Thesis Proposal

UNIVERSITY OF NOTRE DAME

Knowledge Retrieval and Fusion for Knowledge-intensive NLP Tasks

Wenhao Yu 05.13.2022

Advisor:

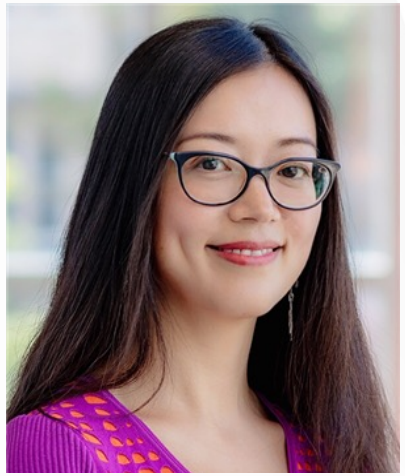
Wenhao Yu
Univ. Notre Dame

Dr. Meng Jiang
Univ. Notre Dame

Committee members (in alphabetic order):

Dr. Nitesh Chawla
Univ. Notre Dame

Dr. David Chiang
Univ. Notre Dame



Dr. Heng Ji
UIUC, Amazon Scholar

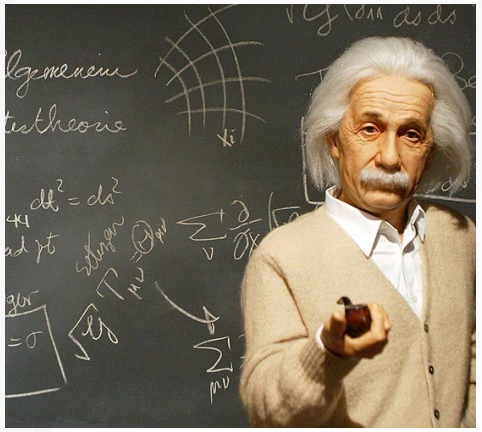
Dr. Scott Yih
Facebook AI Research

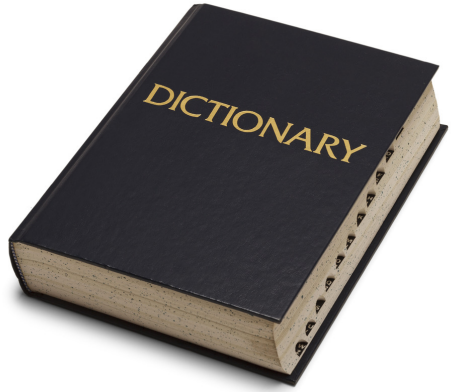
What is knowledge?

UNIVERSITY OF NOTRE DAME

From dictionary: Knowledge is a familiarity or awareness, of someone or something. It can refer to a theoretical or practical understanding of a subject.

Knowledge can be provided in many ways and from many sources





A theoretician / a chef / a language model / an encyclopedia / a dictionary

all of them contain knowledge (theory / skill / language pattens / facts / definitions)

NLP tasks are **NOT** knowledge-intensive

UNIVERSITY OF NOTRE DAME

- **Machine Translation (e.g., English to Chinese)**

Input: Who did Hawaii belong to before 1893?

Output: 夏威夷在 1893 年之前属于谁？ (from Google Translate)

- **Named Entity Recognition (classify named entities)**

Input: Who did Hawaii belong to before 1893?

Output: [location] [year]

NLP tasks are **NOT** knowledge-intensive

UNIVERSITY OF NOTRE DAME

- **Machine Translation (e.g., English to Chinese)**

Inp

Ou

- **K**

These NLP tasks heavily rely on discovering language patterns underlying the input-output pairs.

They need knowledge, but **NOT** intensive.

Input: Who did Hawaii belong to before 1893?

Output: [location] [year]

What are knowledge-intensive NLP tasks?

UNIVERSITY OF NOTRE DAME

- **Machine Translation (e.g., English to Chinese)**
- **Named Entity Recognition (classify named entities)**

- **Open-domain question answering**

Question: Who did Hawaii belong to before 1893?

Do you know the answer?

What are knowledge-intensive NLP tasks?

UNIVERSITY OF NOTRE DAME

(T1) Open domain question answering

- (NQ[16]) Who did Hawaii belong to before 1893? **Hawaiian Kingdom**

Hawaii is the most recent state to join the union, on August 21, 1959 ... **On January 17, 1893**, Queen Liliuokalani was overthrown and replaced by a provisional government. **The United States Minister to the Hawaiian Kingdom conspired with U.S. citizens to overthrow the monarchy.**

Evidence from Wikipedia page "Hawaiian Kingdom"
In 1893, United States Public Law acknowledged that "the overthrow of **Hawaiian Kingdom** occurred with the active participation of agents and citizens of the United States"

[16] Natural Questions: A Benchmark for Question Answering Research. In TACL 2019

What are knowledge-intensive NLP tasks?

UNIVERSITY OF NOTRE DAME

(T2) Commonsense fact checking: given a commonsense claim, and use commonsense knowledge to judge true or false.

- (CREAK[17]) Harry Potter can teach classes on how to fly on a broomstick

OPEN MIND Common Sense

Harry Potter is a wizard.
Harry Potter is **good at riding broomstick.**
Harry Potter is a series of fantasy novels.

ATOMIC

Someone who is **good at something can teach** it.

[17] CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. In Neurips 2021

What are knowledge-intensive NLP tasks?

UNIVERSITY OF NOTRE DAME

(T2) Commonsense fact checking: given a commonsense claim, and use commonsense knowledge to judge true or false.

- (CREAK[17]) Harry Potter can teach classes on how to fly on a broomstick

OPEN MIND Common Sense

Harry Potter is a wizard.
Harry Potter is **good at riding broomstick.**
Harry Potter is a series of fantasy novels.

Knowledge-intensive NLP tasks are challenging that usually require model/human to seek external knowledge

Outline of Proposal

UNIVERSITY OF NOTRE DAME

Knowledge Retrieval and Fusion for Knowledge-intensive NLP Tasks

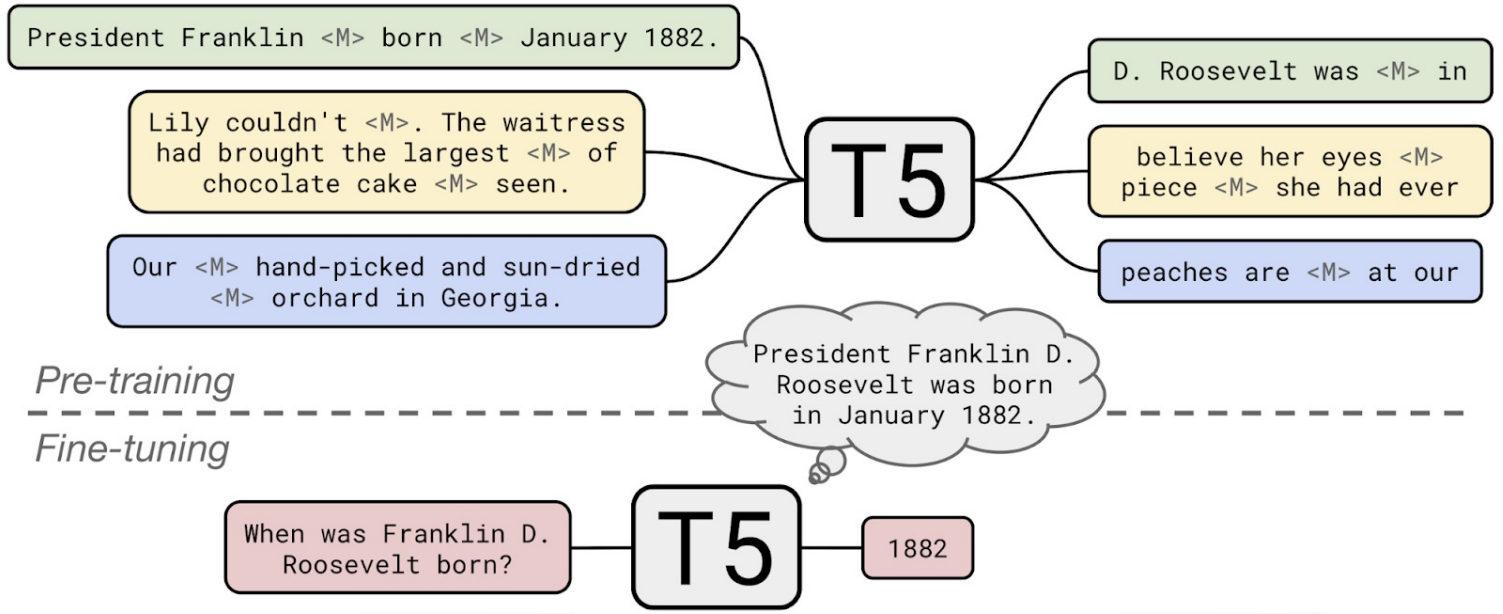
- What is knowledge?
- What are knowledge-intensive NLP tasks?
- How to solve knowledge-intensive NLP tasks?
  - Close-book
  - Open-book

How to solve knowledge-intensive tasks?

UNIVERSITY OF NOTRE DAME

Close-book Models: **Knowledge** (e.g., entity relations, commonsense) is learnt into a language model (LM) **parameters**. During fine-tuning, the LM only encodes the *input text* and make predictions.

The Close-book Methodology



Microsoft
UniLM, Turing-NLG

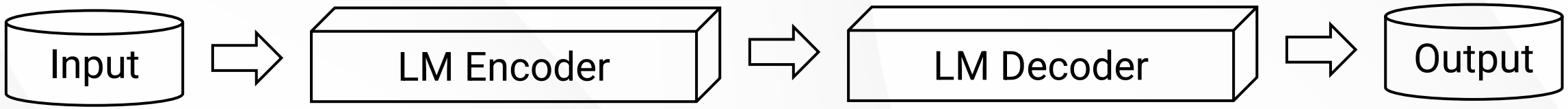
facebook
BART, XLM

OpenAI
GPT-2, GPT-3

NVIDIA
MegatronLM

Google
BERT, T5

Close-book: during fine-tuning, only feed input text into LMs and make predictions



Close-book Models: **Drawbacks**

UNIVERSITY OF NOTRE DAME

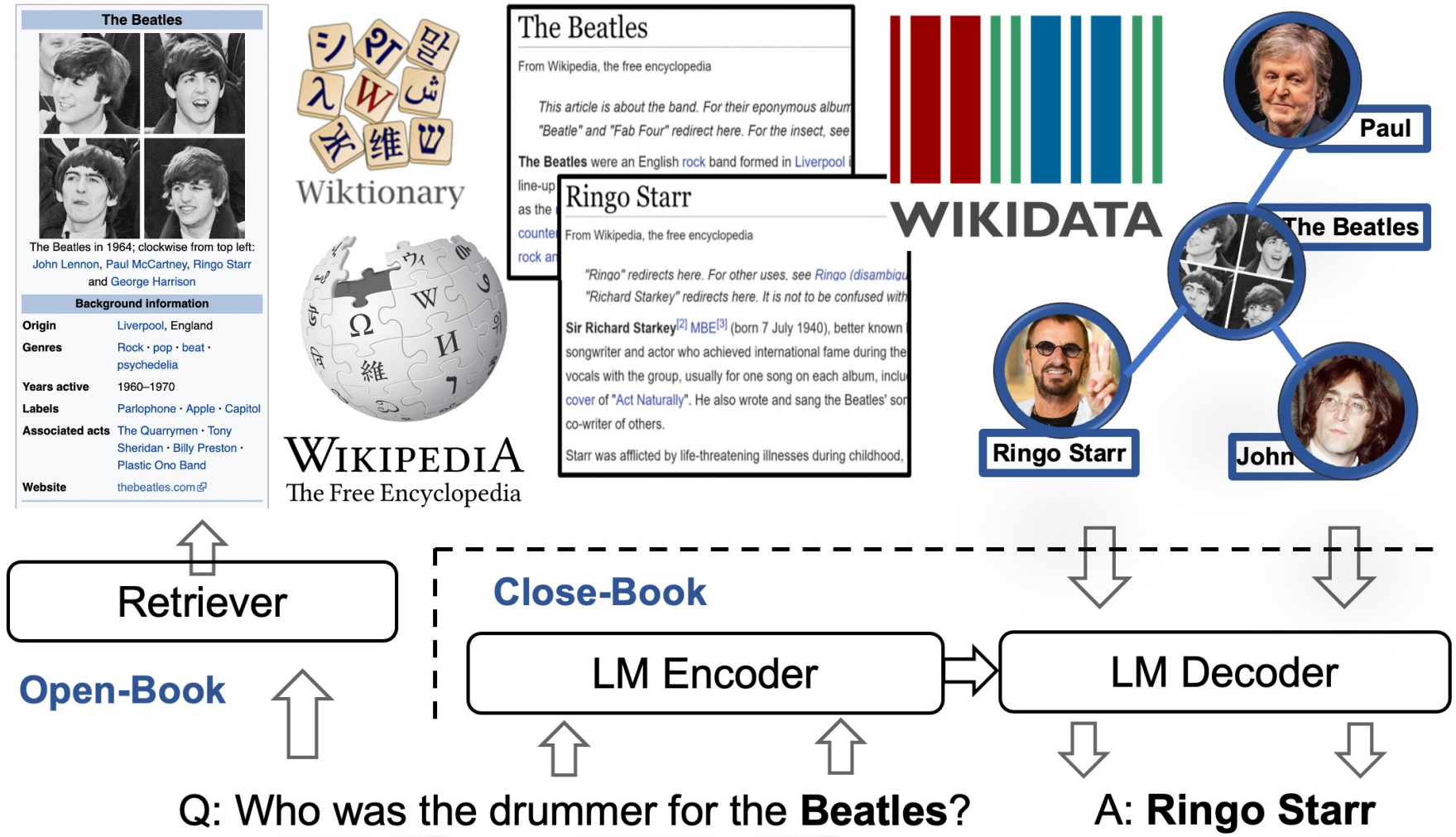
- They make predictions by only “looking up information” stored in its parameters, leading to inferior performance and interpretability.
- They are usually trained offline, rendering the model agnostic to the latest information, e.g., asking BERT (released at 2018) about COVID-19.
- They are mostly trained on general domain corpora, making them less effective on domain-specific tasks, e.g., asking BERT about biology.

How to solve knowledge-intensive tasks?

UNIVERSITY OF NOTRE DAME

Open-book Models: **Knowledge** (e.g., entity relations, commonsense) is **retrieved** based on the input text. The Language model make predictions by *reading and reasoning* over the retrieved texts.

The Open-book Methodology



Open-book: during fine-tuning, not only use input text, but also resort to external knowledge

How to solve knowledge-intensive tasks?

UNIVERSITY OF NOTRE DAME

Open-book Models: **Knowledge** (e.g., entity relations, commonsense) is **retrieved** based on the input text. The Language model make predictions by *reading and reasoning* over the retrieved texts.

(Semi)structured Knowledge (e.g., Wikidata, ConceptNet)	Unstructured knowledge (e.g., Wikipedia, OMCS)
KagNet[26] – USC 2019	DPR[29] – Facebook 2020
QA-GNN[27] – Standford 2020	RAG[30] – Facebook 2020
GreaseLM[28] – Stanford 2022	DensePhrase[31] – Princeton 2021

see references of [26-31] at the last page

Open-book Models: **Advantages**

UNIVERSITY OF NOTRE DAME

Language models (LMs) and Knowledge-aware LMs (e.g., SciBERT[22], Unicorn[24])

Close-book

(Semi-)structured knowledge enhanced methods (e.g., KagNet[26], QA-GNN [27])

Unstructured knowledge enhanced methods (e.g., RAG, FiD, Unik-QA [30,32,33])

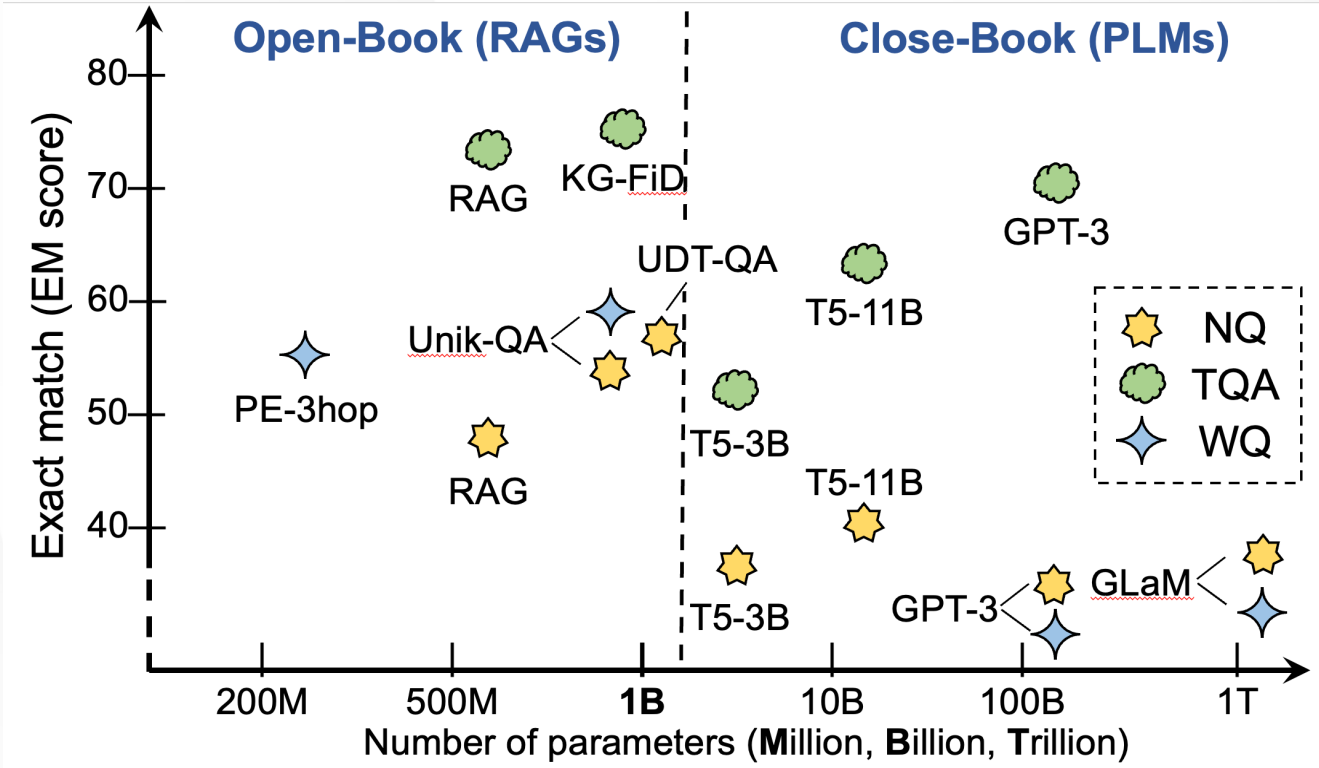
Open-book

- 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 Models: **Advantages** (cont.)

UNIVERSITY OF NOTRE DAME

- 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 Proposal

UNIVERSITY OF NOTRE DAME

Knowledge Retrieval and Fusion for Knowledge-intensive NLP Tasks

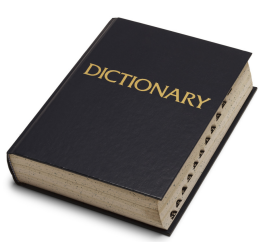
- What is knowledge?
- What are knowledge-intensive NLP tasks?
- How to solve knowledge-intensive NLP tasks?

- Knowledge retrieval – the way to obtain knowledge
- Knowledge fusion -- the way to use knowledge

Knowledge Retrieval

UNIVERSITY OF NOTRE DAME

- **Knowledge Retrieval:** used in open-book, which is the way to obtain external knowledge that is relevant to an information need (i.e., input text) from a collection of resources (e.g., Wikipedia, Wiktionary, Knowledge graph)



· **Key-value retrieval:** find the matching key, return value, e.g., word searching in dictionary

WIKIDATA

· **Entity/concept linking:** assign a unique identity to entity or concept mentioned in the text.

WIKIPEDIA

· **Sparse retrieval (e.g., BM25):** calculating similarity of two texts by sparse vector (TF-IDF)

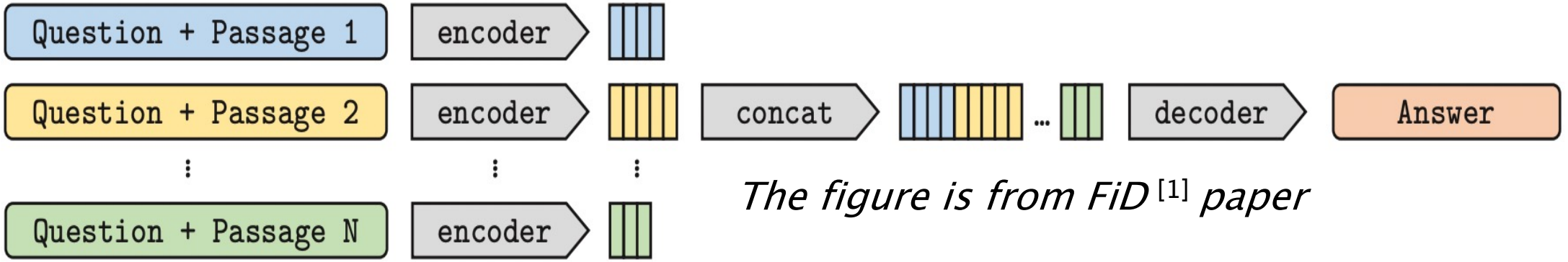
WIKIPEDIA

· **Dense retrieval (e.g., DPR):** calculating similarity of two texts by dense representation

Knowledge Fusion

- **Knowledge Fusion:** used in open-book, which is the way to integrate the retrieved knowledge with the representation learned from text inputs and produce desired outputs.

