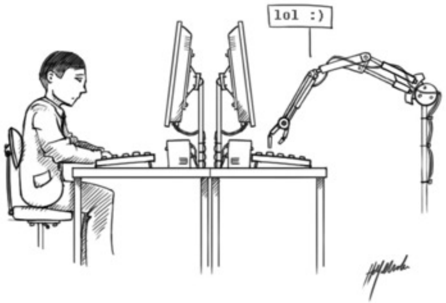
Model Knowledge Stimulation with Prompts for Pre-trained Language Models

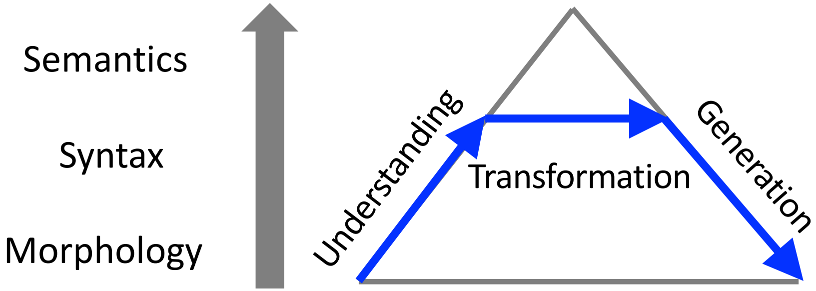
Zhiyuan Liu
Tsinghua University

Background

- NLP is the key to pass Turing Test and Realize AI



Turing Test



Structure Learning for NLP

Alan Turing

(1912 - 1954)

Key founder of CS and AI, proposed Turing test based on language understanding

Dartmouth Conference

(1956)

Proposed AI for the first time and listed NLP as the key research problem

Background

- Deep language understanding requires complicated knowledge

Knowledge

Linguistics

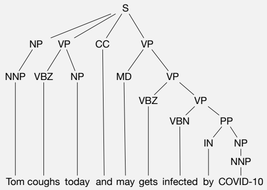
Commonsense

Facts

Expertise

Text

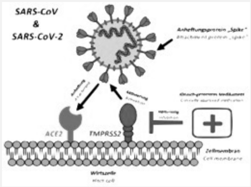
Tom coughs today and may get infected by COVID-19



Subject: Tom
Predicate: cough
Time: today

Cough makes Tom unconformable

Find hospitals and doctors for cure



COVID-19 will cause inflammation and then dry cough

Language understanding requires the ability of **knowledge acquisition, representation and application**

Research Spectrum of NLP

1960

Noam Chomsky

Modern grammar (**Linguistics**) theory proposed in 1950s has been introduced in NLP but **cannot well cover complex language usage**.

Research Spectrum of NLP

Edward Feigenbaum

An **expert system** represents facts and rules with the knowledge base, and **conducts inference based on the knowledge base**

1960

1980

Noam Chomsky

Modern grammar (**Linguistics**) theory proposed in 1950s has been introduced in NLP but **cannot well cover complex language usage**.

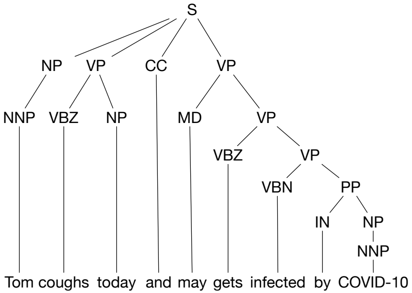
Research Spectrum of NLP

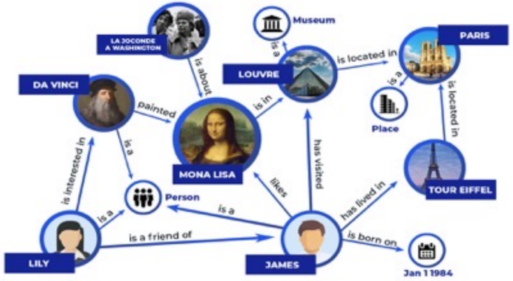
Noam Chomsky

Modern grammar (**Linguistics**) theory proposed in 1950s has been introduced in NLP but **cannot well cover complex language usage**.

Edward Feigenbaum

An **expert system** represents facts and rules with the knowledge base, and **conducts inference based on the knowledge base**





Symboledge (Symbolic Knowledge)

- linguistic rules
- knowledge bases
- ……

human-friendly、discrete、sparse

Research Spectrum of NLP

Edward Feigenbaum

An **expert system** represents facts and rules with the knowledge base, and **conducts inference based on the knowledge base**

1960 1980 **1990**

Noam Chomsky

Modern grammar (**Linguistics**) theory proposed in 1950s has been introduced in NLP but **cannot well cover complex language usage**.

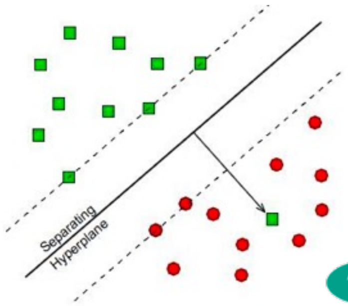
Robert Mercer

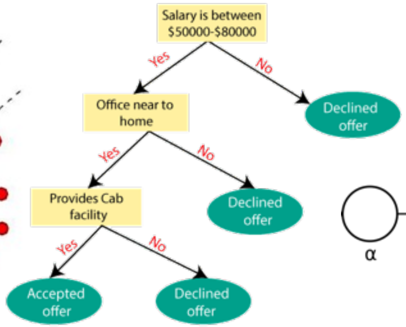
The data-driven **statistical models** proposed in 1990s only **take advantages of shallow lexical information**.

