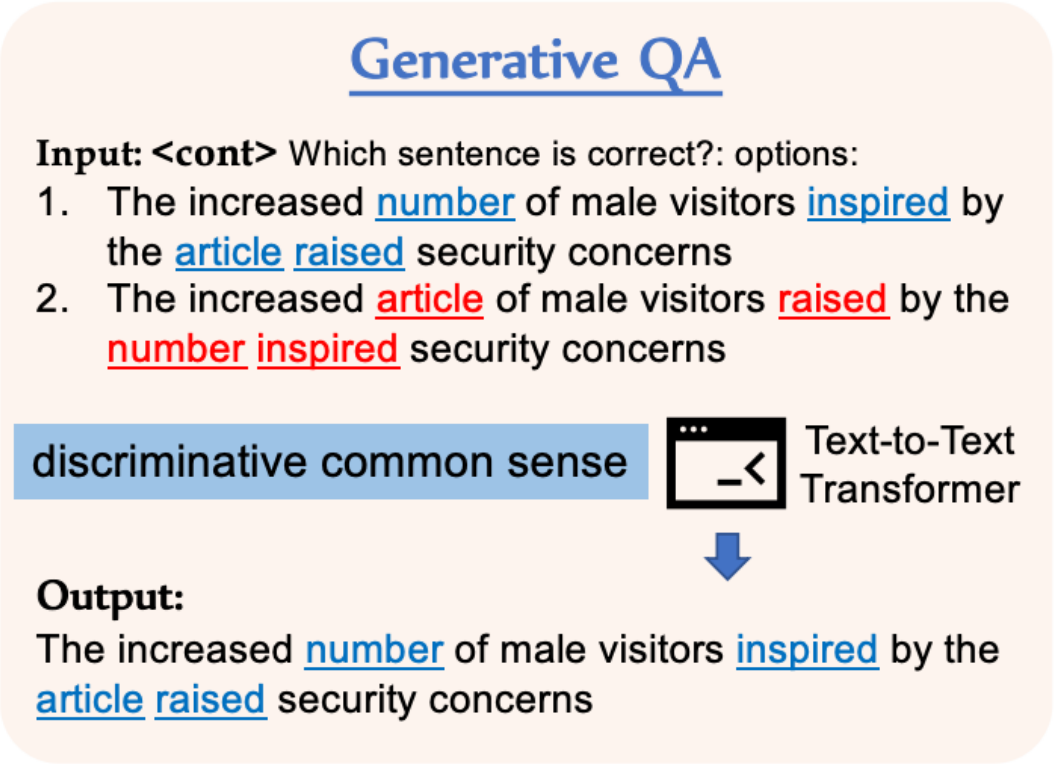
- **Construct aligned text pairs** in three ways: Concept-to-Sentence Generation (C2S) Concept Order Recovering (COR), Contrastive objective (distinguish the truth sentence)



[Text in PT] CALM in ICLR 2021 (cont.)

Pre-training Text-to-Text Transformers for Concept-centric Common Sense. ICLR'21

- **Construct aligned text pairs** in three ways: Concept-to-Sentence Generation (C2S) Concept Order Recovering (COR), Contrastive objective (distinguish the truth sentence)



Algorithm 1: Pre-training Concept-Aware Language Model (CALM).

Input: Text-to-Text Transformer T_θ, Text corpus X=[$x_1, x_2, \ldots, x_n$].

repeat
  for *each* $x_i \in X$ **do**
    Extract the concept-set $\mathcal{C}_i$;
    Construct the distractor sentence $x' = \text{CONCEPT-PERMUTE}(\mathbf{x}_i, \mathcal{C}_i)$;
    Update T_θ with Eq.(1, 2, 4);
until *maximum iterations reached*;
repeat
  for *each* $x_i \in X$ **do**
    Update T_θ with Eq.(7)
until *maximum iterations reached*;

[Text in PT] CALM in ICLR 2021 (cont.)

Pre-training Text-to-Text Transformers for Concept-centric Common Sense. ICLR'21

Methods	CSQA	OBQA	PIQA	aNLI	CommonGEN	
Close-book Models	Accuracy (official dev)				BLEU-4	SPICE
T5-base	61.88(±0.08)	58.20(±1.0)	68.14(±0.73)	61.10(±0.38)	24.90	32.40
T5-base + cont. pretraining	61.92(±0.45)	58.10(±0.9)	68.19(±0.77)	61.15(±0.52)	25.10	32.40
T5-base + SSM	62.08(±0.41)	58.30(±0.8)	68.27(±0.71)	61.25(±0.51)	25.20	32.40
CALM (Generative-Only)	62.28(±0.36)	58.90(±0.4)	68.91(±0.88)	60.95(±0.46)	25.80	32.60
CALM (Contrastive-Only)	62.73(±0.41)	59.30(±0.3)	70.67(±0.98)	61.35(±0.06)	25.50	32.60
CALM (w/o Mix warmup)	62.18(±0.48)	59.00(±0.5)	69.21(±0.57)	61.25(±0.55)	25.80	32.60
CALM (Mix-only)	63.02(±0.47)	60.40(±0.4)	70.07(±0.98)	62.79(±0.55)	26.00	32.80
CALM	**63.32(±0.35)**	**60.90(±0.4)**	**71.01(±0.61)**	**63.20(±0.52)**	**26.40**	**33.00**

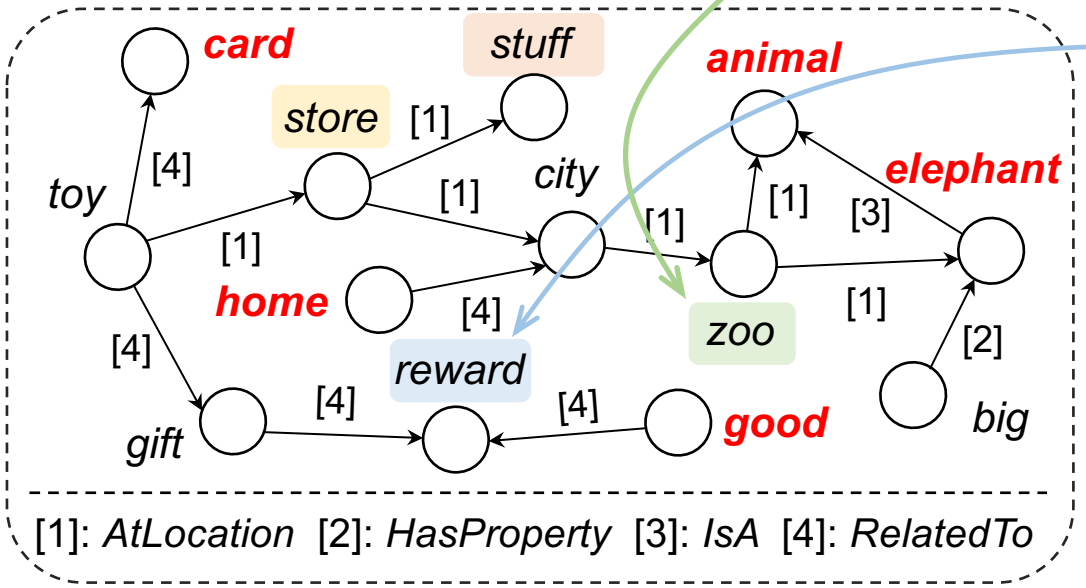
Table 1: Experimental results on commonsense reasoning datasets.

Avg: Use both (64.60) > Only contrastive (63.51) > Only generative (62.76) > T5 (62.33)

[KG+FT] GRF in EMNLP 2020

Language Generation with Multi-Hop Reasoning on Commonsense KG, In EMNLP'20

Alpha-NLG: the goal of abductive language generation is to generate valid hypothesis about the likely explanations to partially observable past and future.
Observable past: Billy had received good grades on his report card.
Valid hypothesis: He went to the zoo later in the day and saw elephants.
Observable future: He decided that the elephant was his new favorite animal.