- E.g., Fusion-in-decoder model[32], Facebook AI Research 2021



The figure is from FiD [1] paper

[32] Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In EACL 2021

My Research Contributions

UNIVERSITY OF NOTRE DAME

Work on Open-book models

- Structured knowledge enhanced models
  - Citation graph for intention classification → WWW 2020
  - Knowledge graph for explanation generation → ACL 2022
  - Knowledge graph for open-domain QA → ACL 2022
- Unstructured knowledge enhanced models
  - Entity type for news generation → EMNLP 2021
  - Dictionary for language understanding and QA → ACL 2022
  - (IBM) Domain text for technical QA → NAACL 2021 Industry

WIKIDATA

WIKIPEDIA

WIKTIONARY

ConceptNet

DBpedia

OPEN MIND Common Sense

Freebase

AMiner

yago select knowledge

Outline of Proposal

UNIVERSITY OF NOTRE DAME

	Paper / Objective	Knowledge	Novelty
Published work 1	"Dict-BERT: Enhancing Language Model with English Dictionary." In ACL 2022	Unstructured (Text)	First work with a novel knowledge fusion method to use dictionary to enhance language model.
Published work 2	"Diversifying Content Generation for Commonsense Reasoning." In ACL 2022	Structured (KG)	A Novel knowledge fusion methods to leverage knowledge graphs to commonsense reasoning
Proposed work 1	Grounding Knowledge-enhanced model across Heterogeneous Knowledge.	Both (Text + KG)	A Novel knowledge retrieval and fusion method to leverage both structured and unstructured knowledge
Proposed work 2	Learning Knowledge Retriever without Ground Truth Supervision	Unstructured (Text)	A novel knowledge retrieval method to obtain knowledge without any ground truth supervision

Published Work 1

UNIVERSITY OF NOTRE DAME

Dict-BERT: Enhancing Language Model Pre-training with Dictionary

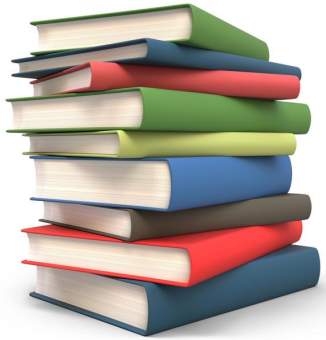
Wenhao Yu , Chenguang Zhu, Yuwei Fang, Donghan Yu,
Shuohang Wang, Yichong Xu, Michael Zeng, Meng Jiang

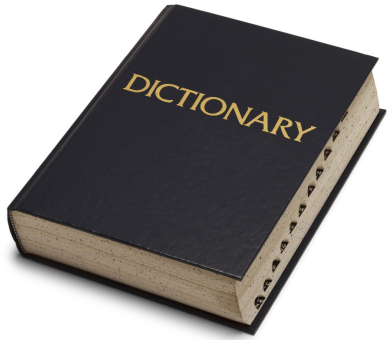
60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)

Motivation: Dictionary in LMs

UNIVERSITY OF NOTRE DAME

- Dictionary is useful **for human to learn a new language.** (similar to training a language model).

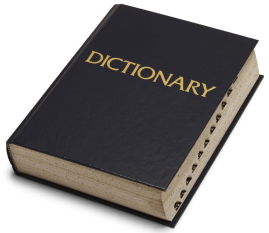




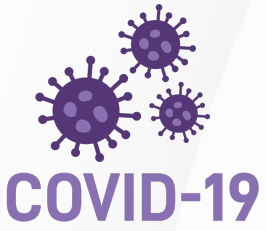
- It is **difficult** of existing language models to **understand rare words** and new words (e.g., using BERT to understand Covid-19) [13, 14].

[13] Wu et al. Taking Notes on the Fly ... ICLR 2021; [14] Schick et al., Rare Words: A Major Problem … AAAI 2020

What does a dictionary have?



- **Wiktionary**
- **Cambridge**

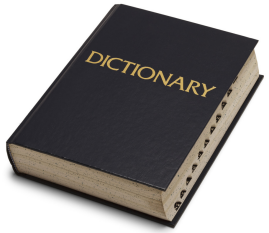


- **New word**
- **Rare word**

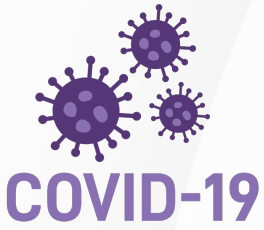
- **Definition:** Short for COVID-19, the disease caused by Severe acute respiratory syndrome-related coronavirus 2.
- **Example1:** I invite you to use the disruption of covid as a chance to disrupt your busyness. That as you navigate your journey through this crisis, you also guide your life onto a path of more color, more impact and more joy.
- **Example2:** We have such grand plans, when covid is over. Some want to get back to packed concert halls, while others want to relax comfortably at our favorite restaurants.

What does a dictionary have?

UNIVERSITY OF NOTRE DAME



- **Wiktionary**
- **Cambridge**



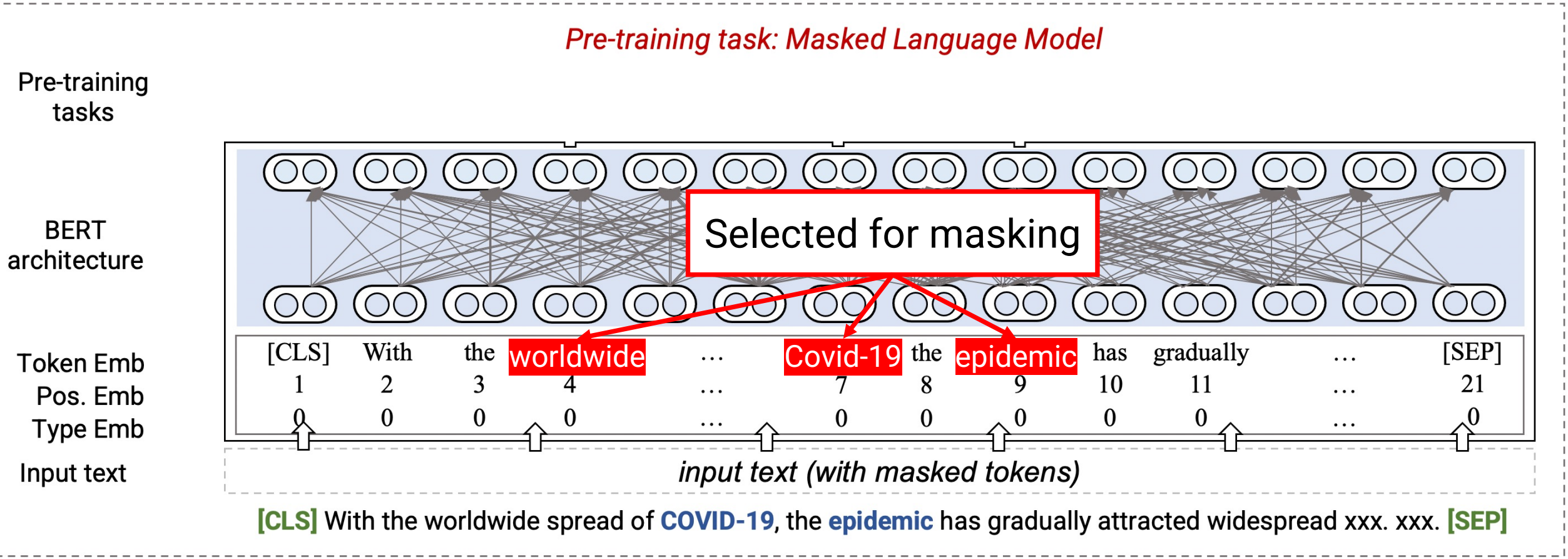
- **New word**
- **Rare word**

- **Definition:** Short for COVID-19, the disease caused by Severe acute respiratory syndrome-related coronavirus 2. **(This is used in our Dict-BERT work)**
- **Example1:** I invite you to use the disruption of covid as a chance to disrupt your busyness. That as you navigate your journey through this crisis, you also guide your life onto a path of more color, more impact and more joy.
- **Example2:** We have such grand plans, when covid is over. Some want to get back to packed concert halls, while others want to relax comfortably at our favorite restaurants.

Background: BERT Pre-training

UNIVERSITY OF NOTRE DAME

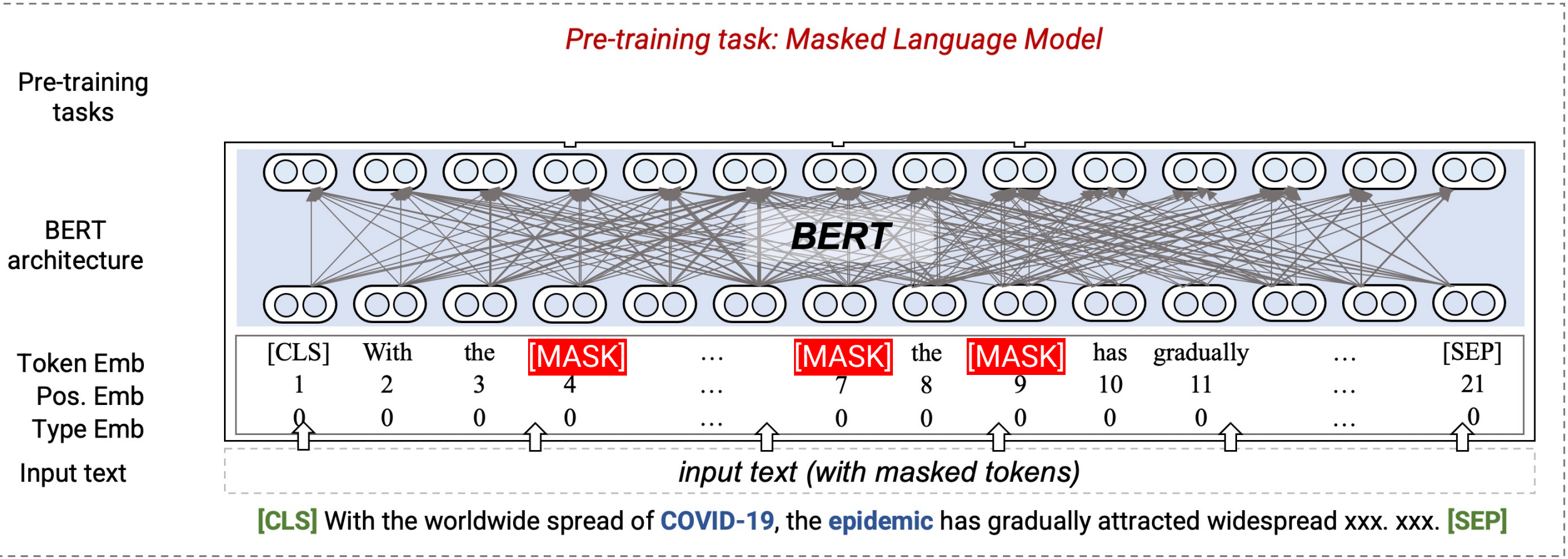
Input text: With the worldwide spread of COVID-19, the epidemic has gradually attracted widespread attention and discussion on social media platforms.



Background: BERT Pre-training

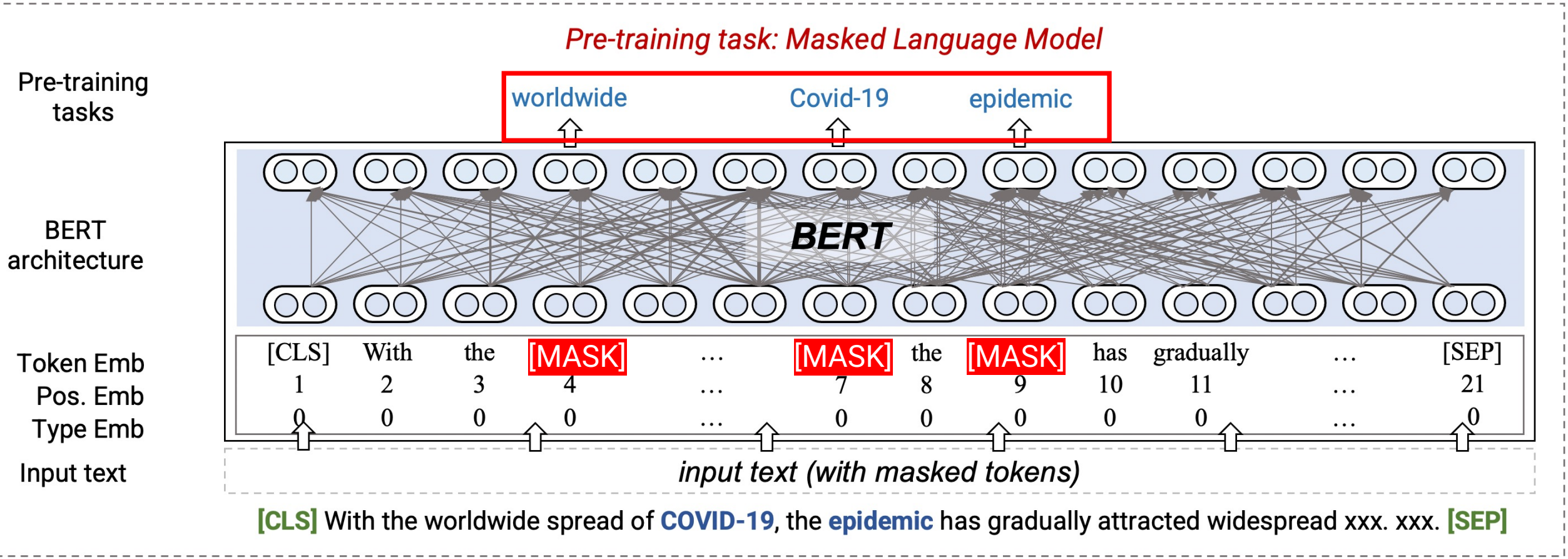
UNIVERSITY OF NOTRE DAME

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



Background: BERT Pre-training

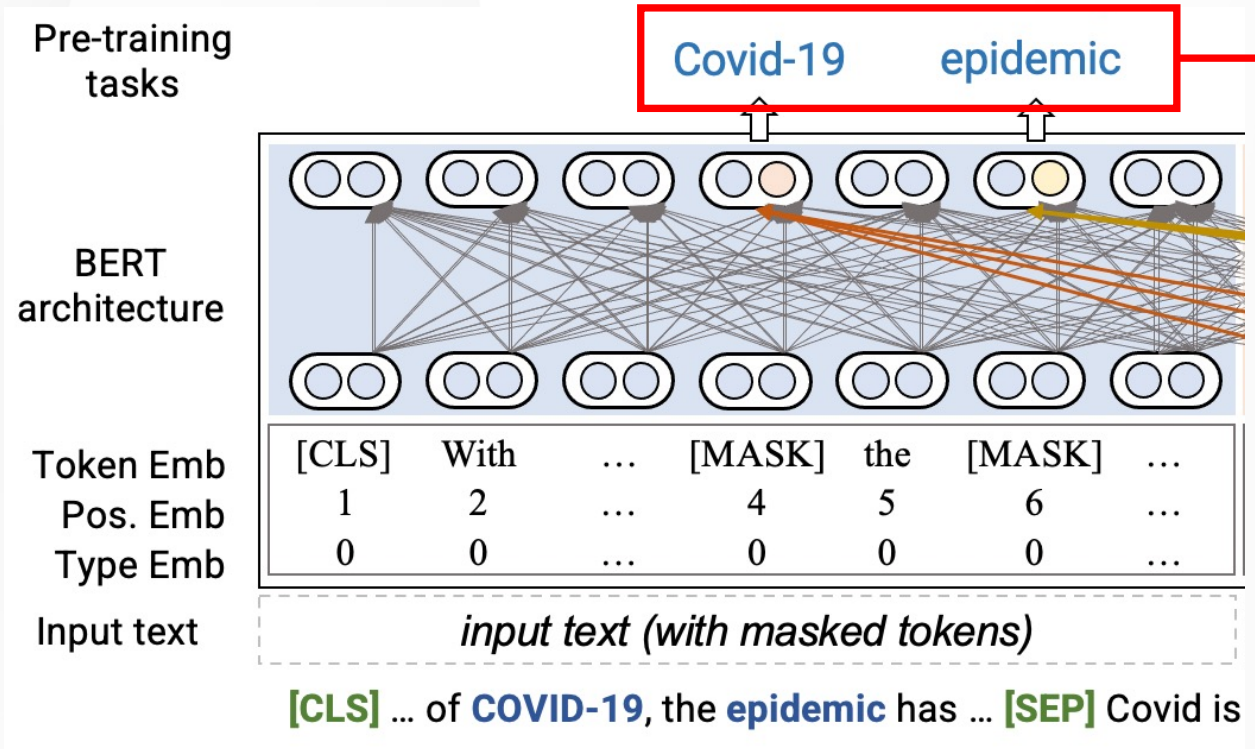
Input text: With the worldwide spread of COVID-19, the epidemic has gradually attracted widespread attention and discussion on social media platforms.



Dictionary-BERT (S1)

UNIVERSITY OF NOTRE DAME

Setting 1: Basic MLM setting **+ dictionary information** (might include negative definition)

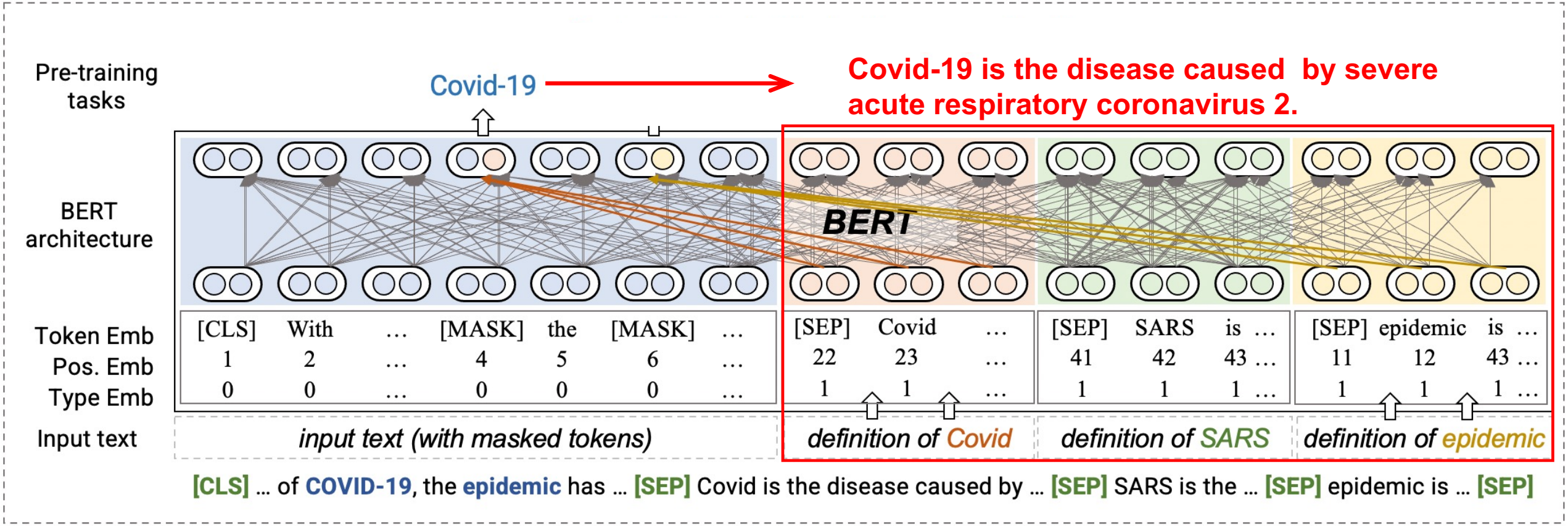


Suppose they are two rare words

Dictionary-BERT (S1)

