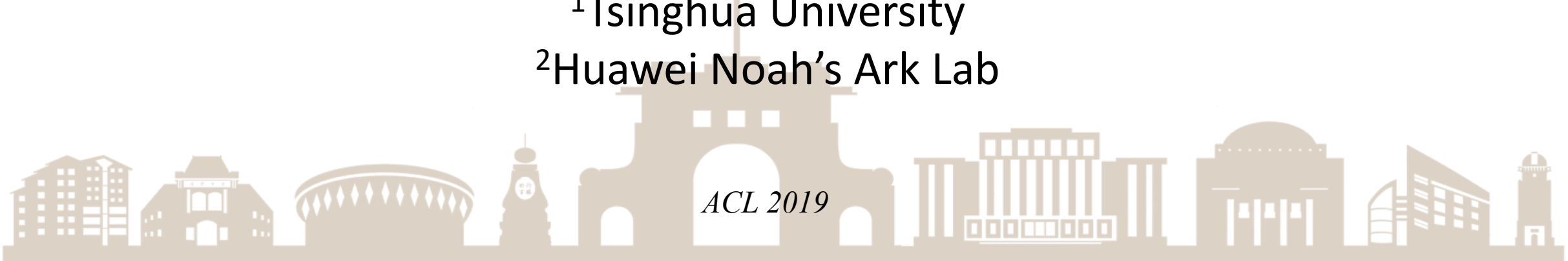
ERNIE: Enhanced Language Representation with Informative Entities

Zhengyan Zhang[1], Xu Han[1], Zhiyuan Liu[1], Xin Jiang[2], Maosong Sun[1], Qun Liu[2]

[1]Tsinghua University

[2]Huawei Noah's Ark Lab



ACL 2019

Pre-trained Language Model

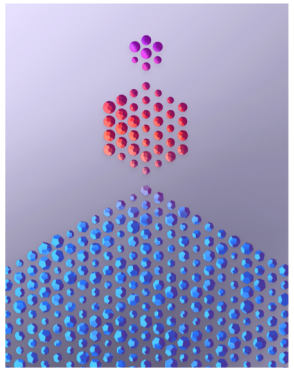
- Pre-trained LMs attract more and more attention
  - ELMo
  - GPT
  - BERT

	Rank	Name	Model	URL	Score
	1	bigbird he	Microsoft D365 AI & MSR AI		81.9
−	2	Jacob Devlin	BERT: 24-layers, 1024-hidden, 16		80.4
			BERT: 12-layers, 768-hidden, 12-l		78.3
	3	Jason Phang	GPT on STILTs		76.9
	4	Alec Radford	Singletask Pretrain Transformer		72.8
+	5	Samuel Bowman	BiLSTM+ELMo+Attn		70.5
	6	GLUE Baselines	BiLSTM+ELMo+Attn		68.9

Leaderboard of GLUE benchmark (2019.1)

