Knowledge Injection & Domain Adaptation

Computer Science Dept, Yonsei University

Seungone Kim

Referenced Papers

- Prerequisites
  - COMET : Commonsense Transformers for Automatic Knowledge Graph Construction (ACL 2019)
  - Ernie : Enhanced Language Representation with Informative Entities (ACL 2019)
  - SciBERT : A Pretrained Language Model for Scientific Text (EMNLP-IJCNLP 2019)

- Key Papers
  - Language Models as Knowledge Bases? (EMNLP-IJCNLP 2019)
  - How Much Knowledge Can You Pack into the Parameters of a Language Model? (EMNLP 2020)
  - Birds Have Four Legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models (EMNLP 2020)
  - Don’t Stop Pretraining : Adapt Language Models to Domains and Tasks (ACL 2020)
  - Unsupervised Domain Adaptation through Language Modeling (NAACL 2021)
  - Improving Question Answering with External Knowledge (ACL 2019 Workshop)
  - Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Rating (EMNLP 2021 Workshop)
  - Knowledge enhanced contextual word representations (EMNLP-IJACI 2019)
  - K-bert : Enabling language representations with knowledge graph (AAAI 2020)
  - LUKE : Deep Contextualized Entity Representations with Entity-aware Self-Attention (EMNLP 2020)
  - KEPLER : A Unified Model for Knowledge Embedding and Pre-trained Language Representation (TACL 2021)

Referenced Papers

- Prerequisites
  - COMET : Commonsense Transformers for Automatic Knowledge Graph Construction (ACL 2019)
  - Ernie : Enhanced Language Representation with Informative Entities (ACL 2019)
  - SciBERT : A Pretrained Language Model for Scientific Text (EMNLP-IJCNLP 2019)

- Key Papers
  - Language Models as Knowledge Bases? (EMNLP-IJCNLP 2019)
  - How Much Knowledge Can You Pack into the Parameters of a Language Model? (EMNLP 2020)
  - Birds Have Four Legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models (EMNLP 2020)
  - Don't Stop Pretraining : Adapt Language Models to Domains and Tasks (ACL 2020)
  - Unsupervised Domain Adaptation through Language Modeling (NAACL 2021)
  - Improving Question Answering with External Knowledge (ACL 2019 Workshop)
  - Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Rating (EMNLP 2021 Workshop)
  - Knowledge enhanced contextual word representations (EMNLP-IJACI 2019)
  - K-bert : Enabling language representations with knowledge graph (AAAI 2020)
  - LUKE : Deep Contextualized Entity Representations with Entity-aware Self-Attention (EMNLP 2020)
  - KEPLER : A Unified Model for Knowledge Embedding and Pre-trained Language Representation (TACL 2021)

Can we inject knowledge / cultivate reasoning abilities to LMs?

=> Keywords :

Commonsense Knowledge / Domain Knowledge / Logical Reasoning

=> Main Proposal :

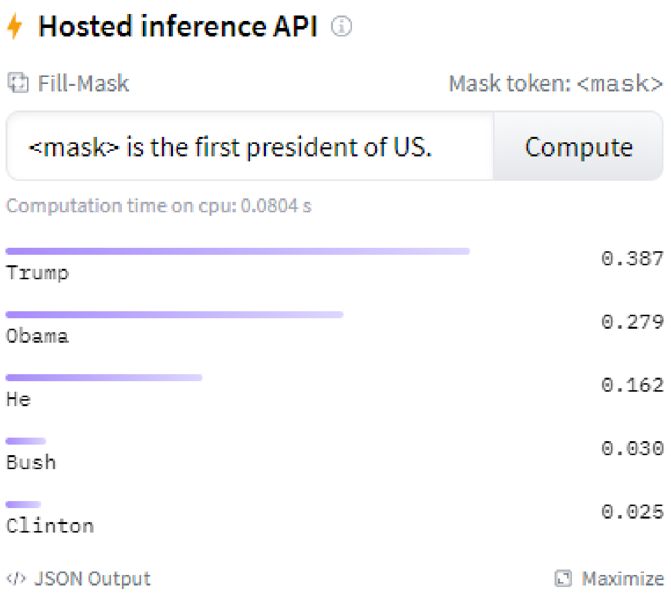
Instead of understanding knowledge and logically reasoning

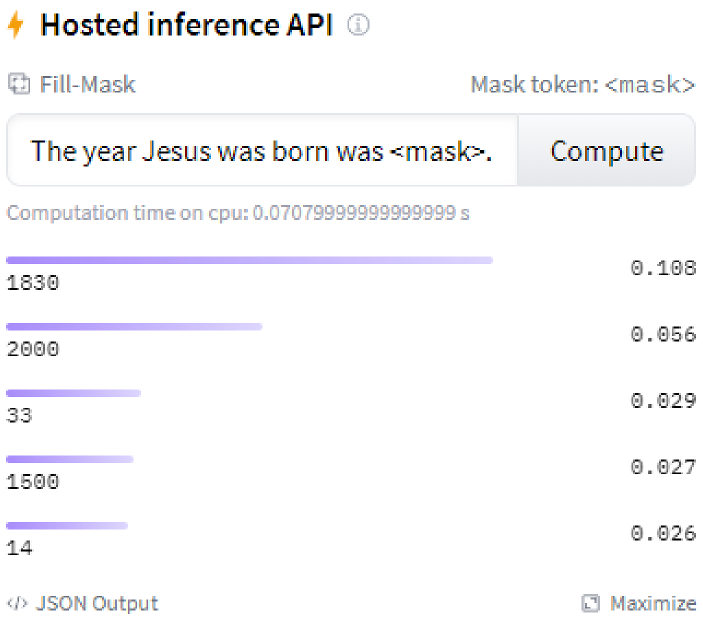
based on what they learned,

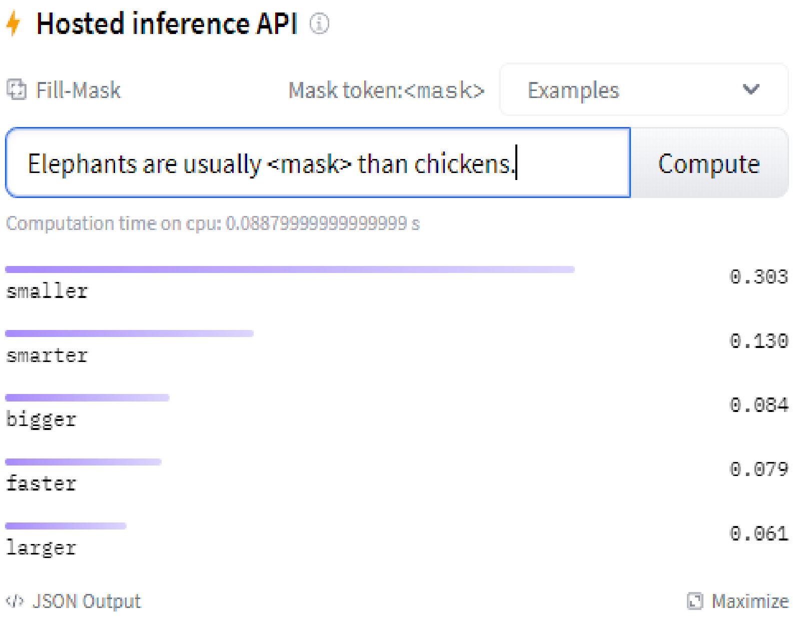
current LMs are just memorizing the patterns they saw during training.

LMs lack Commonsense Knowledge

- The Questions below are questions that require commonsense knowledge.
- The Answers were generated with a pretrained *Roberta-base model*.
- Instead of understanding the information the model saw during training, it is most likely that the model is just memorizing the patterns of the sequence data.

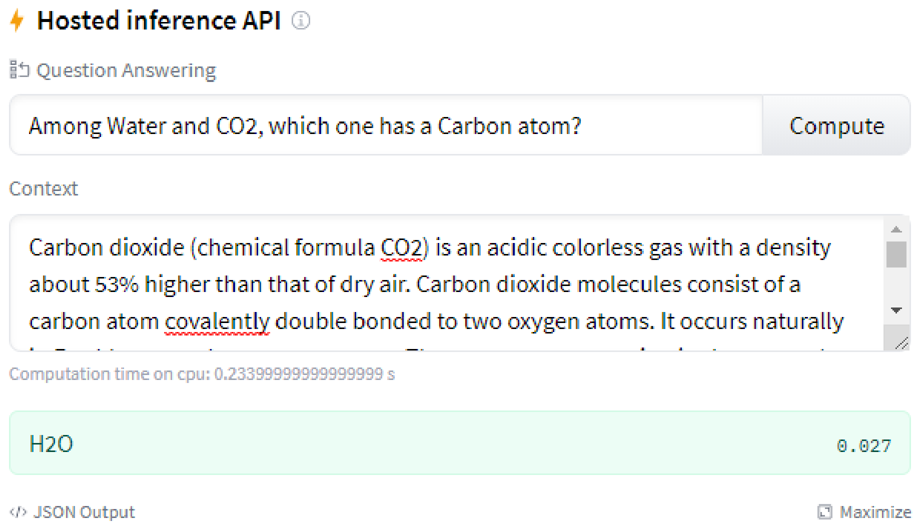






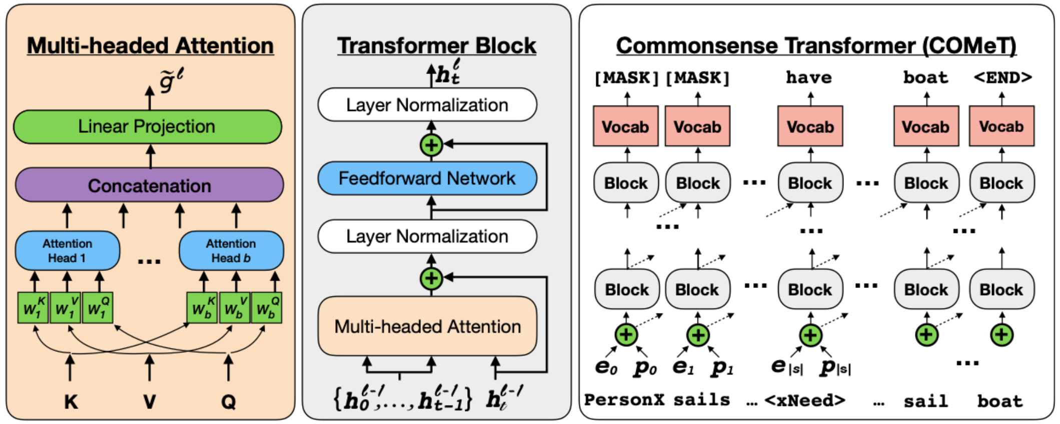
LMs lack ability to learn Domain Knowledge