UNIVERSITY OF NOTRE DAME

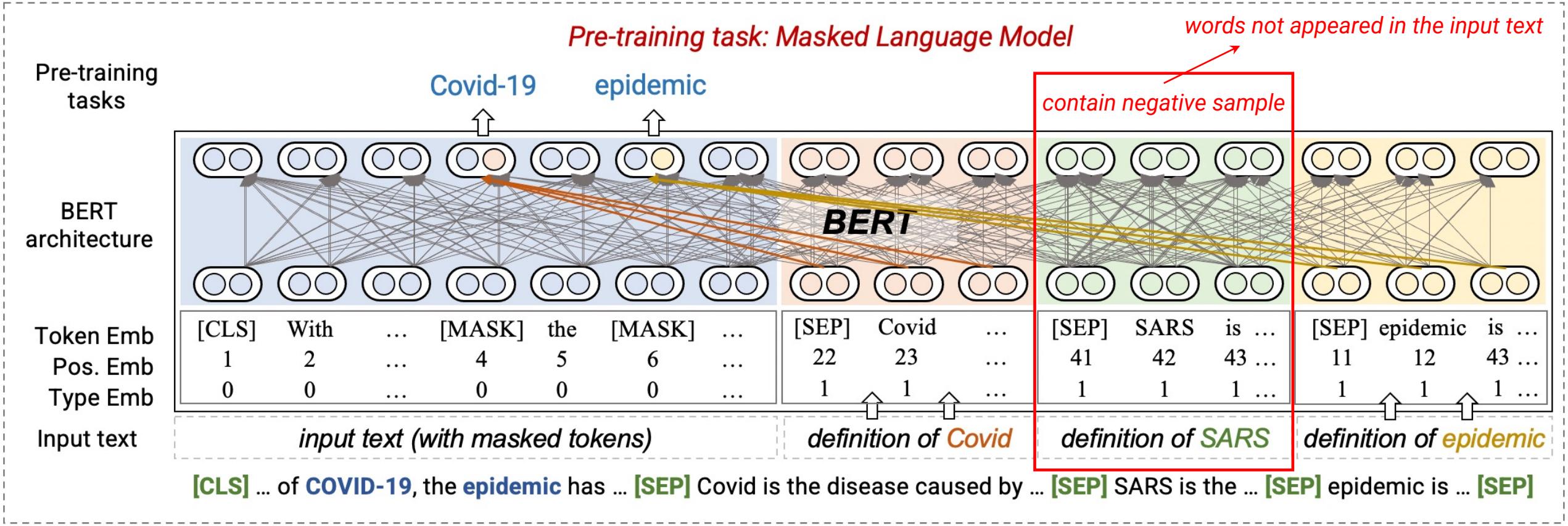
Setting 1: Basic MLM setting **+ dictionary information** (might include negative definition)



Dictionary-BERT (S1)

UNIVERSITY OF NOTRE DAME

Setting 1: Basic MLM setting **+ dictionary information** (might include negative definition)

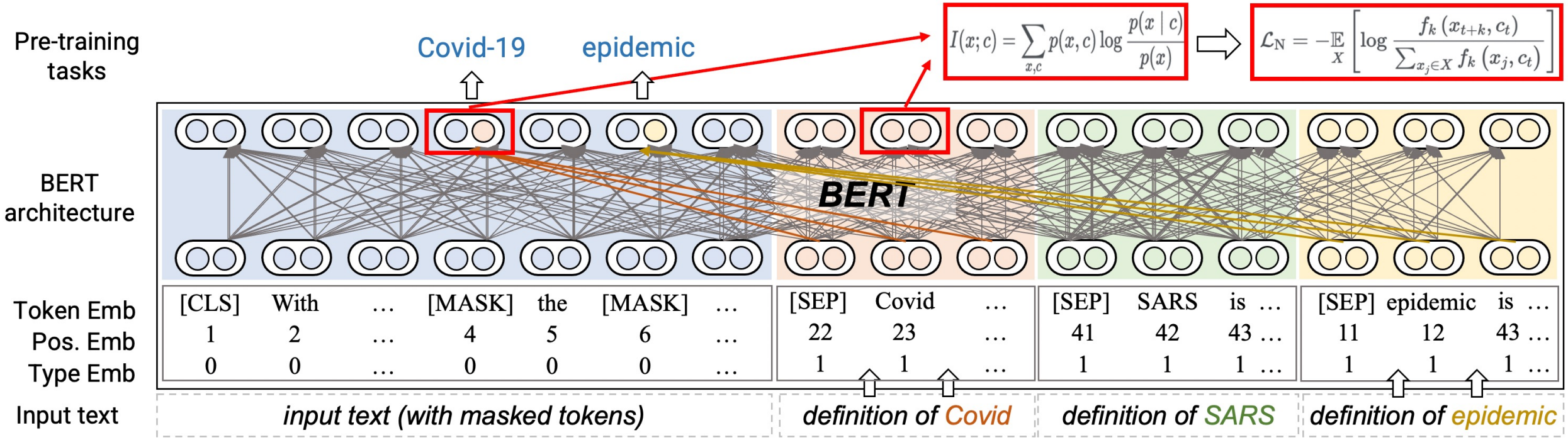


Dictionary-BERT (S2)

UNIVERSITY OF NOTRE DAME

Setting 2: Setting 1 + **word-level mutual information maximization (MIM)**

Pre-training task: Masked Language Model + Mutual Information Maximization between input and dictionary

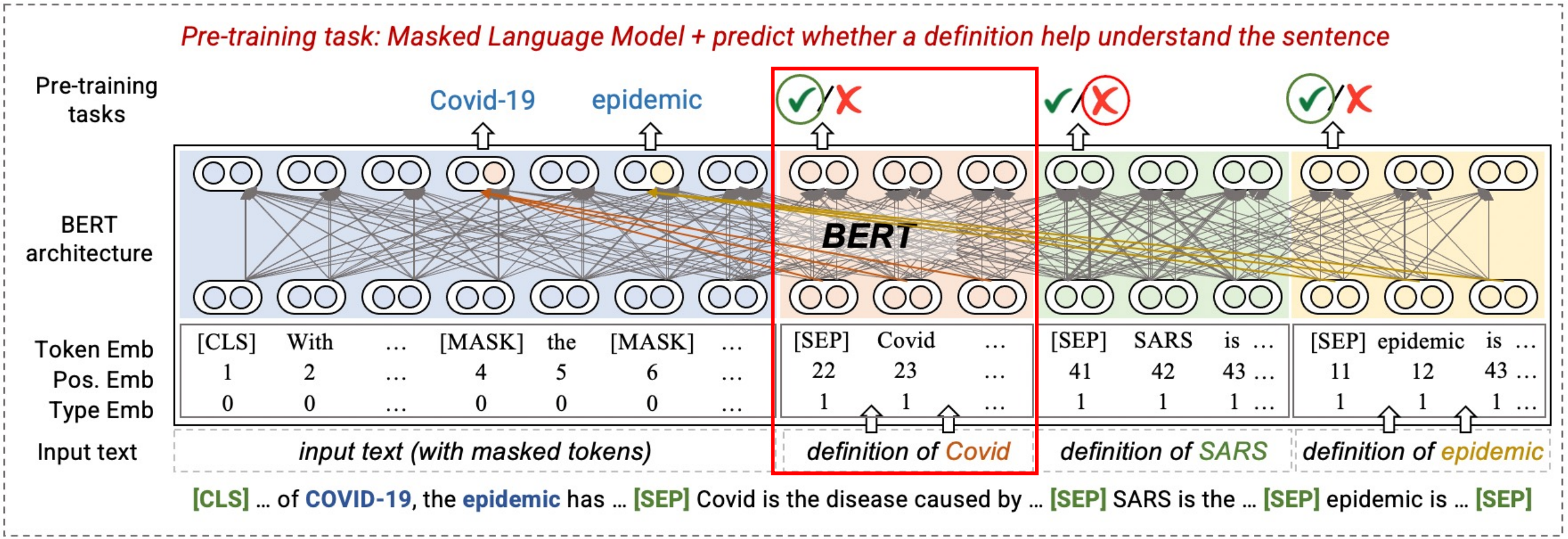


See detailed mathematical formulas are on page 103

Dictionary-BERT (S3)

UNIVERSITY OF NOTRE DAME

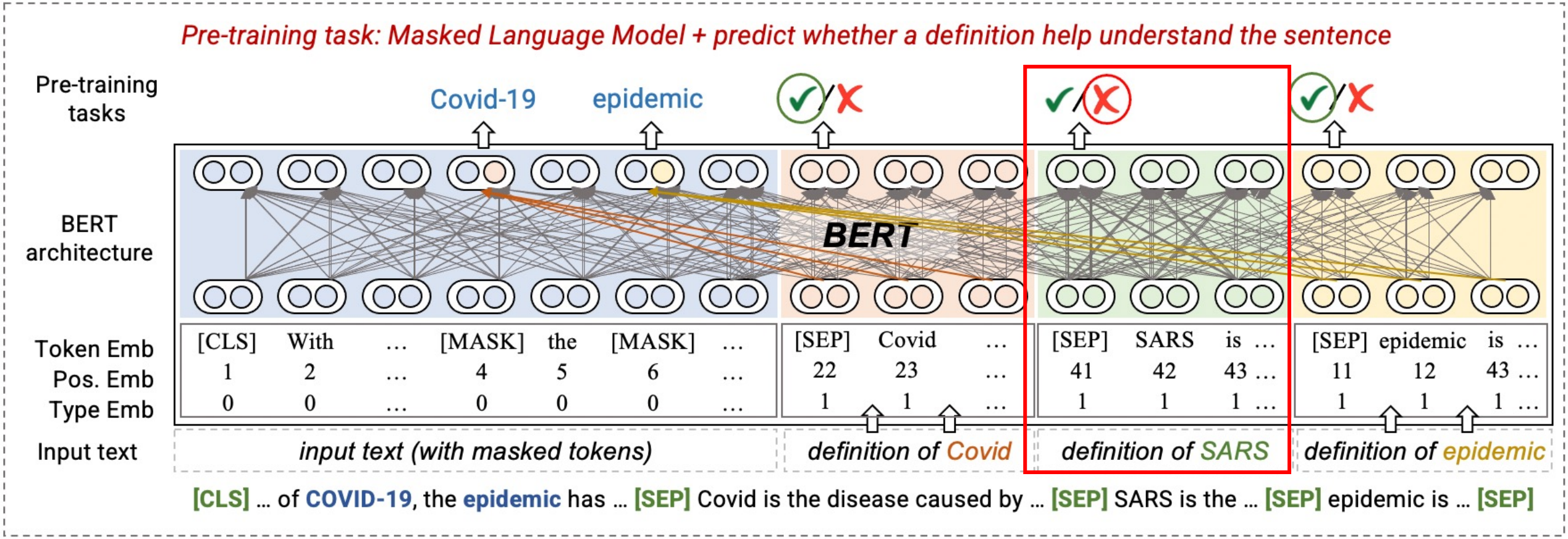
Setting 3: Setting 1 + **sentence-level definition discrimination (DD)**



Dictionary-BERT (S3)

UNIVERSITY OF NOTRE DAME

Setting 3: Setting 1 + **sentence-level definition discrimination (DD)**



Experiments: Tasks and Datasets

UNIVERSITY OF NOTRE DAME

We evaluate Dict-BERT on 16 datasets/tasks.

- **GLUE benchmark:** CoLA (linguistic acceptability); SST-2 (sentiment classification); MRPC (paraphrase prediction); STS-B (sentence similarity); MNLI (language inference); QNLI (question answering); RTE (language inference)
- **Domain adaptation benchmark:** Chemprot (relation classification); RCT (role classification); ACL-ARC (citation intent); SCIERC (relation classification); HyperParisan (partisanship); Agnews (topic prediction); Helpfulness (review helpfulness on Amazon); IMDB (sentiment classification)

Experiments: GLUE Benchmark

UNIVERSITY OF NOTRE DAME

	+DICT	+MIM	+DD	CoLA	MRPC	RTE	**Overall**
Metric				Matthew	Acc.	Acc.	**Acc.**
BERT	n.a.			59.48	90.21	68.67	**82.76**
DictBERT	✓						**83.24**
	✓	✓					**83.33**
	✓		✓	60.78	90.47	**73.34**	**83.84**
	✓	✓	✓	**61.32**	**91.14**	72.78	**83.91**

average improvement on eight sub-tasks (+1.15%)

Observation 1: Both **mutual information maximization** (MIM) and **definition discrimination** (DD) can help learning knowledge from dictionary and improve LM.

Experiments: GLUE Benchmark

UNIVERSITY OF NOTRE DAME

	+DICT	+MIM	+DD	CoLA	MRPC	RTE	Overall
Metric				Matthew	Acc.	Acc.	Acc.
BERT	n.a.			59.48	90.21	68.67	82.76
DictBERT	✓			60.58	89.55	71.27	83.24
	✓	✓		61.06	89.80	71.48	83.33
	✓		✓	60.78	90.47	**73.34**	83.84
	✓	✓	✓	**61.32**	**91.14**	72.78	83.91

These 3 datasets have less than 10K training data

Observation 2: Dict-BERT has made largest improvements on the **RTE** and **CoLA** tasks, indicating our proposed Dict-BERT is particularly useful when fine-tuned on a small datasets.

Experiments: Domain Adaptation

UNIVERSITY OF NOTRE DAME

	+DICT	+MIM	+DD	Biology	CS		New	Review	Avg
Dataset				Chemp.	ARC	SciERC	HyP.	Help.	
Metric				Mi-F1	Ma-F1	Ma-F1	Ma-F1	Ma-F1	--
BERT	n.a.			81.16	64.20	80.40	91.17	69.36	82.65
DAPT	n.a.			83.10	71.45	81.62	**94.72**	69.65	84.57
DictBERT	✓	✓		83.24	72.78	82.54	94.69	**70.43**	85.06
	✓		✓	83.33	72.26	82.70	**94.72**	70.33	85.01
	✓	✓	✓	**83.49**	**74.18**	**83.01**	94.70	70.04	**85.25**

Observation 3: Dict-BERT can achieve better performance on domain adaptation tasks compared with original BERT and **domain adaptive pre-training** (DAPT) proposed in [3].

[3] Gururangan et al., Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. ACL 2020.

Conclusion

UNIVERSITY OF NOTRE DAME

- **First work** to enhance language model with rare word definitions from dictionaries.
- Propose **two novel self-supervised tasks on word-level and sentence-level alignment** between input text and rare word definitions
- Evaluate Dict-BERT on the GLUE benchmark, in which our model can improve accuracy by +1.15% on average over the original BERT.
- Evaluate Dict-BERT on 8 domain adaptation datasets. Our method can improve F1 score by +2.8% on average over the original BERT settings, and by +0.6% on average over DAPT

Published Work 2

UNIVERSITY OF NOTRE DAME

Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts

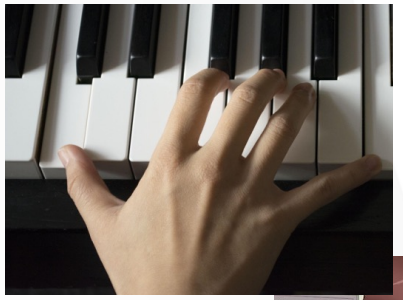
Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao, Meng Jiang

60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)

Diverse Outputs in Commonsense Reasoning

- **Input:** Piano is a kind of sport.
  (a counterfactual statement[5])

- **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)**




three reasons from different perspectives!

[5] Ji et al., Language Generation with Multi-Hop Reasoning on …, EMNLP 2020

Preliminary: Close-book Model

- **Input:** Piano is a kind of sport.
  (a counterfactual statement[5])

Outputs from BART-beam search

- **Output1:** Piano is an instrument.
- **Output2:** Piano is a musical instrument.
- **Output3:** Piano is not a sport.

[5] Ji et al., Language Generation with Multi-Hop Reasoning on …, EMNLP 2020

Diversifying Generation on Knowledge Graph ?

Input: Piano is a kind of sport .

Outputs: *3* different explanations

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

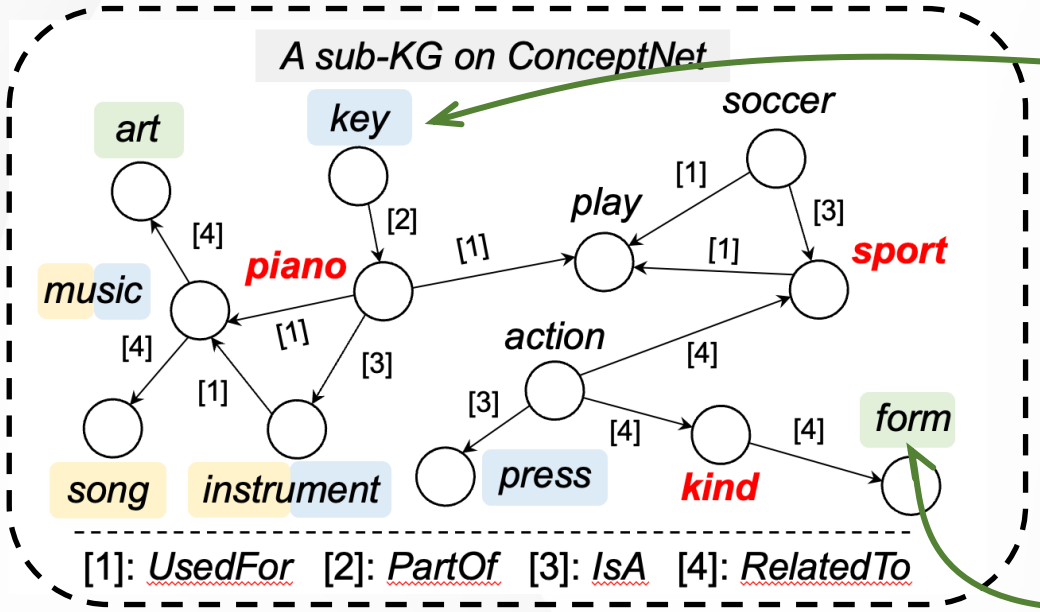
(2) Piano is a musical instrument used in songs to produce different musical tones .

(3) Piano is a kind of art form .

Two observations:

- In three human written outputs, **95%** of concepts can be found on the **knowledge graph** (in shade). *Here we use ConceptNet.*
- **75%** of the concepts in human written outputs were among 2-hop neighbors of the concepts in the input on the **knowledge graph**

Diversifying Generation on Knowledge Graph ?



Input: Piano is a kind of sport .

Outputs: *3* **different explanations**

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

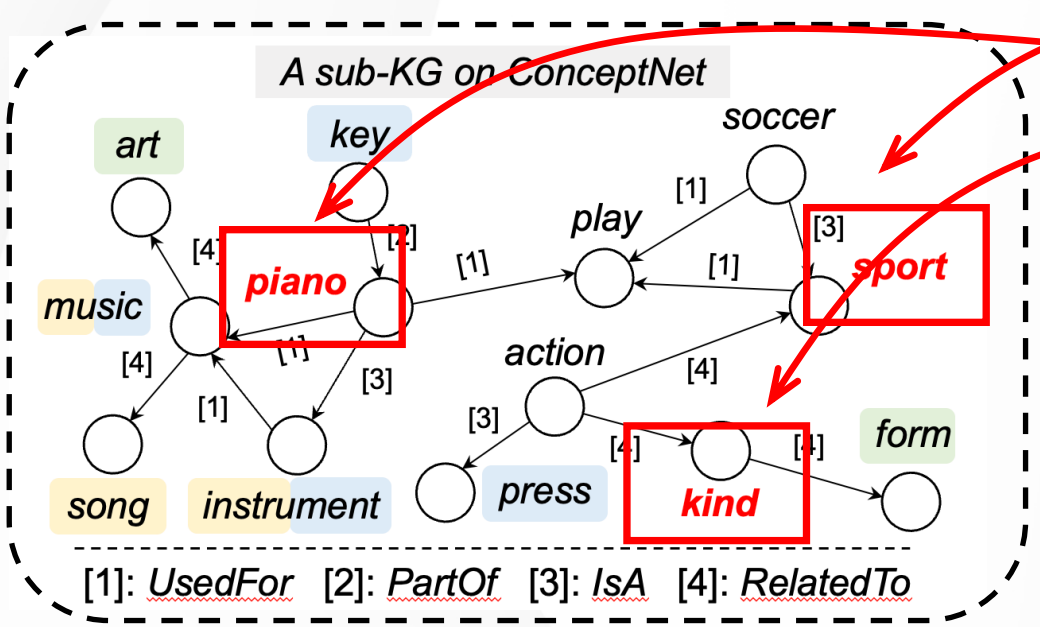
(2) Piano is a musical instrument used in songs to produce different musical tones .

(3) Piano is a kind of art form .

Two observations:

- In three human written outputs, **95%** of concepts can be found on the **knowledge graph** (in shade). *Here we use ConceptNet.*
- **75%** of the concepts in human written outputs were among 2-hop neighbors of the concepts in the input on the **knowledge graph**

Diversifying Generation on Knowledge Graph ?



Input: Piano is a kind of sport .

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 .

Two observations:

- In three human written outputs, **95%** of concepts can be found on the **knowledge graph** (in shade). *Here we use ConceptNet.*
- **75%** of the concepts in human written outputs were among 2-hop neighbors of the concepts in the input on the **knowledge graph**

Background of MoE

UNIVERSITY OF NOTRE DAME

See detailed mathematical formulas are on page 111

- **MoE: mixture of experts**[1,6] are in principle well suited to generating diverse hypotheses which can be achieved through different mixture components.