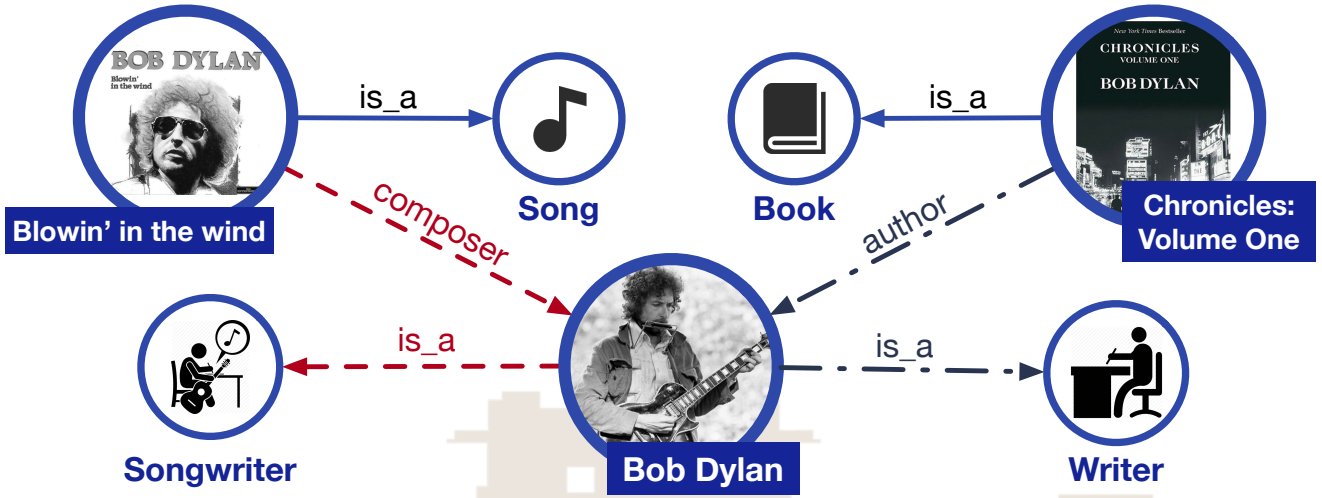
Pre-trained Language Model

- ELMo (CNN+LSTM)
  - 1B Word Benchmark
  - Sentence-level corpus
- GPT (Transformer)
  - BookCorpus dataset
  - Document-level corpus
- BERT (Bidirectional Transformer)
  - BookCorpus + Wikipedia
  - Lager document-level corpus (3x)



External Knowledge Information

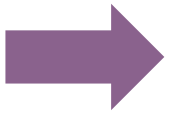
- Intuitively, external knowledge information can effectively benefit language understanding
  - Low resource entities
  - Implicit background knowledge



Bob Dylan wrote ***Blowin' in the Wind*** in 1962, and wrote ***Chronicles: Volume One*** in 2004.

Motivation

- Knowledge graphs can provide rich knowledge information
  - Millions of entities
  - Millions of fact triples
- ERNIE:A pre-trained model on large-scale corpus and knowledge graph



Two Challenges

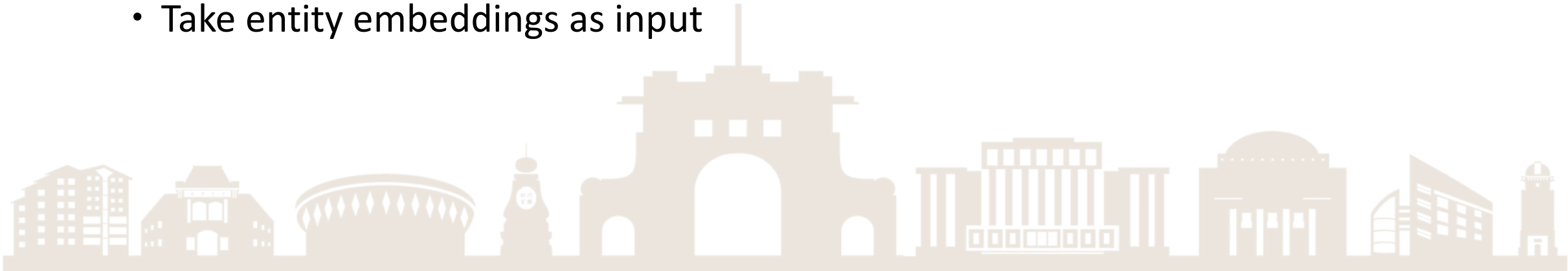
- Structured Knowledge Encoding
  - Extract related facts regarding to given text
  - Represent structured information with low-dimension embedding

- Heterogeneous Information Fusion
  - Lexical information
  - Syntactic information
  - Knowledge information