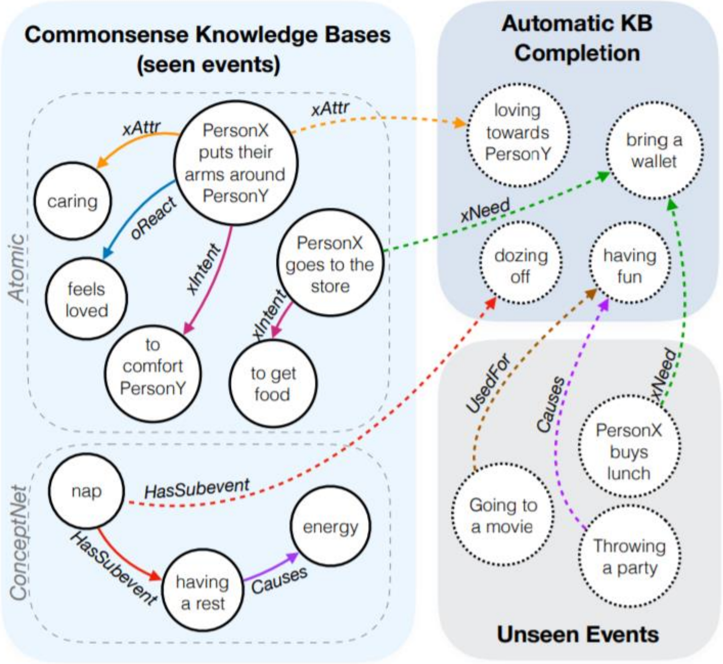
- The Questions below are questions that require domain knowledge.
- The Answers were generated with a finetuned(on SQuAD) *Roberta-base model*.
- Even when the answer could be found within the context(Reading Comprehension), LMs couldn't answer questions if the question is a Domain Specific Query.
- This is because LMs aren't really understanding what they are reading, but instead just memorizing the pattern in which they saw during training. This becomes a more serious problem in Domain Specific tasks than in commonsense.
- For better performance in QA Systems, LMs should understand the entities and relations between them when they are trained on a domain specific corpus. (**<u>Quality over Quantity</u>**)



Commonsense Knowledge Graph

- COMET : Commonsense Transformers for Automatic Knowledge Graph Construction (ACL 2019)
  - COMET is a framework for adapting the weights of language models to learn to produce novel and diverse commonsense knowledge tuples.
  - Using two Commonsense Knowledge Bases, ATOMIC and ConceptNet, COMET produces novel commonsense knowledge.
  - COMET uses GPT as a baseline LM to train on with BLEU-2 as a metric.



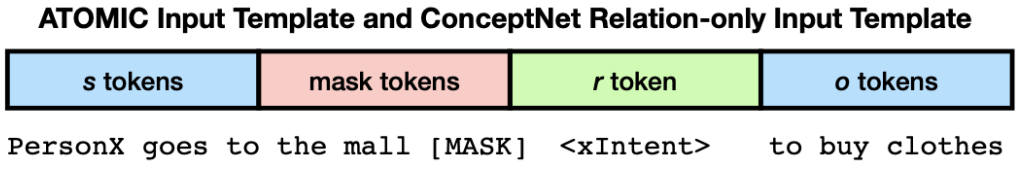


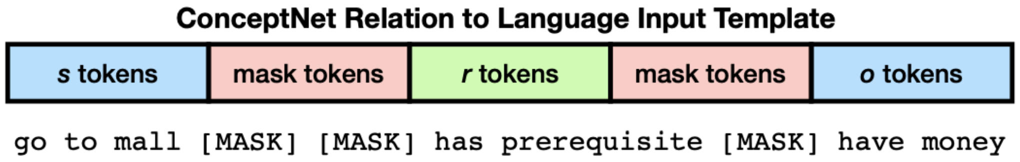
Seed	Relation	Completion	Plausible
piece	PartOf	machine	✓
bread	IsA	food	✓
oldsmobile	IsA	car	✓
happiness	IsA	feel	✓
math	IsA	subject	✓
mango	IsA	fruit	✓
maine	IsA	state	✓
planet	AtLocation	space	✓
dust	AtLocation	fridge	
puzzle	AtLocation	your mind	🤔
college	AtLocation	town	✓
dental chair	AtLocation	dentist	✓
finger	AtLocation	your finger	
sing	Causes	you feel good	✓
doctor	CapableOf	save life	✓
post office	CapableOf	receive letter	✓
dove	SymbolOf	purity	✓
sun	HasProperty	big	✓
bird bone	HasProperty	fragile	✓
earth	HasA	many plant	✓
yard	UsedFor	play game	✓
get pay	HasPrerequisite	work	✓
print on printer	HasPrerequisite	get printer	✓
play game	HasPrerequisite	have game	✓
live	HasLastSubevent	die	✓
swim	HasSubevent	get wet	✓
sit down	MotivatedByGoal	you be tire	✓
all paper	ReceivesAction	recycle	✓
chair	MadeOf	wood	✓
earth	DefinedAs	planet	✓

Commonsense Knowledge Graph

- COMET : Commonsense Transformers for Automatic Knowledge Graph Construction (ACL 2019)
  - Authors use MLM training objective to fill in the o tokens given the s and r tokens.
  - r tokens are learned during fine tuning.
  - Within the ablation studies, it was proved empirically that using pretrained weights outperforms randomly initialized weights.
  - Also, it was proved empirically that using Greedy Decoding outperforms Beam Search or Random Sampling during decoding.
  - Using COMET, we can explicitly extract knowledge from a pretrained LM and represent relations with language.
  - Demo :
  - https://mosaickg.apps.allenai.org/model-comet2020

Model	PPL[5]	BLEU-2	N/T *sro*[6]	N/T *o*	N/U *o*
9ENC9DEC (Sap et al., 2019)	-	10.01	100.00	8.61	40.77
NearestNeighbor (Sap et al., 2019)	-	6.61	-	-	-
Event2(IN)VOLUN (Sap et al., 2019)	-	9.67	100.00	9.52	45.06
Event2PERSONX/Y (Sap et al., 2019)	-	9.24	100.00	8.22	41.66
Event2PRE/POST (Sap et al., 2019)	-	9.93	100.00	7.38	41.99
COMET (- pretrain)	15.42	13.88	100.00	7.25	45.71
COMET	**11.14**	**15.10**	100.00	**9.71**	**51.20**





$$\mathcal{L} = -\sum_{t=|s|+|r|}^{|s|+|r|+|o|} \log P(x_t | x_{<t}) \qquad (11)$$

COMET Decoding method	oEffect	oReact	oWant	xAttr	xEffect	xIntent	xNeed	xReact	xWant	Avg
Top-5 random sampling (n=2500 per relation)	34.60	44.04	35.56	64.56	55.68	58.84	46.68	80.96	58.52	53.27
Top-10 random sampling (n=5000 per relation)	25.20	37.42	27.34	49.20	47.34	47.06	38.24	72.60	48.10	43.61
Beam search - 2 beams (n=1000 per relation)	43.70	54.20	47.60	**84.00**	51.10	73.80	50.70	85.80	78.70	63.29
Beam search - 5 beams (n=2500 per relation)	37.12	45.36	42.04	63.64	**61.76**	63.60	57.60	78.64	68.40	57.57
Beam search - 10 beams (n=5000 per relation)	29.02	37.68	44.48	57.48	55.50	68.32	64.24	76.18	75.16	56.45
Greedy decoding (n=500 per relation)	**61.20**	**69.80**	**80.00**	77.00	53.00	**89.60**	**85.60**	**92.20**	**89.40**	**77.53**
Human validation of gold ATOMIC	84.62	86.13	83.12	78.44	83.92	91.37	81.98	95.18	90.90	86.18