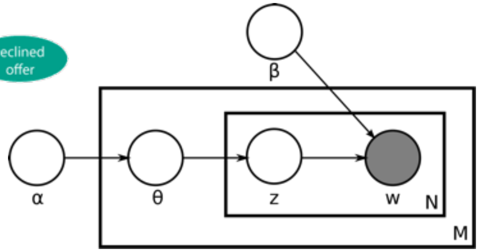
Research Spectrum of NLP

Robert Mercer

The data-driven **statistical models** proposed in 1990s only **take advantages of shallow lexical information**.







Modeledge (Model Knowledge)

- SVM
- Decision Tree
- CRF、LDA

machine-friendly、 discrete/continuous、shallow

Research Spectrum of NLP

Edward Feigenbaum

An **expert system** represents facts and rules with the knowledge base, and **conducts inference based on the knowledge base**

Yoshua Bengio

Neural models are introduced in NLP in 2010s but challenged by **deep understanding with structured knowledge**.

1960 1980 1990 **2010**

Noam Chomsky

Modern grammar (**Linguistics**) theory proposed in 1950s has been introduced in NLP but **cannot well cover complex language usage**.

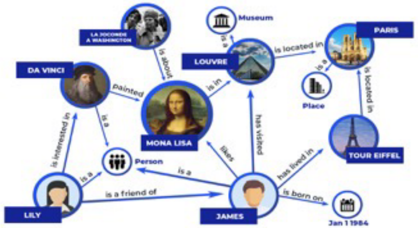
Robert Mercer

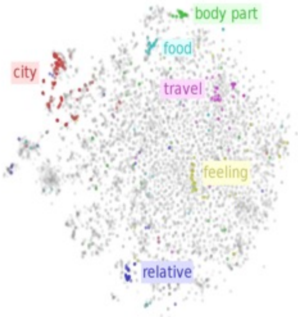
The data-driven **statistical models** proposed in 1990s only **take advantages of shallow lexical information**.

Research Spectrum of NLP

Yoshua Bengio

Neural models are introduced in NLP in 2010s but challenged by **deep understanding with structured knowledge**.





Embeledge (Embedding Knowledge)

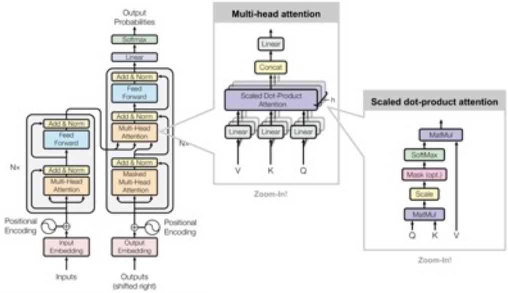
- word embedding
- knowledge graph embedding
- ……

machine-friendly、continuous、shallow

Research Spectrum of NLP

Yoshua Bengio

Neural models are introduced in NLP in 2010s but challenged by deep understanding with structured knowledge.



Embeledge (Embedding Knowledge)

- word embedding
- knowledge graph embedding
- ……

machine-friendly、continuous、shallow

Modeledge (Model Knowledge)

- CNN、RNN、GNN
- BERT、GPT、T5、BART
- ……

machine-friendly、continuous、deep

Research Spectrum of NLP

Edward Feigenbaum

An expert system represents facts and rules with the knowledge base, and conducts inference based on the knowledge base

Yoshua Bengio

Neural models are introduced in NLP in 2010s but challenged by deep understanding with structured knowledge.

1960 1980 1990 2010

Noam Chomsky

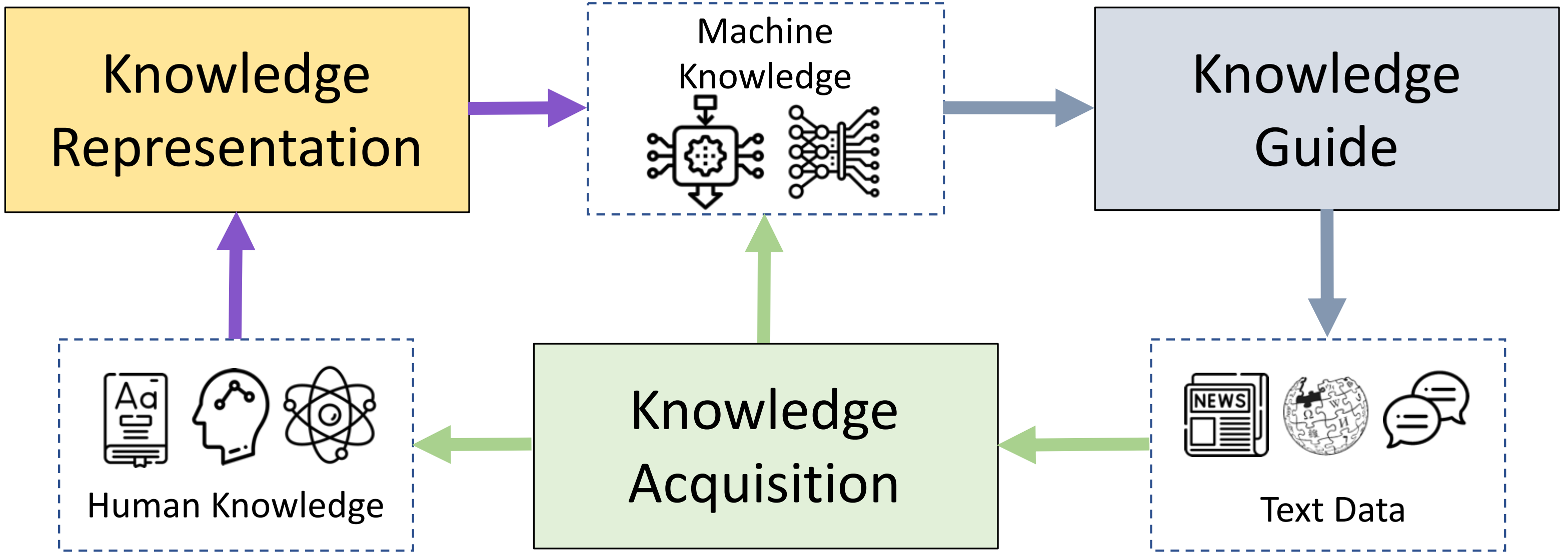
Modern grammar (Linguistics) theory proposed in 1950s has been introduced in NLP but cannot well cover complex language usage.