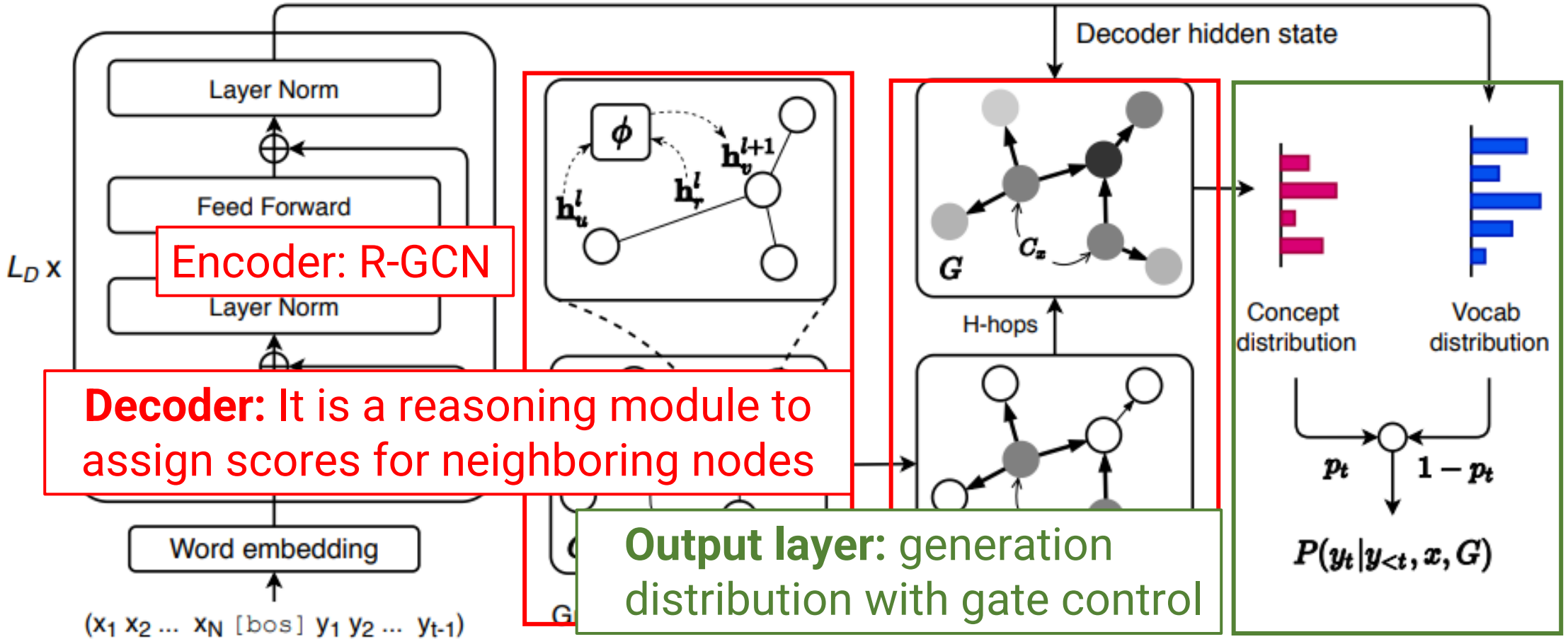


Other valid hypothesis: Billy's parents sent him on an African safari for a reward.
Other valid hypothesis: His mother stopped by the store and bought him a stuffed elephant.

-- 95% of Concepts in the input/output can be found on the commonsense KG (ConceptNet)

[KG+FT] GRF in EMNLP 2020 (cont.)

Language Generation with Multi-Hop Reasoning on Commonsense KG, In EMNLP'20



[KG+FT] GRF in EMNLP 2020 (cont.)

Language Generation with Multi-Hop Reasoning on Commonsense KG, In EMNLP'20

- Dataset: alpha-NLG, EG. Metric: BLEU, METEOR, ROUGE

Models	EG				αNLG			
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
Seq2Seq	6.09	24.94	26.37	32.37	2.37	14.76	22.03	29.09
COMeT-Txt-GPT2	N/A	N/A	N/A	N/A	2.73†	18.32†	24.39†	32.78†
COMeT-Emb-GPT2	N/A	N/A	N/A	N/A	3.66†	19.53†	24.92†	32.67†
GPT2-FT	15.63	38.76	37.32	77.09	9.80	25.82	32.90	57.52
GPT2-OMCS-FT	15.55	38.28	37.53	75.60	9.62	25.83	32.88	57.50
GRF	**17.19**	**39.15**	**38.10**	**81.71**	**11.62**	**27.76**	**34.62**	**63.76**

Table 3: Automatic evaluation results on the test set of EG and αNLG. Entries with N/A mean the baseline is not designated for this task. †: we use the generation results from Bhagavatula et al. (2020).

[KG+FT] GRF in EMNLP 2020 (cont.)

Case study on Explanation Generation (ComVE)

Goal: generate explanation given a counterfactual statement for sense-making

Statement: Coke is made of alcohol.
GPT2-FT: Coke is a drink. ✗
GPT2-OMCS-FT: Coke is not a liquid. ✗
GRF: Coke is made from corn. ✓

Statement: She cut up a blanket.
GPT2-FT: A blanket is not sharp enough to cut. ✗
GPT2-OMCS-FT: A blanket is too small to be cut. ✗
GRF: Blankets are too soft to be cut. ✓

[KG+FT] MoKGE in ACL 2022

Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL'22

- **Jbqvy!**Piano is a kind of sport.
  (a counterfactual statement)


- **Output1:** You can produce music when pressing keys on the piano, so it is an instrument. **(usage)**

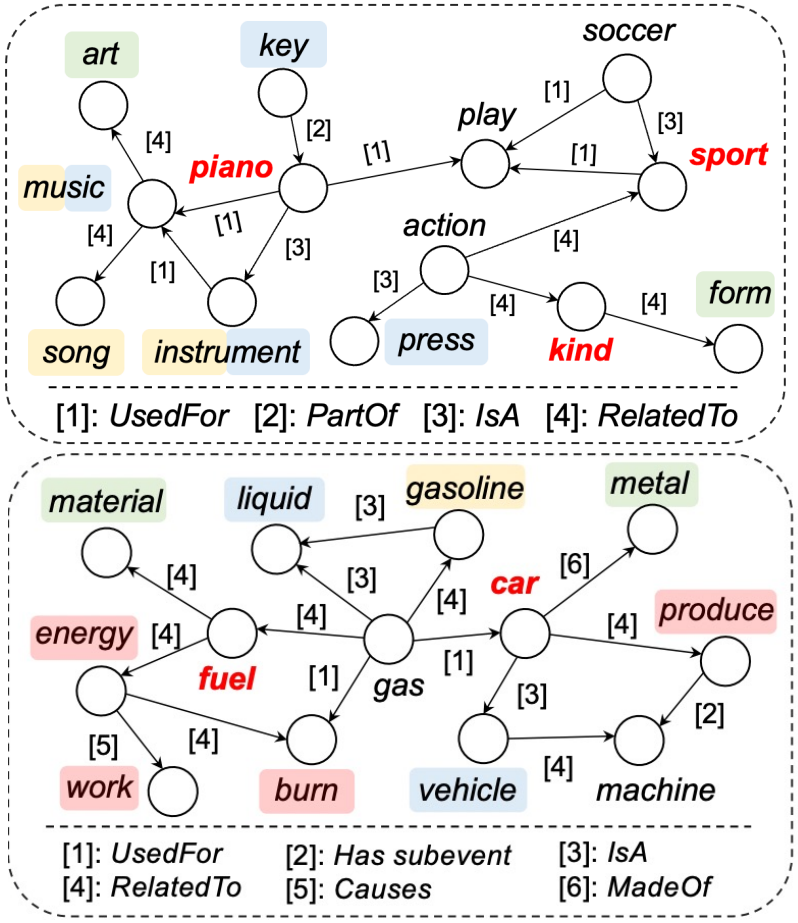
- **Output2:** Piano is a musical instrument used in songs to produce different musical tones. **(effect)**

- **Output3:** Piano is a kind of art form. **(taxonomy)**

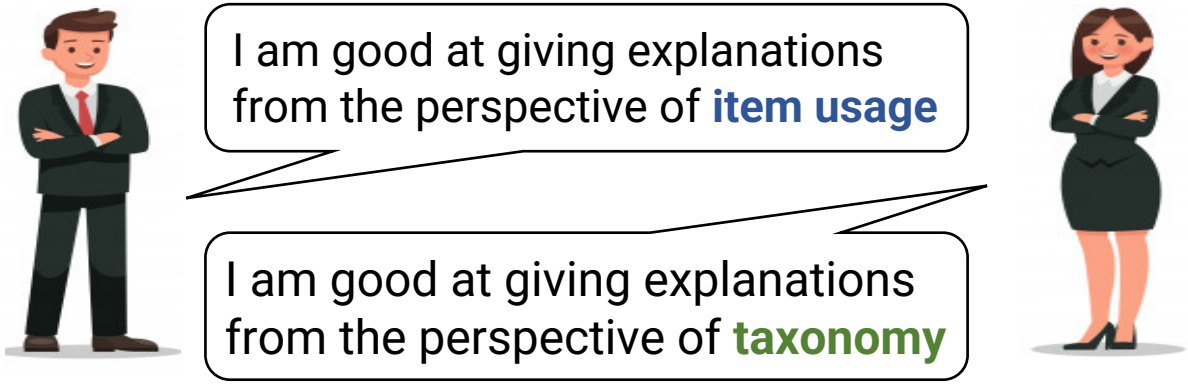
Many plausible reasons from different perspectives!

[KG+FT] MoKGE in ACL 2022 (cont.)

Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL'22



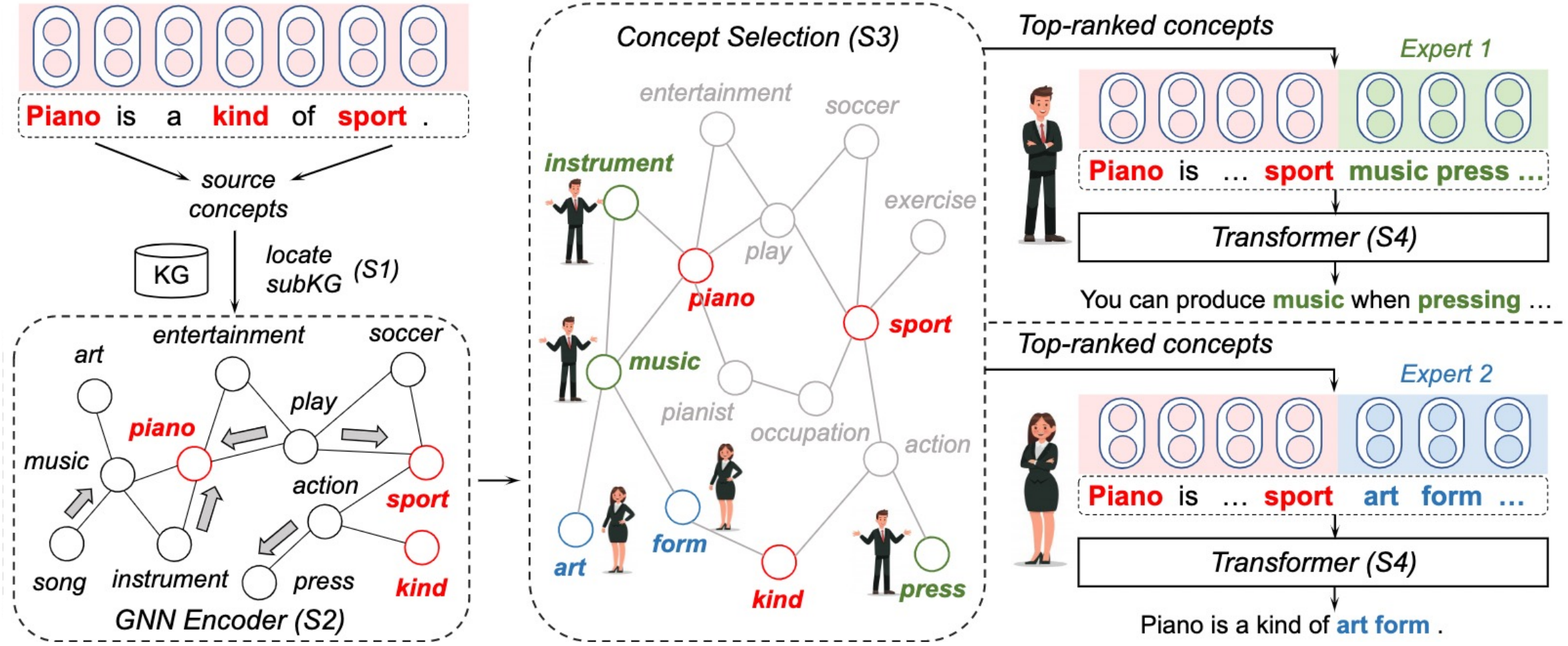
Input: Piano is a kind of sport .
Output (usage): You can produce music when pressing keys on the piano, so it is an instrument .
Output (taxonomy): Piano is a kind of art form .



Input: Cars are made of fuel.
Output (usage): Cars burn fuel to produce energy and work.
Output (taxonomy): Cars are made of metal.

[KG+FT] MoKGE in ACL 2022 (cont.)

Diversifying Content Generation with Mixture of Knowledge Graph Experts, ACL '22



Step4: generate the outputs by integrating the input sequence and the top-ranked entities

[KG+FT] MoKGE in ACL 2022

ComVE

Methods	Model Variant	Pairwise diversity		Corpus diversity		Quality	
		SB-3 (⇓)	SB-4 (⇓)	D-2(⇑)	E-4(⇑)	B-4 (⇑)	R-L (⇑)
Baseline methods							
CVAE	z = 16	$66.66_{0.4}$	$62.83_{0.5}$	$33.75_{0.5}$	$9.13_{0.1}$	$16.67_{0.3}$	$41.52_{0.3}$
	z = 32	$59.20_{1.3}$	$54.30_{1.5}$	$32.86_{1.1}$	$9.07_{0.5}$	$17.04_{0.2}$	$42.17_{0.5}$
	z = 64	$55.02_{0.8}$	$49.58_{1.0}$	$32.55_{0.5}$	$9.07_{0.2}$	$15.54_{0.4}$	$41.03_{0.3}$
Truncated sampling	k = 5	$74.20_{0.2}$	$71.38_{0.2}$	$31.32_{0.4}$	$9.18_{0.1}$	$16.44_{0.2}$	$40.99_{0.2}$
	k = 20	$64.47_{2.1}$	$60.33_{2.4}$	$33.69_{0.6}$	$9.26_{0.1}$	$17.70_{0.2}$	$42.58_{0.5}$
	k = 50	$61.39_{2.4}$	$56.93_{2.8}$	$34.80_{0.3}$	$9.29_{0.1}$	$17.48_{0.4}$	$42.44_{0.5}$
Nucleus sampling	p = .5	$77.66_{0.8}$	$75.14_{0.9}$	$28.36_{0.6}$	$9.05_{0.3}$	$16.09_{0.6}$	$40.95_{0.5}$
	p = .75	$71.41_{2.5}$	$68.22_{2.9}$	$31.21_{0.3}$	$9.16_{0.1}$	$17.07_{0.5}$	$41.88_{0.7}$
	p = .95	$63.43_{3.4}$	$59.23_{3.8}$	$34.17_{0.3}$	$9.27_{0.2}$	$17.68_{0.4}$	$42.60_{0.8}$
MoE	embed	33.6			$9.31_{0.1}$	$18.66_{0.5}$	$\underline{43.72}_{0.2}$
	prompt	$33.42_{0.3}$	$28.40_{0.3}$	$46.93_{0.2}$	$9.60_{0.2}$	$18.91_{0.4}$	$43.71_{0.5}$
MoKGE							
MoKGE (ours)	embed	35.[illegible]$_{1.1}$	29.[illegible]$_{1.2}$	47.[illegible]$_{0.4}$	$\underline{9.[illegible]}_{0.1}$	$\mathbf{19.13}_{0.1}$	$43.70_{0.1}$
	prompt	$\mathbf{30.93}_{0.9}$	$\mathbf{25.30}_{1.1}$	$\mathbf{48.44}_{0.2}$	$\mathbf{9.67}_{0.2}$	$\underline{19.01}_{0.1}$	$\mathbf{43.83}_{0.3}$
Human		$12.36_{0.0}$	$8.01_{0.0}$	$63.02_{0.0}$	$9.55_{0.0}$	$100.0_{0.0}$	$100.0_{0.0}$

Best baseline performance

[KG+FT] MoKGE in ACL 2022

- Independent scoring: 1 to 5 based on *diveristy*, *quality*, *flency* and *grammar*

Methods	ComVE			α-NLG		
	Diversity	Quality	Flu. & Gra.	Diversity	Quality	Flu. & Gra.
Truncated samp.	2.15±0.76	2.22±1.01	3.47±0.75	2.31±0.76	2.63±0.77	3.89±0.36
Nucleus samp.	2.03±0.73	**2.29**±1.03	**3.52**±0.70	2.39±0.73	**2.67**±0.72	**3.91**±0.28
MoKGE (ours)	**2.63**±0.51*	2.10±0.99	3.46±0.81	**2.66**±0.51*	2.57±0.71	3.87±0.34
Human Ref.	2.60±0.59	3.00	4.00	2.71±0.57	3.00	4.00