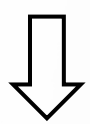
- Formally, given a source sentence $\boldsymbol{x}$ and reference $\boldsymbol{y}$, a mixture model introduces a multinomial latent variable $\boldsymbol{z} \in \{\mathbf{1}, \cdots, \boldsymbol{K}\}$, and decomposes the marginal likelihood as:

$$p(y|x;\theta) = \sum_{z=1}^{K} p(y,z|x;\theta) = \sum_{z=1}^{K} p(z|x;\theta)p(y|z,x;\theta)$$

parameter

an expert

Suppose we have two experts,



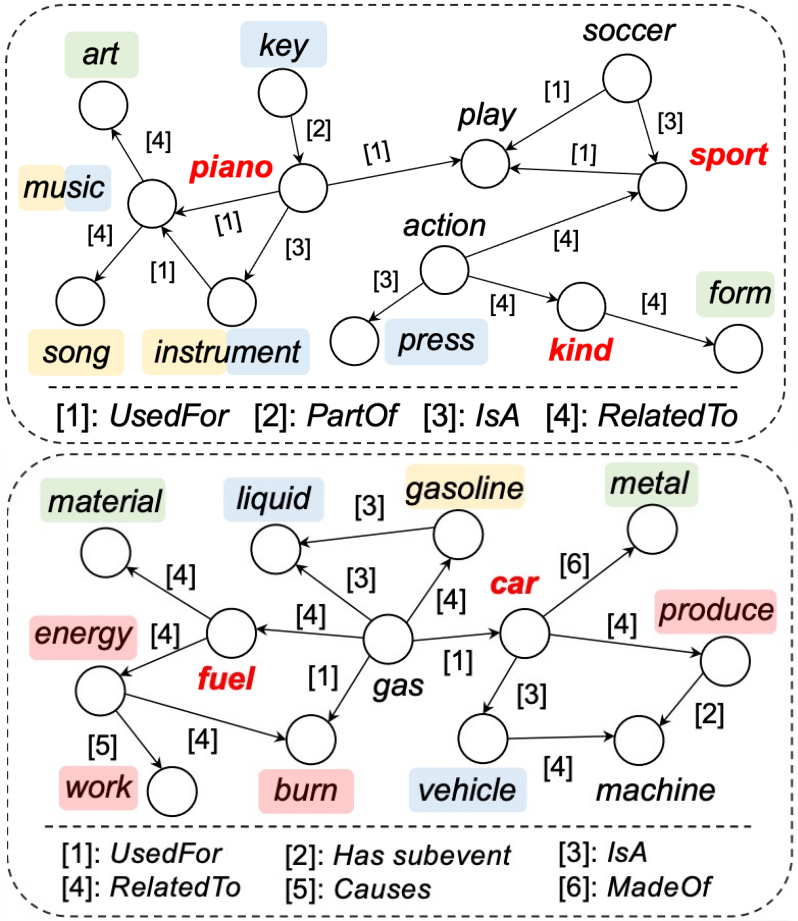
$C_z^{(i)}$ measures the contribution of expert i

If $C_z^{(i)}$ =  , **most of**  parameters are updated, **only a few of**  parameters of are updated.

If $C_z^{(i)}$ =  , **most of**  parameters are updated, **only a few of**  parameters of are updated.

MoE on Knowledge Graph

Goal: generate multiple reasonable explanations given a counterfactual statement.



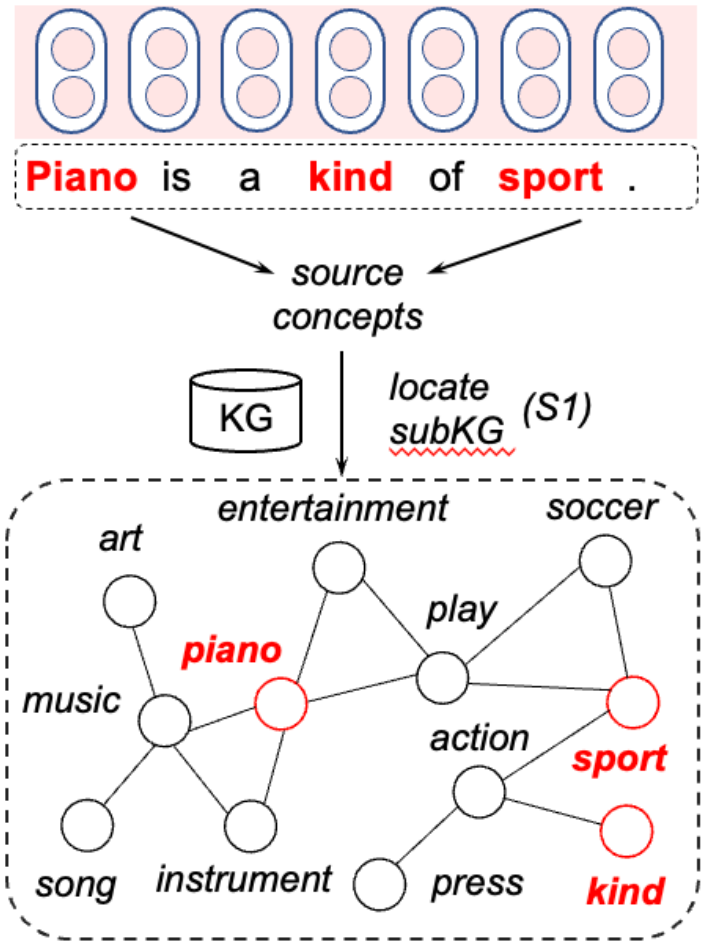
Input: Piano is a kind of sport .
Output (usage): You can produce music when pressing keys on the piano, so it is an instrument .
Output (taxonomy): Piano is a kind of art form .

I am good at giving explanations from the perspective of item usage

I am good at giving explanations from the perspective of taxonomy

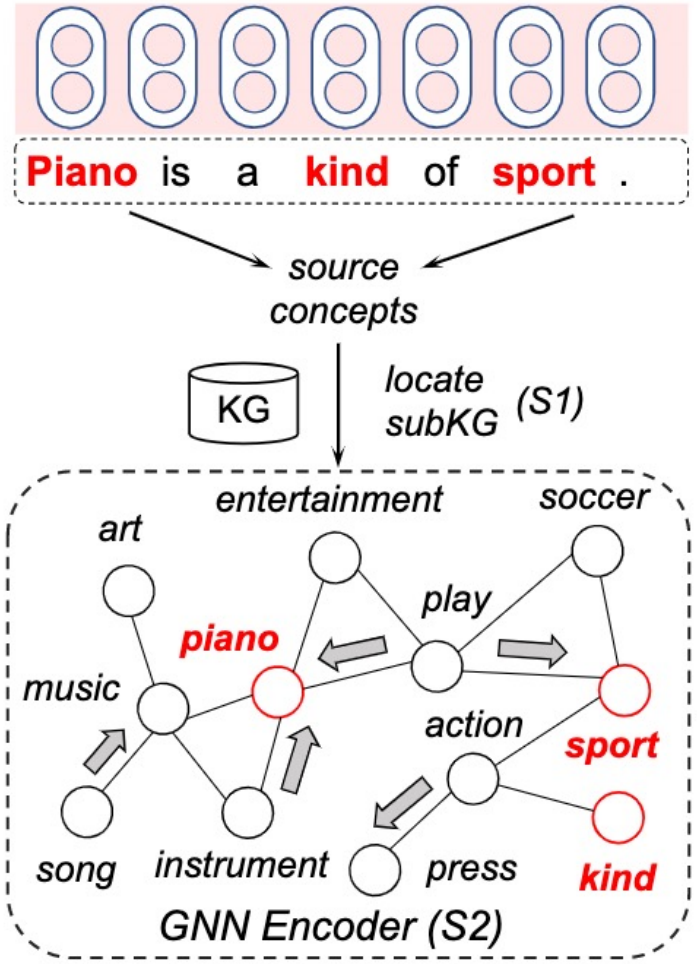
Input: Cars are made of fuel.
Output (usage): Cars burn fuel to produce energy and work.
Output (taxonomy): Cars are made of metal.

Proposed Method: MoKGE (1/4)



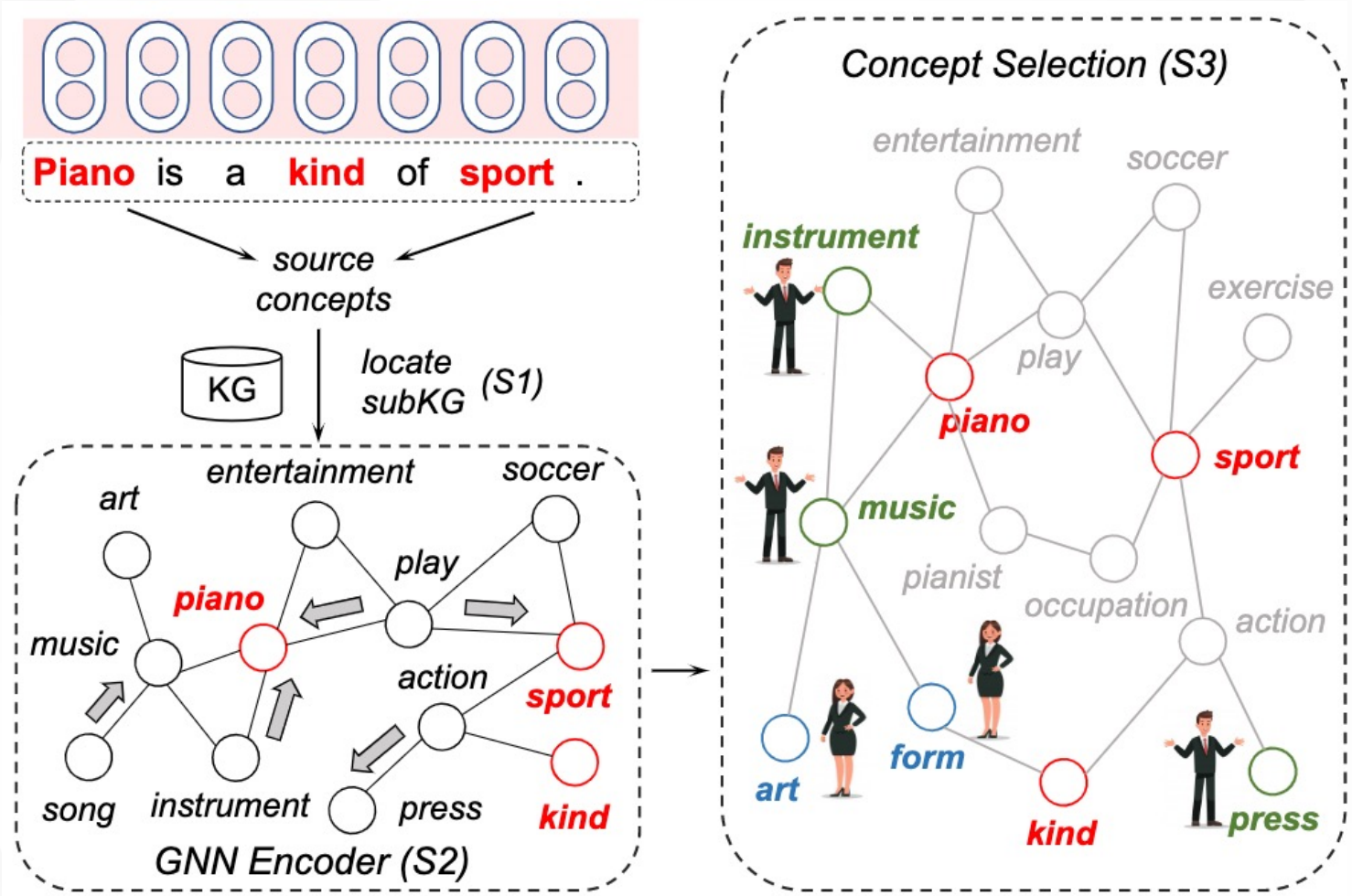
Step1: Construct a sequence-associated subgraph from the commonsense KG

Proposed Method: MoKGE (2/4)



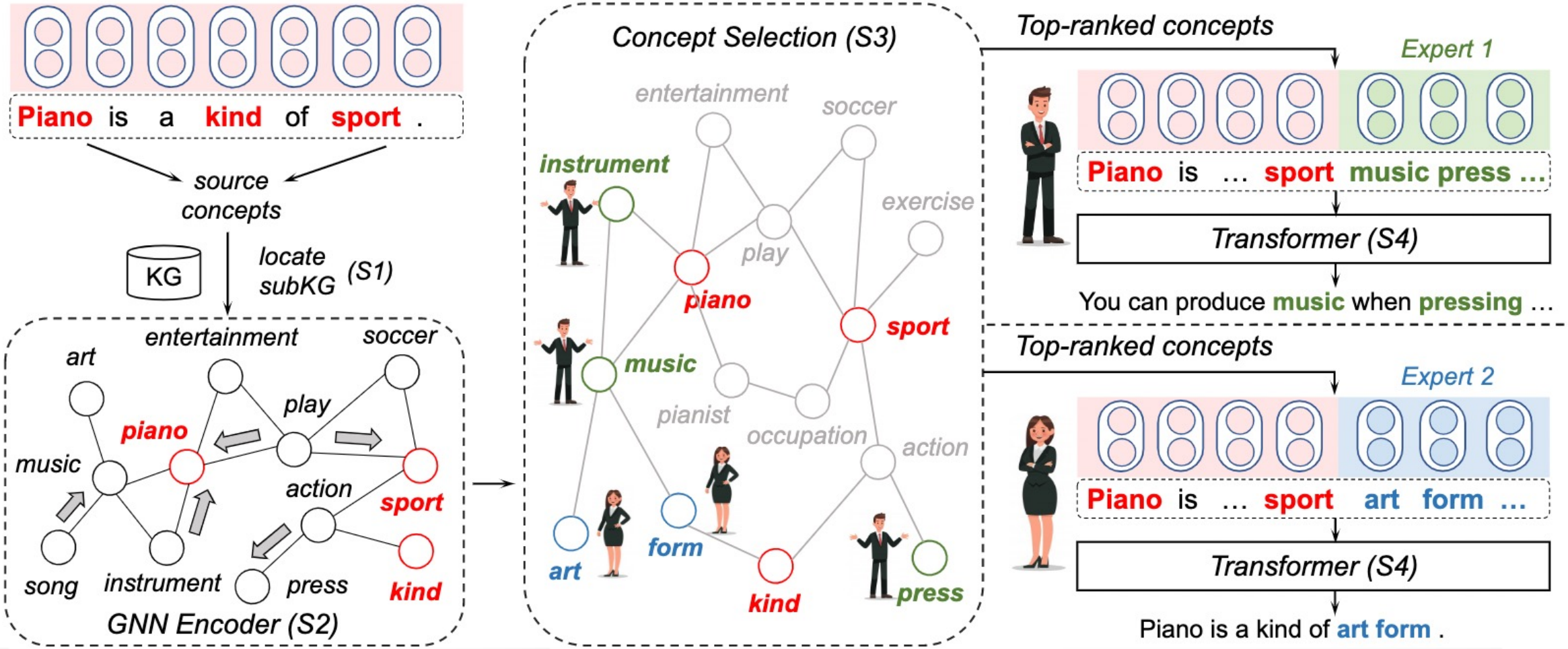
Step2: A relational-GCN iteratively updates the **representation** of a concept node by aggregating information from its neighboring nodes and edges

Proposed Method: MoKGE (3/4)



Step3: Each knowledge expert selects different salient concepts on the sub-graph

Proposed Method: MoKGE (4/4)



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

Experiments: Tasks and Datasets

UNIVERSITY OF NOTRE DAME

We evaluate MoKGE on 2datasets/tasks.

- **Counterfactual explanation generation:** generate explanation given a counterfactual statement for sense-making[7]
- **Abductive generation:** generate valid hypothesis about the likely explanations to partially observable past and future[

We evaluate using two primary metrics.

- **Self-BLEU (lower the better):** whether generated outputs are diverse
- **BLEU (higher the better):** whether generated outputs are accurate

[7] Wang et al., Does it make sense? And why? a pilot study …, In EMNLP 2019
[8] Bhagavatula et al., Abductive Commonsense Reasoning, In ICLR 2020

Experimental Results (ComVE)

UNIVERSITY OF NOTRE DAME

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 =						$60_{0.8}$
MoE	emb						$72_{0.2}$
	pron						$71_{0.5}$
Ours							
MoKGE (ours)	emb						$70_{0.1}$
	pron						**83**$_{0.3}$
Human		$12.36_{0.0}$	$8.01_{0.0}$	$63.02_{0.0}$	$9.55_{0.0}$	$100.0_{0.0}$	$100.0_{0.0}$

	SBLEU (↓)	BLEU (↑)
SoTA	28.4	18.9
Ours	**25.3**	**19.1**

Experimental Results (Alpha-NLG)

UNIVERSITY OF NOTRE DAME

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	$67.89_{0.4}$	$64.72_{0.5}$	$26.27_{0.2}$	$10.34_{0.0}$	$13.64_{0.1}$	$37.96_{0.1}$
	z = 32	$62.08_{0.2}$	$58.25_{0.3}$	$26.67_{0.1}$	$10.36_{0.0}$	$13.35_{0.1}$	$37.73_{0.1}$
	z = 64	$57.87_{0.4}$	$53.61_{0.4}$	$24.91_{0.1}$	$10.21_{0.1}$	$11.77_{0.1}$	$36.35_{0.2}$
Truncated sampling	k= 5	$67.09_{1.0}$	$63.82_{1.1}$	$25.47_{0.3}$	$10.44_{0.1}$	$13.33_{0.2}$	$38.07_{0.2}$
	k= 20	$54.65_{2.1}$	$50.36_{2.4}$	$29.30_{0.5}$	$10.62_{0.2}$	$14.12_{0.7}$	$38.76_{0.6}$
	k= 50	$52.11_{3.7}$	$47.75_{4.2}$	$30.08_{0.3}$	$10.64_{0.1}$	$14.01_{0.8}$	$\mathbf{38.98}_{0.6}$
Nucleus sampling	p= .5	$73.34_{0.3}$	$71.01_{0.3}$	$25.49_{0.0}$	$10.46_{0.0}$	$11.71_{0.1}$	$36.53_{0.2}$
	p= .75	$64.49_{0.4}$	$61.45_{0.5}$	$27.72_{0.1}$	$10.54_{0.1}$	$12.63_{0.0}$	$37.48_{0.1}$
							$38.42_{0.3}$
MoE							$\underline{38.91}_{0.2}$
							$38.78_{0.4}$
Ours							
MoKGE (ours)							$38.06_{0.2}$
							$38.82_{0.7}$
Human		$10.36_{0.0}$	$6.04_{0.0}$	$53.57_{0.0}$	$10.84_{0.0}$	$100.0_{0.0}$	$100.0_{0.0}$

	SBLEU (↓)	BLEU (↑)
SoTA	23.2	**14.2**
Ours	**22.4**	**14.2**

Experimental Results (Human)

UNIVERSITY OF NOTRE DAME

- 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

Experimental Results (Human)

UNIVERSITY OF NOTRE DAME

- 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

- 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

Conclusion & Future Work

UNIVERSITY OF NOTRE DAME

- Aim to produce **diverse contents** for the commonsense reasoning (CR) generation task.
- Propose a novel Mixture of Knowledge Graph Experts (MoKGE) to enhance diversity.
- Experiments on two CR tasks demonstrated the effectiveness of our proposed MoKGE that can not only generate accurate outputs but also diverse **(i.e., cover different aspects)**

Outline of Proposal

UNIVERSITY OF NOTRE DAME

Knowledge Retrieval and Fusion for Knowledge-intensive NLP Tasks

Published work 1	[1] Dict-BERT: Enhancing Language Model with English Dictionary. ACL 2022 Unstructured knowledge	Language understanding Question answering
Published work 2	[2] Diversifying Content Generation for Commonsense Reasoning. ACL 2022 Structured knowledge	Commonsense reasoning Language generation
Proposed work 1	[1] Grounding Knowledge Methods across Heterogeneous Knowledge. Heterogeneous knowledge	Commonsense reasoning Open-domain Question answering
Proposed work 2	[2] Learning Knowledge Retriever without Ground Truth Supervision. Unstructured knowledge	Commonsense reasoning Language generation

Proposed Work 1

UNIVERSITY OF NOTRE DAME

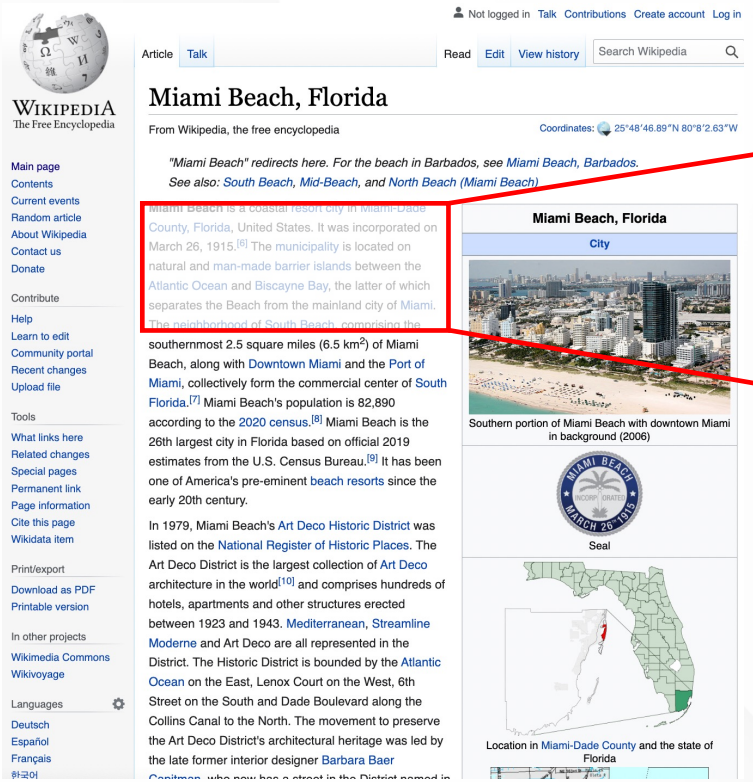
Grounding Knowledge-enhanced model across Heterogeneous Knowledge Sources

Proposed Work 1

UNIVERSITY OF NOTRE DAME

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? **Atlantic Ocean**



Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

Wikipedia
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons
Wikivoyage

Languages
Deutsch
Español
Français
한국어

Miami Beach, Florida

From Wikipedia, the free encyclopedia

Coordinates: 25°48′46.89″N 80°8′2.63″W

"Miami Beach" redirects here. For the beach in Barbados, see Miami Beach, Barbados.
See also: South Beach, Mid-Beach, and North Beach (Miami Beach)

Miami Beach is a coastal resort city in Miami-Dade County, Florida, United States. It was incorporated on March 26, 1915.[6] The municipality is located on natural and man-made barrier islands between the Atlantic Ocean and Biscayne Bay, the latter of which separates the Beach from the mainland city of Miami. The neighborhood of South Beach, comprising the southernmost 2.5 square miles (6.5 km²) of Miami Beach, along with Downtown Miami and the Port of Miami, collectively form the commercial center of South Florida.[7] Miami Beach's population is 82,890 according to the 2020 census.[8] Miami Beach is the 26th largest city in Florida based on official 2019 estimates from the U.S. Census Bureau.[9] It has been one of America's pre-eminent beach resorts since the early 20th century.

