

Neuro-Symbolic Representations for IR

Part 2.3

Infusion of Symbolic Knowledge into Text Representation

Jian-Yun Nie

University of Montreal

Outline

1. Part 1: Knowledge Graphs and Entities

1. Welcome & Motivation (Dietz)
2. Knowledge Graphs and GPT (Bast)
3. Entity Linking (Bast)

2. Part 2: Neuro-Symbolic Foundations

1. Ranking Wikipedia Entities / Aspects (Chatterjee)
2. Neural Text Representations and Semantic Annotations (Dietz)
3. Infusion of Symbolic Knowledge into Text Representation (Nie)

← We are here

3. Part 3: Reasoning, Robustness, and Relevance

1. Denoising Dense Representations with Symbols (Nogueira)
2. Reasoning about Relevance (Dalton)
3. From PRF to Retrieval Enhanced Generation (Dietz)

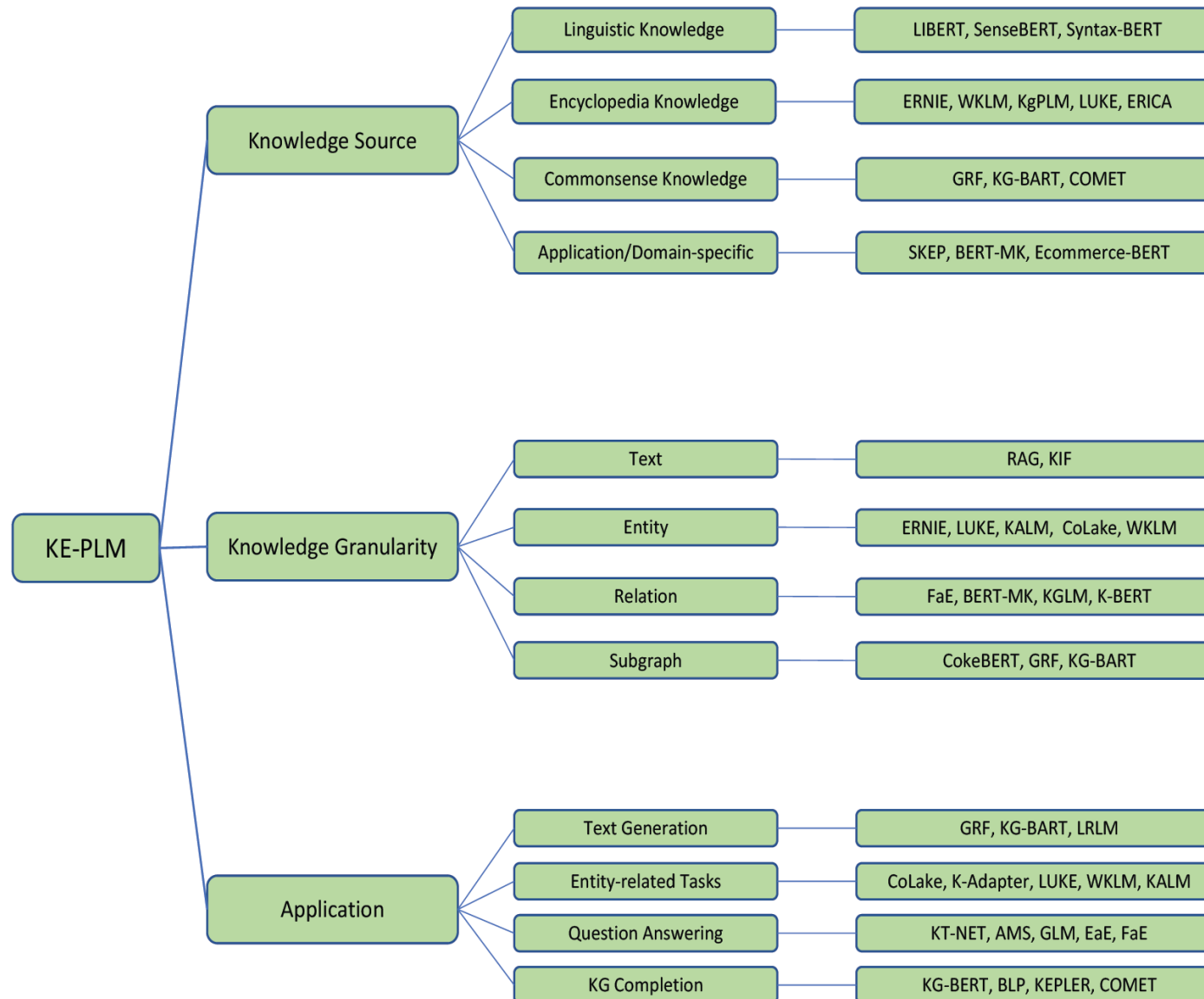
4. Part 4: Emerging Topics

1. Conclusion and Outlook
2. Panel Discussion

Motivation to infuse knowledge into LM

- Texts contain much world knowledge useful for applications
- But can be limited
 - A piece of knowledge may be infrequent in text
 - Much of our common knowledge is not stated in texts
 - E.g. A father is a male person
- Much of the knowledge implicitly captured by LM may be noise
 - True knowledge vs spurious knowledge is hard to distinguish
- Knowledge graph: Another form of structured knowledge often crafted by experts
 - From expert's domain knowledge
 - Synthetic knowledge
- Goal: Build better representations of texts for end tasks
 - Enhance true knowledge (conformity)
 - Extend the representation by what the knowledge implies

A taxonomy of knowledge-enhanced LM



Credit: Wei et al. Knowledge Enhanced Pretrained Language Models: A Comprehensive Survey, 2021

Text and knowledge graph: How to integrate?


- Naive approaches

- Text → representation
 - Knowledge graph → representation
 - E.g. Jeong et al. A Context-Aware Citation Recommendation Model with BERT and Graph Convolutional Networks. arXiv:1903.06464 (2019)
- 
- Fusion (concatenation) of representations

- More sophisticated approaches

1. Interactions between text and knowledge graph representations
2. Joint objective: LM + knowledge graph objectives
3. Enriching raw text with knowledge

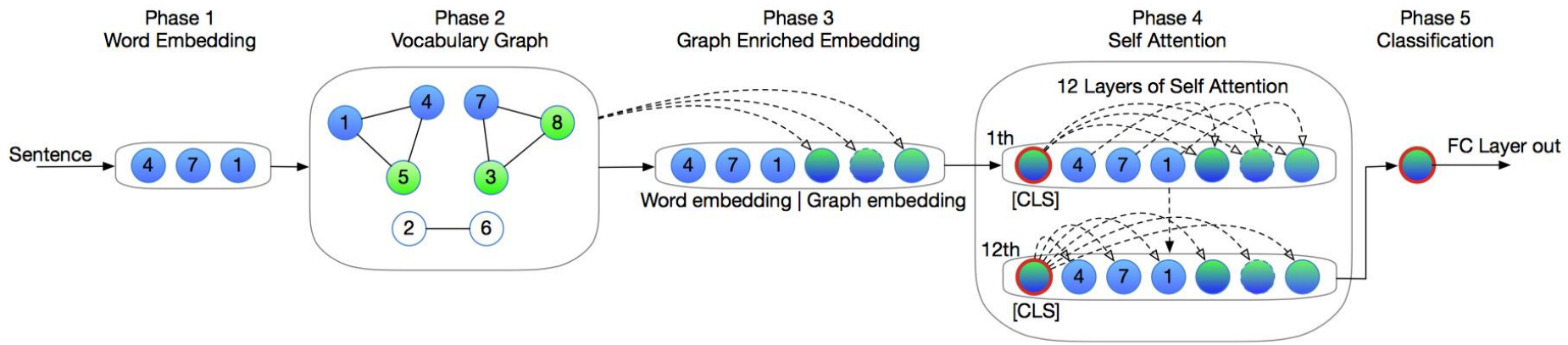
1. Interactions between text and graph representations