Language Models & Knowledge Bases

- *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

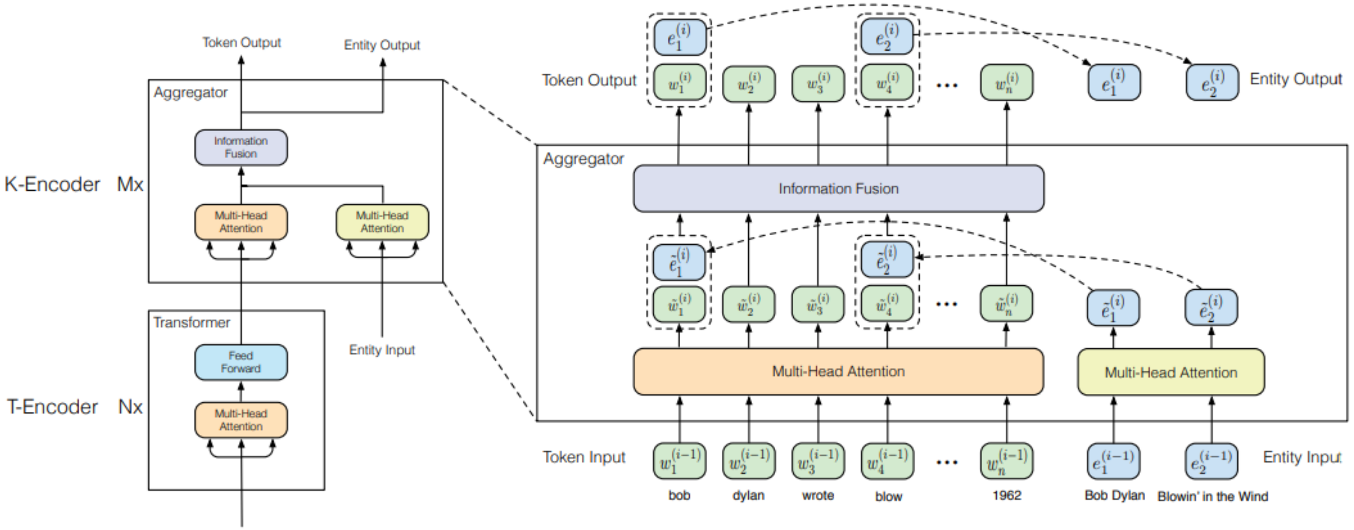
ERNIE: Enhanced Language Representation with Informative Entities

Zhengyan Zhang[1,2,3*], Xu Han[1,2,3*], Zhiyuan Liu[1,2,3†], Xin Jiang[4], Maosong Sun[1,2,3], Qun Liu[4]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Institute for Artificial Intelligence, Tsinghua University, Beijing, China
[3]State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China
[4]Huawei Noah's Ark Lab
{zhangzhengyan14,hanxu17}@mails.tsinghua.edu.cn

ERNIE : Enhanced Language Representations with Informative Entities

- **Incorporate knowledge information into language representation models**
  - For existing pretrained language representation models, "Bob Dylan wrote Blowin' in the wind" -> "UNK wrote UNK in UNK"

- **Underlying Textual Encoder(e.g. BERT) + Upper Knowledgeable Encoder(Aggregator; proposed in paper)**
  - **Underlying Textual Encoder** captures basic **lexical** and **syntactic** information from input tokens
  - **Upper Knowledgeable Encoder** integrates token-oriented **knowledge** information into textual information from underlying layer
  - Instead of using graph-based facts in KGs, **encode graph structure of KGs** with knowledge embedding algorithms like **TransE**
- **Tasks**
  - **Entity Typing**(*Upper Table*), **Relation Classification**(*Lower Table*), **GLUE**(*similar performance with BERT*)



Model	P	R	F1
NFGEC (LSTM)	68.80	53.30	60.10
UFET	77.40	60.60	68.00
BERT	76.37	70.96	73.56
ERNIE	**78.42**	**72.90**	**75.56**

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

Domain Adaptation

- *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.*

SciBERT: A Pretrained Language Model for Scientific Text

Iz Beltagy **Kyle Lo** **Arman Cohan**
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,kylel,armanc}@allenai.org

SciBERT : A Pretrained Language Model for Scientific Text

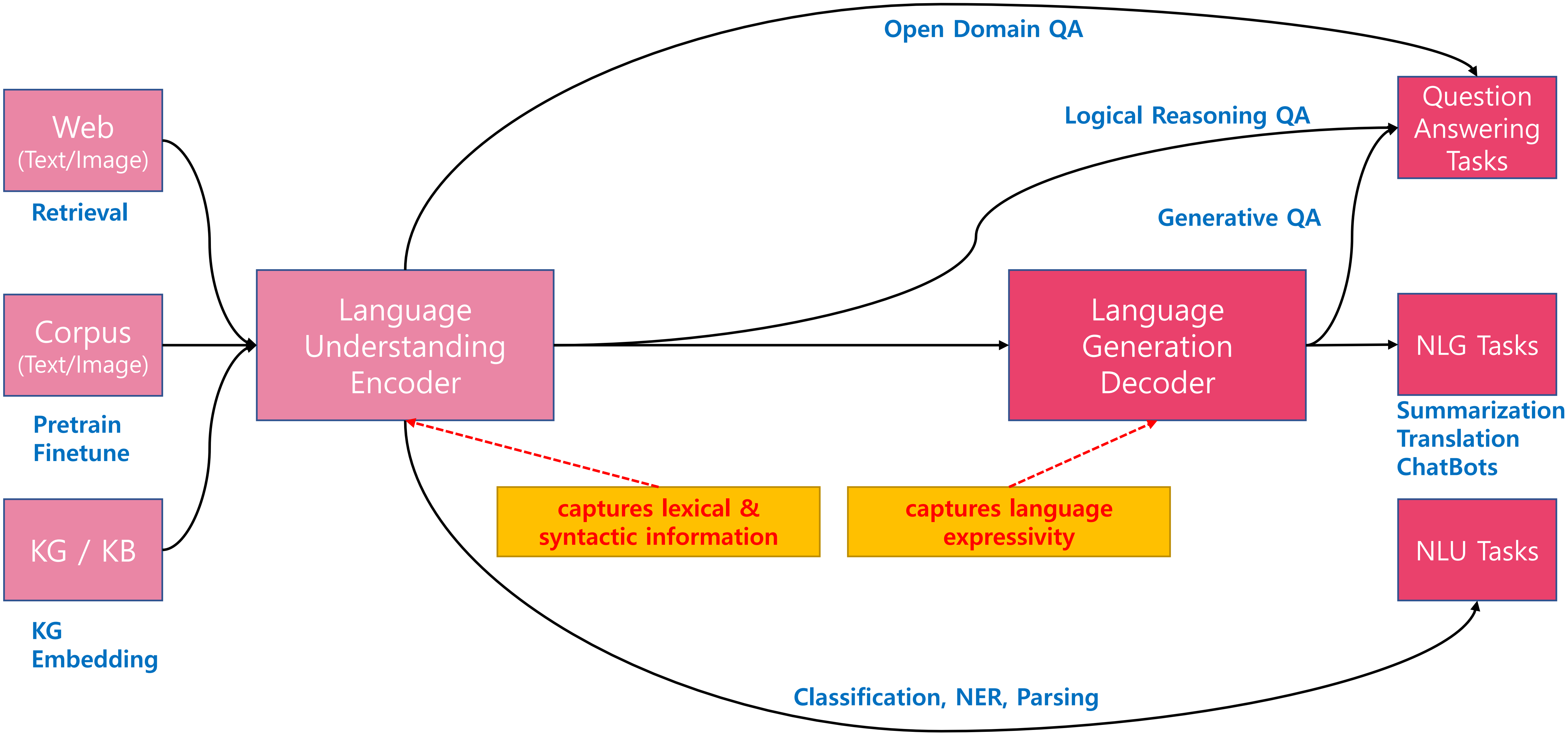
- **Perform Unsupervised Pretraining on a large multi-domain corpus of scientific publications**
  - In scientific domains, annotated data is difficult and expensive to collect due to expertise required for quality annotation
  - Similar approach with BioBERT, but SciBERT obtains better results

- **SciVocab vs BERTVocab**
  - Token overlap is 42%, illustrating substantial difference in frequently used words

- **Tasks**
  - **Named Entity Recognition**(NER)
  - **PICO Extraction**(PICO) : similar with NER, Spans are from medical domain
  - **Text Classification**(CLS)
  - **Relation Classification**(REL) : similar with CLS, Predicting relationship
  - **Dependency Parsing**(DEP)

Field	Task	Dataset	SOTA	BERT-Base		SciBERT	
				Frozen	Finetune	Frozen	Finetune
Bio	NER	BC5CDR (Li et al., 2016)	88.85[7]	85.08	86.72	88.73	**90.01**
		JNLPBA (Collier and Kim, 2004)	**78.58**	74.05	76.09	75.77	77.28
		NCBI-disease (Dogan et al., 2014)	**89.36**	84.06	86.88	86.39	88.57
	PICO	EBM-NLP (Nye et al., 2018)	66.30	61.44	71.53	68.30	**72.28**
	DEP	GENIA (Kim et al., 2003) - LAS	**91.92**	90.22	90.33	90.36	90.43
		GENIA (Kim et al., 2003) - UAS	**92.84**	91.84	91.89	92.00	91.99
	REL	ChemProt (Kringelum et al., 2016)	76.68	68.21	79.14	75.03	**83.64**
CS	NER	SciERC (Luan et al., 2018)	64.20	63.58	65.24	65.77	**67.57**
	REL	SciERC (Luan et al., 2018)	n/a	72.74	78.71	75.25	**79.97**
	CLS	ACL-ARC (Jurgens et al., 2018)	67.9	62.04	63.91	60.74	**70.98**
Multi	CLS	Paper Field	n/a	63.64	65.37	64.38	**65.71**
		SciCite (Cohan et al., 2019)	84.0	84.31	84.85	**85.42**	**85.49**
Average				73.58	77.16	76.01	79.27

Current Approaches



Referenced Papers

- Prerequisites
  - COMET : Commonsense Transformers for Automatic Knowledge Graph Construction (ACL 2019)
  - Ernie : Enhanced Language Representation with Informative Entities (ACL 2019)
  - SciBERT : A Pretrained Language Model for Scientific Text (EMNLP-IJCNLP 2019)