Robert Mercer

The data-driven statistical models proposed in 1990s only take advantages of shallow lexical information.

Acquisition, Representation and Application of knowledge for language understanding

Closed-Loop of Knowledge in NLP



1 Knowledge Acquisition

How to extract accurate knowledge from noisy text

2 Knowledge Representation

How to represent complex knowledge in machine

3 Knowledge Guide

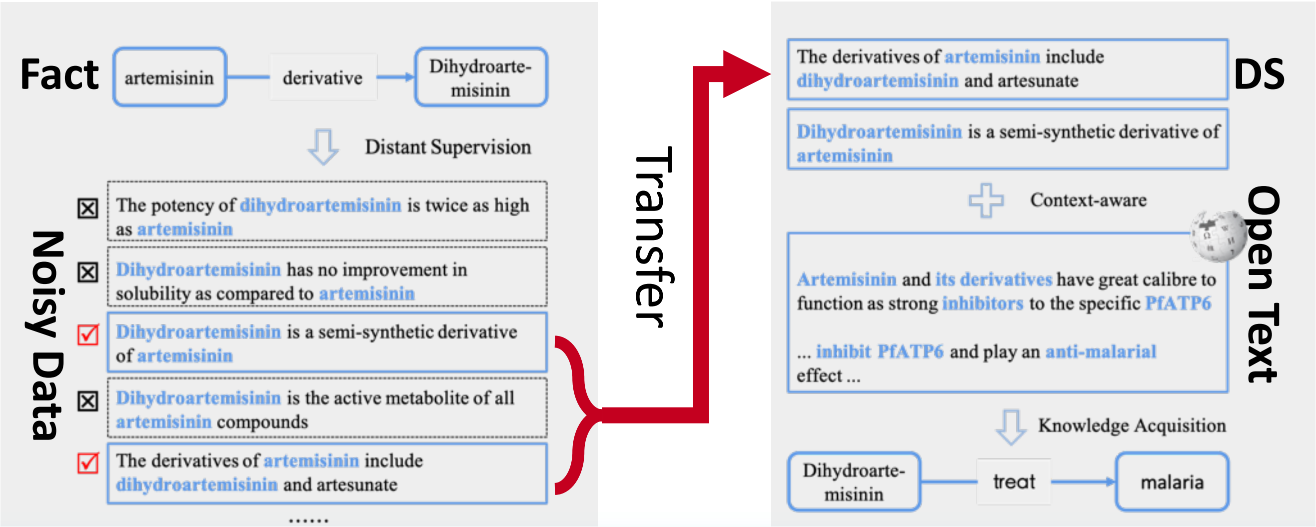
How to improve NLP models by incorporating knowledge