- Early work based on word embedding
 - Knowledge (relations/triples) as constraint to build word embeddings
 - Retrofitting (Faruqui et al. 2015): modify word embeddings to make embeddings of related words closer
 - Relation as part of the loss (Yu and Dredze, 2015)
 - Application to medical IR (Liu, Nie and Sordoni, 2016)
 - More recent work
 - BERT representation of text
 - Graph representation
- 
- Interact to create richer representations

VGCN-BERT: Combining GCN and BERT

(Lu et al. ECIR 2020)

- BERT encodes local information from a text
- Knowledge graph: global information encoded by GCN on vocabulary graph
- Attention mechanism on joint representations

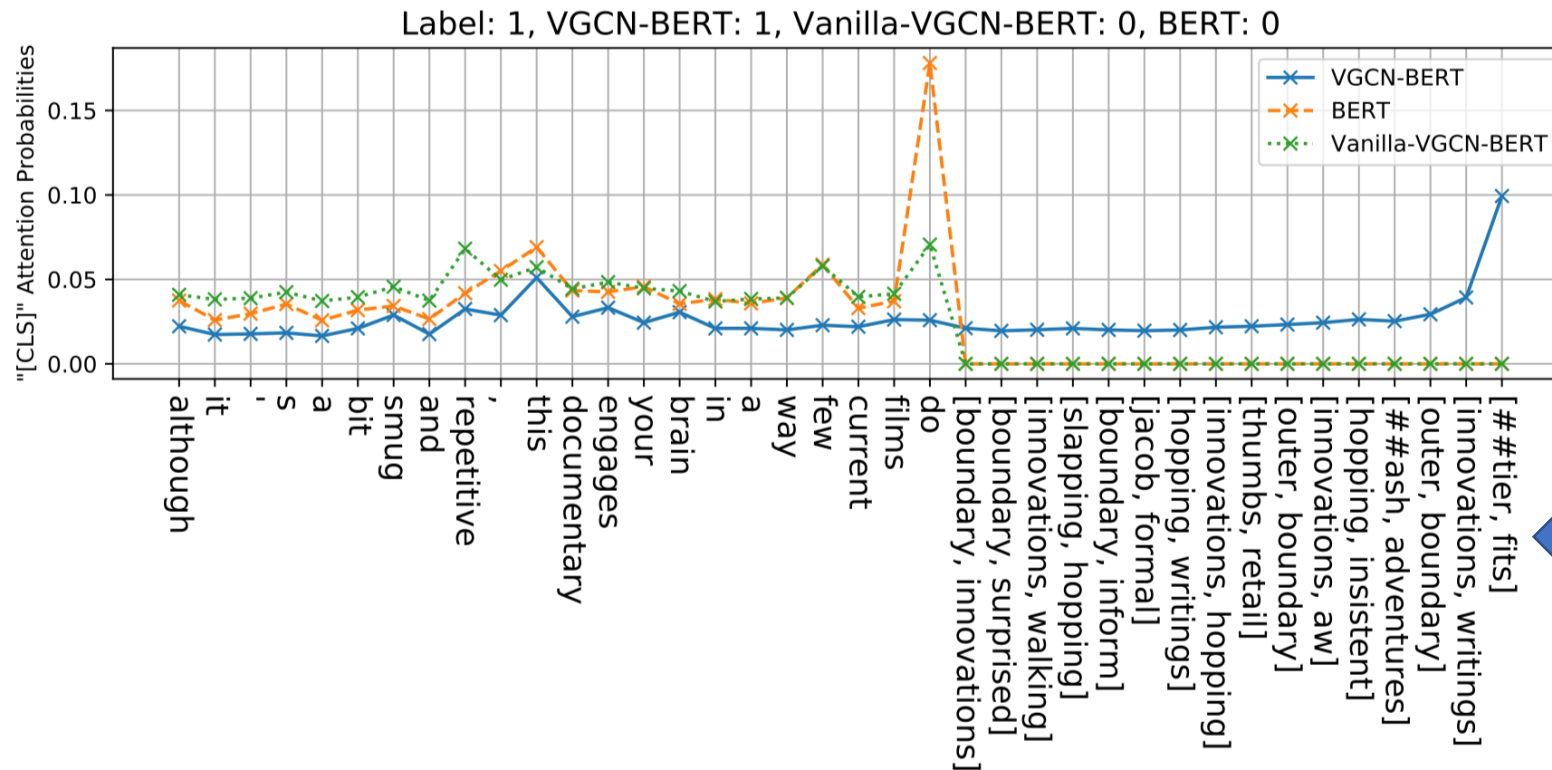


Visualization for sentiment classification

❖ “Although it’s a bit **smug** and **repetitive**, this documentary engages your brain a way few current films do.” (movie review SST-2)

- Negative by BERT

❖ “a way few current films do” in general language -> “innovation” **positive**



Meaning of graph
embedding
dimensions

Importing entity embedding into text

- Knowledge graph embedding:
 - Incorporating information of neighbor nodes
- Text embedding:
 - Incorporate information from context words
- Inject entity embedding of graph into text
- ERNIE (Zhang et al. 2019)
 - Linking entities in text to entity nodes in graph
 - Aggregate entity embedding with token embedding

ERNIE

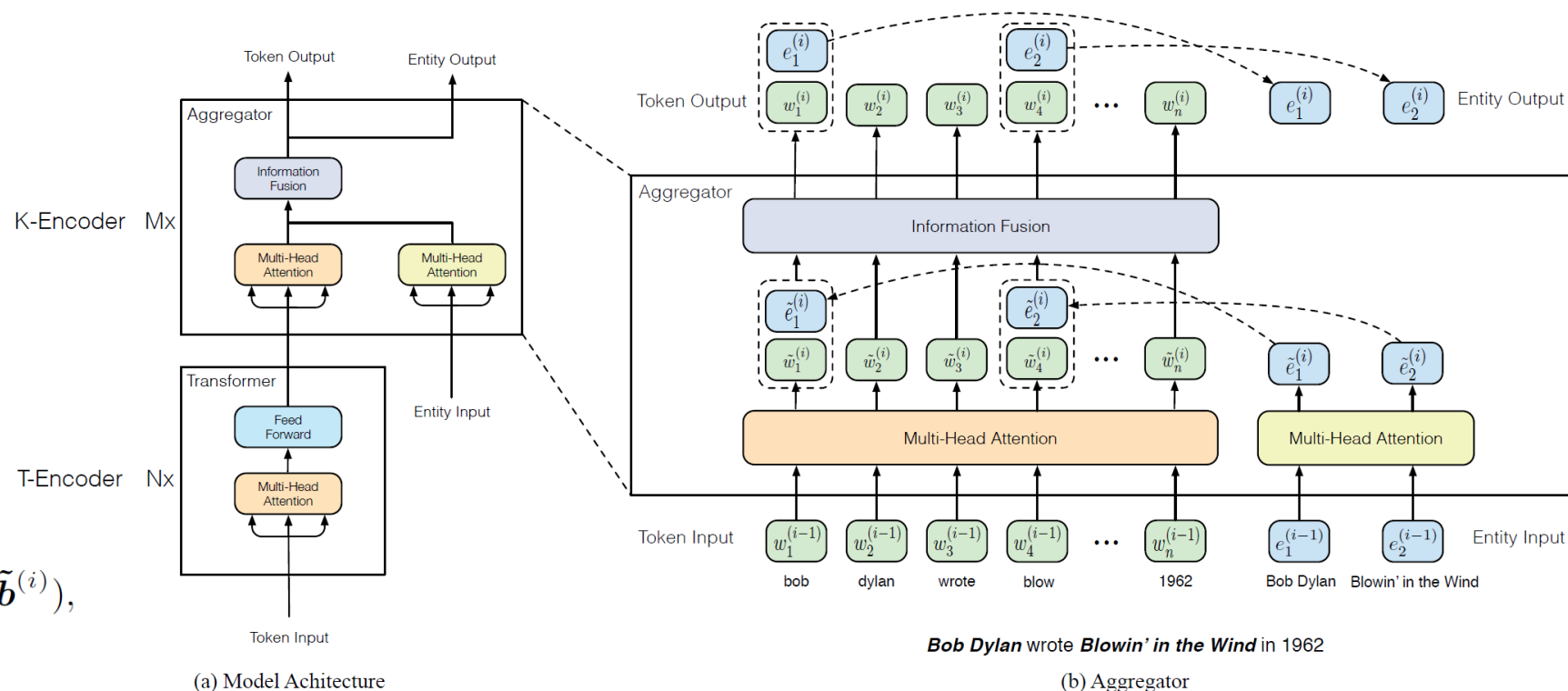
- Text embedding
- Entity embedding
 - TransE (minimize $|e_h + e_r - e_t|$)