In 1979, Miami Beach's Art Deco Historic District was listed on the National Register of Historic Places. The Art Deco District is the largest collection of Art Deco architecture in the world[10] and comprises hundreds of hotels, apartments and other structures erected between 1923 and 1943. Mediterranean, Streamline Moderne and Art Deco are all represented in the District. The Historic District is bounded by the Atlantic Ocean on the East, Lenox Court on the West, 6th Street on the South and Dade Boulevard along the Collins Canal to the North. The movement to preserve the Art Deco District's architectural heritage was led by the late former interior designer Barbara Baer

Miami Beach, Florida
City

Southern portion of Miami Beach with downtown Miami in background (2006)

Seal

Location in Miami-Dade County and the state of Florida

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, […]

Proposed Work 1

UNIVERSITY OF NOTRE DAME

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?

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, […]

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

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

Knowledge can be found in different sources.

Proposed Work 1

UNIVERSITY OF NOTRE DAME

- **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

Proposed Work 1

UNIVERSITY OF NOTRE DAME

-- 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	**Text**	**0.86**[1]	**0.85**	**0.83**

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

To increase the coverage,

(1) increase the number of entries of a single knowledge

(2) leverage heterogeneous knowledge sources (e.g., structured)

[1]Performance is based on Hit@100 by using DPR retriever

Related Work 1

UNIVERSITY OF NOTRE DAME

-- 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: Final performance evaluation. This is different from the previous table.

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.

Related Work 1

UNIVERSITY OF NOTRE DAME

-- **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
Wikitable	**Table**	**0.53**	**0.66**	**0.64**
Wikidata	**Graph**	**0.10**	**0.16**	**0.22**
Wiktionary	**Text**	**0.12**	**0.27**	**0.23**
Wikipedia	**Text**	**0.86**	**0.85**	**0.83**
All (Oracle)		**0.92**	**0.89**	**0.88**

Table: Coverage evaluation. Heterogeneous knowledge increases coverage.

Related Work 1

UNIVERSITY OF NOTRE DAME

Observations from existing works:

-- it is not wise to improve the coverage by expanding the number of entries in single knowledge.

-- Heterogeneous knowledge is needed for knowledge-intensive NLP tasks, even combined.

-- When answering a question, we human beings often seek to various kinds of knowledge learned from different sources.

Proposed Solutions -- Outline

UNIVERSITY OF NOTRE DAME

Proposed solution 1: Homogenize different knowledge sources to a unified knowledge representation

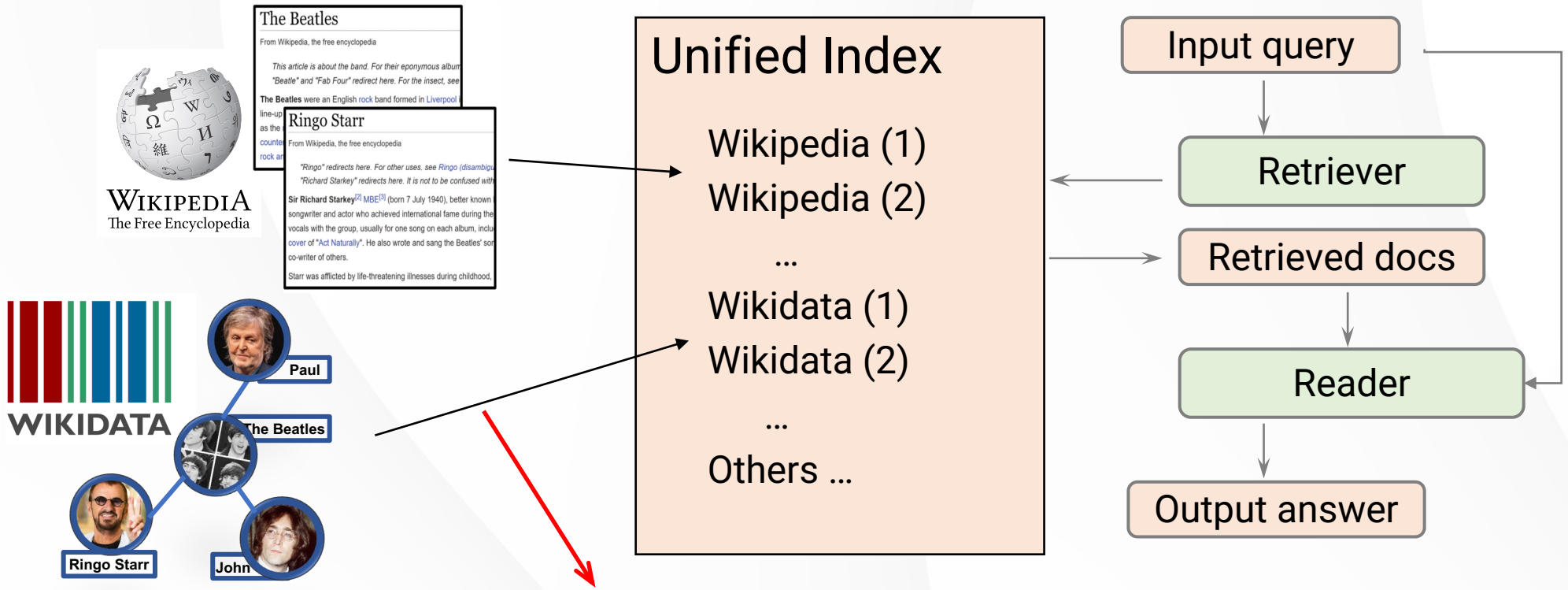
Proposed solution 2: Reasoning over retrieved documents using structured knowledge

Proposed solution 3: Commonsense reasoning over encyclopedic knowledge – a new perspective of heterogenous knowledge

Proposed Solution 1

UNIVERSITY OF NOTRE DAME

- **Solution:** Homogenize different knowledge sources to a unified knowledge representation (e.g., covert knowledge graph triple into natural language)

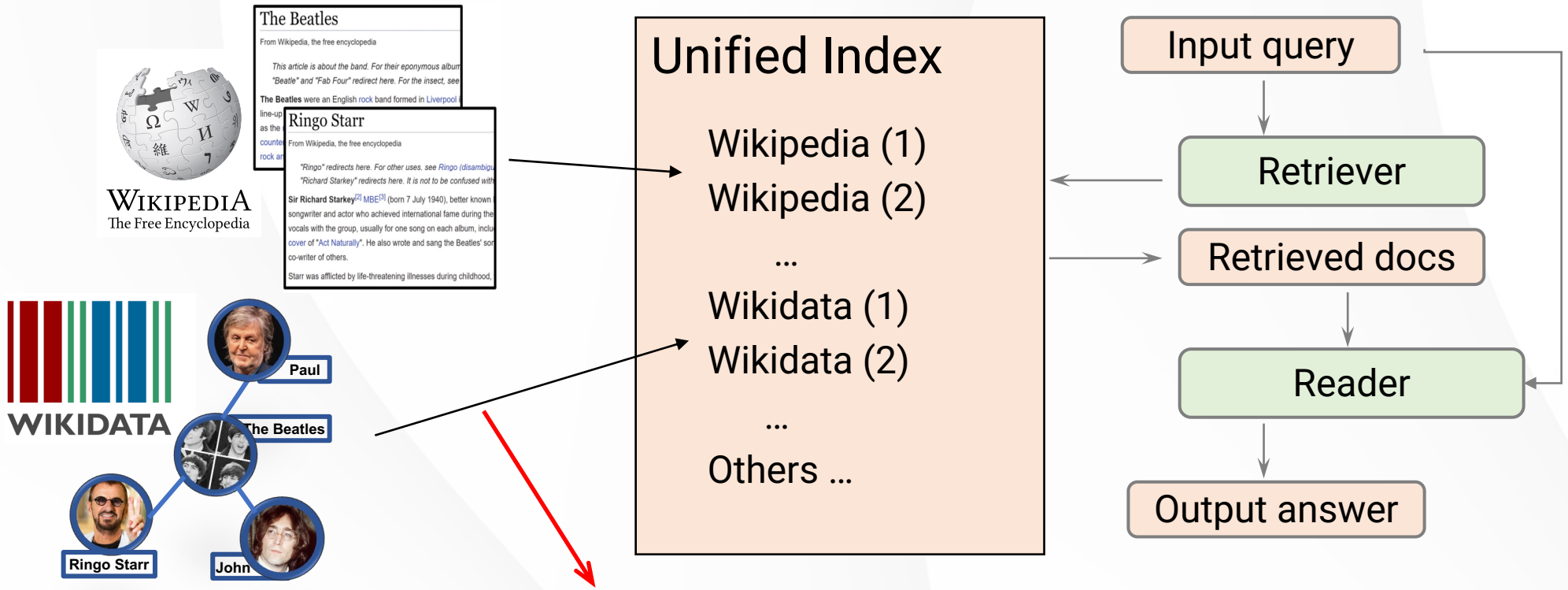


(1) Retrieve raw knowledge graph triple is hard. -- UDT-QA[25]

(2) Template methods cause incorrect semantics. – Unik-QA[26]

Proposed Solution 1

- **Solution:** Homogenize different knowledge sources to a unified knowledge representation (e.g., covert knowledge graph triple into natural language)



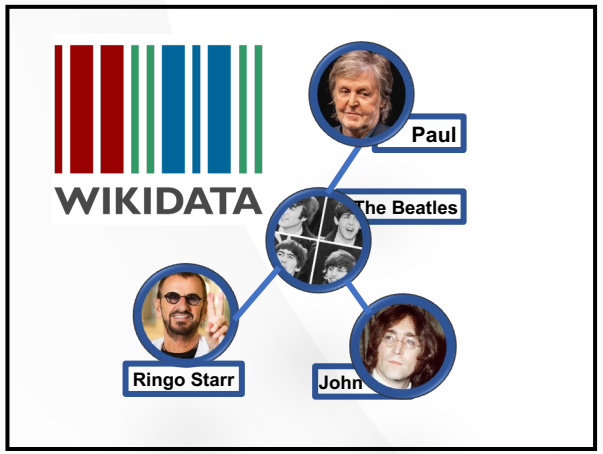
My solution: using neural data/graph-to-text methods

Proposed Solution 1

UNIVERSITY OF NOTRE DAME

- **Preliminary study:** Trained a T5-large model on WikiGraphs[1] (20k training data), then made graph-to-text generation on all Wikidata (14M sentences)

Graph



T5

Text

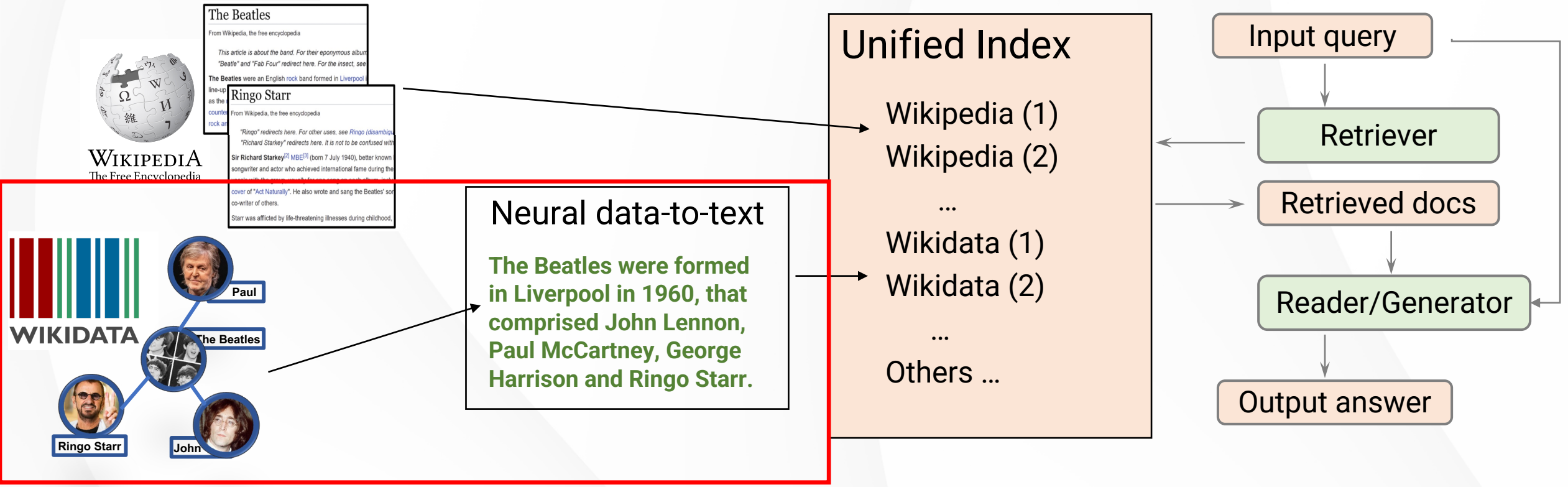
The Beatles were formed in Liverpool in 1960, that comprised John Lennon, Paul McCartney, George Harrison and Ringo Starr.

After training, we convert 15M Wikidata triples into natural language text

[1] WikiGraphs: A Wikipedia Text - Knowledge Graph Paired Dataset. arXiv 07/16/2021. DeepMind

Proposed Solution 1

- **Solution:** Homogenize different knowledge sources to a unified knowledge representation (e.g., covert knowledge graph triple into natural language)



[1] WikiGraphs: A Wikipedia Text - Knowledge Graph Paired Dataset. arXiv 07/16/2021. DeepMind

Proposed Solution 1

UNIVERSITY OF NOTRE DAME

- **Preliminary study:** Trained a T5-large model on WikiGraphs (20k training data), then made graph-to-text generation on all Wikidata (14M sentences)

K-source	Size	NQ		TriviaQA		WebQ	
		Hit@20	Hit@100	Hit@20	Hit@100	Hit@20	Hit@100
Wikipedia	21M	0.78	0.86	0.78	0.84	0.75	0.83
Using template to covert KG triples to plain text – **Unik-QA NAACL 2022**[33]							
Only Wikidata (**Unik-QA**[33] template)	15M	**0.06**	**0.09**	**0.11**	**0.16**	**0.13**	**0.22**
Merge with Wikipedia	36M	0.78	0.86	0.78	0.84	0.76	0.84
Using neural data-to-text to covert KG triples to plain text							
Wikidata (data-to-text)	15M	**0.20**	**0.31**	**0.30**	**0.41**	**0.34**	**0.50**
Merge with Wikipedia	36M	0.80	0.87	0.80	0.86	0.78	0.86

Proposed Solutions -- Outline

UNIVERSITY OF NOTRE DAME

Proposed solution 1: Homogenize different knowledge sources to a unified knowledge representation

Proposed solution 2: Reasoning over retrieved documents using structured knowledge

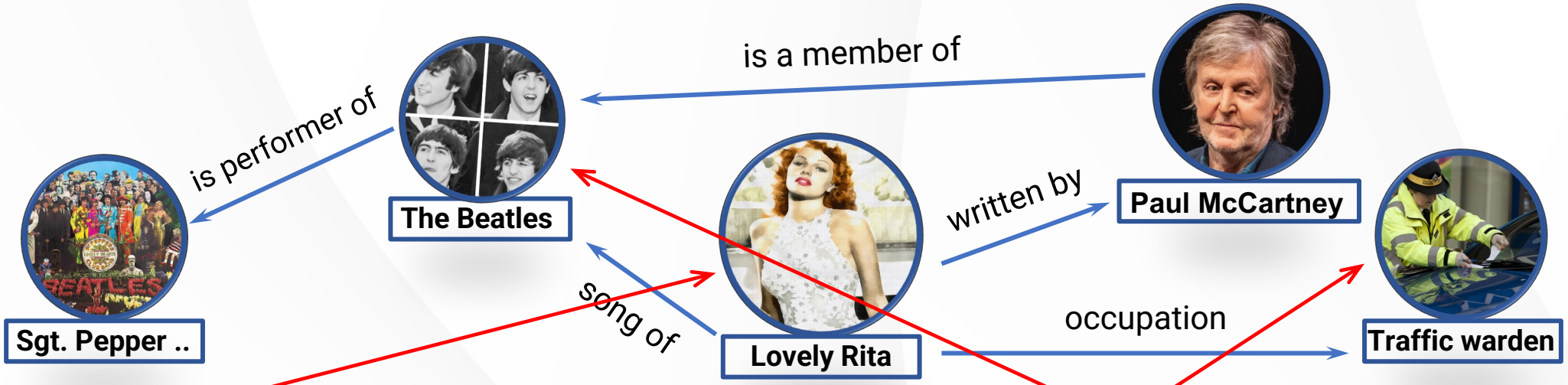
Proposed solution 3: Commonsense reasoning over encyclopedic knowledge – a new perspective of heterogenous knowledge

Proposed Solution 2

UNIVERSITY OF NOTRE DAME

- **Motivation:** Entity relations aren't reflected in unstructured text, making the representation of important entities in the retrieved documents undermined.
- **Solution:** Reasoning over retrieved documents using structured knowledge

Query: What was the occupation of **Lovely Rita** according to the song by **the Beatles**?

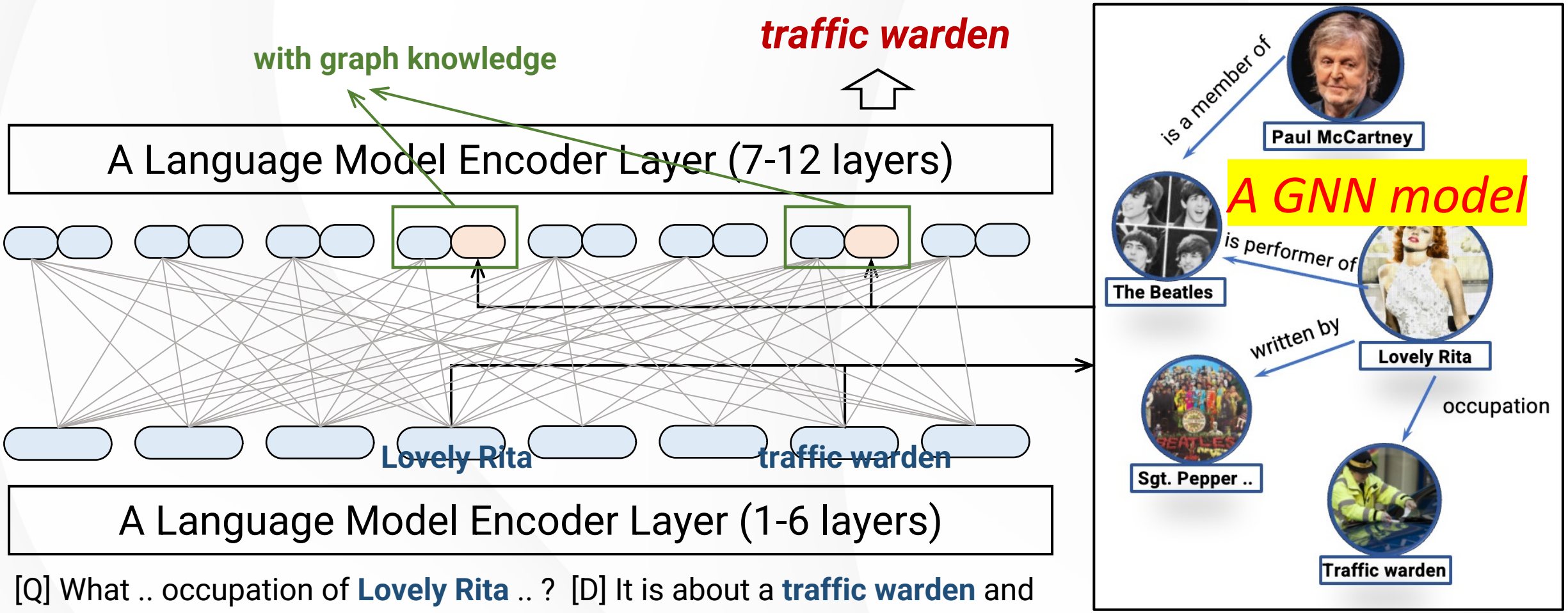


Document: **Lovely Rita** is a song by the English rock band **the Beatles** from their 1967 album **Sgt. Pepper's Lonely Hearts Club Band**. It was written and sung by **Paul McCartney** and credited to **Lennon-McCartney**. It is about a female **traffic warden** and the narrator's affection for her

Proposed Solution 2

UNIVERSITY OF NOTRE DAME

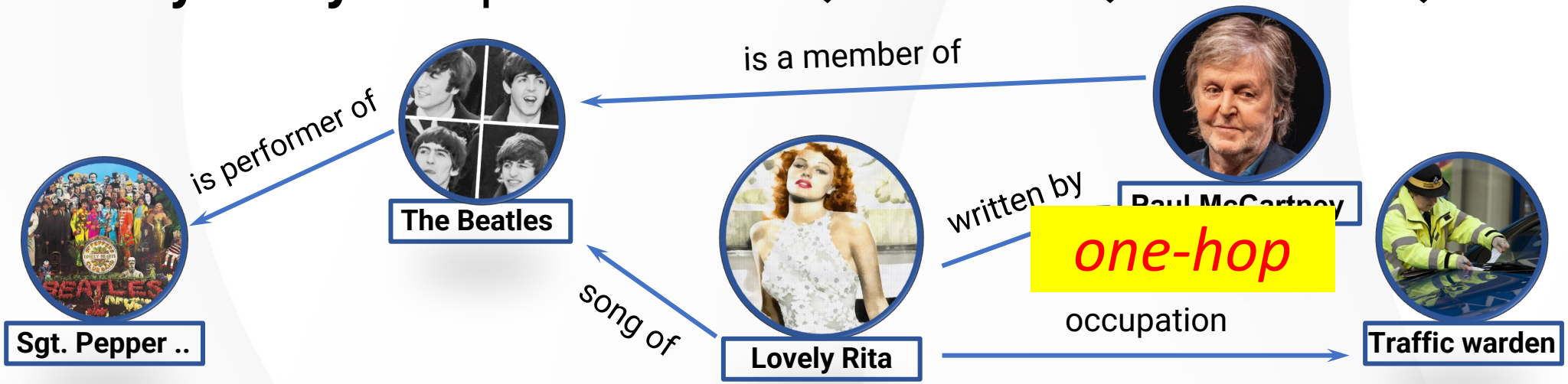
- **Framework:** Add node representation from Wikidata to language model



Proposed Solution 2

UNIVERSITY OF NOTRE DAME

- **Preliminary study** on open-domain QA – TriviaQA and WebQ



Wikidata	TriviaQA	WebQ
One-hop	34.5	51.2
Two-hop	56.8	60.0

Table: Proportion of question-answer entities that are connected on Wikidata

Model	TriviaQA	WebQ
FiD (SoTA)	65.10	43.05
GNNReader	**66.25**	**44.78**

Table: Add node representation of entity from Wikidata improves odqa performance

Proposed Solutions -- Outline

UNIVERSITY OF NOTRE DAME

Proposed solution 1: Homogenize different knowledge sources to a unified knowledge representation

Proposed solution 2: Reasoning over retrieved documents using structured knowledge

Proposed solution 3: Commonsense reasoning over encyclopedic knowledge – a new perspective of heterogenous knowledge

Proposed Solution 3

UNIVERSITY OF NOTRE DAME

- **Motivation:** A new perspective to define heterogeneous knowledge sources.