Knowledge Extraction from Open Text

- Challenge：From noisy text to accurate knowledge

  Filtering - Instance-level attention to remove noise

  Context - Use rich context to improve accuracy & coverage



Instance-Level Attention

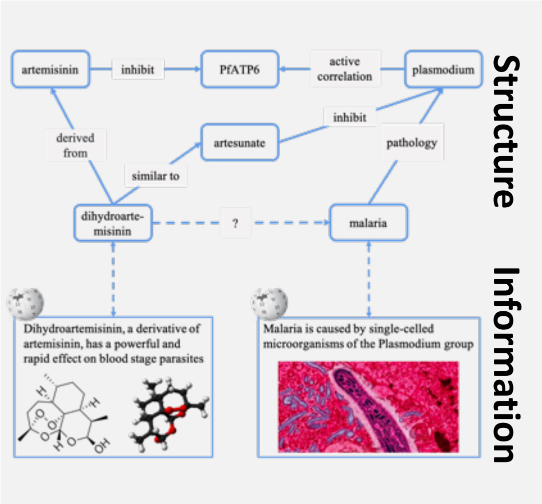
Relation Extraction from Open Text

Representation Learning for Complex Knowledge

- Challenge: Efficient knowledge representation for machine

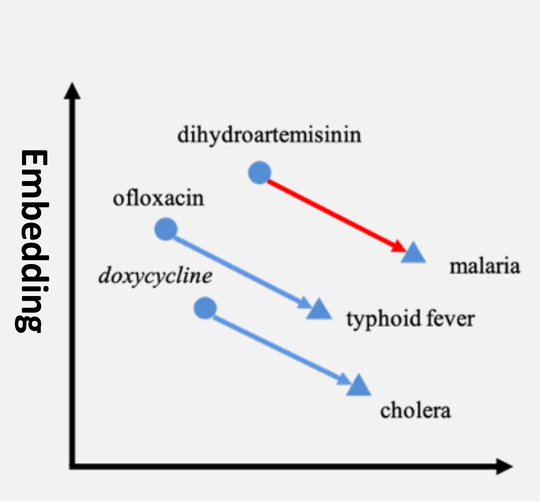
  Fusion – Consider internal and external information of KGs

  Unified - Build unified knowledge embeddings



Human Knowledge

Unified Representation



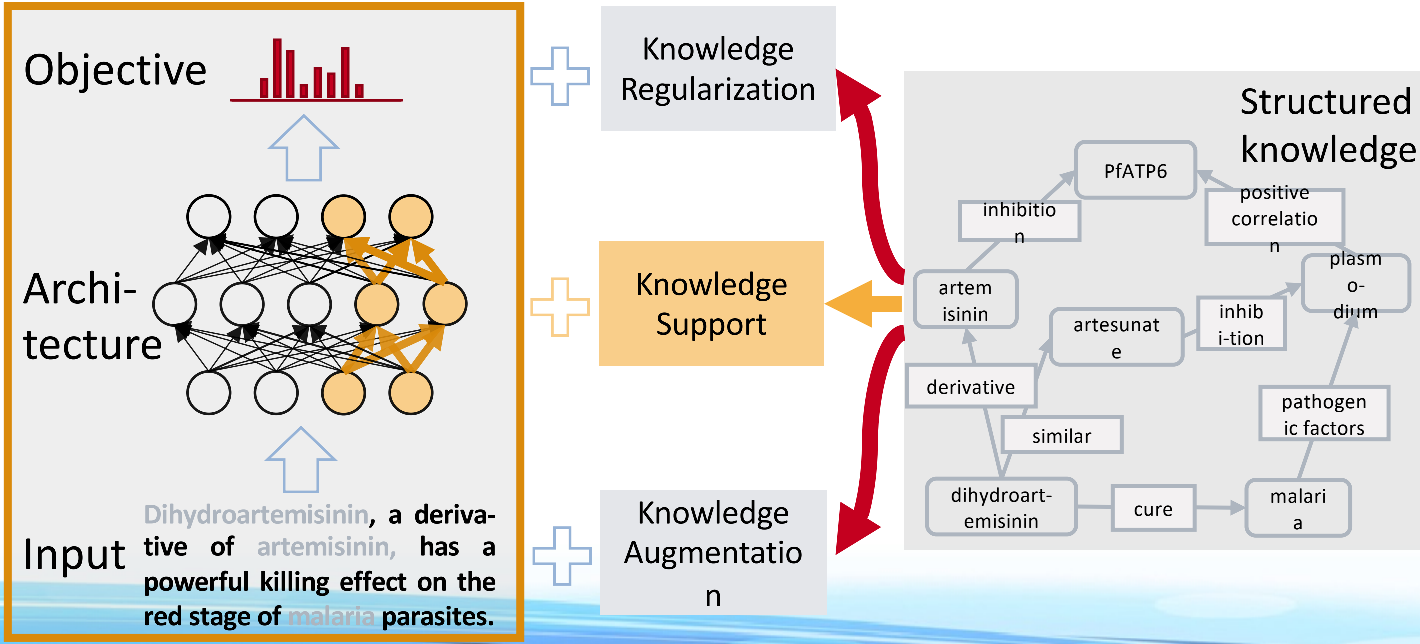
Machine Knowledge

Knowledge-Guided NLP Models

- Challenge: Incorporate knowledge in heterogeneous models

  Arch – Design learning architecture with knowledge

  In/Out – Design inputs and objectives with knowledge



Pre-trained Language Models as Advanced Model Knowledge

Pretrained Language Model as a Breakthrough in 2018

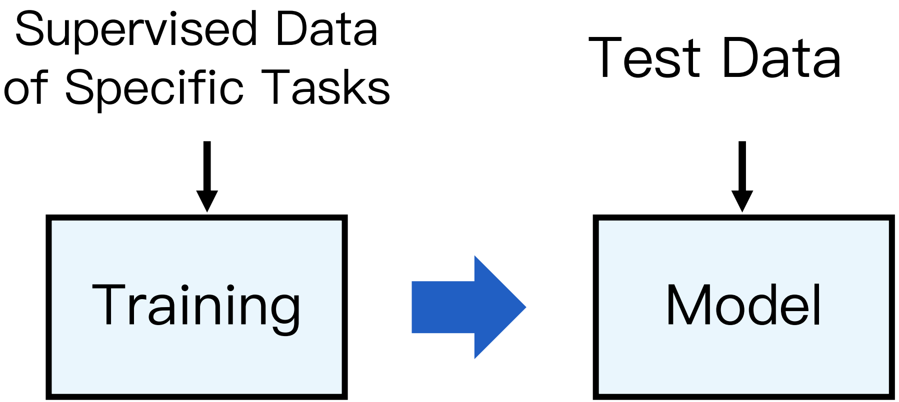
- Impressive progress of deep learning on unsupervised text corpora