Structured Knowledge Encoding

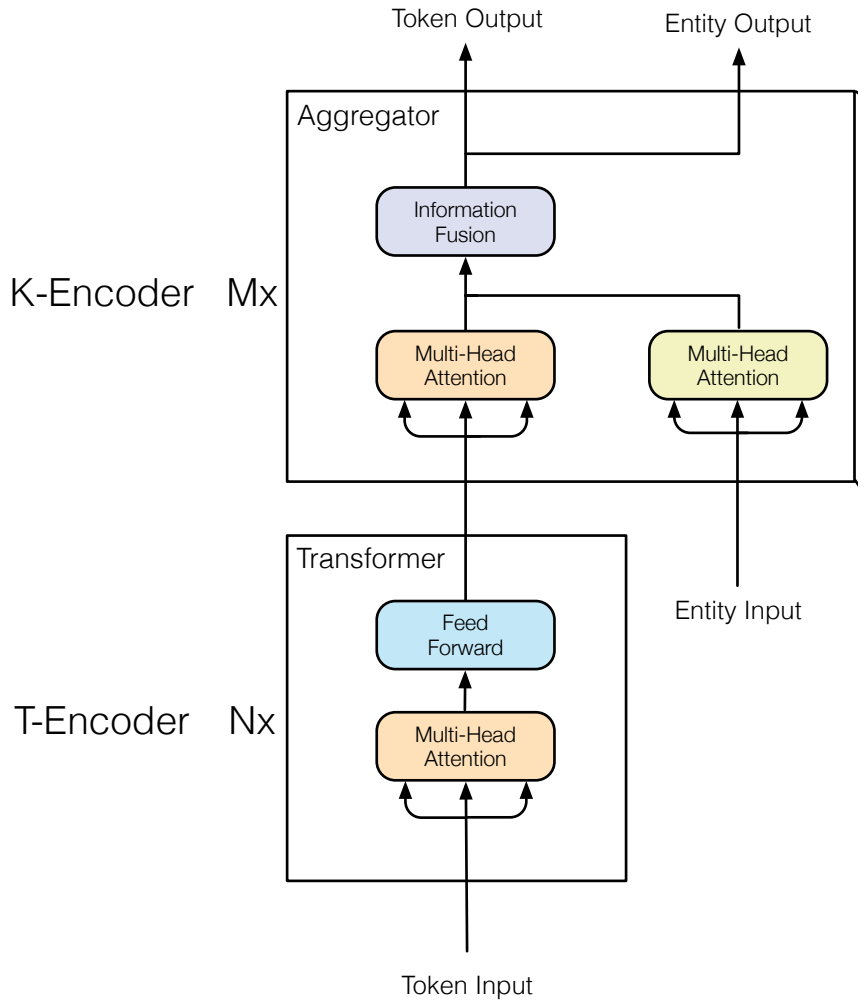
- Extract related facts regarding to given text
  - Link named entity mentions to entities in KGs by TAGME
  - Introduce knowledge information into the pre-trained model by informative entities

- Represent structured information
  - Encode the graph structure of KGs with knowledge embedding algorithms
  - Take entity embeddings as input