(structured v.s. unstructured → Encyclopedic Knowledge v.s. Commonsense Knowledge)

- **Target task:** Commonsense reasoning over entities (e.g., CREAK, CSQA2.0)

Structure Knowledge (i.e., knowledge graph)

Unstructured Knowledge (i.e., grounded document)

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

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

Proposed Solution 3

UNIVERSITY OF NOTRE DAME

- **Target task:** Commonsense reasoning over entities (e.g., CREAK, CSQA2.0)

Not only need entity knowledge from encyclopedia, but also need commonsense knowledge

- **The goal of task**: given a commonsense claim, predict true or false
- **Task example:**

Claim: Harry Potter can teach classes on how to fly on a broomstick. **TRUE**

Harry Potter is a wizard ...
He plays Quidditch while riding on a broomstick.

+

Someone who's good at something can teach it.

Claim: One can drive La Jolla to New York City in less than two hours. **FALSE**

La Jolla is in California.
NYC is in New York.

+

It takes 5h with airplane to fly from California to New York.

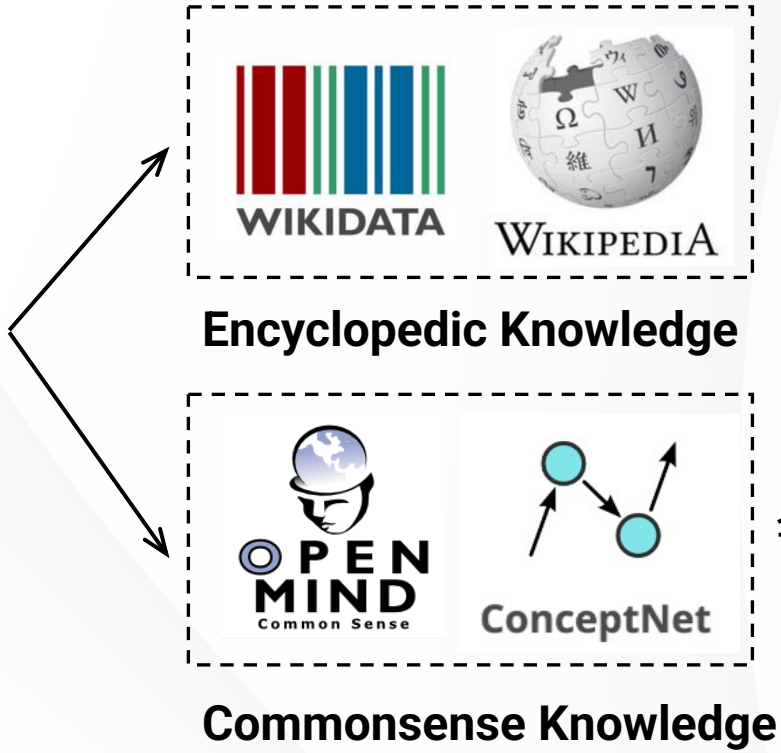
Proposed Solution 3

UNIVERSITY OF NOTRE DAME

- Close-book v.s. Only Wikipedia v.s. Only Commonsense v.s. Both

(test on two commonsense reasoning over entity datasets – CSQA2.0 and CREAK)

Harry Potter can teach classes on how to fly on a broomstick → True



Knowledge	CSQA2.0	CREAK
Wikipedia	57.8	84.5

Knowledge	CSQA2.0	CREAK
Wikipedia + Common	**61.7**	**85.76**

Knowledge	CSQA2.0	CREAK
Common	59.5	83.7

Baseline

Knowledge	CSQA2.0	CREAK
None	53.8	77.5

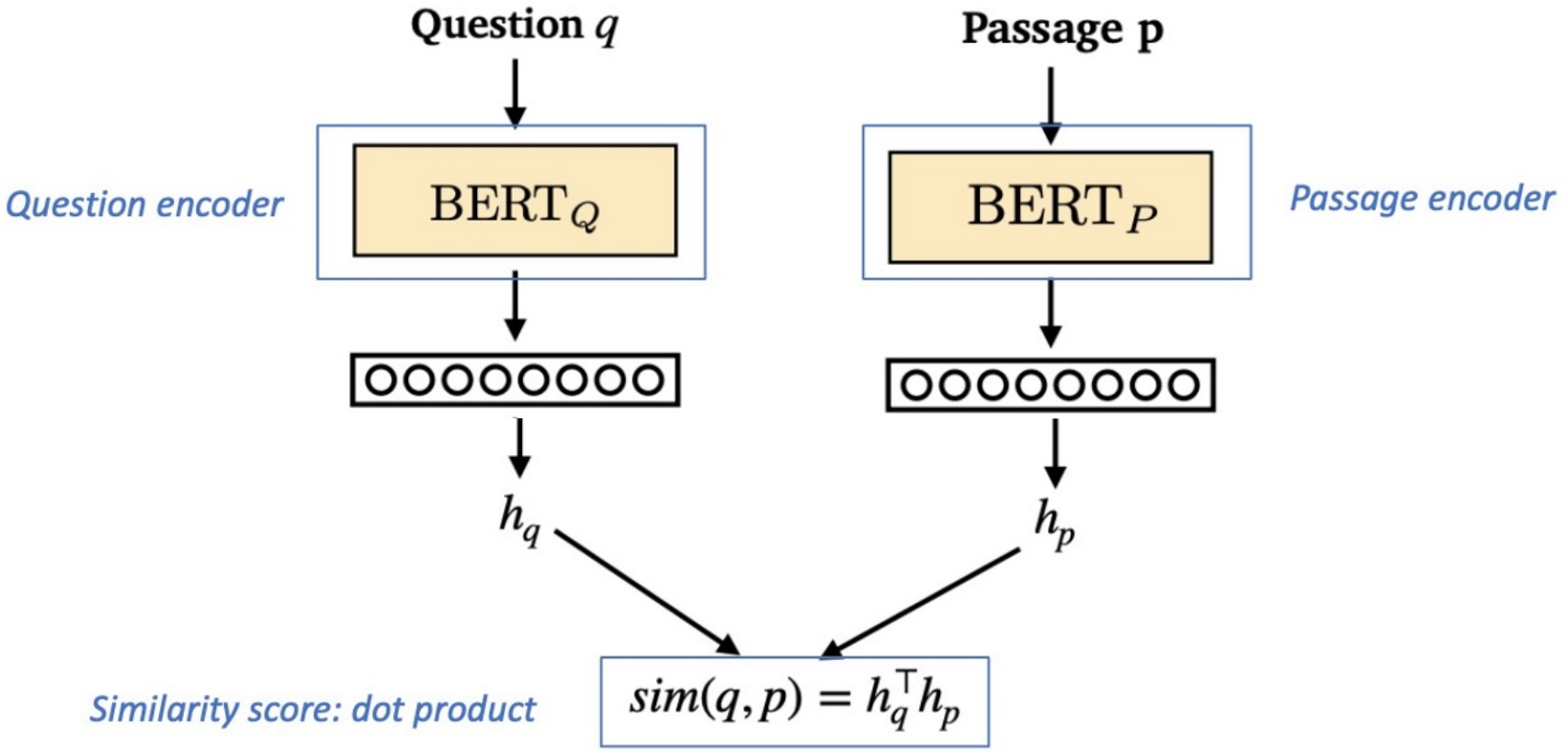
Proposed Work 2

UNIVERSITY OF NOTRE DAME

Learning Knowledge Retriever without Ground Truth Supervision

Proposed Work 2

- **Train a dense retriever (e.g., DPR):** employs two independent encoders like BERT to encode the query and the document separately, and estimates their relevance by computing a similarity score between two representations.



Proposed Work 2

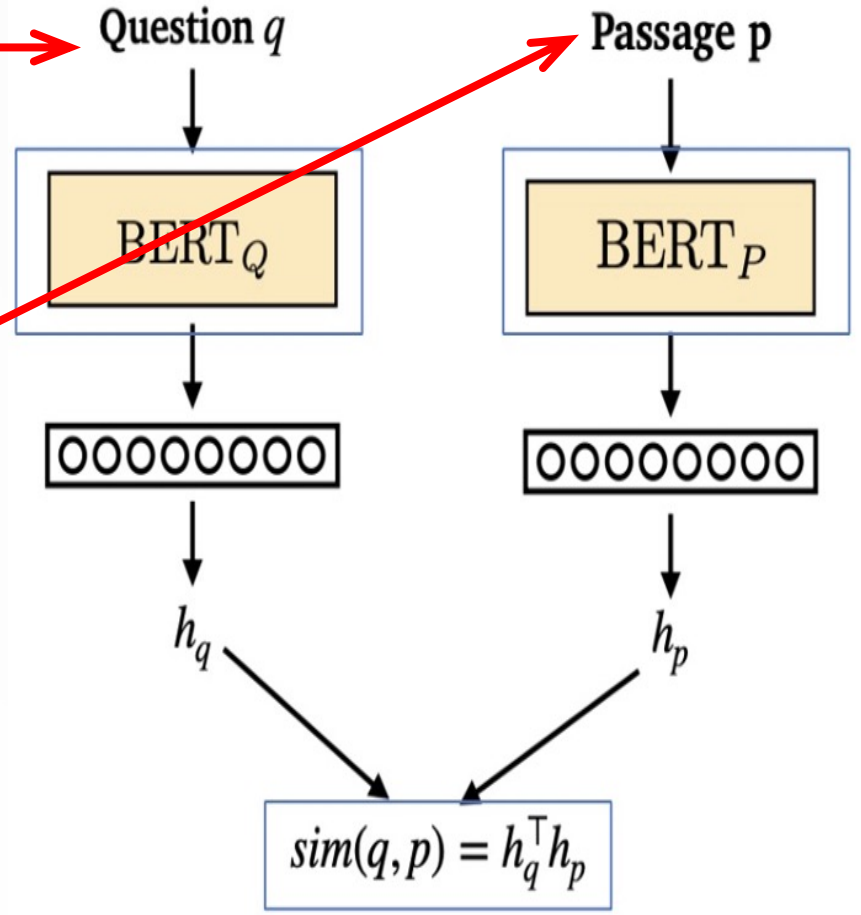
UNIVERSITY OF NOTRE DAME

- **Positive document:** a Wikipedia passage containing the correct answer

(Q) Miami Beach in Florida borders which ocean?

(D) Miami Beach is a coastal resort city […] The municipality is located on natural and barrier islands between the **Atlantic Ocean** and Biscayne Bay.

(D) 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 **Atlantic** Ocean, […]



Proposed Work 2

UNIVERSITY OF NOTRE DAME

- However, such training pairs might not be applicable on many other NLP tasks, such as commonsense reasoning.

(CREAK) Harry Potter can teach classes on how to fly on a broomstick.
(A good doc from OMCS) Harry Potter is good at riding broomstick.
-- The task is to predict **True**/**False** so passage does not contain answer.

- Besides, manually collecting query-document pairs is costly as the labeling process requires significant effort from domain experts.

Proposed Solution 1

UNIVERSITY OF NOTRE DAME

- Hypothesis 1: A document similar with an explanation for correct choice can help

  Commonsense QA 1.0: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what? **(A) bank** (B) library (c) shopping mall (D) new York city (E) department store

  Positive document: at a bank, a revolving door which serves security measure is convenient for two direction travel.

- Hypothesis 2: A document similar with golden output can help generation

  Counterfactual Explanation Generation: Piano is a kind of sport.

  Ground truth as a positive document: You can produce music when pressing keys on the piano, so it is an instrument.

  Inference after training: Piano is an instrument to produce music.

Proposed Solution 1

UNIVERSITY OF NOTRE DAME

- **Preliminary study:** for a commonsense reasoning task, given an input, retrieve commonsense knowledge from a collection of commonsense corpus, fuse the representations of retrieved documents and input text to produce outputs.

- A collection of commonsense corpus: OMCS, ATOMIC, Wiktionary, ARC

Model	#Paras	CSQA1.0	OBQA	CREAK	CSQA2.0	ComGen	ComVE
		Acc.	Acc.	Acc.	Acc.	SPICE	BLEU-4
BM25	3 Billion	78.25	67.23	72.58	59.95	36.57	26.78
DPR(Wiki)	3.2 Billion	74.12	62.44	74.93	61.25	33.58	25.35
DPR(Com)	**3.2 Billion**	**83.26**	**83.45**	**88.15**	**66.01**	**38.10**	**27.51**

Commonsense DPR can outperform BM25 and DPR trained with Wikipedia data.

Proposed Solution 1

UNIVERSITY OF NOTRE DAME

- **Preliminary study:** for a commonsense reasoning task, given an input, retrieve commonsense knowledge from a collection of commonsense corpus, fuse the representations of retrieved documents and input text to produce outputs.

Model	#Paras	CSQA1.0	OBQA	CREAK	CSQA2.0	ComGen	ComVE
		Acc.	Acc.	Acc.	Acc.	SPICE	BLEU-4
T5-3B	3 Billion	79.69	80.20	85.60	60.24	34.19	25.61
T5-11B	11 Billion	83.25	**87.15**	87.95	**67.80**	35.25	**29.14**
KG-SoTA	-	82.80	84.80	-	-	36.93	22.87
Ours	3.2 Billion	**83.26**	83.45	**88.15**	66.01	**38.10**	27.51

Our commonsense retrieval augmented method can outperform close-book model (T5-3B/11B) and KG enhanced model (KG-SoTA) counterparts

Proposed Solution 2

UNIVERSITY OF NOTRE DAME

- **Motivation:** a useful or a good document can help the reader/generator predict the desired outputs.
- **Solution:** distill knowledge from reader to retriever

(CREAK) Harry Potter can teach classes on how to fly on a broomstick – **TURE**

A close-book T5 model gives [0.7, 0.3] probability of True/False

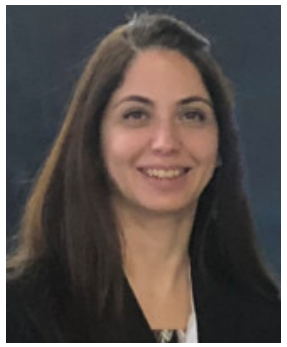
Suppose the retriever obtains 4 documents	*Reader prediction*	*Rank*
Harry Potter is a wizard.	[0.68, 0.32]	3 ✗
Harry Potter is **good at riding broomstick.**	[0.91, 0.09]	1 ✓
Harry Potter plays Quidditch.	[0.75, 0.25]	2 ✓
Harry Potter is a series of fantasy novels.	[0.60, 0.40]	4 ✗

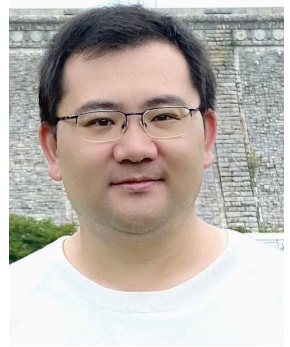
OPEN MIND Common Sense

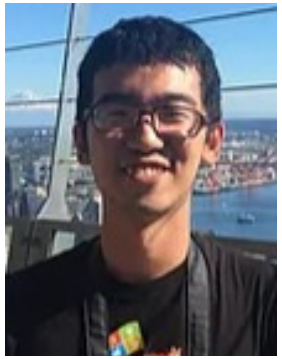
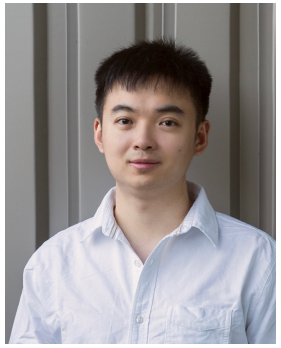
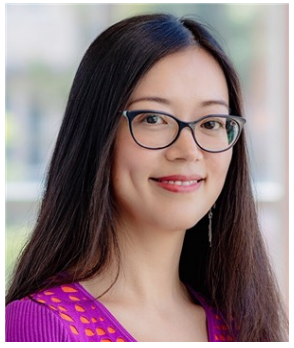
Acknowledgements








Open-source/Tutorials/Publications

UNIVERSITY OF NOTRE DAME

- 600+ stars for open-source code on GitHub
- Two conference tutorials at ACL 2022[1] and EMNLP 2021[2]
- A 44-parge comprehensive survey paper in ACM Computing survey[3]
- 10 papers (6 first-author) in NLP conferences[4-12]
  - 4 in ACL, 4 in EMNLP, 2 in NAACL

- [1] Zhu et al., Knowledge-augmented Methods for Natural Language Processing. ACL'22
- [2] Yu et al., Knowledge-enriched Methods for Natural Language Generation. EMNLP'21
- [3] Yu et al., A Survey of Knowledge-Enhanced Text Generation. CSUR'22
- [4] Yu et al., Crossing Variational Autoencoders for Answer Retrieval. ACL'20
- [5] Yu et al., A Technical Question Answering System with Transfer Learning. EMNLP'20
- [6] Zeng et al., Automatic Pre-Fine Tuning between Pre-Training and Fine-Tuning for SciNER. EMNLP'20
- [7] Yu et al., Technical Question Answering across Tasks and Domains. NAACL'21
- [8] Yu et al., Sentence-Permuted Paragraph Generation, EMNLP'21
- [9] Dong et al., Injecting Entity Types into Entity-Guided Text Generation. EMNLP'21
- [10] Yu et al., Dict-BERT: Enhancing Language Model Pre-training with Dictionary. ACL'22
- [11] Yu et al., KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain QA. ACL'22
- [12] Yu et al., Diversifying Content Generation for Commonsense Reasoning. ACL'22

References

- [1] Shen et al., Mixture Models for Diverse Machine Translation: Tricks of the Trade, ICML 2019
- [2] Yu et al., Sentence-Permuted Paragraph Generation, EMNLP 2021
- [3] Cho et al., Mixture content selection for diverse sequence generation, EMNLP 2019
- [4] Qian et al., Exploring Diverse Expressions for Paraphrase Generation, EMNLP 2019
- [5] Ji et al., Language Generation with Multi-Hop Reasoning on Knowledge Graph, EMNLP 2020
- [6] Jacobs et al., Adaptive mixtures of local experts, Neural computation 1991
- [7] Wang et al., Does it make sense? A pilot study for sense making and explanation, EMNLP 2019
- [8] Bhagavatula et al., Abductive Commonsense Reasoning, ICLR 2020
- [9] Yu et al., A Survey of Knowledge-Enhanced Text Generation, ACM Computing Survey 2022
- [10] Zhang et al., Grounded Conversation Generation as Guided Traverses in commonsense …, ACL 2020
- [11] Liu et al., KG-BART: Knowledge Graph-Augmented BART for Generative Commonsense …, AAAI 2021
- [12] Yu et al., KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain QA, ACL 2022
- [13] Wu et al. Taking Notes on the Fly Helps BERT Pre-training. ICLR 2021
- [14] Schick et al., Rare Words: A Major Problem for Contextualized Embeddings and How to Fix it by Attentive Mimicking. AAAI 2020
- [15] Gururangan et al., Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. ACL 2020.