• Aggregation:

$$h_j = \sigma(\tilde{W}_t^{(i)} \tilde{w}_j^{(i)} + \tilde{W}_e^{(i)} \tilde{e}_k^{(i)} + \tilde{b}^{(i)}),$$

$$w_j^{(i)} = \sigma(W_t^{(i)} h_j + b_t^{(i)}),$$

$$e_k^{(i)} = \sigma(W_e^{(i)} h_j + b_e^{(i)}).$$



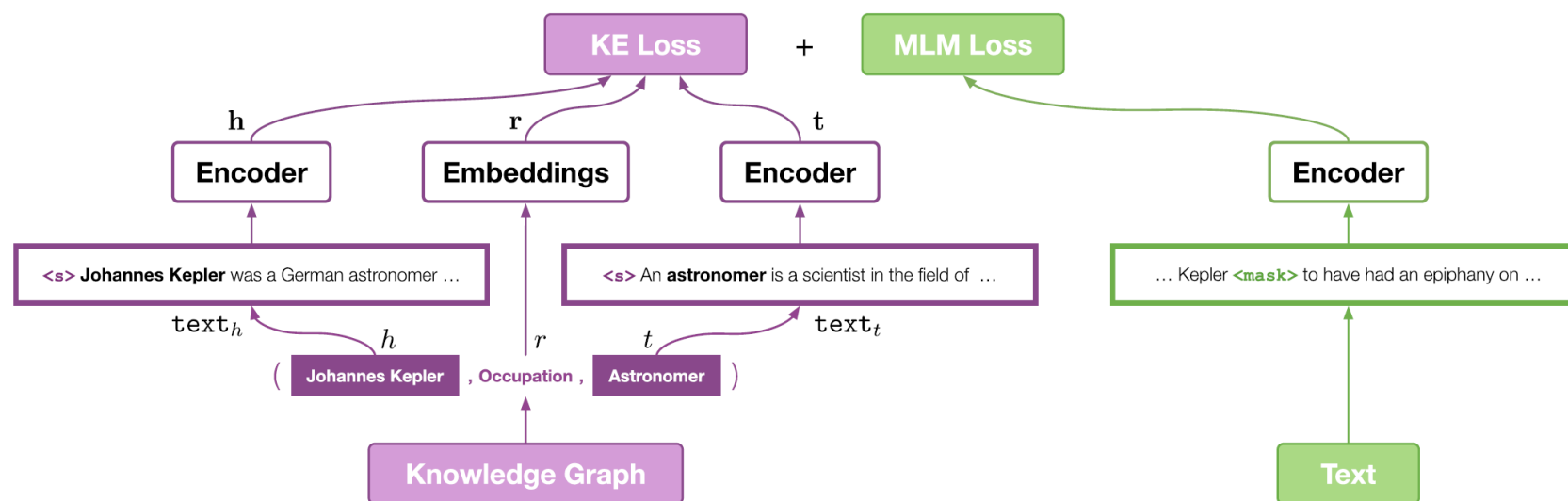
2. Knowledge infusion by multi-task learning

- Language modeling objective
 - Masked Language Modeling: predict the masked token
 - Auto regressive: predict the next token
- Knowledge objective
 - Related entity: head + relation \rightarrow tail (e.g. TransE)
 - Relation between entities (Relation classification)
 - Linguistic knowledge: dependency
 - Sense: predict the super sense of a token (SenseBERT)
- Combine multiple objectives in multi-task learning

KEPLER: joint training of MLM and TransE

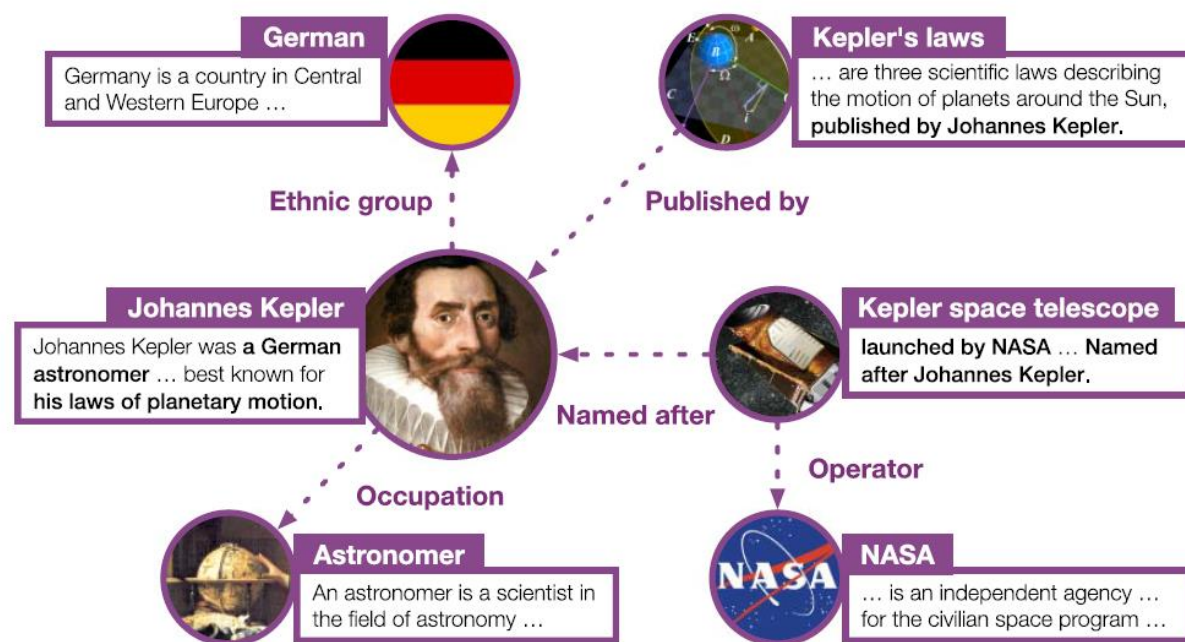
(Wang et al., KEPLER, Trans. ACL, 2021)

- Optimize a joint objective of text encoding (MLM) and graph encoding (TransE): $\mathcal{L} = \mathcal{L}_{KE} + \mathcal{L}_{MLM}$,
- Architecture of KEPLER