2001	Neural language models
2008	Multi-task learning
2013	Word embeddings
2013	Neural networks for NLP
2014	Sequence-to-sequence models
2015	Attention
2015	Memory-based networks
2018	Pretrained language models

Sebastian Ruder http://ruder.io/a-review-of-the-recent-history-of-nlp/

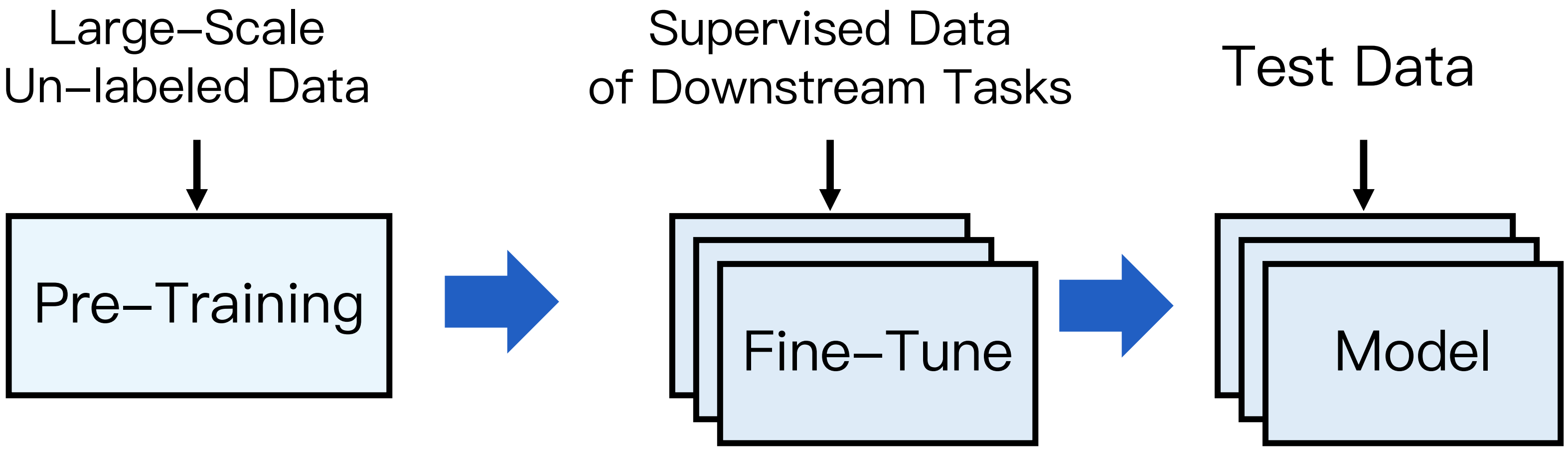
Challenge of Deep Learning in NLP

- Deep Learning has achieved the best performance in most NLP tasks
- Challenges: require large-scale supervised training