References

UNIVERSITY OF NOTRE DAME

- [16] Natural Questions: A Benchmark for Question Answering Research. In TACL 2019
- [17] CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. In Neurips 2021
- [18] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL 2019
- [19] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In ACL 2020
- [20] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR 2020
- [21] Language Models are Few-Shot Learners. In Neurips 2020.
- [22] SciBERT: A Pretrained Language Model for Scientific Text. In EMNLP 2019.
- [23] Don‘t Stop Pretraining: Adapt Language Models to Domains and Tasks. In ACL 2020.
- [24] UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. In AAAI 2021
- [25] ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning. In ICLR 2022.
- [26] KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In EMNLP 2019.
- [27] QA-GNN: Reasoning with Language Models and Knowledge Graphs for QA. In NAACL 2021.
- [28] GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. In ICLR 2022.
- [29] Dense Passage Retrieval for Open-Domain Question Answering. In EMNLP 2019.

References

- [30] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Neurips 2020.
- [31] Learning Dense Representations of Phrases at Scale. In ACL 2021.
- [32] Leveraging Passage Retrieval with Generative Models for Open Domain QA. In EACL 2021.
- [33] UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain Question Answering. In NAACL 2022.

Appendix -- Terminology

UNIVERSITY OF NOTRE DAME

- **Pre-trained language model (PLMs):** e.g., BERT, T5, GPT-3
- **Fine-tune:** e.g., fine-tune BERT on a QA dataset
- **Close-book:** during fine-tuning, only feed *input text* into PLMs and make predictions
- **Open-book:** during fine-tuning, not only use *input text,* but also resort to external knowledge (KG, Wikipedia and etc.)
- **Knowledge retrieval:** used in open-book, which is the way to obtain external knowledge, e.g., String matching, BM25
- **Knowledge fusion:** used in open-book, the way to integrate external knowledge with input text for making predictions

Appendix – Evaluation Metrics

UNIVERSITY OF NOTRE DAME

-- Retrieval performance

- **Hit@K：** For each question, search K documents and the proportion of K documents that contain at least one answer
- **SoTA：** Around 80 to 85 for Hit@100

-- Reader performance

- **ExactMatch / Accuracy：** True Positive / All
- **SoTA：** Around 50 to 60

Appendix – Close-book examples

UNIVERSITY OF NOTRE DAME

Close-book Models: **Knowledge** (e.g., entity relations, commonsense) is learnt into a language model (LM) **parameters**. During fine-tuning, only feed *input text* into LMs and make predictions.

Pre-trai
(only tex

BERT[18] – G
BART[19] – Facebook 2020
T5[20] – Google 2020
GPT-3[21] – OpenAI 2020

-tune on
dge corpus

Tsinghua 2019
DAPT[23] – AI2 2020
Unicorn[24] – AI2 2021
ExT5[25] – Google 2021

Unicorn loaded T5 and further continue pretrained on six commonsense reasoning datasets.

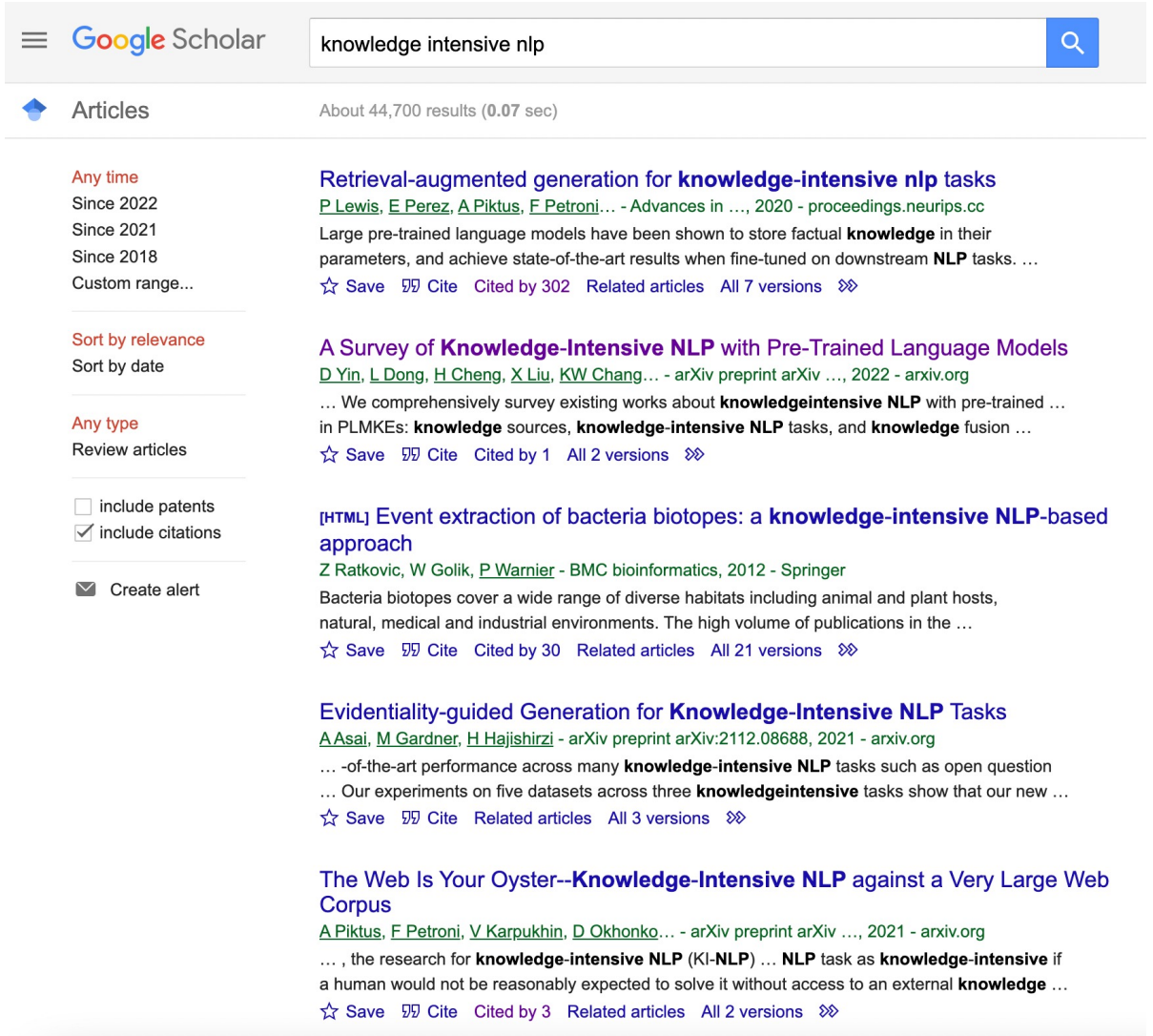
see references of [18-25] at the last page

Appendix – KIT definition

- **In total 15 papers**

(<2019: 2, 2020: 2, 2021: 5, 2022: 3)

- Open-domain question answering: 11
- (Open-ended) commonsense reasoning: 2
- Dialogue system: 4
- Jeopardy question generation: 1
- Long-form question answering: 4
- Fact checking: 3
- Slot filling: 3



Appendix – KIT definition

UNIVERSITY OF NOTRE DAME

1. knowledge-intensive tasks – task that humans could not reasonably be expected to perform without access to an external knowledge source
   -- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Neurips 2020

2. Challenging problems such as open-domain question answering require access to large, external knowledge sources. Solving knowledge-intensive tasks requires–even for humans–access to a large body of information.
   -- KILT: a Benchmark for Knowledge Intensive Language Tasks. NAACL 2021

3. Knowledge-intensive NLP tasks are served as testbeds to evaluate the capability of NLP models to solve problems that require external knowledge.
   -- A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models. arXiv 2022.

Appendix KIT v.s. Non-KIT

UNIVERSITY OF NOTRE DAME

- **NLP tasks that are NOT knowledge-intensive:**

  Summarization, Machine Translation, Information Retrieval,
  Style Transfer, Sentiment classification, Tagging, Chunking

- **NLP tasks that are knowledge-intensive:**

  Open-domain question answering, fact checking
  Commonsense reasoning, Creative generation

Appendix – Motivation of DictBERT

UNIVERSITY OF NOTRE DAME

- **Setting:** Language model pretraining (e.g., BERT, RoBERTa)

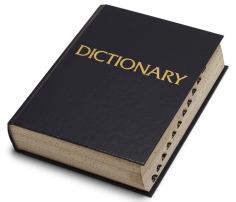
What kind of corpus do we use for pre-training a language model ?


Bookcorpus ✓


Wikipedia ✓

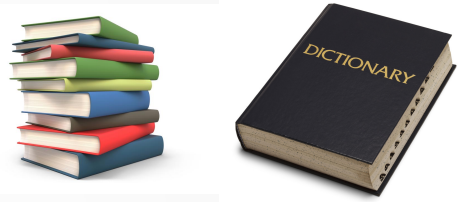
CC-news ✓

OpenWeb ✓



Dictionary ✗

What do humans do for learning a new language?



A lot of books and some dictionaries ✓

- **Problem:** Can we use dictionary to improve language model pretraining?
- **Motivations:**
  - (1) Dictionary is useful **for human to learn a new language** (like training a model from scratch).
  - (2) It is **difficult** of existing language models to **understand rare words** and new words (e.g., using BERT to understand Covid-19), making the learning process hard and slow [1, 2].

[1] Wu et al. Taking Notes on the Fly ... ICLR 2021; [2] Schick et al., Rare Words: A Major Problem … AAAI 2020

Appendix – MIM Task

UNIVERSITY OF NOTRE DAME

- maximize the mutual information between a rare word x_i in the input sequence and its well-defined meaning in the dictionary $c^{(i)}$, with joint density $p(x_i, c^{(i)})$and marginal densities $p(x_i)$ and $p(c^{(i)})$, is defined as the Kullback–Leibler (KL) divergence between the joint and the product of the marginals,

$$I(x_i; c^{(i)}) = D_{KL}\left(p(x_i, c^{(i)}) || p(x_i)p(c^{(i)})\right)$$

- In order to approximate the mutual information, we adopted InfoNCE, which is one of the most commonly used estimators in the representation learning literature, defined as

$$I(x_i; c^{(i)}) \geq \mathbb{E}[\sum_{i=1}^{K} \log \frac{e^{f_{\text{MI}}(h_i, h^{(i)})}}{\sum_{j=1}^{K} \mathbb{1}_{[j \neq i]} e^{f_{\text{MI}}(h_i, h^{(j)})}}]$$
$$\triangleq I_{NCE}(x_i; c^{(i)}),$$

Appendix – Downstream FT

UNIVERSITY OF NOTRE DAME

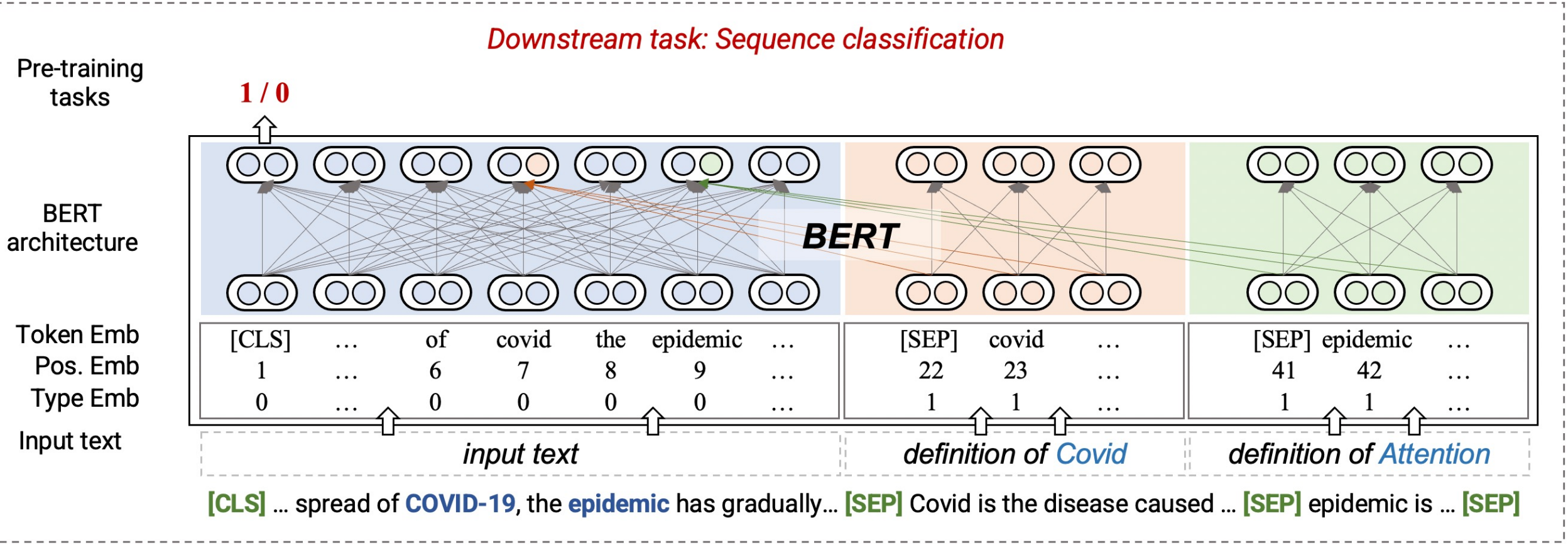
Three options:

- Only use in the input text itself. (Assume no rare words in the downstream tasks)
- Use rare word definitions from downstream. Only use the traditional attention methods
- Use rare word definitions from downstream. Use a knowledge-visible attention mechanism.

Appendix -- DictBERT (FT)

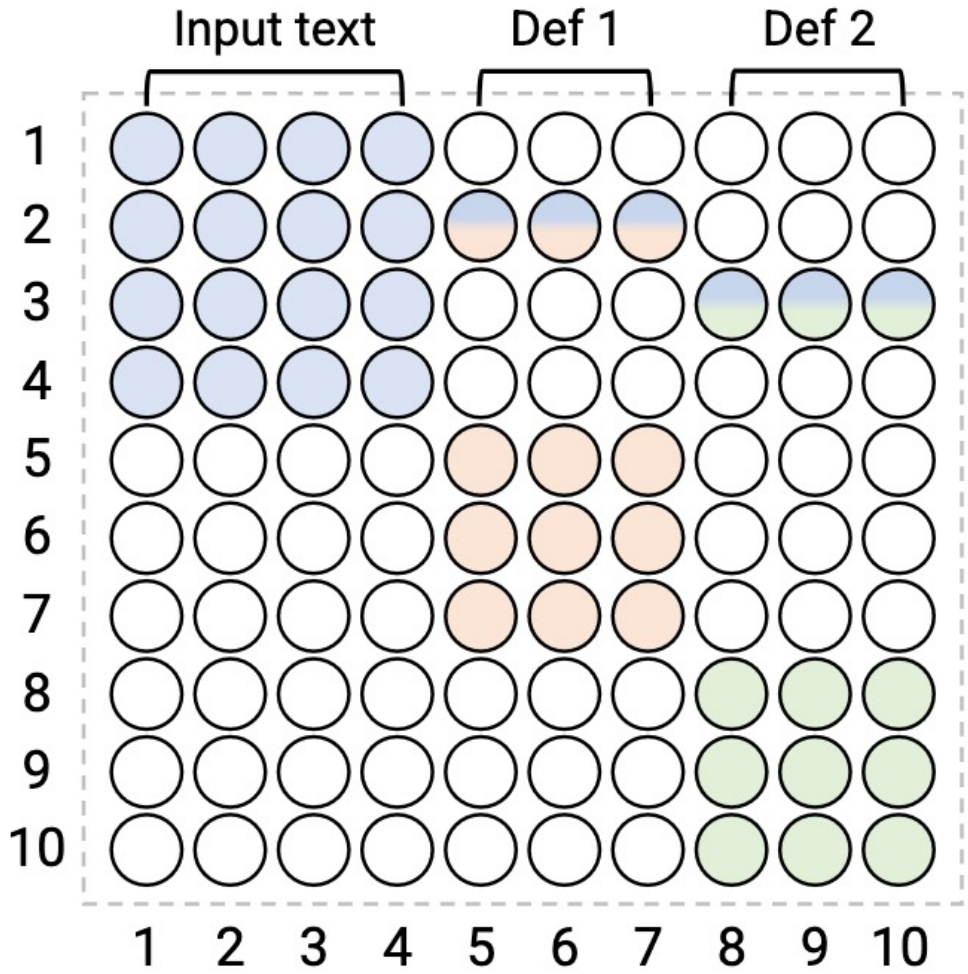
UNIVERSITY OF NOTRE DAME

Fine-tuning： Knowledge attention (definition is only visible to dictionary word)



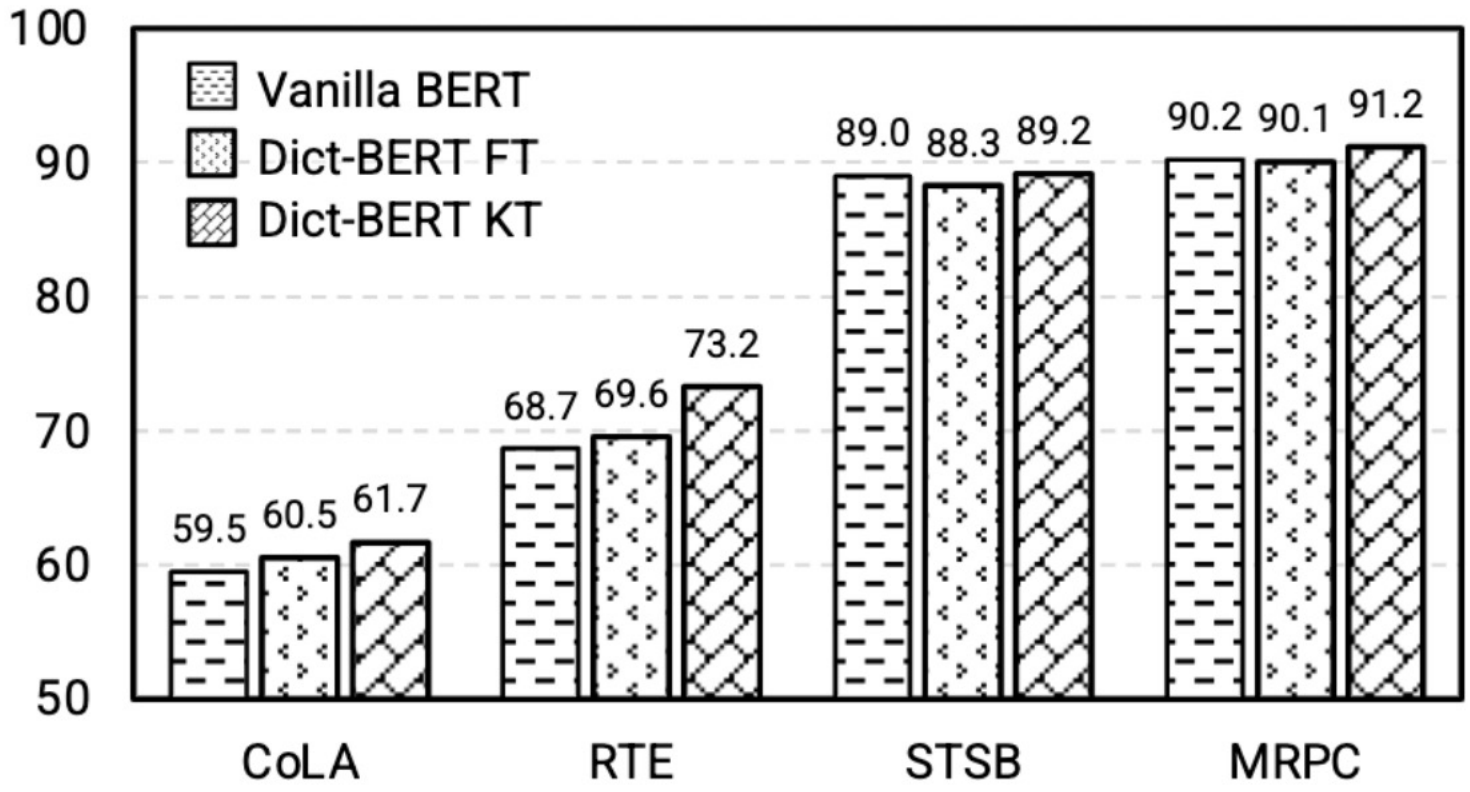
Appendix -- DictBERT (FT)

Fine-tuning：Knowledge attention (definition is only visible to dictionary word)



Appendix -- DictBERT (FT)