KEPLER

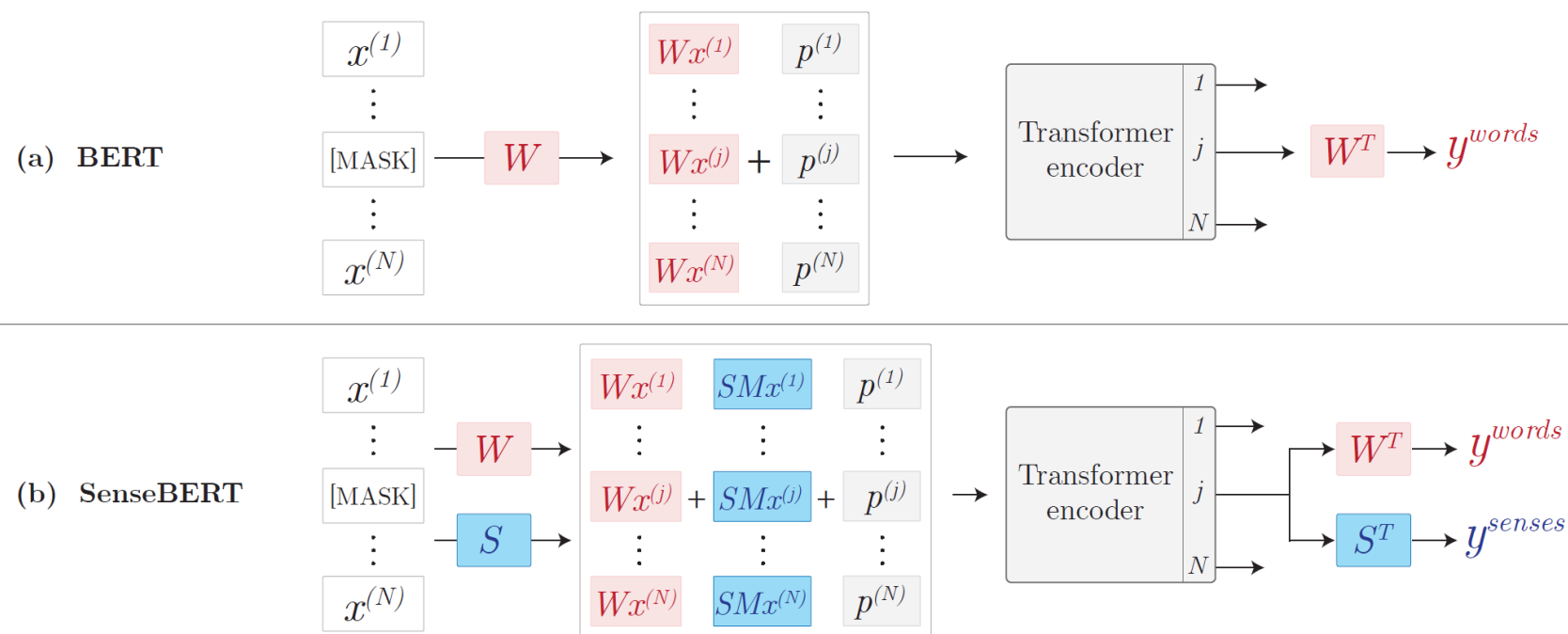
- Use description of entities and relations to create better embedding for entities and relations



SenseBERT: LM + sense

- Sense = higher synset in WordNet
- LM objective + Sense prediction objective

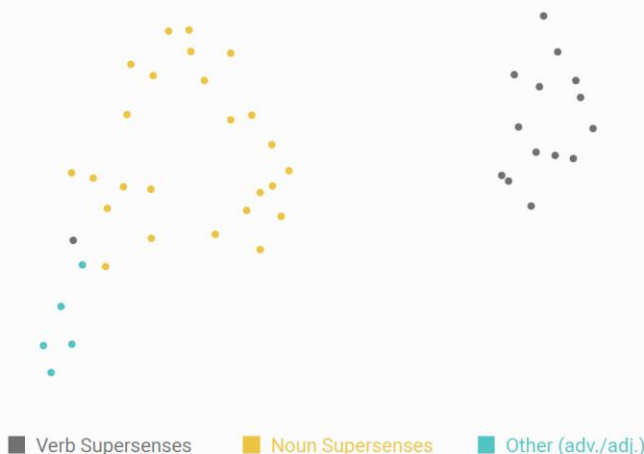
1. “This **bass** is delicious”
(supersenses: noun.food, noun.artifact, *etc.*)
2. “This **chocolate** is delicious”
(supersenses: noun.food, noun.attribute, *etc.*)
3. “This **pickle** is delicious”
(supersenses: noun.food, noun.state, *etc.*)



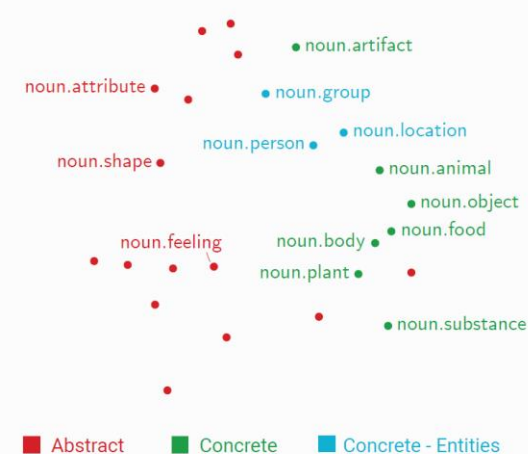
Effect of SenseBERT

- The model has a better capability to project tokens of the same super-sense closer

(a) All Supersenses



(b) Noun Supersenses



(a)

SemEval-SS

The team used a battery of the newly developed “gene probes”

Ten shirt-sleeved ringers stand in a circle, one foot ahead of the other in a prize-fighter's stance

BERT

noun.artifact

noun.quantity

SenseBERT

noun.group

noun.body

(b)

WiC

Sent. A:
The kick must be synchronized with the arm movements.

Sent. A:
Plant bugs in the dissident's apartment.

Sent. B:
A sidecar is a smooth drink but it has a powerful kick.

Sent. B:
Plant a spy in Moscow.