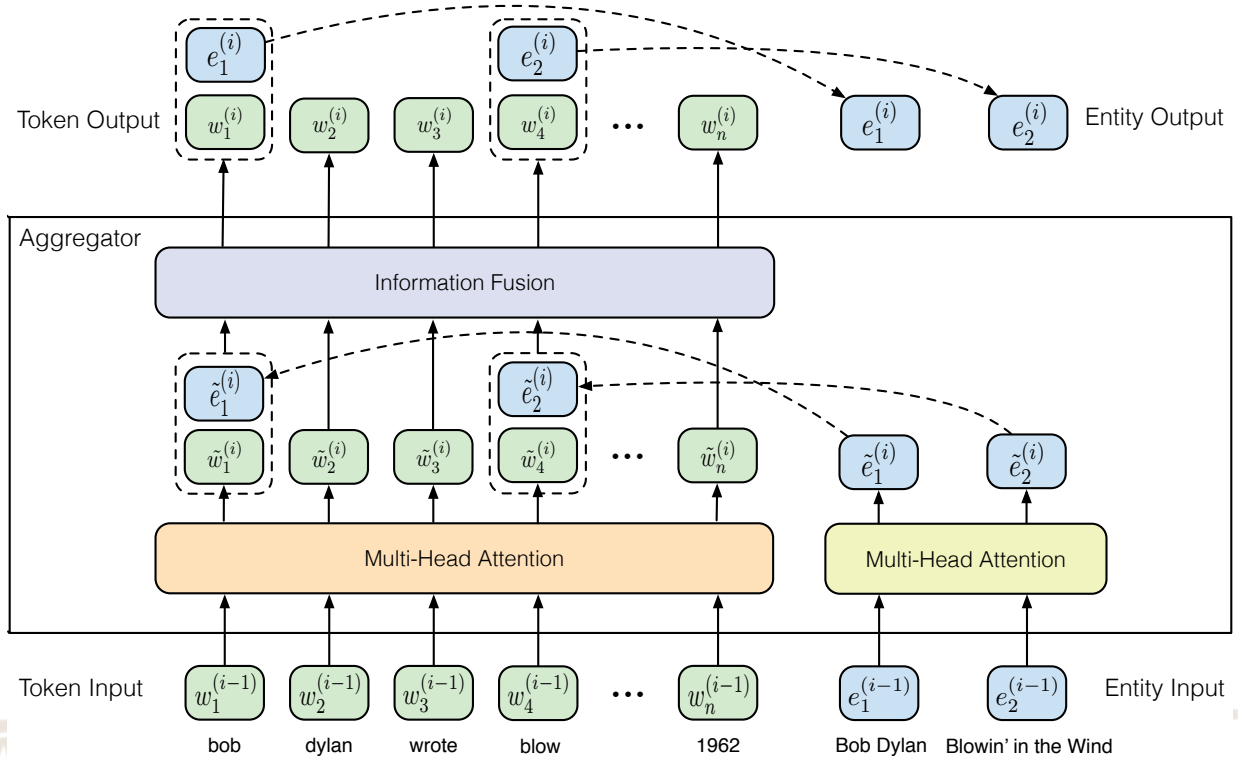
Structured Knowledge Encoding

- The architecture of ERNIE
  - Lower layers for text
  - Higher layers for knowledge integration



Structured Knowledge Encoding

- For the integration of knowledge information, we design the aggregator layer for ERNIE



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

Heterogeneous Information Fusion

- The information fusion layer in the aggregator
- Take two kinds of input
- A token with its aligned entity

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

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

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

- A token without its aligned entity

$$h_j = \sigma(\widetilde{W}_t^{(i)}\ \widetilde{w}_j^{(i)} + \tilde{b}^{(i)})$$

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

Heterogeneous Information Fusion

- New pre-training task
  - Denoising Entity Auto-encoder (dEA)
- Require the system to predict entities based on the input

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

- Add some noise for better generalization
  - Replace 5% entities with another random entities
  - Mask 15% token-entity alignments
  - Keep 80% token-entity alignments unchanged

Experiments

- How to fine-tune on the knowledge-driven tasks
  - Focus on the entity mention in the given text

Mark Twain wrote ***The Million Pound Bank Note*** in 1893.

Input for Common NLP tasks:

Input for Entity Typing:

Input for Relation Classification:

Experiments

- Relation classification

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**

Experiments

- Entity Typing

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

Table 2: Results of various models on FIGER.

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**

Table 3: Results of various models on Open Entity.

Conclusion

- Incorporate knowledge in pre-trained language model
  - Extract related information from KGs
- ERNIE with better fusion of heterogeneous information
  - Knowledgeable aggregator
  - Pre-training task dEA

Q&A

Paper

Code

Email: zhangzhengyan14@mails.tsinghua.edu.cn