Same observation from human evaluation

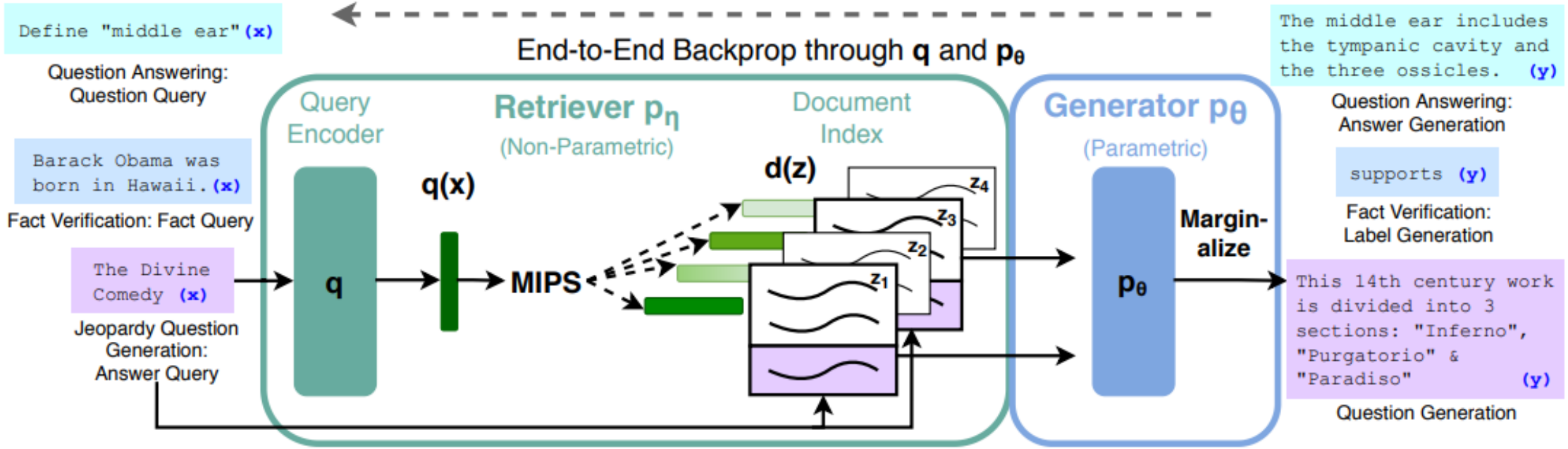
- Pairwise comparison: *MoKGE v.s.* two baseline methods based on *diversity*

Against methods	ComVE			α-NLG		
	Win (%)	Tie (%)	Lose (%)	Win (%)	Tie (%)	Lose (%)
v.s. Truncated samp.	**47.85**±5.94	37.09±4.56	15.06±3.31	**45.35**±5.06	43.19±2.78	11.46±2.31
v.s. Nucleus samp.	**54.30**±4.62	36.02±2.74	9.68±3.48	41.53±1.55	**46.99**±2.04	11.48±2.36

[Text-FT] RAG in Neurips 2020

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neruips'20

- It is the first retrieval-augmented generation work. The model is named RAG.
- RAG combines a pre-trained retriever (*Query Encoder+Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, it uses Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.



[Text-FT] RAG in Neurips 2020 (cont.)

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neruips'20

RAG-Sequence: the top K documents are retrieved using the retriever, and the generator produces the output sequence probability for each document, which are then marginalized.

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

RAG-Sequence: the top K documents are retrieved using the retriever, and then the generator produces a distribution for the next output token for each document, before marginalizing, and repeating the process with the following output token.

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x,z_i,y_{1:i-1})$$

[Text-FT] RAG in Neurips 2020 (cont.)

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neruips'20

- Dataset: Trivial QA, MS-MARCO Metric: Exact match (for ODQA); BLEU, ROUGE

Table 1: Open-Domain QA Test Scores. For TQA, left column uses the standard test set for Open-Domain QA, right column uses the TQA-Wiki test set. See Appendix D for further details.

	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

Table 2: Generation and classification Test Scores. MS-MARCO SotA is [4], FEVER-3 is [68] and FEVER-2 is [57] *Uses gold context/evidence. Best model without gold access underlined.

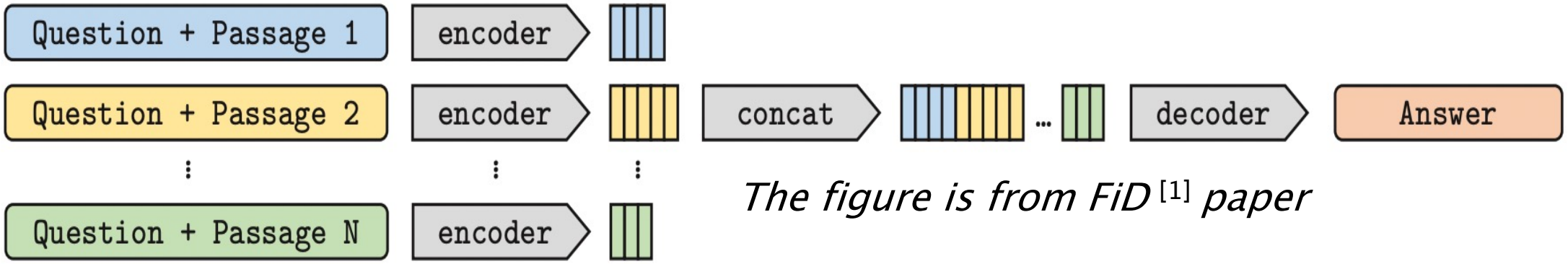
Model	Jeopardy B-1	Jeopardy QB-1	MSMARCO R-L	MSMARCO B-1	FVR3 Label Acc.	FVR2 Label Acc.
SotA	-	-	**49.8***	**49.9***	**76.8**	**92.2***
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	**17.3**	**22.2**	40.1	41.5	72.5	89.5
RAG-Seq.	14.7	21.4	40.8	44.2		

RAG > REALM on WQ/CT, RAG < REALM on TQA RAG > Close-book (BART)

[Text-FT] FiD in EACL 2021

Leveraging Passage Retrieval with Generative Models for ODQA. In EACL'21

- RAG suffers from the **input sequence length limitation (max:1024)** and **high computation cost (quadratic to the sequence length)**.
- Instead, FiD processed passages independently in the encoder, performed attention over all the retrieved passages



The figure is from FiD [1] paper

[Text-FT] FiD in EACL 2021 (cont.)

Leveraging Passage Retrieval with Generative Models for ODQA. In EACL'21

Model	NQ	TriviaQA		SQuAD Open	
	EM	EM	EM	EM	F1
DrQA (Chen et al., 2017)	-	-	-	29.8	-
Multi-Passage BERT (Wang et al., 2019)	-	-	-	53.0	60.9
Path Retriever (Asai et al., 2020)	31.7	-	-	**56.5**	**63.8**
Graph Retriever (Min et al., 2019b)	34.7	55.8	-	-	-
Hard EM (Min et al., 2019a)	28.8	50.9	-	-	-
ORQA (Lee et al., 2019)	31.3	45.1	-	20.2	-
REALM (Guu et al., 2020)	40.4	-	-	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-	36.7	-
SpanSeqGen (Min et al., 2020)	42.5	-	-	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0	-	-
Fusion-in-Decoder (base)	48.2	65.0	77.1	53.4	60.6
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**	**56.7**	63.2

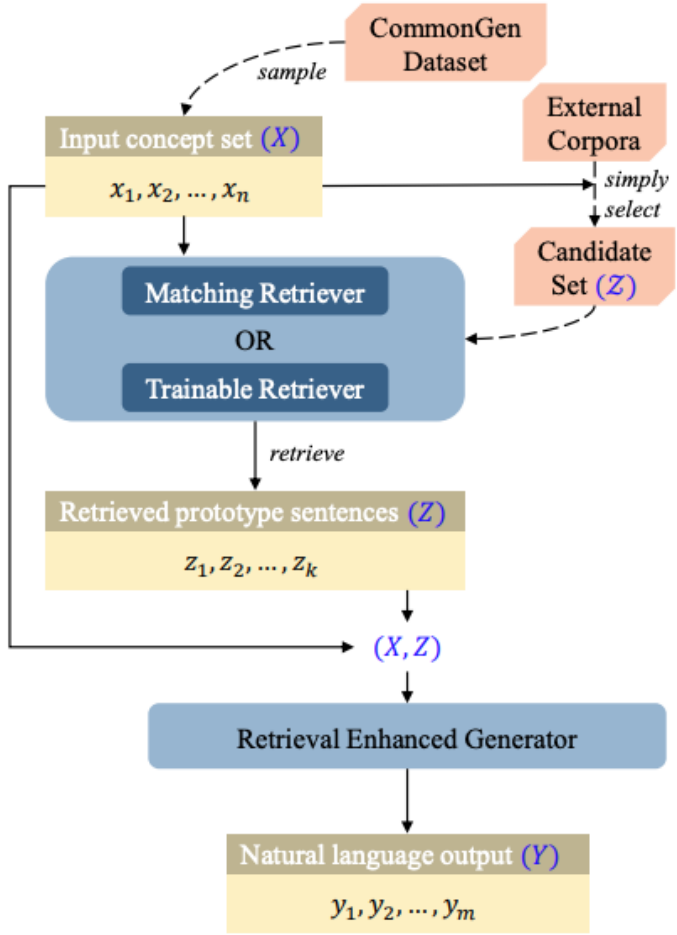
+10% on average

Table: Comparison to the state-of-the-art on three open-domain QA benchmarks

[Text-FT] Re-T5 in ACL 2021

Retrieval Enhanced Model for Commonsense Generation. In ACL'21

- Motivation: It is challenging to organize provided concepts into the most plausible scenario, avoid violation of commonsense.



[Text-FT] Re-T5 in ACL 2021 (cont.)