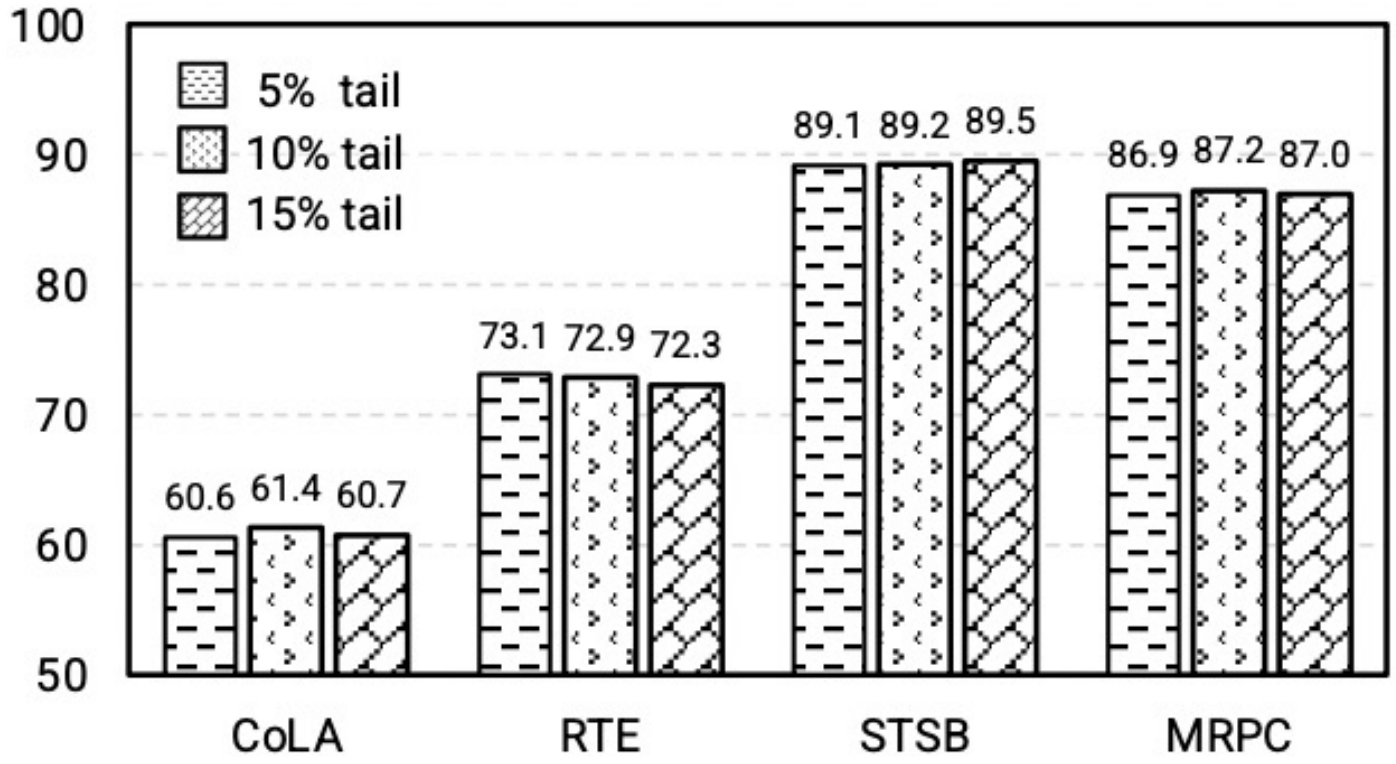
Fine-tuning： Knowledge attention (definition is only visible to dictionary word)



(a) Full attn. (FT) v.s. Knowledge attn. (KT)

Appendix -- DictBERT (FT)

Fine-tuning: selecting different rare word ratios on the downstream tasks



(b) Rare word ratios (5% v.s. 10% v.s. 15%)

Appendix – Experiments on WNLaMPro

- WNLaMPro: Probing of rare word understanding

Key	Rel.	Targets
new	ANT	old
general	ANT	specific
book	HYP	product, publication, ...
basketball	HYP	game, ball, sport, ...
samosa	COH+	pizza, sandwich, salad, ...
harmonium	COH+	brass, flute, sax, ...
simluation	COR	simulation
chepmistry	COR	chemistry

Example entries from WNLaMPro

ANTONYM	HYPERNYM
<W> is the opposite of __ .	<W> is a __ .
<W> is not __ .	a <W> is a __ .
someone who is <W> is not __ .	"<W>" refers to a __ .
something that is <W> is not __ .	<W> is a kind of __ .
"<W>" is the opposite of "__" .	a <W> is a kind of __ .
CORRUPTION	COHYPONYM+
"<W>" is a misspelling of "__" .	<W> and __ .
"<W>" . did you mean "__" ?	"<W>" and "__" .

Patterns for all relations of WNLaMPro.

	Variants		Rare (0, 10) in Wikipedia			Frequent (100, +∞) in Wikipedia			Overall (0, +∞)		
	MMI	DD	MRR	P@3	P@10	MRR	P@3	P@10	MRR	P@3	P@10
BERT	n.a.		**0.117**	**0.053**	**0.036**	**0.357**	**0.179**	**0.116**	**0.266**	**0.130**	**0.084**
DictBERT	✓		0.141	0.065	0.040	0.359	0.180	0.118	0.272	0.133	0.086
		✓	0.144	0.067	**0.041**	**0.360**	**0.183**	**0.118**	**0.274**	**0.137**	**0.088**
	✓	✓	**0.145**	**0.068**	**0.041**	0.359	0.182	**0.118**	**0.274**	**0.137**	**0.088**

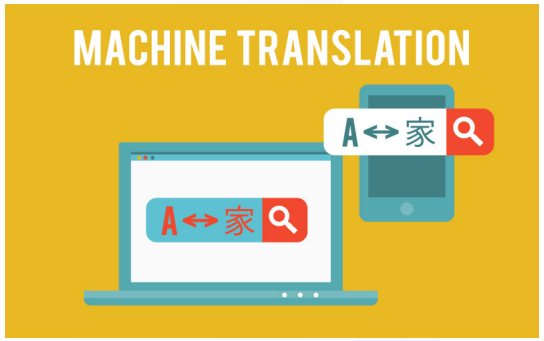
Appendix -- Motivation of MoKGE

UNIVERSITY OF NOTRE DAME

Diversity

Given the same source, an NLG model is expected to create a variety of outputs in terms of content, semantic style, and word variability[1-4].

[1]



Machine Translation
(Chinese <-> English)

尽管 → although / despite

[2]

Story Generation
(Title, keywords -> Story)

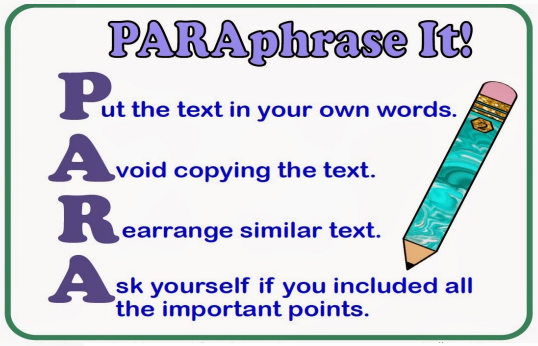
One start → Happy end / Sad end

[3]

Question Generation
(Context -> Question)

Answer → why? / how?

[4]



Paraphrase Generation
(Original -> Paraphrase)

One sent → Para-1 / Para-2

[1] Shen et al., Mixture Models for …, ICML 2019
[2] Yu et al., Sentence-Permuted … , In EMNLP 2020
[3] Cho et al., Mixture Content … , In EMNLP 2019
[4] Qian et al., Exploring Diverse … , In EMNLP 2019

Appendix -- Background of MoE (1/2)

UNIVERSITY OF NOTRE DAME

- **MoE: mixture of experts**[1,6] are in principle well suited to generating diverse hypotheses which can be achieved through different mixture components.

- Formally, given a source sentence $\boldsymbol{x}$ and reference $\boldsymbol{y}$, a mixture model introduces a multinomial latent variable $\boldsymbol{z} \in \{\boldsymbol{1}, \cdots, \boldsymbol{K}\}$, and decomposes the marginal likelihood as:

$$p(y|x;\theta) = \sum_{z=1}^{K} p(y,z|x;\theta) = \sum_{z=1}^{K} p(z|x;\theta)p(y|z,x;\theta)$$

parameter

an expert

- Goal: find $\boldsymbol{\theta}$ that maximizes the log likelihood. So, for each example we compute the gradient:

$$\nabla_\theta \log p(y^{(i)}|x^{(i)};\theta) = \sum_z p(z|x^{(i)},y^{(i)};\theta) \cdot \nabla_\theta \log p(y^{(i)},z|x^{(i)};\theta)$$

[1] Shen et al., Mixture Models for Diverse Machine Translation … , In ICML 2019
[6] Jacobs et al., Adaptive mixtures of local experts, In Neural computation 1991

Appendix -- Background of MoE (2/2)

UNIVERSITY OF NOTRE DAME

- **Train MoE via EM algorithm by iteratively applying the following two steps:**
  - **E-step:** estimate the responsibilities of each expert

  $r_z^{(i)} \leftarrow 1\,[z = argmax_z\, p(y^{(i)}, z'|x^{(i)}; \theta)]$ using the current parameters θ.
  - **M-step:** update θ for **expert *i*** with gradients $\nabla\theta \log p(y^{(i)}, z^{(i)}|x^{(i)}; \theta)$

Suppose we have two experts, ⇩ 

If $r_z^{(i)}$ =  , **most of**  parameters are updated, **only a few of**  parameters of are updated.

If $r_z^{(i)}$ =  , **most of**  parameters are updated, **only a few of**  parameters of are updated.

* We refer the parameter sharing schema to section 3.2 in our paper.

Appendix -- Experimental Settings

UNIVERSITY OF NOTRE DAME

- **Dataset and Tasks:**
  - **ComVE:** generate explanation given a counterfactual statement for sense-making[7]
  - **Alpha-NLG:** generate valid hypothesis about the likely explanations to partially observable past and future[8]

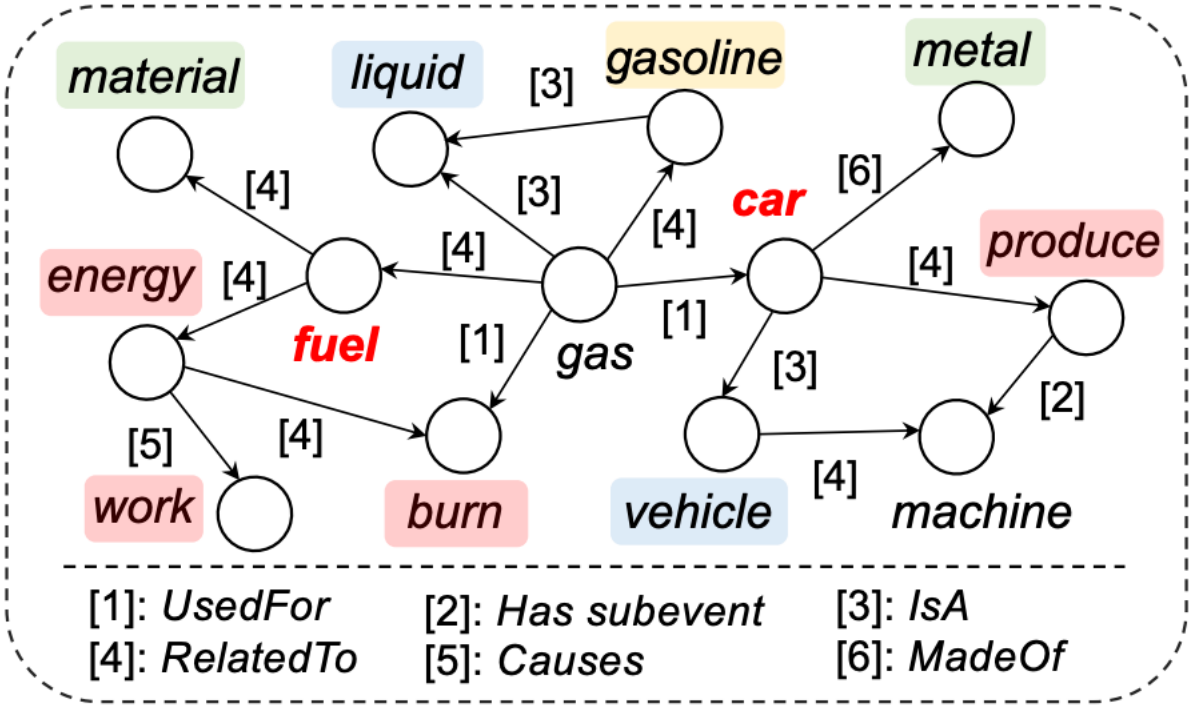
- **Evaluation Methods:** accuracy (e.g., BLEU) and concept/pairwise/corpus diversity
  - **Pairwise diversity:** it measures the within-distribution similarity, often refereed as self-, e.g., BLEU score between all pairwise combinations of hypotheses
  - **Corpus diversity:** it's widely used in dialogue evaluation, e.g., distinct-k measures unique k-grams normalized by the total number of generated k-gram tokens

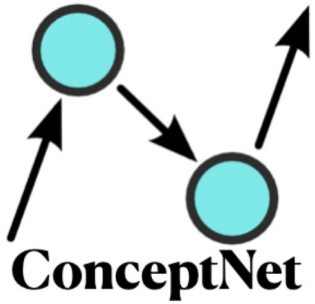
[7] Wang et al., Does it make sense? And why? a pilot study …, In EMNLP 2019
[8] Bhagavatula et al., Abductive Commonsense Reasoning, In ICLR 2020

Appendix – ComVE exmaple

Eg., Counterfactual explanation generation: generate an explanation given a counterfactual statement for sense-making.

- (ComVE) **Cars** are made of **fuel**.





- **KG triples**

(Car, made of, metal)

(Fuel, used for, burn)

(Burn, related to, energy)

Human references

(1) Cars are not made of fuel.

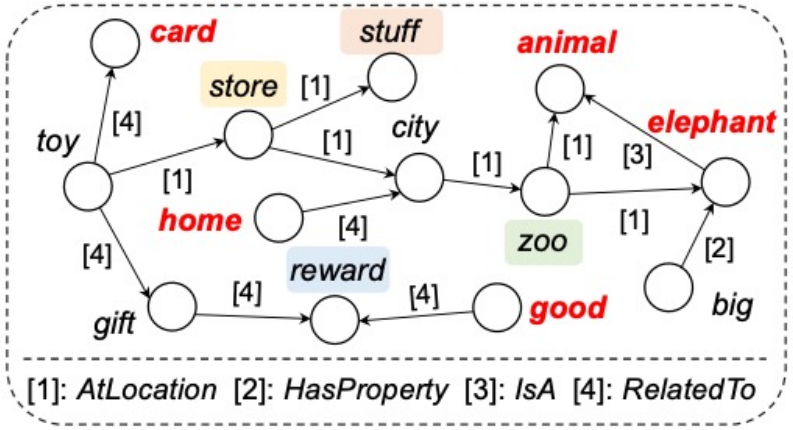
(2) Cars burn fuel to produce energy and work.

(3) Fuel is a liquid which cannot make cars.

Appendix – MoKGE case

UNIVERSITY OF NOTRE DAME

α-NLG -- Input: Billy had received **good** grades on his report **card**. [?]. He decided as he got **home** that **elephants** were his new favorite **animal**.



Nucleus sampling

(1) Billy wanted to go to the zoo and see elephants.
(2) Billy was excited to go on his trip to the zoo.
(3) Billy went to the zoo to see the animals.

MoE (Shen et al.,)

(1) Billy went to the zoo to see the animals.
(2) Billy was excited to go to the zoo with his friends.
(3) Billy's parents took him to the zoo to see elephants.

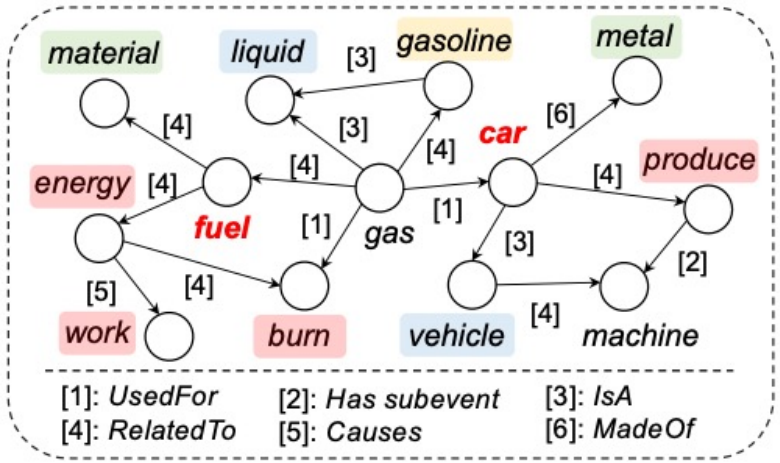
MoKGE (ours)

(1) Billy's parents took him to the zoo as a reward.
(2) Billy wanted to go to the zoo. He saw elephants.
(3) Billy went to the store and bought an elephant.

Human references

(1) Billy's parents sent him on an African safari for a reward.
(2) He went to the zoo later in the day and saw elephants.
(3) His mother stopped by the store and bought him a stuffed elephant.

ComVE -- Input: ***Cars*** *are made of* ***fuel***. **Goal (explanation for sense-making):** [?].



Nucleus sampling

(1) Cars are made of rubber. Fuel is not used to make cars.
(2) Cars are made of aluminum, which is not fuel.
(3) Cars are powered by electric motors and not by fuel.

MoE (Shen et al.,)

(1) Cars are made of metal. but not fuel.
(2) Cars are made of aluminum, not made by fuel.
(3) Fuel is used to make cars more efficient, not less so.

MoKGE (ours)

(1) Fuel is not a vehicle material.
(2) Fuel is not used to make cars. They use gasoline.
(3) Cars are not made of fuel. They are made of metal.

Human references

(1) Cars are not made of fuel.
(2) Cars burn fuel to produce energy and work.
(3) Fuel is a liquid which cannot make cars.

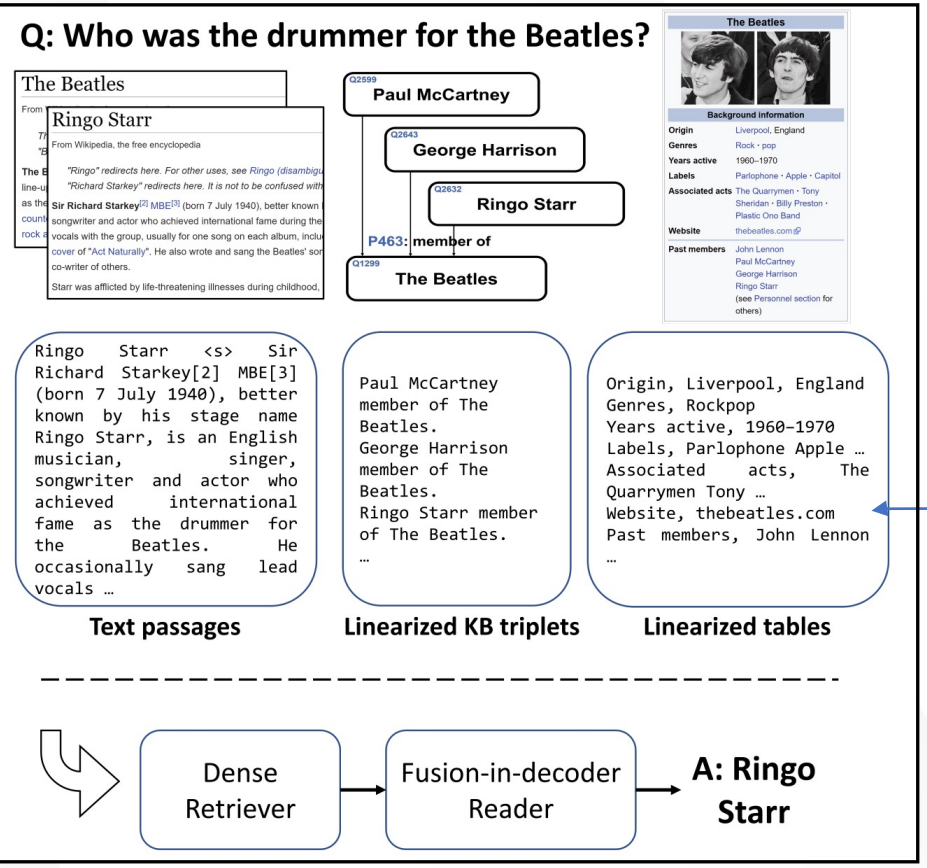
Appendix -- Unik-QA

UNIVERSITY OF NOTRE DAME

-- Existing work: add tables / KG triples into retrieval corpus

Work [3]: Wikipedia -> Wikipedia + Wikitable / Wikidata (converted to text using **template**)

Model	NQ	WebQ	Trivia	TREC	Avg.
SoTA	51.4[1]	55.1[3]	67.6[1]	55.3[2]	57.3
Retrieval-free	28.5[4]	30.6[4]	28.7[4]	-	-
Per-dataset models					
Text	49.0	50.6	64.0	54.3	54.5
Tables	36.0	41.0	34.5	32.7	36.1
KB	27.9	55.6	35.4	32.4	37.8
Text + tables	**54.1**	50.2	**65.1**	53.9	55.8
Text + tables + KB	54.0	**57.8**	64.1	**55.3**	**57.8**
Multi-dataset model					
Text	50.3	45.0	62.6	45.7	50.9
Tables	34.2	38.4	33.7	31.1	34.4
KB	25.9	43.3	34.2	38.0	35.4
Text + tables	**54.6**	44.3	**64.0**	48.7	52.9
Text + tables + KB	53.7	**55.5**	63.4	**51.3**	**56.0**



[3] UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain QA. NAACL 2022. Meta AI.

Acknowledgement

- Funding agency:

UNIVERSITY OF NOTRE DAME
College of Engineering