Same

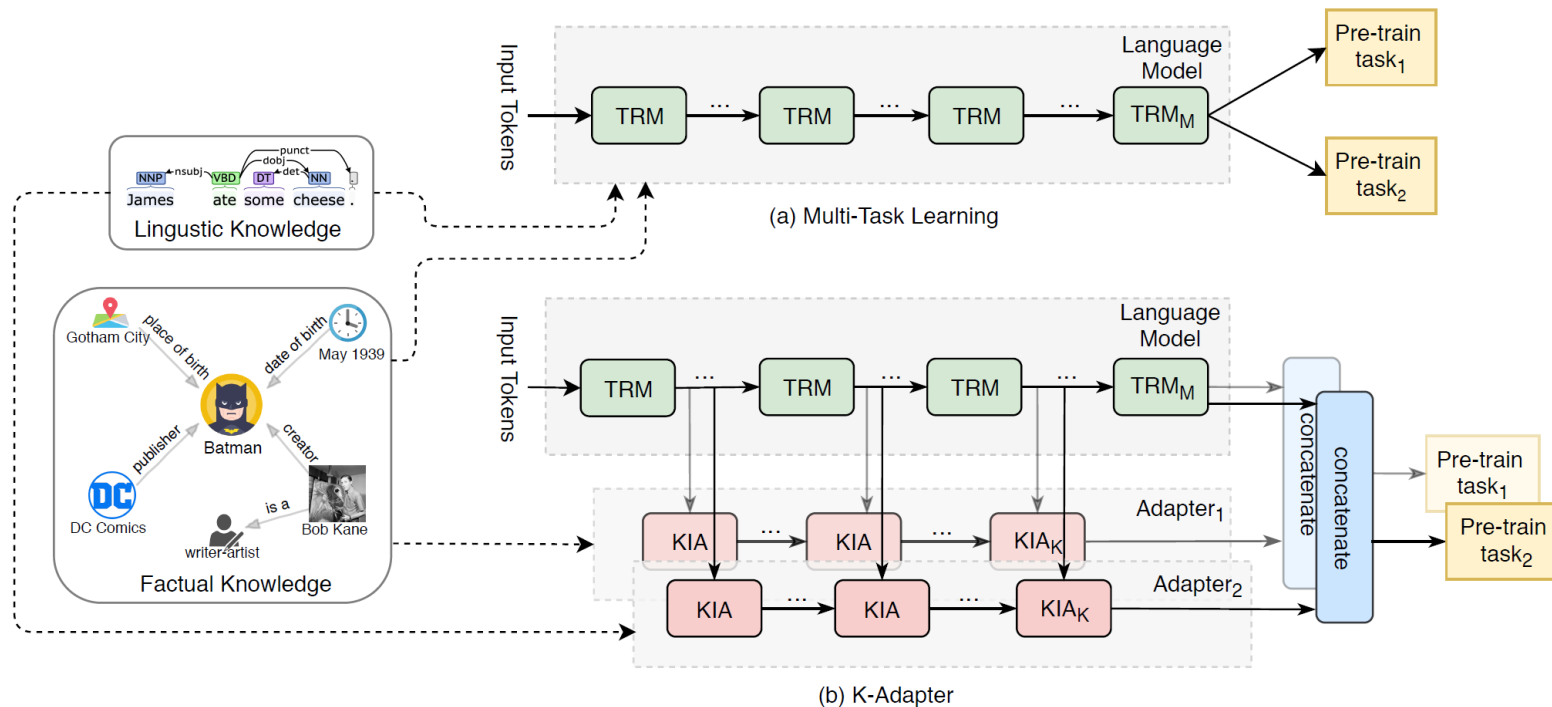
Different

Different

Same

K-ADAPTER: adding an adapter aside

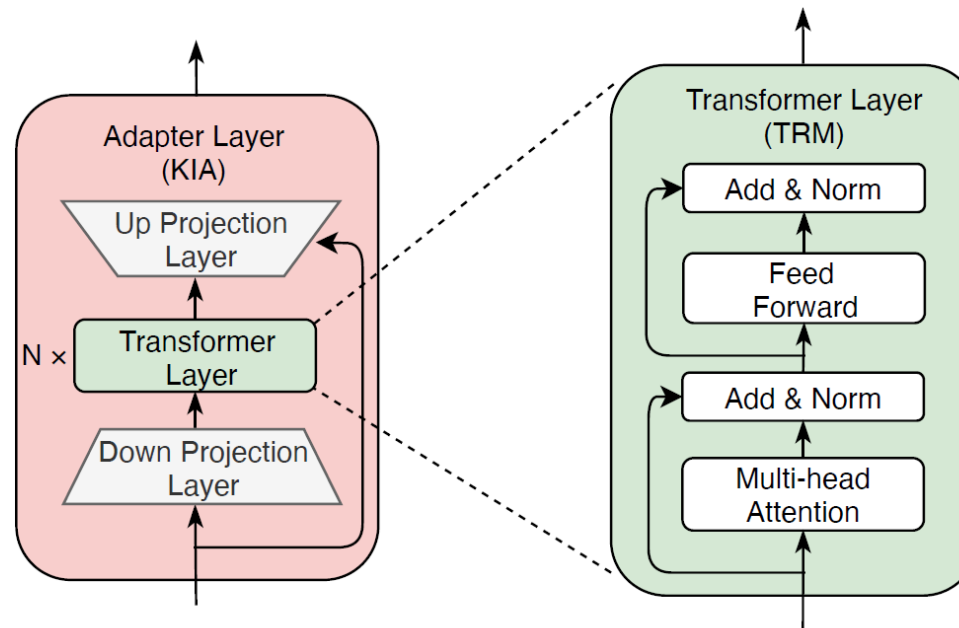
- Idea:
 - Injecting knowledge into a language model may lead to catastrophic forgetting
 - Using adapter instead
- Adapter: a small model that complements a large LM
- Concatenate the outputs of transformer and adapter



Credit: Wang et al.
K-ADAPTER, ACL-
IJCNLP 2021

K-ADAPTER

- Adapter Architecture
 - Several layers of transformer
 - + projections



K-ADAPTER: Some results

Model	OpenEntity			FIGER		
	P	R	Mi-F ₁	Acc	Ma-F ₁	Mi-F ₁
NFGEC (Shimaoka et al., 2016)	68.80	53.30	60.10	55.60	75.15	71.73
BERT-base (Zhang et al., 2019)	76.37	70.96	73.56	52.04	75.16	71.63
ERNIE (Zhang et al., 2019)	78.42	72.90	75.56	57.19	75.61	73.39
KnowBERT (Peters et al., 2019)	78.60	73.70	76.10	-	-	-
KEPLER (Wang et al., 2021)	77.20	74.20	75.70	-	-	-
WKLM (Xiong et al., 2020)	-	-	-	60.21	81.99	77.00
RoBERTa	77.55	74.95	76.23	56.31	82.43	77.83
RoBERTa + multitask	77.96	76.00	76.97	59.86	84.45	78.84
K-ADAPTER (w/o knowledge)	74.47	74.91	76.17	56.93	82.56	77.90
K-ADAPTER (F)	79.30	75.84	77.53	59.50	84.52	80.42
K-ADAPTER (L)	80.01	74.00	76.89	61.10	83.61	79.18
K-ADAPTER (F+L)	78.99	76.27	77.61	61.81	84.87	80.54

Table 2: Results on two entity typing datasets OpenEntity and FIGER.

Model	SearchQA		Quasar-T		CosmosQA
	EM	F ₁	EM	F ₁	
BiDAF (Seo et al., 2017)	28.60	34.60	25.90	28.50	-
AQA (Buck et al., 2018)	40.50	47.40	-	-	-
R ³ (Wang et al., 2018a)	49.00	55.30	35.30	41.70	-
DSQA (Lin et al., 2018)	49.00	55.30	42.30	49.30	-
Evidence Agg. (Wang et al., 2018b)	57.00	63.20	42.30	49.60	-
BERT (Xiong et al., 2020)	57.10	61.90	40.40	46.10	-
WKLM (Xiong et al., 2020)	58.70	63.30	43.70	49.90	-
WKLM + Ranking (Xiong et al., 2020)	61.70	66.70	45.80	52.20	-
BERT-FT _{RACE+SWAG} (Huang et al., 2019)	-	-	-	-	68.70
RoBERTa	59.01	65.62	40.83	48.84	80.59
RoBERTa + multitask	59.92	66.67	44.62	51.17	81.19
K-ADAPTER (F)	61.85	67.17	46.20	52.86	80.93
K-ADAPTER (L)	61.15	66.82	45.66	52.39	80.76
K-ADAPTER (F+L)	61.96	67.31	46.32	53.00	81.83