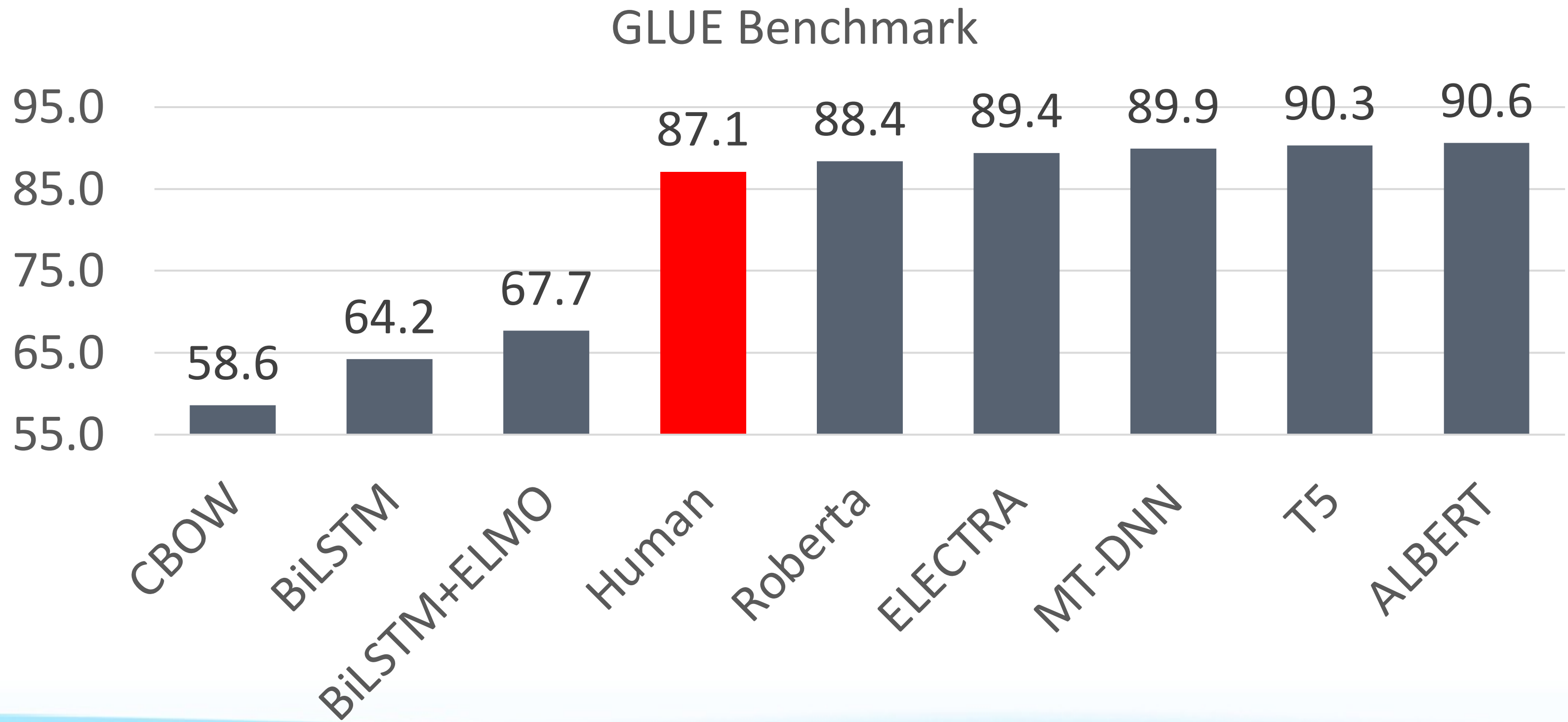


Pretrained Language Models

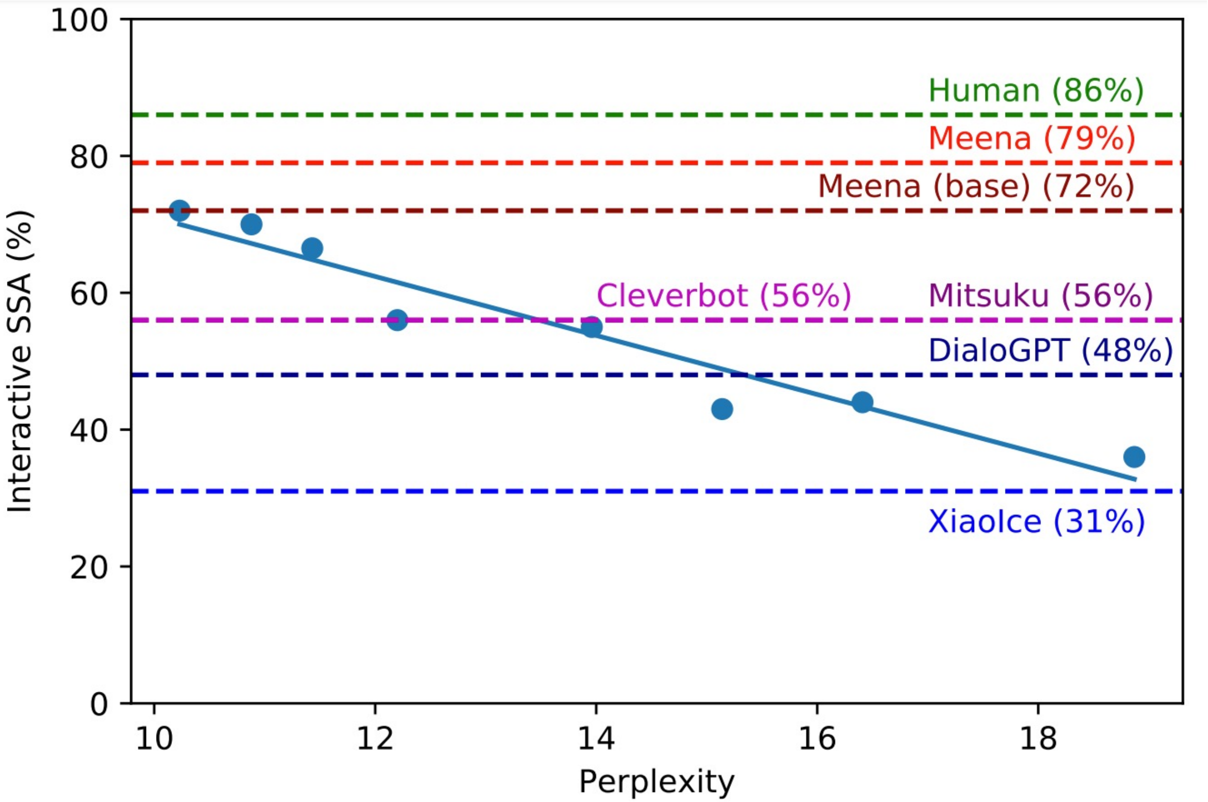
- Pre-trained Language Models (PLMs) can learn language patterns from large-scale un-labeled data, and improve the performance on downstream tasks by fine-tuning parameters