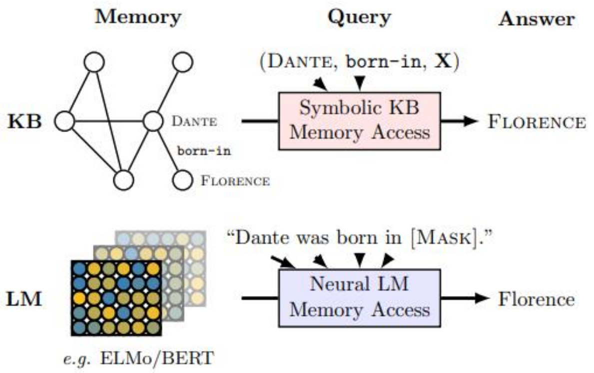
- Key Papers
  - Language Models as Knowledge Bases? (EMNLP-IJCNLP 2019)
  - How Much Knowledge Can You Pack into the Parameters of a Language Model? (EMNLP 2020)
  - Birds Have Four Legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models (EMNLP 2020)
  - Don't Stop Pretraining : Adapt Language Models to Domains and Tasks (ACL 2020)
  - Unsupervised Domain Adaptation through Language Modeling (NAACL 2021)
  - Improving Question Answering with External Knowledge (ACL 2019 Workshop)
  - Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Rating (EMNLP 2021 Workshop)
  - Knowledge enhanced contextual word representations (EMNLP-IJACI 2019)
  - K-bert : Enabling language representations with knowledge graph (AAAI 2020)
  - LUKE : Deep Contextualized Entity Representations with Entity-aware Self-Attention (EMNLP 2020)
  - KEPLER : A Unified Model for Knowledge Embedding and Pre-trained Language Representation (TACL 2021)

P1) Language Models : Do they store knowledge?

- *Petroni, Fabio, et al. "Language Models as Knowledge Bases?." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.*

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com

- While Pretrained Language Models learn linguistic knowledge, they may also be storing relational knowledge present in data.
- Language Models have many advantages over structured knowledge bases; they require no schema engineering, allow practitioners to query about an open class of relations, are easy to extend more data, and require no human supervision to train.
- Even without any fine-tuning, these models recall factual knowledge, demonstrating their potential as unsupervised open-domain QA systems.



P1) Language Models : Do they store knowledge?

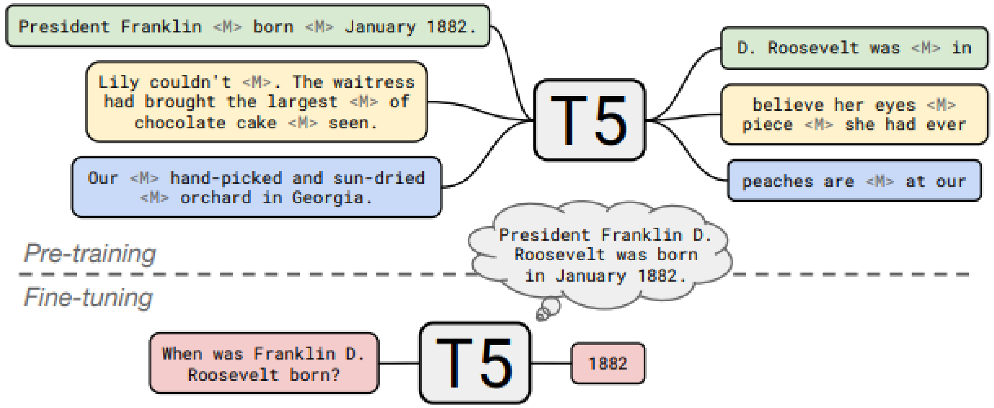
- *Roberts, Adam, Colin Raffel, and Noam Shazeer. "How Much Knowledge Can You Pack into the Parameters of a Language Model?." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.*

Adam Roberts* Google adarob@google.com

Colin Raffel* Google craffel@gmail.com

Noam Shazeer Google noam@google.com

- Pretrained Language Models can implicitly store and retrieve knowledge using natural language queries.
- Unlike providing a context or retrieving information, the authors determine how much knowledge is stored in parameters by measuring performance on a "Closed-book QA Task".
- By fine-tuning T5 model with Salient Span Masking, the authors show NN alone could obtain competitive results compared to Open Domain QA Systems.

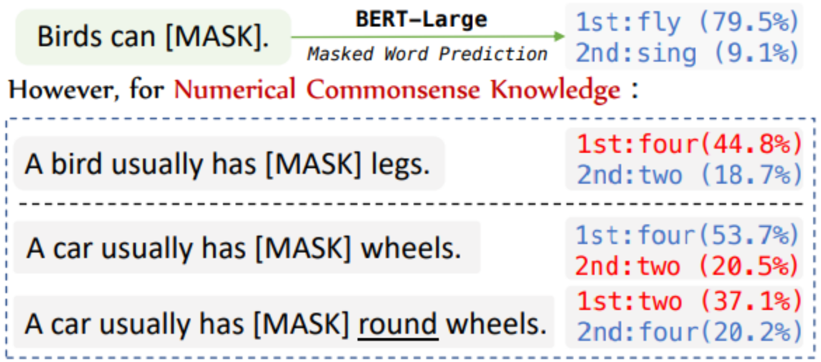


P1) Language Models with Common Sense?

- *Lin, Bill Yuchen, et al. "Birds Have Four Legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.*

Bill Yuchen Lin   **Seyeon Lee**   **Rahul Khanna**   **Xiang Ren**
{yuchen.lin,seyeonle,rahulkha,xiangren}@usc.edu
Department of Computer Science,
University of Southern California

- Pretrained Language Models are known to possess certain commonsense and factual knowledge.
- It is very promising to use Pretrained Language Models as "neural knowledge bases".
- However, in the paper, the authors claim that this may not work for numerical commonsense knowledge.

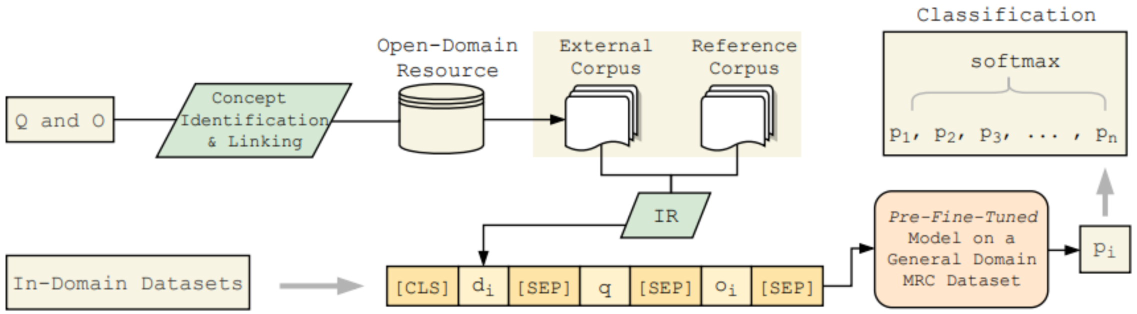


P2) Does External Unstructured Knowledge help?

- *Pan, Xiaoman, et al. "Improving Question Answering with External Knowledge." Proceedings of the 2nd Workshop on Machine Reading for Question Answering. 2019.*
- Focus on Multiple-choice QA in subject areas that require both broad background knowledge and the facts from the given subject-area.

- Identify concepts in question & answer options and link these potentially ambiguous concepts to an open-domain resource.
- Each concept mention is disambiguated and linked to corresponding concept(page) in Wikipedia. (e.g. Mercury => Mercury_(planet)) (*Pan et al, 2015*)

- Perform information retrieval based on the enriched corpus instead of the original one to form a document for answering a question. (*Lucene*)
- Compare settings where, 1) original reference corpus of each dataset is independent; 2) original reference corpora are integrated to further leverage external in-domain knowledge.
- ***Observe consistent gains by introducing knowledge from Wikipedia, employing additional in-domain training data is not uniformly helpful.***



Question: a magnet will stick to __?

A.	a belt buckle. ✓	**B.**	a wooden table.
C.	a plastic cup.	**D.**	a paper plate.

To correctly answer the question in Table 1, for example, scientific facts[1] from the provided reference corpus — {*"a magnet attracts magnetic metals through magnetism"* and *"iron is always magnetic"*}, as well as general world knowledge extracted from an external source such as {*"a belt buckle is often made of iron"* and *"iron is metal"*}

Question: Mercury, the planet nearest to the Sun, has extreme surface temperatures, ranging from 465°C in sunlight to −180°C in darkness. Why is there such a large range of temperatures on Mercury?

A. The planet is too small to hold heat.
B. The planet is heated on only one side.
C. The planet reflects heat from its dark side.
D. The planet lacks an atmosphere to hold heat. ✓

P2) Does External Unstructured Knowledge help?