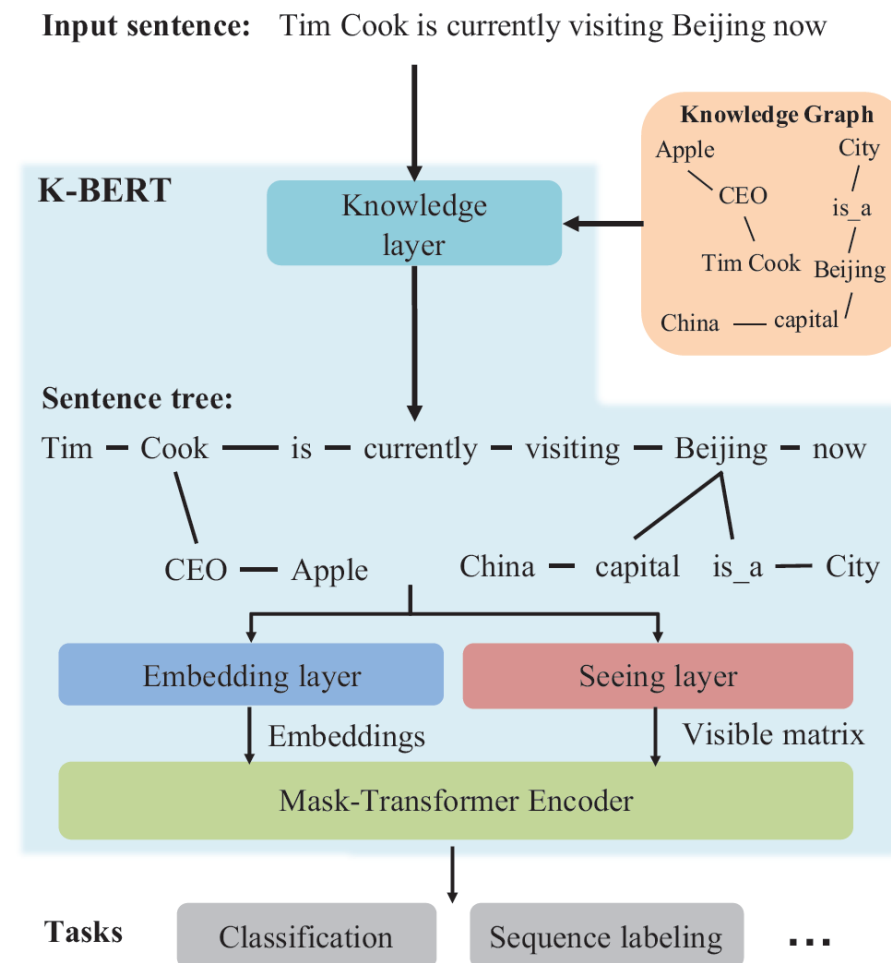
Table 3: Results on question answering datasets including: CosmosQA, SearchQA and Quasar-T.

3. Enriching raw text by knowledge (K-BERT)

- Input text → mapping to entities of KG
- Infusion: Add links to these tokens
- Visible matrix: what tokens can be seen – a way to define paths
- Embedding training: masking as in BERT

➔ a way to exploit BERT training

Credit: Liu et al. K-BERT: Enabling Language Representation with Knowledge Graph, AAAI 2020

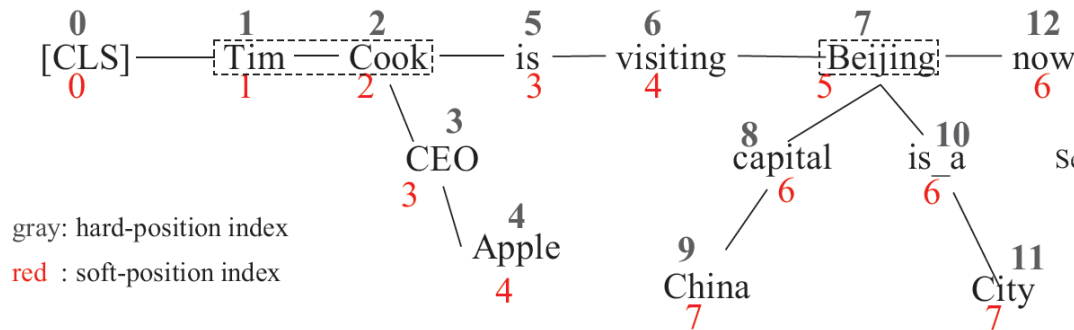


K-BERT: Infusing triples into text

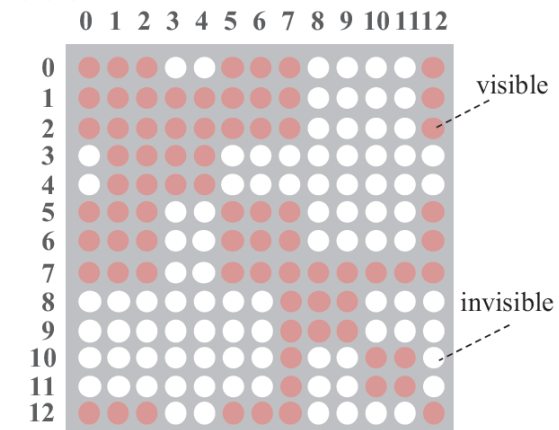
Embedding Representation

Token embedding	[CLS]	Tim	Cook	CEO	Apple	is	visiting	Beijing	capital	China	is_a	City	now
	+	+	+	+	+	+	+	+	+	+	+	+	+
Soft-position embedding	0	1	2	3	4	3	4	5	6	7	6	7	6
	+	+	+	+	+	+	+	+	+	+	+	+	+
Segment embedding	A	A	A	A	A	A	A	A	A	A	A	A	A

Sentence Tree



Visible Matrix



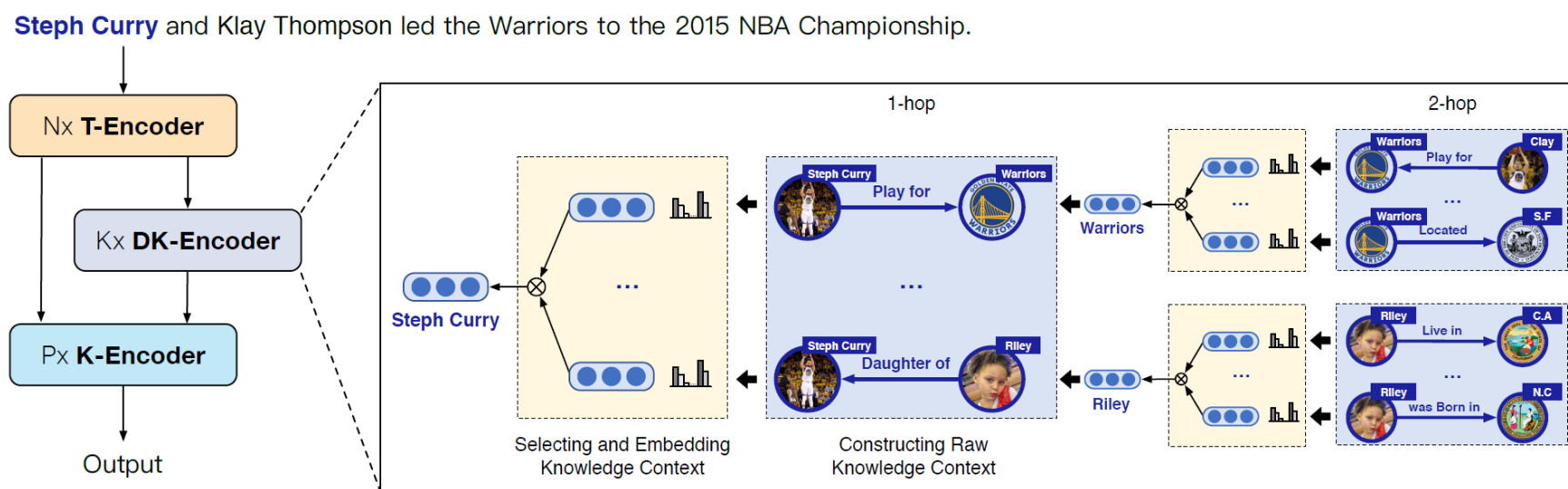
Relevance of knowledge:

What pieces of knowledge to use?