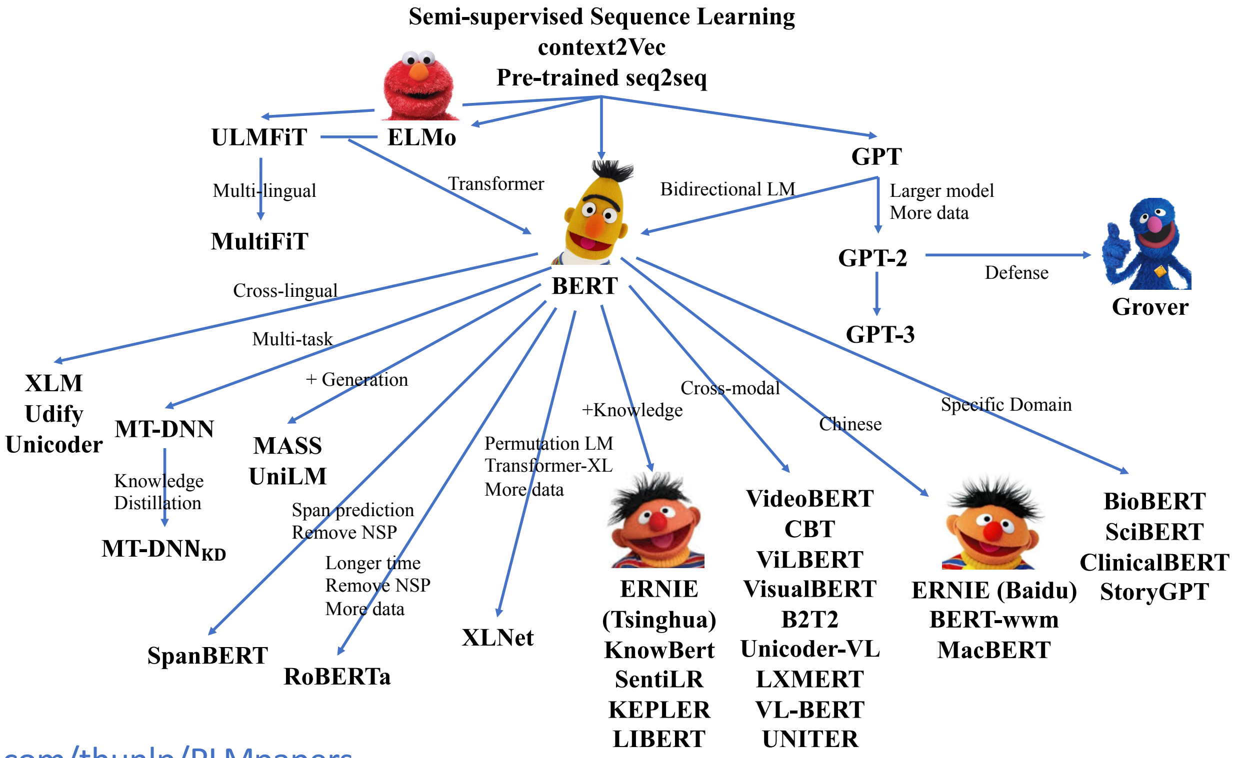
Superior Performance on Language Understanding