- *Pan, Xiaoman, et al. "Improving Question Answering with External Knowledge." Proceedings of the 2nd Workshop on Machine Reading for Question Answering. 2019.*

Question	Answer Options	Sentence(s) From Wikipedia
What boils at the boiling point?	A. ***Kool-Aid***. ✓ B. Cotton. C. Paper Towel. D. Hair.	***Kool-Aid*** is known as Nebraska's official soft drink. Common types of drinks include plain drinking ***water***, milk, coffee, tea, hot chocolate, juice and ***soft drinks***.
Forest fires occur in many areas due to ***drought conditions***. If the drought conditions continue for a long period of time, which might cause the repopulation of trees to be threatened?	A. a decrease in the ***thickness of soil***. ✓ B. a decrease in the amount of erosion. C. an increase in the bacterium population. D. an increase in the production of oxygen and fire.	It is highly resistant to ***drought conditions***, and provides excellent fodder; and has also been used in controlling ***soil erosion***, and as revegetator, often after ***forest fires***.
Juan and LaKeisha roll a few objects down a ramp. They want to see which object rolls the farthest. What should they do so they can repeat their ***investigation?***	A. Put the objects in groups. B. Change the height of the ramp. C. Choose different objects to roll. D. ***Record*** the details of the ***investigation***. ✓	The use of measurement developed to allow ***recording*** and comparison of ***observations*** made at different times and places, by different people.
Which statement best explains why the sun appears to ***move across the sky*** each day?	A. The sun revolves around Earth. B. Earth rotates around the sun. C. The sun revolves on its axis. D. ***Earth rotates*** on its ***axis***. ✓	***Earth's rotation*** about its ***axis*** causes the fixed stars to apparently ***move across the sky*** in a way that depends on the observer's latitude.

Method	**ARC-E**	**ARC-C**	**OBQA**
IR (Clark et al., 2018)	62.6	20.3	–
Odd-One-Out (Mihaylov et al., 2018)	–	–	50.2
DGEM (Khot et al., 2018)	59.0	27.1	24.4
KG^2 (Zhang et al., 2018)	–	31.7	–
AIR (Yadav et al., 2018)	58.4	26.6	–
NCRF++ (Musa et al., 2018)	52.2	33.2	–
TriAN++ (Zhong et al., 2018)	–	33.4	–
Two Stage Inference (Pirtoaca et al., 2019)	61.1	26.9	–
ET-RR (Ni et al., 2019)	–	36.6	–
GPT^{II} (Radford et al., 2018; Sun et al., 2019)	57.0	38.2	52.0
RS^{II} (Sun et al., 2019)	66.6	40.7	55.2
Our BERT-Based Implementations			
Setting 1			
Reference Corpus (RC) (i.e., $BERT^{II}$)	71.9	44.1	64.8
External Corpus (EC)	65.0	39.4	62.2
RC + EC	73.3	45.0	65.2
Setting 2			
Integrated Reference Corpus (IRC)	73.2	44.8	65.0
Integrated External Corpus (IEC)	68.9	40.1	63.0
IRC + IEC	**74.7**	46.1	67.0
IRC + MD	69.4	50.7	67.4
IRC + IEC + MD	72.3	**53.7**	**68.0**
Human Performance	–	–	91.7

P2) Does External Unstructured Knowledge help?

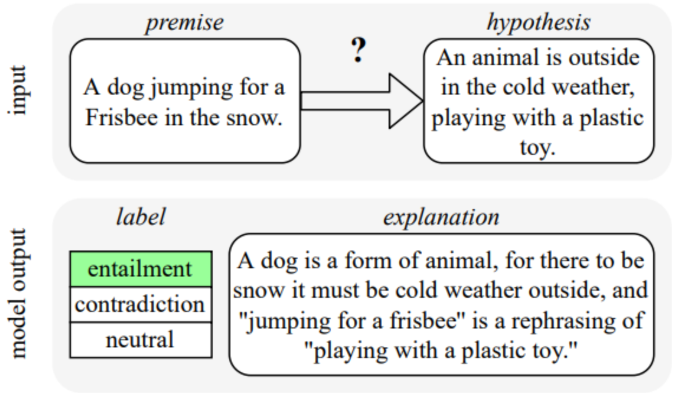
- *Schuff, Hendrik, et al. "Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Ratings." Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. 2021.*
- Natural Language Inference(NLI) is closely related to real-world applications, such as fact checking.
- Solving the task requires models to not only reason over provided information but also to link it with commonsense knowledge.

- Following a model's reasoning process is valuable to ML engineers as well as end users.
- Former can use insights to improve models and latter can base their decision on them whether to trust the system or not.
- One approach to gain insight into a system is to train it to generate explanations as an additional output.

- Q1) Does the positive effect of external knowledge on the inference ability transfer to the generation of explanations?
- Q2) How effective is the implicit commonsense knowledge of LMs compared to symbolic sources of knowledge, such as knowledge base triplets?
- Q3) How do humans perceive explanation quality of SOTA NLI models?

Type	Model	Label Acc.	BLEU	BLEURT
non-LM	PRED-EXPL	84.21	19.77	-0.871
	VANILLA	89.20	19.71	-0.820
	COMET	88.97	18.84	-0.822
	CONT	89.02	20.1	-0.799
	COMET+CONT	89.07	19.66	-0.809
LM-based	GPT-EF	87.89	21.70	-0.624
	GPT-LF	89.70	26.90	-0.577
	ENSEMBLE	90.24	27.10	-0.576
	FILTERED ENS	90.24	27.09	-0.577
	NILE:POST-HOC	91.49	26.26	-0.577
	WT5-11B	**92.3**	**29.01**	**-0.511**

			Competence Test		Distraction Test			Noise Test
Type	Model	Total	Antonymy	Numerical	Word Overlap	Length Mismatch	Negation	Spelling
non-LM	PRED-EXPL	48.69	36.36	36.55	47.17	53.44	45.31	52.42
	VANILLA	56.94	37.94	32.24	55.46	65.21	52.03	62.90
	COMET	57.05	34.54	35.48	57.31	64.15	52.85	62.33
	CONT	57.09	32.50	**40.28**	52.10	64.35	**53.38**	62.77
	COMET+CONT	56.26	44.43	34.16	51.34	64.39	49.36	63.03
LM-based	GPT-EF	52.74	51.81	31.33	55.91	60.97	38.44	58.20
	GPT-LF	**59.28**	**54.84**	28.80	**64.06**	**68.72**	42.82	67.07
	ENSEMBLE	59.19	37.97	34.03	58.13	67.45	52.51	65.92
	FILTERED-ENS	58.99	52.53	28.54	63.70	68.02	42.18	**67.10**



P3) Domain Adaptation & Task Adaptation

- *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.*

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

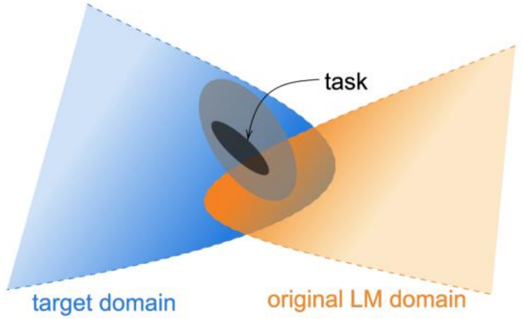
Suchin Gururangan[†] Ana Marasović[†◇] Swabha Swayamdipta[†]
Kyle Lo[†] Iz Beltagy[†] Doug Downey[†] Noah A. Smith[†◇]

[†]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
{suching,anam,swabhas,kylel,beltagy,dougd,noah}@allenai.org

Don’t Stop Pretraining : Adapt Language Models to Domains and Tasks

- **DAPT(Domain Adaptive Pretraining) & TAPT(Task Adaptive Pretraining)**
  - **(DAPT) Continue pretraining** RoBERTa on a large corpus of **unlabeled domain-specific text**
  - Four domains – biomedical papers(BIO), computer science papers(CS), news text(NEWS), amazon reviews(REVIEWS)

  - **(TAPT) Pretraining** on the **unlabeled training set for a given task**
  - Task data is a narrowly-defined subset of the broader domain
  - **Methods to retrieve unlabeled text that aligns with task distribution(kNN) boosts up performance***
  - Applying both leads to best performance



- **Like SciBERT and BioBERT, build a new domain vocabulary**
  - While similar with NEWS and REVIEWS, CS and BIO are far more dissimilar

- **Dataset / Task**
  - Every task is a classification task and measured with F1 scores
  - **CHEMPROT** : Relation Classification
  - **RCT(PubMed)** : Abstract Sentence Roles
  - **ACL-ARC** : Citation Intent
  - **SCIERC** : Relation Classification
  - **HyperPartisan** : Partisanship
  - **AGNews** : topic
  - **HELPFULNESS** : review helpfulness
  - **IMDB** : review sentiment

Domain	Task	RoBERTa	Additional Pretraining Phases: DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

P3) Domain Adaptation & Task Adaptation

- *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

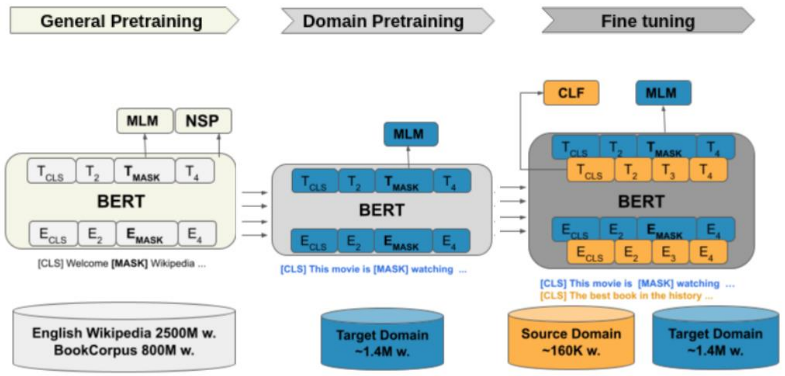
UDALM: Unsupervised Domain Adaptation through Language Modeling

Constantinos Karouzos[1], Georgios Paraskevopoulos[1,4], Alexandros Potamianos[1,2,3]