- Usually rely on entity linking: entities in text → related triples/1-hop subgraph
- Knowledge selection is key
 - A wrong piece of knowledge will lead to wrong inference direction
 - A useless (irrelevant) piece of knowledge will add noise
- E.g. Obama visited Japan and met with the Primary minister.
 - Obama – function → president of US ✓
 - Obama – spouse → Michele ✗
- Knowledge selection by attention/relevance

Coke-BERT: Selecting sub-graph with context (Su et al. 2020)

- Text encoding: Using BERT to encode input text
- Knowledge selection: Attention of text context to neighbors of entity
 - Neighbor entities are aggregated according to their attention weights
- Fusion: concatenate text embedding and knowledge embedding as in ERNIE

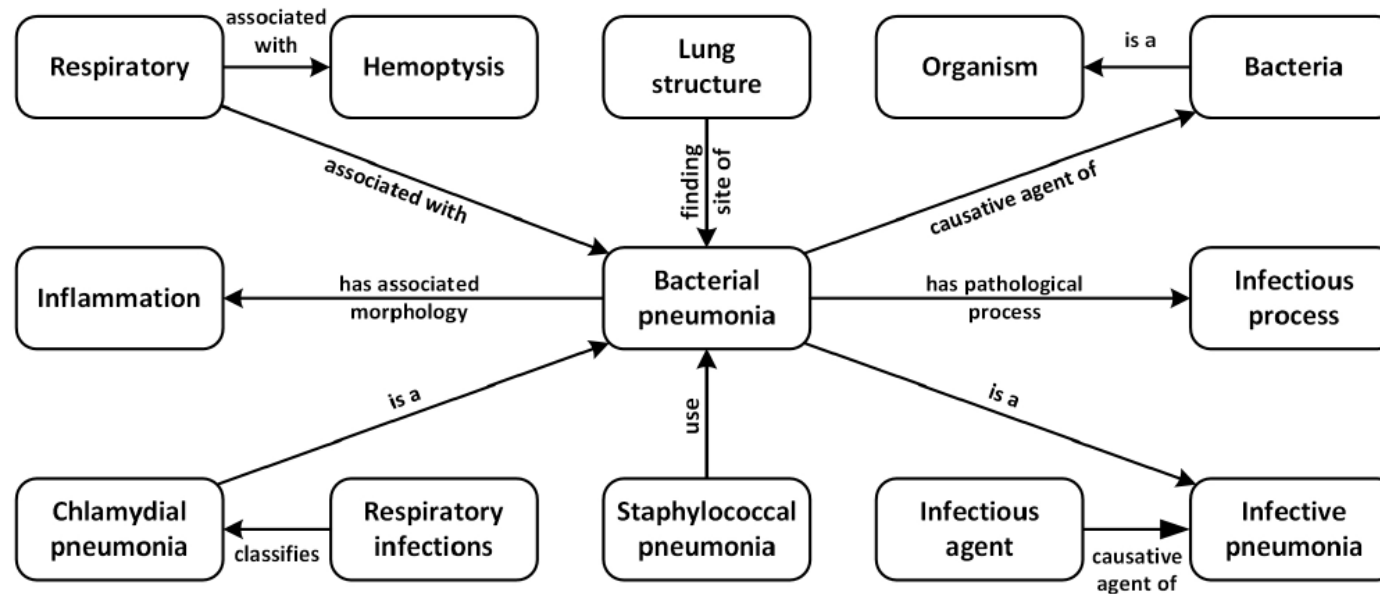


Selecting knowledge according to usefulness (Zhu, Nie et al. SIGIR 2021)

- Given a context (in dialogue), should a piece of knowledge (triple) be used?
- Attention \sim similarity (implicit in end-to-end training)
- ➔ More explicit learning of usefulness/relevance of a piece of knowledge
- ➔ Use a piece of knowledge only when it fits the context – useful for the task
- Problem: No annotated data about usefulness/relevance of knowledge
- Pseudo labeling:
 - Raw training data = (context, response)
 - Testing if a triple is useful to connect context to response? \rightarrow usefulness label
 - Extended training data = (context, response, knowledge)
- Train an explicit knowledge selection network using extended training data
 - Better than implicit selection by attention

Application in medical domain: a special case

- Medical domain is rich of expert knowledge
- Much of such knowledge cannot be extracted from texts



Applications in medical domain

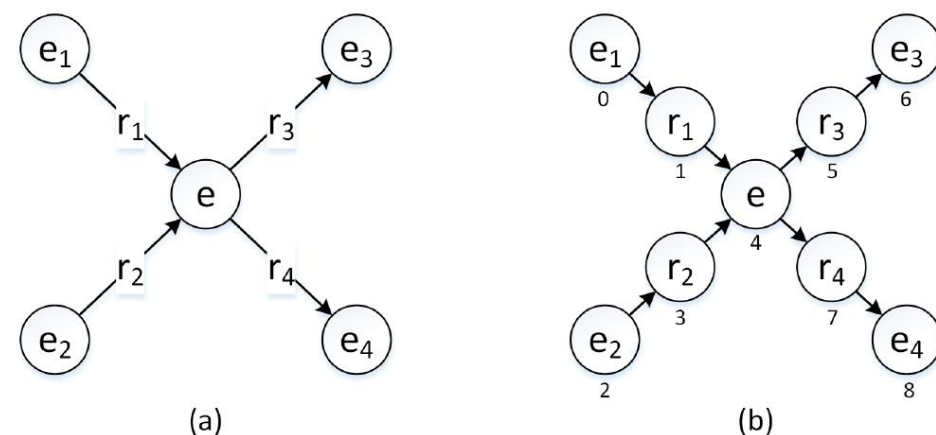
- Rich medical domain knowledge
 - The International Classification of Diseases (ICD)
 - Medical Subject Headings (MeSH)
 - The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT)
 - The Unified Medical Language System (UMLS)
 - ...
- Tasks
 - Semantic Annotation of Medical Texts
 - Entities/concepts in the document
 - Relationships (implicit or explicit) between entities
 - Relationship between an entity and a document
 - Medical IR

Applications to medical IR

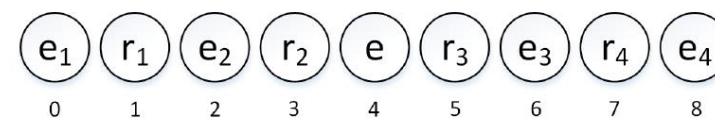
- Main approaches: query expansion using knowledge graph
 - Add related terms/concepts into query
- Neural representations
 - Specifically trained medical LM: BioBERT, ClinicalBERT
 - Doc2vec: creating an embedding for a medical text
 - Extending word embedding by medical knowledge
 - Liu et al. 2016: Adjusting word embedding to fit medical knowledge, improved medical IR
- Some limited approaches to infuse medical knowledge

BERT-MK: Applying medical knowledge graph

- Extract subgraph from entities in a text (1-hop)
- Converting the subgraph to sequence (special position index)
- BERT encoding



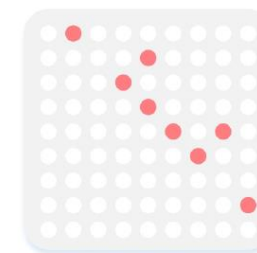
Node sequence



Node position indexes

0	1	4
2	3	4
4	5	6
4	7	8

Adjacent matrix



(c)

Credit: He et al. BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models, EMNLP 2020

BERT-MK: training process

- Train knowledge subgraph encoding to restore the relations

$$\mathcal{L} = \sum_{\mathbf{t} \in \mathbf{T}} \max\{d(\mathbf{t}) - d(f(\mathbf{t})) + \gamma, 0\}$$

- Maximize the distance between $d(\mathbf{t})$ – TransE distance of a valid triple and $f(\mathbf{t})$ – TransE distance of an invalid triple (tail entity replaced)
- Integration with LM: Similar to ERNIE – concatenating text embedding and graph embedding of entities

BERT-MK: experiments

- Knowledge graph: subset of UMLS

Table 1: Statistics of UMLS.