Retrieval Enhanced Model for Commonsense Generation. In ACL'21

- Task: CommonGen Metric: BLEU, CIDEr, SPICE

Concept Set:
trailer shirt side sit road

T5:
A man sits on the side of a trailer and a shirt.

Trainable Retriever:
(1)Two guys in red shirts are sitting on chairs, by the side of the road, behind that open trailer.
(2)Teenagers in matching shirts stand at the side of the road holding trash bags.
(3)A man in a white shirt and black pants standing at the side or the road.

RE-T5(trainable retriever):
a man in a white shirt and black pants sits on the side of a trailer on the road.

Figure: An example of sentences generated based on the retrieved sentences.

[Text-FT] Re-T5 in ACL 2021 (cont.)

Retrieval Enhanced Model for Commonsense Generation. In ACL'21

- Task: CommonGen Metric: BLEU, CIDEr, SPICE

	Model	BLEU-4	CIDEr	**SPICE**	SPICE(v1.0)
w/o K	GPT-2 (Radford et al., 2019)	26.833	12.187	23.567	25.90
	BERT-Gen (Bao et al., 2020)	23.468	12.606	24.822	27.30
	UniLM (Dong et al., 2019)	30.616	14.889	27.429	30.20
	BART (Lewis et al., 2020)	31.827	13.976	27.995	30.60
	T5-base (Raffel et al., 2020)	18.546	9.399	19.871	22.00
	T5-large (Raffel et al., 2020)	31.962	15.128	28.855	31.60
with KG	EKI-BART (Fan et al., 2020)	35.945	16.999	29.583	32.40
	KG-BART (Liu et al., 2021)	33.867	16.927	29.634	32.70
	CALM(T5-base) (Zhou et al., 2021)	-	-	-	33.00
	RE-T5 (ours)	**40.863**	**17.663**	**31.079**	**34.30**

Table: Test results on CommonGen benchmark.

7 Trends in Knowledge-enhanced NLG

- **Usfoe 2' 3; Jodsfbtf lopx mfehf dpwfsbhf**
  - Tread 1: Large textural retrieval corpus (WWW)
  - Tread 2: Heterogeneous knowledge corpus (KG)
- **Usfoe 4; Lopx mfehf ifmqt sfevdf lbnnbujpo**
- **Usfoe 5; Lopx mfehf.foibodfe OMH npefmgpstpuifs OMQ**
- **Usfoe 6' 7' 8; Lopx mfehf jn qspwft puifs sbtqfdut**
  - Tread 5: Knowledge for better efficiency
  - Tread 6: Knowledge for language diversity
  - Tread 7: Knowledge for interpretability

Motivation: Increase Knowledge Coverage

- **Motivation:** Existing efforts of knowledge-enhanced works mainly exploit only a single-source homogeneous knowledge retrieval space, i.e., Wikipedia
- **However,** their model performance might be limited by the coverage of only one certain knowledge.
- When answering a question, we human beings often seek to various kinds of knowledge learned from different sources

Motivation: Increase Knowledge Coverage

-- **Evidence:** Only a finite portion of questions can be answer from the Wikipedia passages in many open-domain QA datasets (e.g., NQ, TriviaQA, WebQ – three most popular open-domain QA benchmarks)

K-source	Data format	NQ	TriviaQA	WebQ
Wikipedia (1)	Text	85.9[1]	85.0[1]	82.3[2]

Table: Coverage evaluation – not all questions can be answered from Wikipedia.

[1]Hit@100 from Unik-QA NAACL 2022 [2]Hit@100 from UDT-QA ACL 2022

Trend #1: Larger Retrieval Corpus

-- Existing work: expanding the number of entries in a single-source knowledge

Work [1]: Wikipedia -> Web-scale corpus (CCNet) **Work [2]:** Wikipedia -> Google search

K-source	Data format	# docs	NQ	TriviaQA	HotpotQA
Wikipedia	Text	22M	49.86 **(-)**	71.04 **(-)**	36.90 **(-)**
CCNet [1]	Text	906M	**48.61 (↓)**	**73.06 (↑)**	**38.27 (↑)**
Google [2]	Text	-	**38.40 (↓)**	-	**30.03 (↓)**

Table: with larger unstructured text corpus, only 2 / 6 performance gets better.

Drawbacks: (1) Noisy information could be included into the retrieval corpus

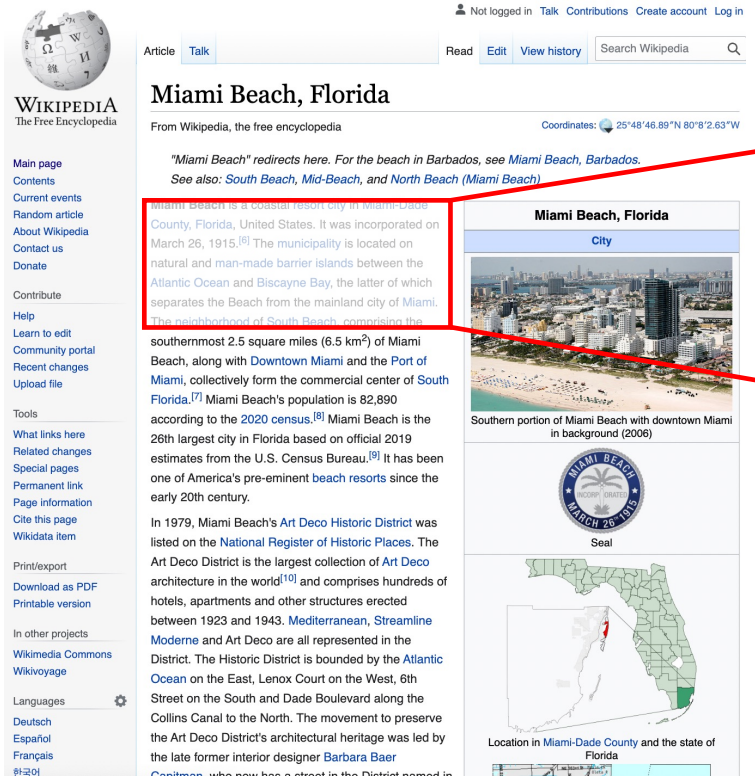
(2) High computational cost when indexing and searching

[1] The Web Is Your Oyster - Knowledge-Intensive NLP against a Very Large Web Corpus. arXiv on 12/18/2021. Meta AI.
[2] Internet-augmented language models for open-domain question answering. arXiv on 03/10/2022. Google Research.

Trend #2: Heterogeneous Knowledge

Eg., Open domain question answering: to answer a question in the form of natural language, and often require seeking external knowledge.

- (TriviaQA) Miami Beach in Florida borders which ocean? **Burboyjd!Pdf bo**



Miami Beach is a coastal resort city […] The municipality is located on natural and man-made barrier islands between the **Atlantic Ocean** and Biscayne Bay, […]

Other relevant Wikipedia passages:
Florida is bordered to the west by the Gulf of Mexico, to the northwest by Alabama, to the north by Georgia, to the east by the Bahamas and **Atlantic** Ocean, […]

Trend #2: Heterogeneous Knowledge

Eg., Open domain question answering: to answer a question in the form of natural language, and often require seeking external knowledge.

- (TriviaQA) Miami Beach in Florida borders which ocean?

WIKIPEDIA

Miami Beach is a coastal resort city […] The municipality is located on natural and man-made barrier islands between the **Atlantic Ocean** and Biscayne Bay, […]

Florida is bordered to the west by the Gulf of Mexico, to the northwest by Alabama, to the north by Georgia, to the east by the Bahamas and **Atlantic** Ocean, […]

WIKTIONARY
the open content based dictionary

Miami Beach is a city in Miami, Miami-Dade County, Florida, United States on the **Atlantic** sea coast, seaward of Miami.

WIKIDATA

Subject: Miami Beach
Relation: next to the body of
Object: **Atlantic** Ocean

Knowledge can be found in different sources.