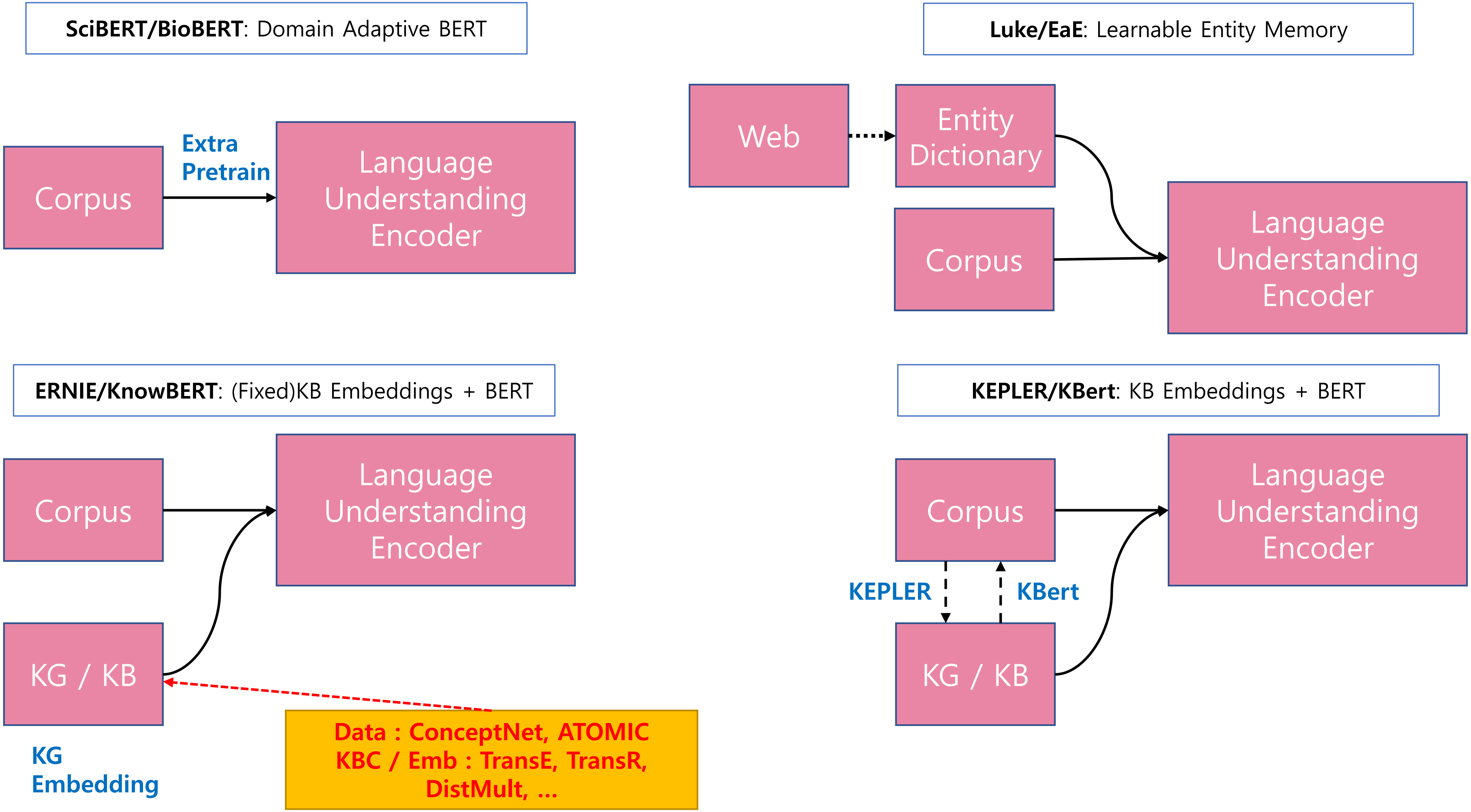
[1] School of ECE, National Technical University of Athens, Athens, Greece
[2] Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA, USA
[3] Behavioral Signal Technologies, Los Angeles, CA, USA
[4] Institute for Language and Speech Processing, Athena Research Center, Athens, Greece
ckarouzos@gmail.com, geopar@central.ntua.gr, potam@central.ntua.gr

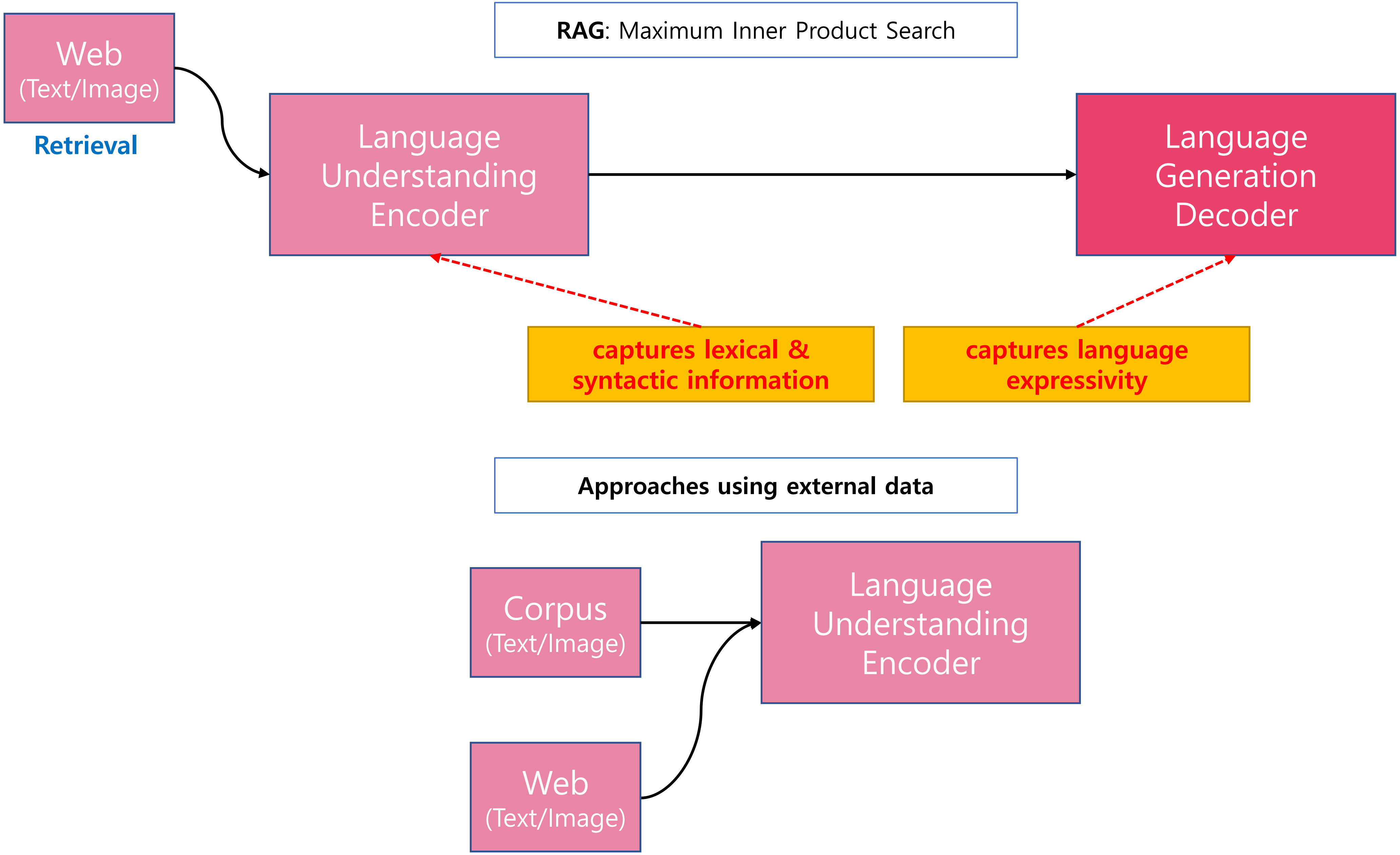
UDALM : Unsupervised Domain Adaptation through Language Modeling

- **Unsupervised Domain Adaptation(UDA)**
  - Model is trained with data from a specific domain, and then **optimized for use in new setting**
  - **Simultaneously learn** task from labeled data in source distribution, while adapting to target distribution with **multi-task learning**
  - $\mathcal{L}(s,t) = \lambda\mathcal{L}_{CLF}(s) + (1-\lambda)\mathcal{L}_{MLM}(t),\ where\ \lambda = \frac{n}{n+m},\ n = (\#\ labeled\ source\ data),\ m = (\#\ unlabeled\ target\ data)$
- **Approaches for Unsupervised Domain Adaptation**
  - Pseudo-labeling Approaches
  - Domain Adversarial Training Approaches (Base Line Model : BERT-DAAT, *Du et al., 2020*)
  - Pivot-based Approaches (Base Line Model : R-PERL, *Ben-David et al., 2020*)
  - Self Training Approaches (XLM-R based p+CFd, *Ye et al., 2020*)
- **Dataset / Task**
  - Amazon reviews multi-domain sentiment dataset with 4 domains (B: Books, D: DVDs, E: Electronics, K: Kitchen appliances)