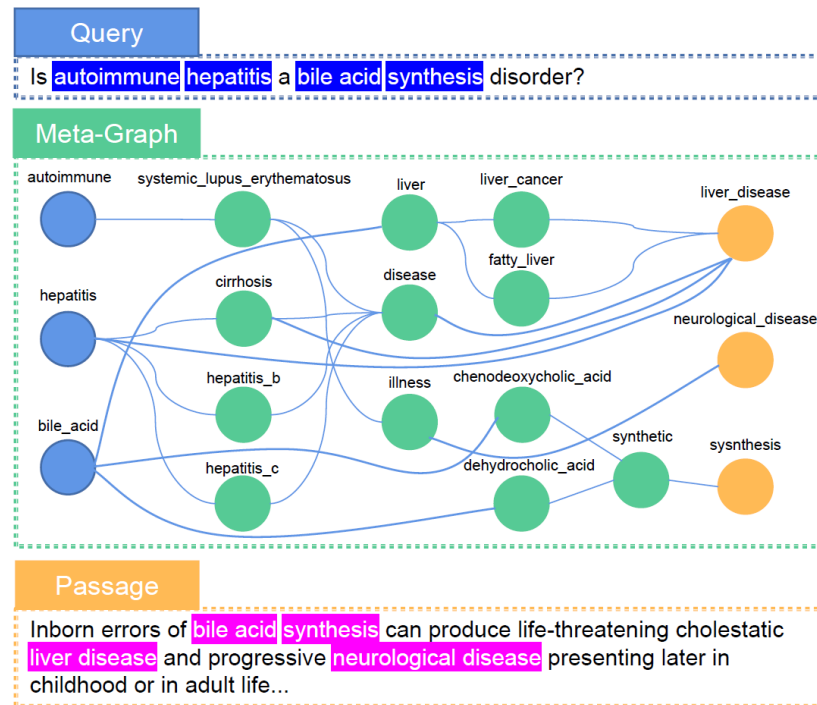
# Entities	# Relations	# Triples
2,842,735	874	13,555,037
In-degree	Out-degree	Median degree
5.05	5.05	4

- Results on entity typing and relation classification

Task	Dataset	Metrics	E-SVM	CNN-M	BERT-Base	BioBERT	SCIBERT	BERT-MK
Entity Typing	2010 i2b2/VA	Acc	-	-	96.76	97.43	97.74	97.70
	JNLPBA	Acc	-	-	94.12	94.37	94.60	94.55
	BC5CDR	Acc	-	-	98.78	99.27	99.38	99.54
Relation Classification	2010 i2b2/VA	P	-	<u>73.1</u>	72.6	76.1	74.8	77.6
		R	-	<u>66.7</u>	65.7	71.3	71.6	72.0
		F	-	<u>69.7</u>	69.2	73.6	73.1	74.7
	GAD	P	<u>79.21</u>	-	<u>74.28</u>	<u>76.43</u>	77.47	81.67
		R	<u>89.25</u>	-	<u>85.11</u>	<u>87.65</u>	85.94	92.79
		F	<u>83.93</u>	-	<u>79.33</u>	<u>81.66</u>	81.45	86.87
	EU-ADR	P	-	-	<u>75.45</u>	<u>81.05</u>	78.42	84.43
		R	-	-	96.55	<u>93.90</u>	90.09	91.17
		F	-	-	<u>84.71</u>	<u>87.00</u>	85.51	87.49

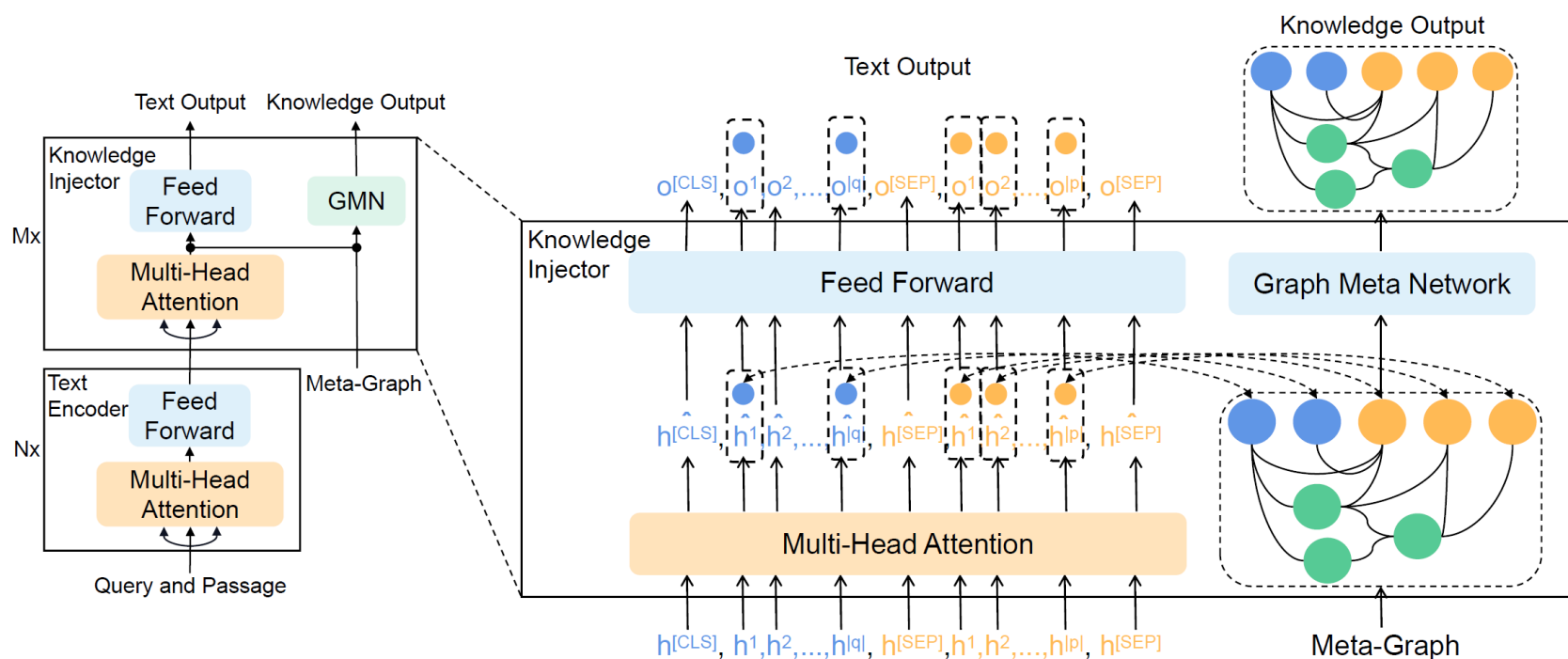
KERM for passage re-ranking (Dong et al. SIGIR 2022)

- Selecting meta-graph that connects query entities to document entities



KERM

- Text encoding
- Graph encoding (TransE)
- Knowledge aggregation



KERM on medical IR

- Subset of MS MARCO and OHSUMED

	MARCO Dev Queries		Ohsumed Queries	
	General	Bio-Medical		
	MRR@10	MRR@10	MRR@10	MAP@10
BM25	19.2	16.4	69.6	9.5
ERNIE	38.5	30.5	79.7	10.1
KERM	39.7	32.1	81.2	11.0

Remaining questions

- Improvements mainly on entity and relation classification, also on IR
- More experimental evidence of knowledge infusion for IR to be made
- Main observations:
 - Knowledge infusion can improve text representation for tasks requiring knowledge
 - No standard way yet for knowledge infusion
- Key question: What knowledge to use
 - Relying on entity linking
 - Related entities/triples/subgraph?
 - Relatedness by Attention / usefulness?

Slides available at

<https://github.com/laura-dietz/neurosymbolic-representations-for-IR/>