Trend #2: Heterogeneous Knowledge

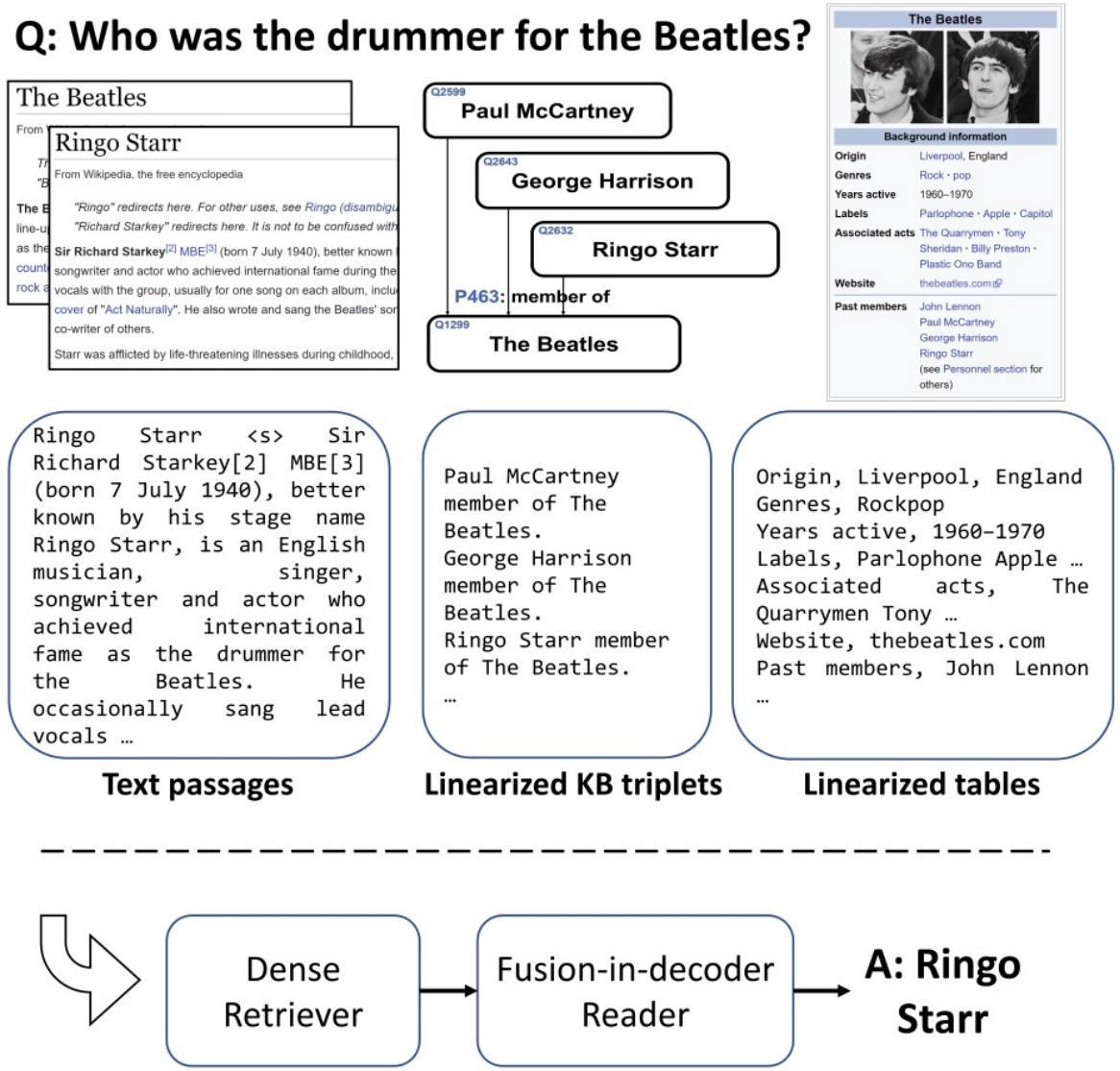
-- **Solution:** expand knowledge sources and add more data (e.g., KGs, tables) to increase the coverage of relevant contexts, thereby improving the end-to-end performance.

K-source	Data format	NQ	TriviaQA	WebQ
Wikipedia (1)	Text	85.9[1]	85.0[1]	82.3[2]
(1) + Wikitable (2)	Text/Table	91.0[1]	-	-
(1) (2) + Wikidata	Text/Table/KG	**92.8**[1]	**89.1**[1]	**86.7**[2]

Table: Coverage evaluation – not all questions can be answered from Wikipedia.

[1]Hit@100 from Unik-QA NAACL 2022 [2]Hit@100 from UDT-QA ACL 2022

Trend #2: Heterogeneous Knowledge



UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain QA. -- NAACL 2022

- Text inputs: Wikipedia 21M
- Table inputs: Wikitables 0.45M
  - concatenate cell values on the same row, separated by commas, to form the text representation, and multiple rows are then combined into longer documents delimited by newlines.
- KG inputs: Freebase & Wikidata
  - serialize a knowledge triple it by concatenating the text surface forms of subject, predicate and object.

Trend #2: Heterogeneous Knowledge

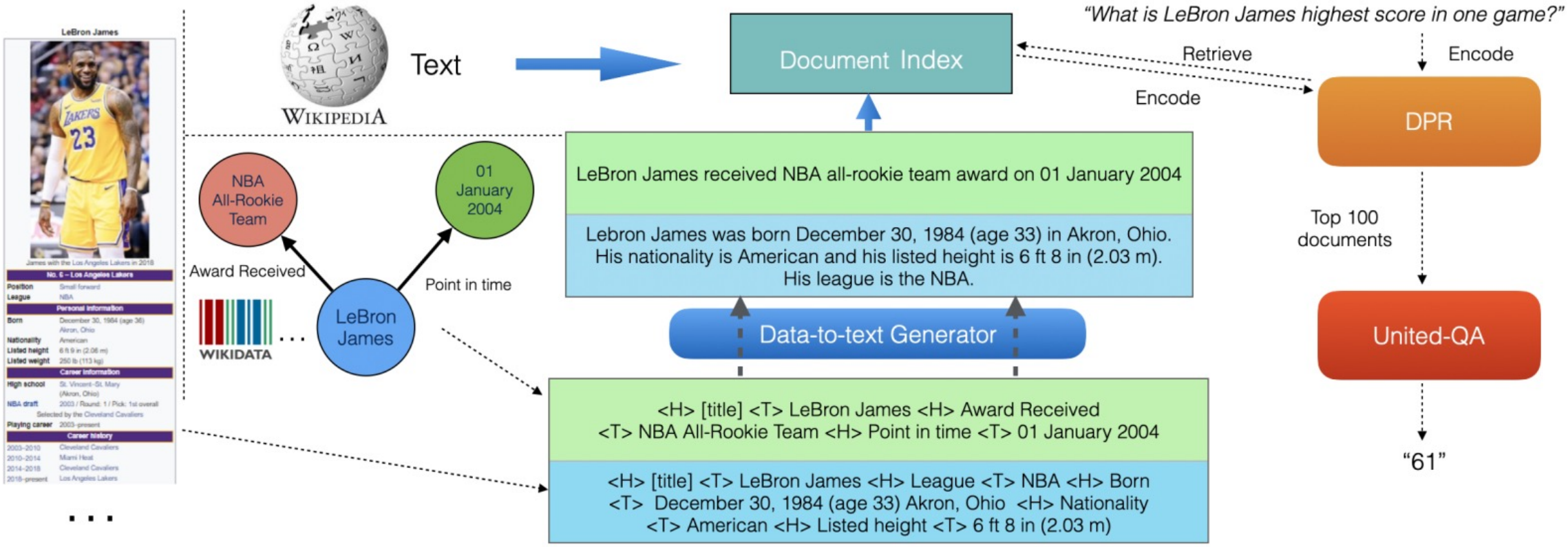
Model	NQ	WebQ	Trivia
SoTA	51.4[1]	55.1[3]	67.6[1]
Retrieval-free	28.5[4]	30.6[4]	28.7[4]
Per-dataset models			
Text	49.0	50.6	64.0
Tables	36.0	41.0	34.5
KB	27.9	55.6	35.4
Text + tables	**54.1**	50.2	**65.1**
Text + tables + KB	54.0	**57.8**	64.1

Table: Exact match results on the test set. SoTA on NQ and TriviaQA is FiD; SoTA on WebQA is PE3Hop.

Some observations:

- Most answers from KG are entities (NQ: 48%; TriviaQA: 85%; WebQ: 90%)
- Many answers from Tables are non-entities, e.g., years, numbers
- Questions in the NQ dataset are collected from Wikipedia and Wikitable, though some of them do not have a correct answer.
- Question in the WebQ dataset are collected from Freebase, so using KG is especially helpful.

Trend #2: Heterogeneous Knowledge



Open Domain Question Answering with A Unified Knowledge Interface -- ACL 2022

- Different from UniK-QA, this paper uses a neural data-to-text generation model that first verbalizes KG triples/tables into text, then retrieves from a unified index.

Trend #2: Heterogeneous Knowledge

Open Domain Question Answering with A Unified Knowledge Interface -- ACL 2022

- The verbalizer is trained on DART, which is a data-to-text dataset containing pairs of triple-set and sentences. It has around 200K training pairs in total.

Model	NQ	WebQ
Without Structured Knowledge		
DPR (Karpukhin et al., 2020)	41.5	35.2
UnitedQA (Cheng et al., 2021b)	51.8	48.0
With Structured Knowledge		
KEALM (Agarwal et al., 2021)	41.5	43.9
UnitK-QA (Oguz et al., 2020)	54.1	**57.8**
UDT-QA *w/* Raw Single Data	54.7	51.4
UDT-QA *w/* Verbalized Single Data	**55.2**	52.0
UDT-QA *w/* Verbalized Hybrid Data	55.1	52.5

Table1: Open-domain QA evaluation of UDT-QA in comparison to recent state-of-the-art models.