	R-PERL	DAAT	p+CFd	SO BERT	DAT BERT	DPT BERT	UDALM
$B \to D$	87.8	90.9	87.7	89.51 ± 0.76	87.31 ± 2.14	90.49 ± 0.38	**90.97** ± 0.22
$B \to E$	87.2	88.9	91.3	90.51 ± 0.51	86.91 ± 2.71	90.38 ± 1.59	**91.69** ± 0.31
$B \to K$	90.2	88.0	92.5	91.75 ± 0.28	90.59 ± 1.17	92.66 ± 0.43	**93.21** ± 0.22
$D \to B$	85.6	89.7	**91.5**	90.26 ± 0.64	86.30 ± 3.10	91.02 ± 0.75	91.00 ± 0.42
$D \to E$	89.3	90.1	91.6	88.71 ± 1.48	87.85 ± 1.24	91.03 ± 0.82	**92.30** ± 0.47
$D \to K$	90.4	88.8	92.5	91.22 ± 0.69	89.95 ± 1.53	92.30 ± 0.42	**93.66** ± 0.37
$E \to B$	90.2	89.6	88.7	87.96 ± 0.89	85.65 ± 1.91	88.52 ± 0.55	**90.61** ± 0.30
$E \to D$	84.8	**89.3**	88.2	87.37 ± 0.64	83.99 ± 1.31	87.85 ± 0.47	88.83 ± 0.61
$E \to K$	91.2	91.7	93.6	93.30 ± 0.50	92.45 ± 1.35	94.39 ± 0.72	**94.43** ± 0.24
$K \to B$	83.0	**90.8**	89.8	88.15 ± 0.64	85.07 ± 1.03	88.83 ± 0.81	90.29 ± 0.51
$K \to D$	85.6	**90.5**	87.8	87.23 ± 0.49	84.11 ± 0.62	88.52 ± 0.69	89.54 ± 0.59
$K \to E$	91.2	93.2	92.6	93.23 ± 0.34	92.07 ± 0.24	93.42 ± 0.40	**94.34** ± 0.26
Average	87.50	90.12	90.63	89.93 ± 0.65	87.68 ± 1.53	90.78 ± 0.67	**91.74** ± 0.38


SciBERT/BioBERT: Domain Adaptive BERT
Corpus
Extra Pretrain
Language Understanding Encoder
Luke/EaE: Learnable Entity Memory
Web
Entity Dictionary
Corpus
Language Understanding Encoder
ERNIE/KnowBERT: (Fixed)KB Embeddings + BERT
Corpus
Language Understanding Encoder
KG / KB
KG Embedding
Data : ConceptNet, ATOMIC
KBC / Emb : TransE, TransR, DistMult, ...
KEPLER/KBert: KB Embeddings + BERT
Corpus
Language Understanding Encoder
KEPLER
KBert
KG / KB



RAG: Maximum Inner Product Search
Web
(Text/Image)
Retrieval
Language
Understanding
Encoder
Language
Generation
Decoder
captures lexical &
syntactic information
captures language
expressivity
Approaches using external data
Corpus
(Text/Image)
Web
(Text/Image)
Language
Understanding
Encoder

P4) Does External Structured Knowledge help?

- *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.*

Knowledge Enhanced Contextual Word Representations

Matthew E. Peters[1], Mark Neumann[1], Robert L. Logan IV[2], Roy Schwartz[1,3], Vidur Joshi[1], Sameer Singh[2], and Noah A. Smith[1,3]

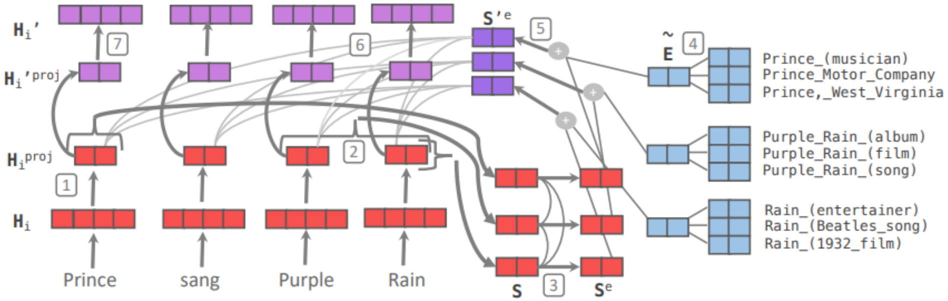
[1]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[2]University of California, Irvine, CA, USA
[3]Paul G. Allen School of Computer Science & Engineering, University of Washington
{matthewp,markn,roys,noah}@allenai.org
{rlogan,sameer}@uci.edu

Knowledge Enhanced Word Representations

- **(Key Idea)** Explicitly model entity spans in the input text and use an entity linker to retrieve relevant entity embeddings from a KB to form knowledge enhanced entity-span representations
- Knowledge Attention and Recontextualization(KAR) mechanism
- The entire KAR is inserted between two layers in the middle of a pretrained model such as BERT

- Important to align entity embeddings with pretrained BERT contextual representations.
- $\tilde{e}_m = \sum_k \tilde{\psi}_{mk}$ *where* $\psi_{mk} = MLP(p_{mk}, s_m^e \cdot e_{mk})$ *and* $\tilde{\psi}_{mk}$ *is softmax of* ψ_{mk} *candidates.*
- ψ_{mk} acts as a score to choose between candidates, and is also used in loss function.

- Candidate selector uses a rule-based lemmatizer.
- Both KnowBERT-Wiki & KnowBERT-WordNet insert KB between layer 10,11.
- To test ability to recall facts from KBs, extract 90K tuples from Wikidata for 17 different relationships written in natural language.



System	PPL	Wikidata MRR	# params. masked LM	# params. KAR	# params. entity embed.	Fwd. / Bwd. time
$\text{BERT}_{\text{BASE}}$	5.5	0.09	110	0	0	0.25
$\text{BERT}_{\text{LARGE}}$	4.5	0.11	336	0	0	0.75
KnowBert-Wiki	4.3	0.26	110	2.4	141	0.27
KnowBert-WordNet	4.1	0.22	110	4.9	265	0.31
KnowBert-W+W	**3.5**	**0.31**	110	7.3	406	0.33

P4) Does External Structured Knowledge help?

- ***Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 03. 2020.***

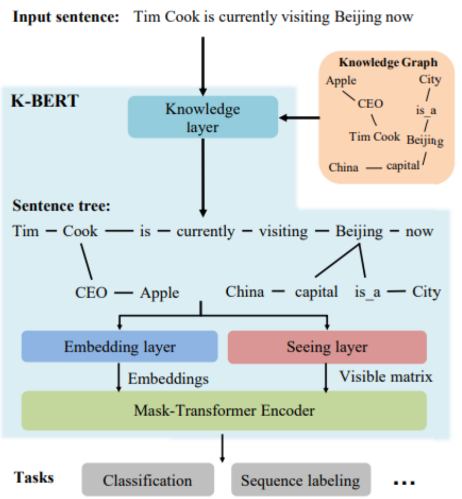
K-BERT: Enabling Language Representation with Knowledge Graph

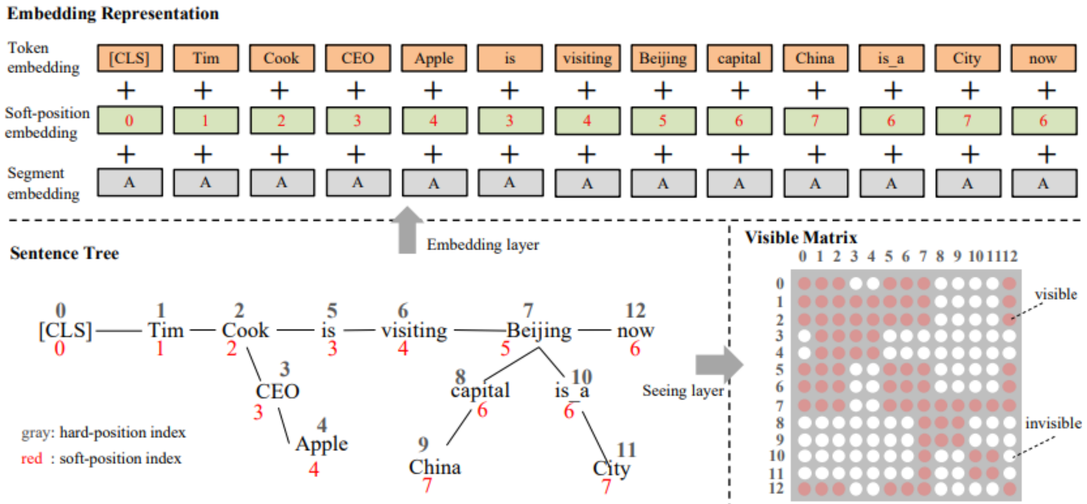
Weijie Liu,[1] Peng Zhou,[2] Zhe Zhao,[2] Zhiruo Wang,[3] Qi Ju,[2,*]
Haotang Deng,[2] Ping Wang[1,*]
[1]Peking University, Beijing, China
[2]Tencent Research, Beijing, China
[3]Beijing Normal University, Beijing, China
{dataliu, pwang}@pku.edu.cn, SherronWang@gmail.com,
{rickzhou, nlpzhe, zhiruowang, damonju, haotangdeng}@tencent.com

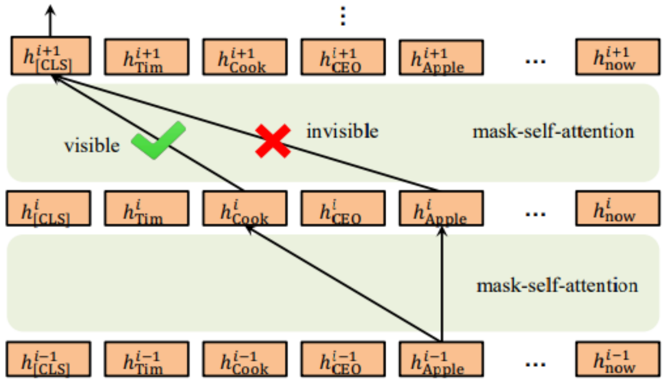
Enabling Language Representations with KGs

- **(Key Idea)** Instead of only utilizing knowledge graph embeddings, use the knowledge graph triplets to inject sentences as LM input, and using soft-position and visible matrix, limit the impact of knowledge.
- Two issues occur, 1)Heterogeneous Embedding Space(HES); 2) Knowledge Noise(KN) Issue.
- Empirical results demonstrate that KG is especially helpful for knowledge-driven specific-domain tasks.

- For input sentence, knowledge layer first injects relevant triples into it from a KG, transforming sentence into a knowledge-rich sentence tree.
- Sentence Tree is simultaneously fed into the embedding layer and the seeing layer and then converted to a token-level embedding representation and a visible matrix.
- Knowledge tree can have multiple branches, but its depth is fixed with a hyperparameter.
- Visible matrix is used to control the visible area of each token, preventing changing the meaning of the original sentence due to injection.







P4) Does External Structured Knowledge help?

- *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.*

LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention

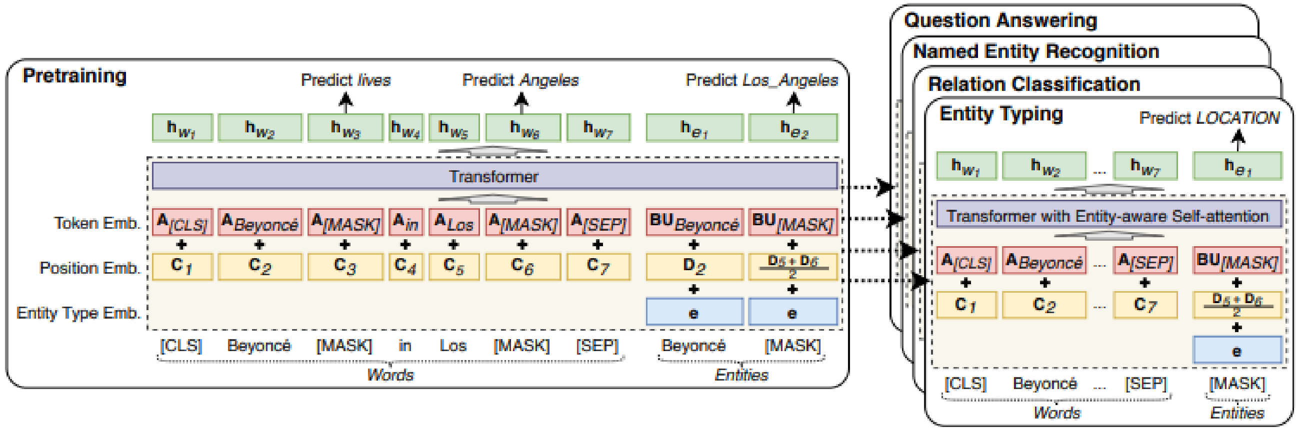
Ikuya Yamada[1,2] ikuya@ousia.jp
Akari Asai[3] akari@cs.washington.edu
Hiroyuki Shindo[4,2] shindo@is.naist.jp
Hideaki Takeda[5] takeda@nii.ac.jp
Yuji Matsumoto[2] matsu@is.naist.jp

[1]Studio Ousia [2]RIKEN AIP [3]University of Washington
[4]Nara Institute of Science and Technology [5]National Institute of Informatics

Deep Contextualized Entity Representations with Entity-aware Self-attention

- **(Key Idea)** LUKE treats not only words, but also entities as independent tokens, and computes intermediate and output representations for all tokens using the transformer architecture.
- Utilize entity-annotated corpus obtained from Wikipedia
- Contextualized Word Representations do not output span-level representations of entities, so they need to learn to compute the representations.
- It is difficult to perform reasoning about relationships between entities with self-attention because many entities are split into multiple tokens in the model.

- Instead of the Self-attention that most transformer-based architectures use, LUKE sets up a new Entity-aware Self-attention.
- Computational costs of original mechanism and proposed mechanism are identical except the additional cost of computing gradients and updating the parameters of the additional query matrices at training time.
- Outperforms baseline models in tasks that require reasoning based on relationships between entities because model easily focus on capturing the relationship between entities.



$$\mathbf{y}_i = \sum_{j=1}^{k} \alpha_{ij} \mathbf{V} \mathbf{x}_j$$

$$e_{ij} = \frac{\mathbf{K}\mathbf{x}_j^\top \mathbf{Q}\mathbf{x}_i}{\sqrt{L}}$$

$$\alpha_{ij} = \mathrm{softmax}(e_{ij})$$

$$e_{ij} = \begin{cases} \mathbf{K}\mathbf{x}_j^\top \mathbf{Q}\mathbf{x}_i, & \text{if both } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are words} \\ \mathbf{K}\mathbf{x}_j^\top \mathbf{Q}_{w2e}\mathbf{x}_i, & \text{if } \mathbf{x}_i \text{ is word and } \mathbf{x}_j \text{ is entity} \\ \mathbf{K}\mathbf{x}_j^\top \mathbf{Q}_{e2w}\mathbf{x}_i, & \text{if } \mathbf{x}_i \text{ is entity and } \mathbf{x}_j \text{ is word} \\ \mathbf{K}\mathbf{x}_j^\top \mathbf{Q}_{e2e}\mathbf{x}_i, & \text{if both } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are entities} \end{cases}$$

P4) Does External Structured Knowledge help?

- *Transactions of the Association for Computational Linguistics 9 (2021): 176-194.*

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

Xiaozhi Wang[1], **Tianyu Gao**[3], **Zhaocheng Zhu**[4,5], **Zhengyan Zhang**[1], **Zhiyuan Liu**[1,2*], **Juanzi Li**[1,2], **Jian Tang**[4,6,7*]
[1]Department of CST, BNRist; [2]KIRC, Institute for AI, Tsinghua University, Beijing, China
{wangxz20,zy-z19}@mails.tsinghua.edu.cn
{liuzy,lijuanzi}@tsinghua.edu.cn
[3]Department of Computer Science, Princeton University, Princeton, USA
tianyug@princeton.edu
[4]Mila - Québec AI Institute; [5]Univesité de Montréal; [6]HEC, Montréal, Canada
zhaocheng.zhu@umontreal.ca, jian.tang@hec.ca
[7]CIFAR AI Research Chair

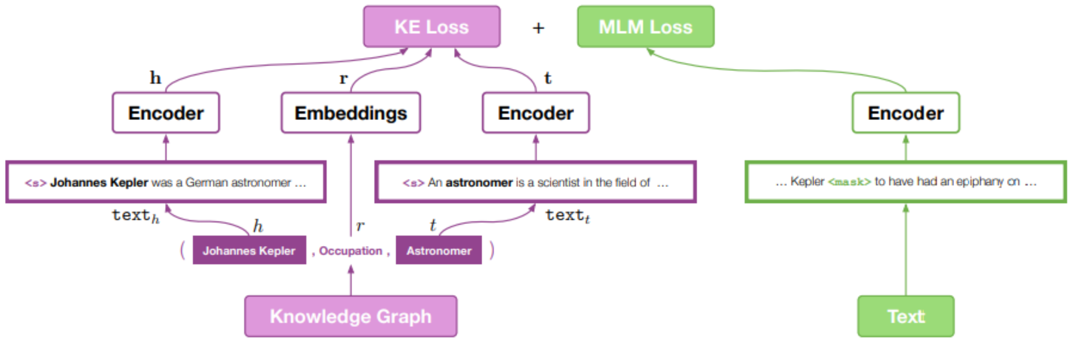
A Unified Model for Knowledge Embedding and Pre-trained Language Representation

- **(Key Idea)** Encode textual entity descriptions with a PLM as their embeddings, then jointly optimize the KE and language modeling objectives.
- While PLMs cannot capture factual knowledge from text, KGs can effectively represent the relational facts in KGs.
- However, conventional KE models cannot take full advantage of the abundant textual information.

- For the scoring function, choose to follow TransE.

- As a PLM, KEPLER is able to integrate factual knowledge into language representations with the supervision from KG by the KE objective.
- Models like Ernie that directly integrate fixed entity embeddings into PLMs, KE cannot be easily aligned with the language representation space.

- As a KE model, KEPLER can take full advantage of the abundant information from entity descriptions with the help of the MLM objective.
- While conventional KE methods are inherently transductive, KEPLER can produce embeddings for unseen entities from their descriptions.



$$\mathcal{L} = \mathcal{L}_{\mathrm{KE}} + \mathcal{L}_{\mathrm{MLM}},$$

$$\mathcal{L}_{\mathrm{KE}} = -\log \sigma(\gamma - d_r(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^{n} \frac{1}{n} \log \sigma(d_r(\mathbf{h}'_i, \mathbf{t}'_i) - \gamma),$$

$$d_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_p,$$

$$\mathbf{h} = \mathrm{E}_{<s>}(\texttt{text}_h),$$
$$\mathbf{t} = \mathrm{E}_{<s>}(\texttt{text}_t),$$
$$\mathbf{r} = \mathbf{T}_r,$$
$$\hat{\mathbf{r}} = \mathrm{E}_{<s>}(\texttt{text}_r),$$

Model	MR	MRR	HITS@1	HITS@3	HITS@10
DKRL (Xie et al., 2016)	78	23.1	5.9	32.0	54.6
RoBERTa	723	7.4	0.7	1.0	19.6
Our RoBERTa	1070	5.8	1.9	6.3	13.0
KEPLER-KE	138	17.8	5.7	22.9	40.7
KEPLER-Rel	35	33.4	15.9	43.5	66.1
KEPLER-Wiki	32	35.1	15.4	46.9	71.9
KEPLER-Cond	**28**	**40.2**	**22.2**	**51.4**	**73.0**

(b) Inductive results on Wikidata5M (% except MR).

ANY QUESTIONS?