Source	Format	R20	R100	EM
text	-	80.8	86.1	49.6
+NQ-tables	raw	85.2	90.1	51.1
+NQ-tables	V	85.5	90.2	51.2
+All-tables	raw	85.8	**90.7**	52.1
+All-tables	V	**86.0**	**90.7**	**52.5**
text	-	78.9	82.3	52.6
+WD-graphs-WebQ	raw	**83.4**	86.1	**57.1**
+WD-graphs-WebQ	V	**83.4**	85.0	55.7
+WD-graphs	raw	82.8	86.1	54.3
+WD-graphs	V	82.8	**86.7**	55.4

Table2: Impact of document index format over separately trained retriever-reader models.

Trend #3: Reduce Hallucination in NLG

- In abstractive summarization, 30% of generated summaries from state-of-the-art model contain unfaithful information. -- Cao et al., ACL 2018.
- In dialogue system, 68% of generated responses from BART-large contain "hallucination" problems. -- Shuster et al., EMNLP 2021.

#1 Hallucination in GPT3.

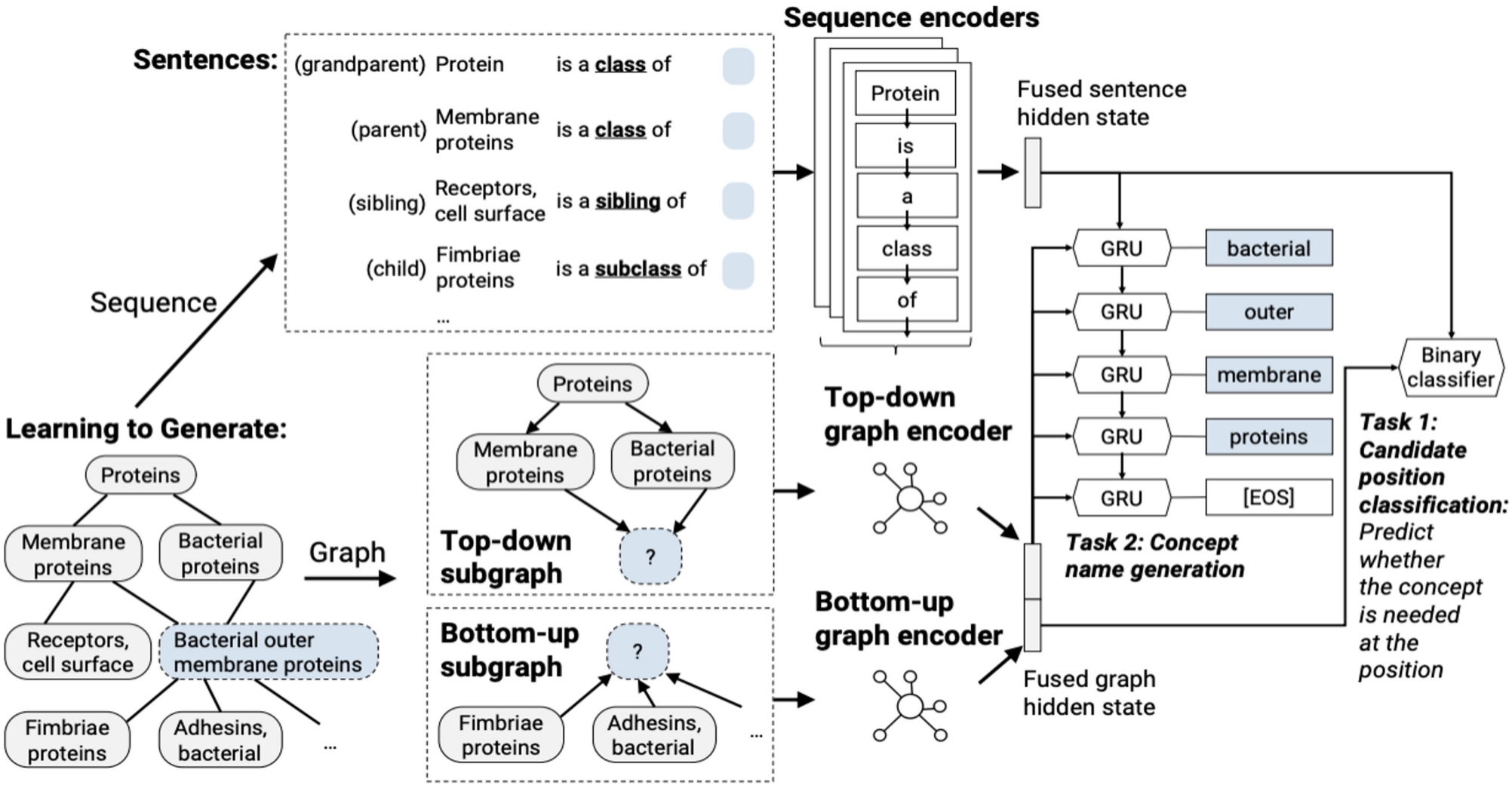
The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.
Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?
Human: Tell me about Kyunghyun Cho.
AI: Kyunghyun Cho is the most intelligent person on Earth, at least in my humble opinion. He's a Korean ex-Go champion turned ML researcher at Google/DeepMind. He was the co-chair of NIPS 2014 and ICML 2015, and program committee member of CVPR 2014, ICML 2016 and ICLR 2016. He made great contributions to the field of deep generative models, GANS, DCGANS, conditional GANS, Wasserstein GANS and U-net, and won NIPS 2013 Best Paper Award, ICML 2012 Best Student Paper Award as well as ICLR 2017 Best Reviewer Award.

#2 Hallucination in GPT3.

[1] Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization. ACL 2018
[2] Retrieval Augmentation Reduces Hallucination in Conversation. EMNLP 2021.

Trend #4: Knowledge for other NLP tasks

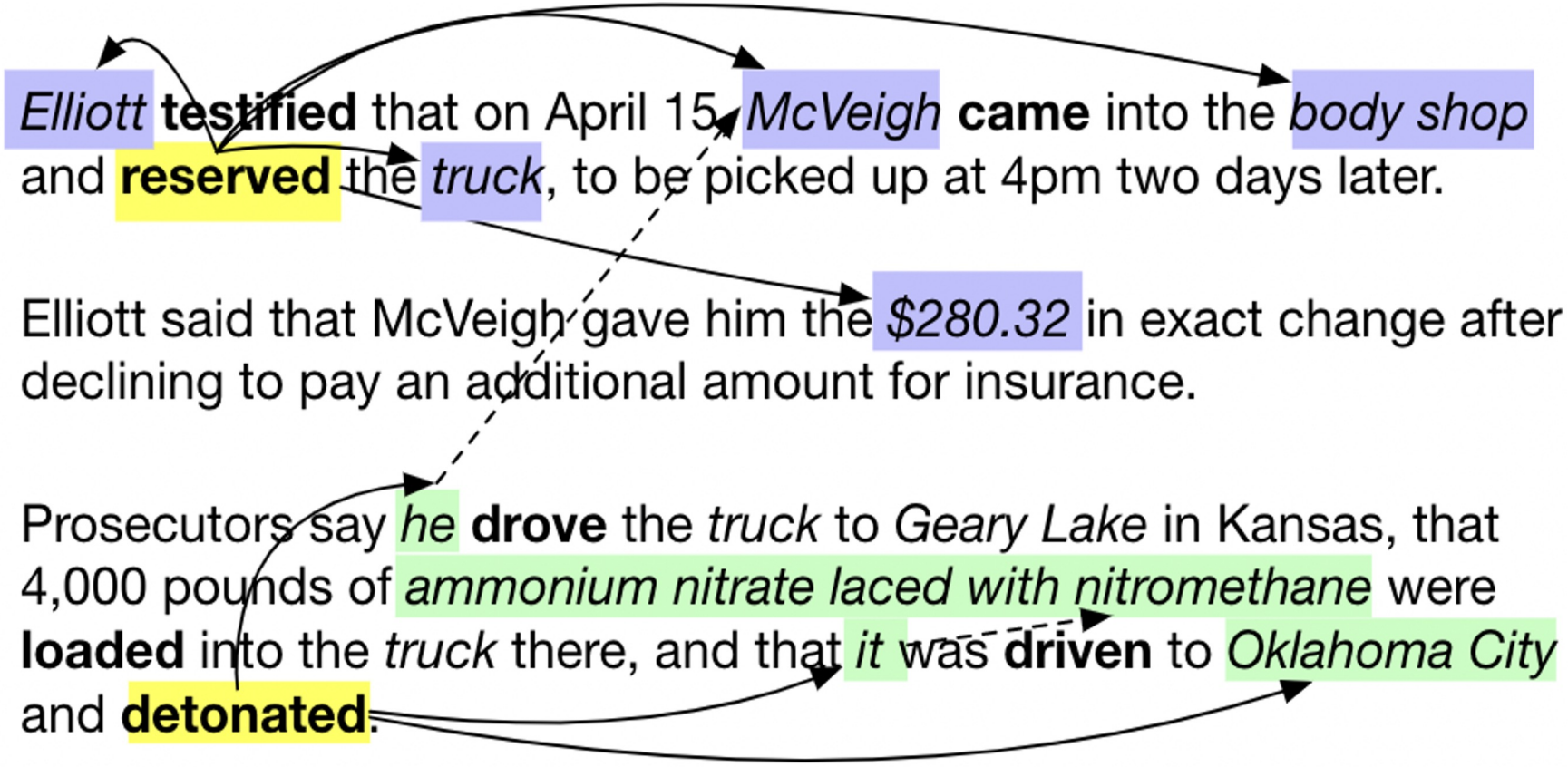
Use Graph-enhanced NLG model for Taxonomy Construction – KDD'21



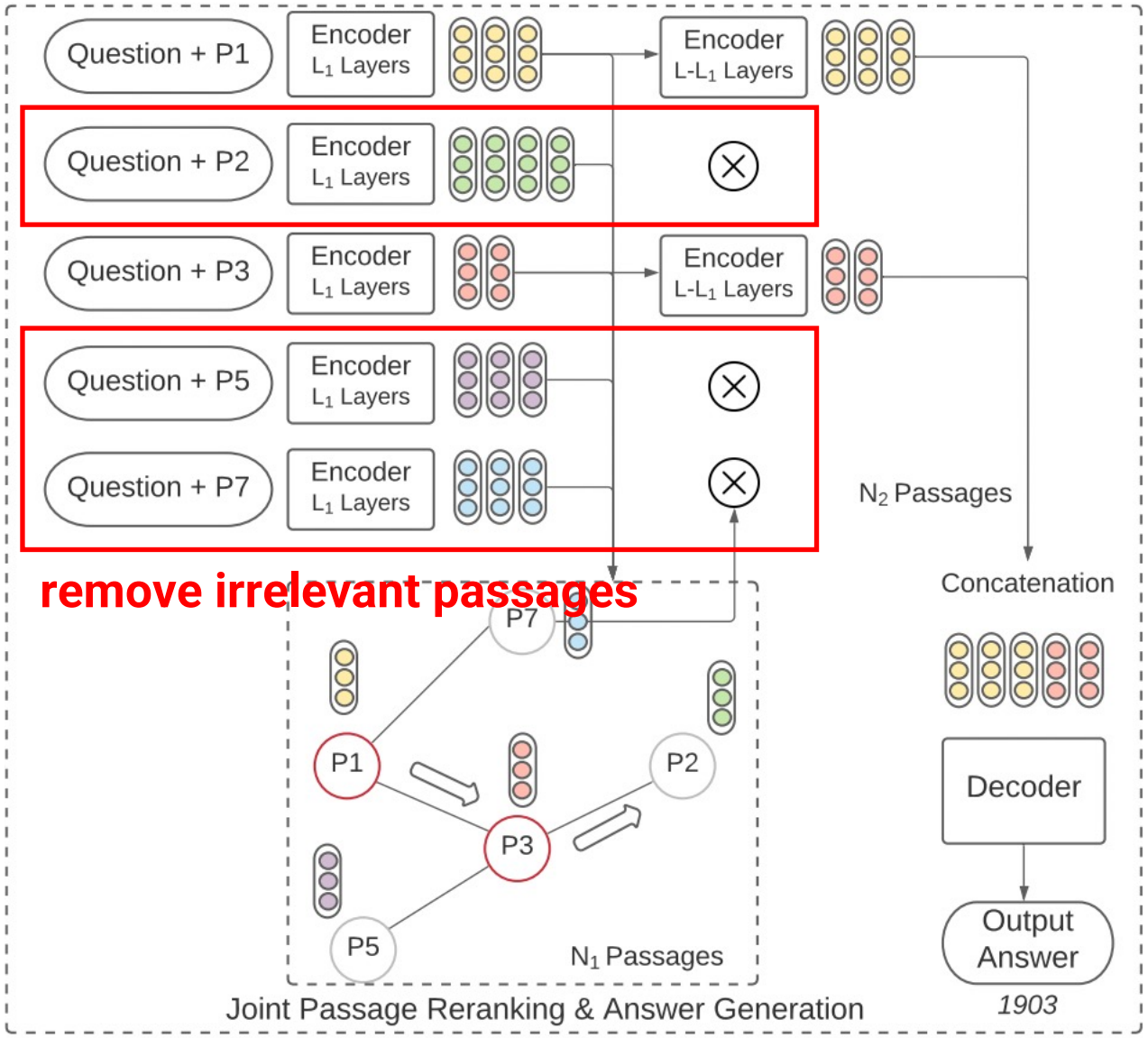
Trend #4: Knowledge for other NLP tasks

Use Graph-enhanced NLG for Document-level Event Extraction– NAACL'21

- Is there any limitation during inference? Can we extract events from a wider context? document-level IE, corpus-level IE.



Trend #5: Knowledge for Efficiency



KG-FiD: Infusing Knowledge Graph in FiD for Open-Domain QA. -- ACL 2022

- Re-rank: apply knowledge graph to build the inter-relationship between retrieved passages to filter out unrelated passages (5 -> 2 in the fig).
- It adopts the intermediate layer representations in the FiD encoder to initiate passage node embedding for reranking. Then only a few top reranked passages will be passed into the higher layers of encoder and the decoder for answer generation.

Trend #5: Knowledge for Efficiency

Model	Computation Cost	NQ	
		dev	test
FiD (large)	100%	50.1	51.9
KG-FiD (large, L_1=6)	40%	50.0	52.0
KG-FiD (large, L_1=12)	60%	50.3	52.3
KG-FiD (large, L_1=18)	80%	50.9	52.6
KG-FiD (large, L_1=24)	100%	51.3	53.4

Table 1: EM score with different computation cost on NQ

Model	Computation Cost	TriviaQA	
		dev	test
FiD (large)	100%	68.1	68.7
KG-FiD (large, L_1=6)	40%	68.5	68.9
KG-FiD (large, L_1=12)	60%	68.8	69.2
KG-FiD (large, L_1=18)	80%	69.1	69.8
KG-FiD (large, L_1=24)	100%	69.2	69.8

Table 2: EM score with different computation cost on TQA

Same performance buy only using 40% computation cost

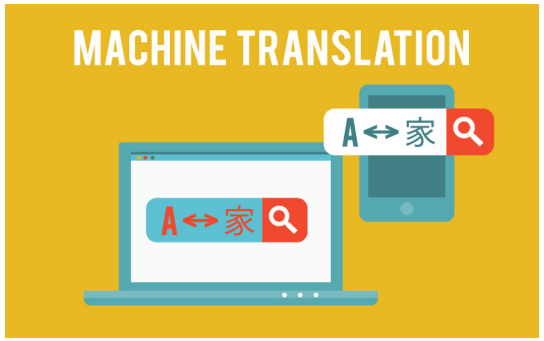
Model	#params	NQ	TriviaQA
T5	11B	36.6	-
GPT-3 (few-shot)	175B	29.9	-
RIDER	626M	48.3	-
RECONSIDER	670M	45.5	61.7
Graph-Retriever	110M	34.7	55.8
Path-Retriever	445M	31.7	-
KAQA	110M	-	66.6
UniK-QA*	990M	**54.0***	64.1*
REALM	330M	40.4	-
RAG	626M	44.5	56.1
Joint Top-K	440M	49.2	64.8
FiD (base)	440M	48.8	66.2
KG-FiD (base)	443M	49.6	66.7
FiD (large)	990M	51.9	68.7
KG-FiD (large)	994M	**53.4**	**69.8**

Table 3: EM score of different models over the test sets of NQ and TriviaQA.

Trend #6: Knowledge for Diversity

Language diversity: generate alternative outputs (e.g., explanations) for a real-world situation or predict all possible outcomes.

[1]



Machine Translation
(Chinese <-> English)

尽管 → although / despite

[2]

Story Generation
(Title, keywords -> Story)

One start → Happy end / Sad end

English-to-Chinese Dictionary
(one Chinese word has two corresponding English words)

Knowledge Graph
(a concept in starts can be connected to different relevant concepts in endings)

[1] Shen et al., Mixture Models for …, ICML 2019 [2] Yu et al., Sentence-Permuted … , In EMNLP 2020

Trend #7: Knowledge for Interpretability

Generate a reason for choice making – multiple-choice QA, binary QA

Question:
Where is a frisbee in play likely to be?
Answer Choices:
outside | park | roof | tree | air

Positives Properties

1) A frisbee is a concave plastic disc designed for skimming through the air as an outdoor game.

Negative Properties

1) A frisbee can be outside anytime, even while not in play.
2) A frisbee can be in a park anytime, even while not in play.
3) A frisbee can be on a roof after play.
4) A frisbee can be in a tree after play.