Superior Performance on Language Generation



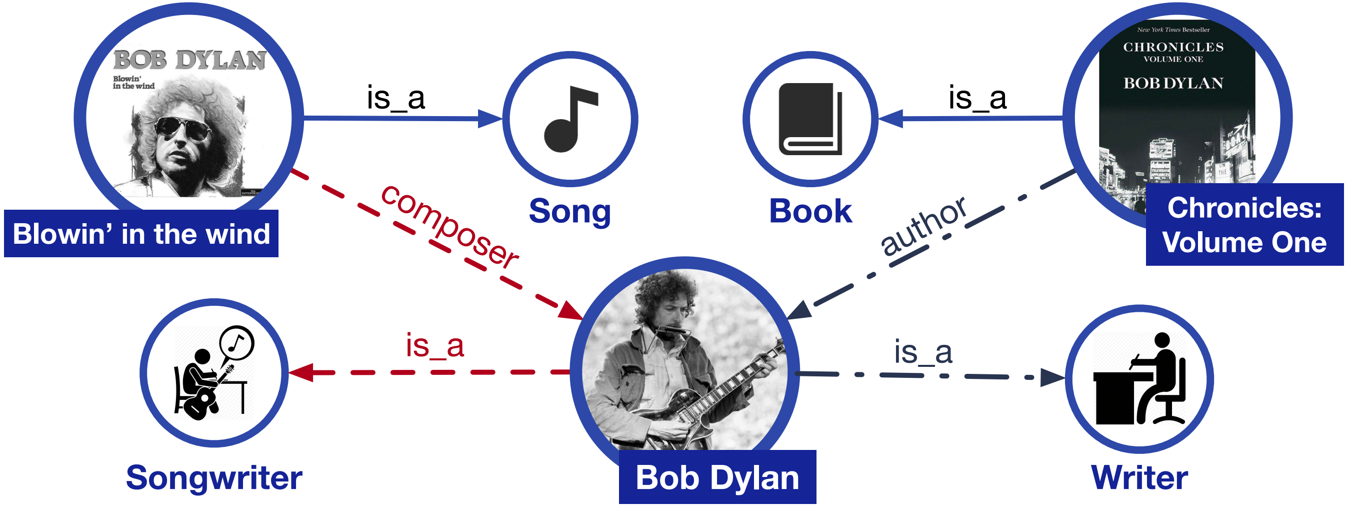
Contests of Pretrained Language Models



https://github.com/thunlp/PLMpapers

Knowledgeable PLM

- Incorporate external symbolic knowledge with model knowledge of PLMs



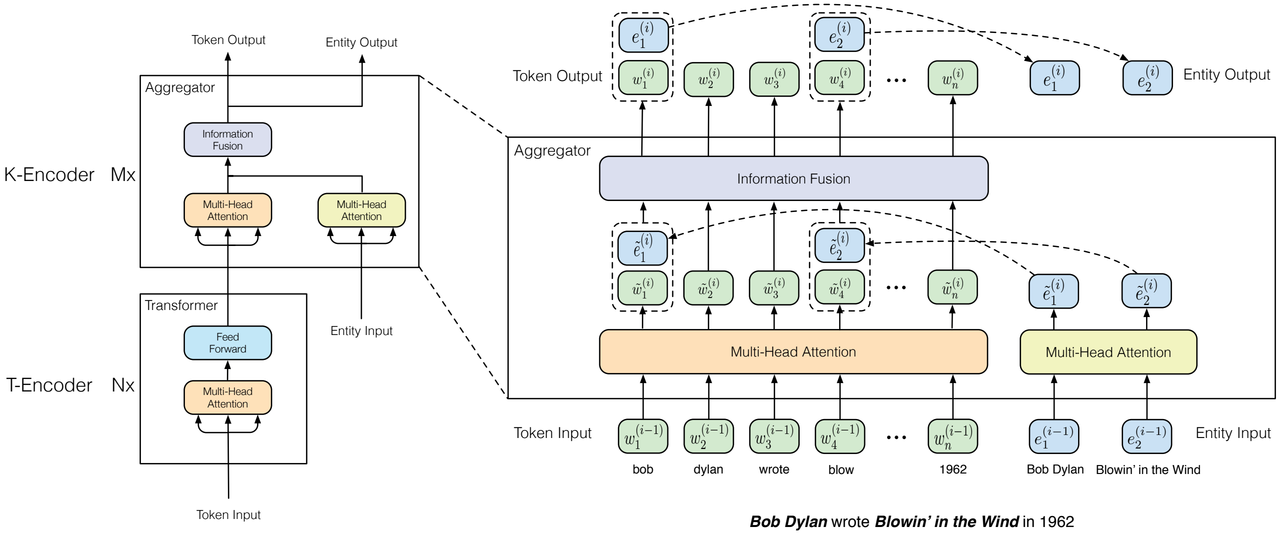
Bob Dylan wrote ***Blowin' in the Wind*** in 1962, and wrote ***Chronicles: Volume One*** in 2004.

How to Make PLMs Knowledgeable

- **Knowledgeable Input**: input augmentation as extra features
- **Knowledgeable Tasks**: knowledge-guided pre-training tasks
- **Knowledgeable Framework**: knowledge-guided neural architecture

Knowledgeable Input

- ERNIE: Enhanced Language Representation with Informative Entities
  - Lower layers for text, and higher layers for knowledge integration
  - Link Prediction Objective with MLM

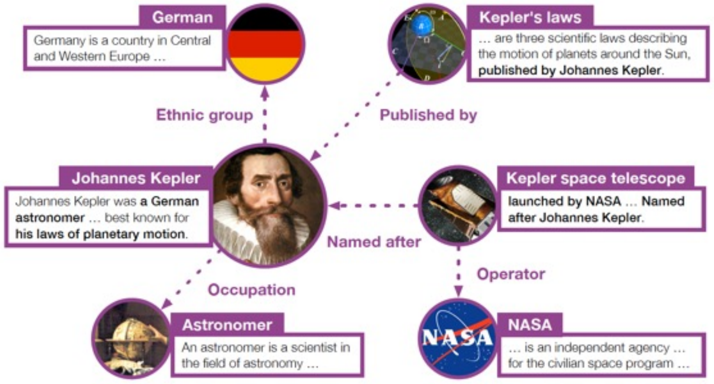


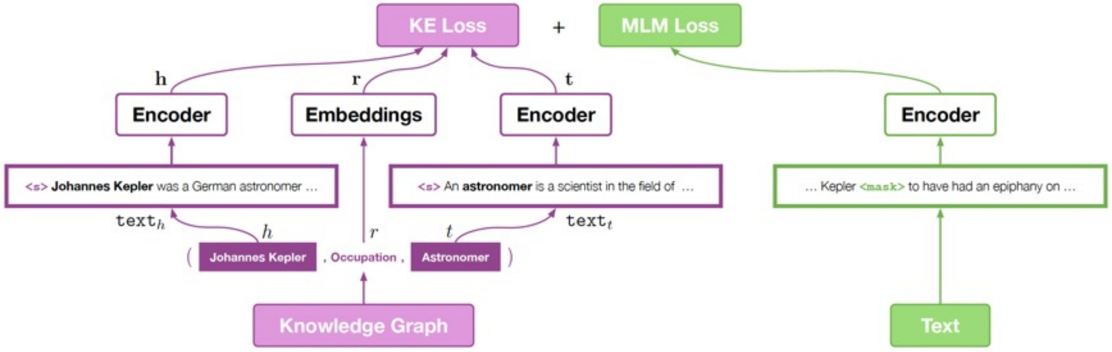
(a) Model Achitecture (b) Aggregator

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu. ERNIE: Enhanced Language Representation with Informative Entities. ACL 2019.

Knowledgeable Tasks

- KEPLER: Joint learning of knowledge and language modeling
- Unify knowledge embedding and language representation into the same semantic space





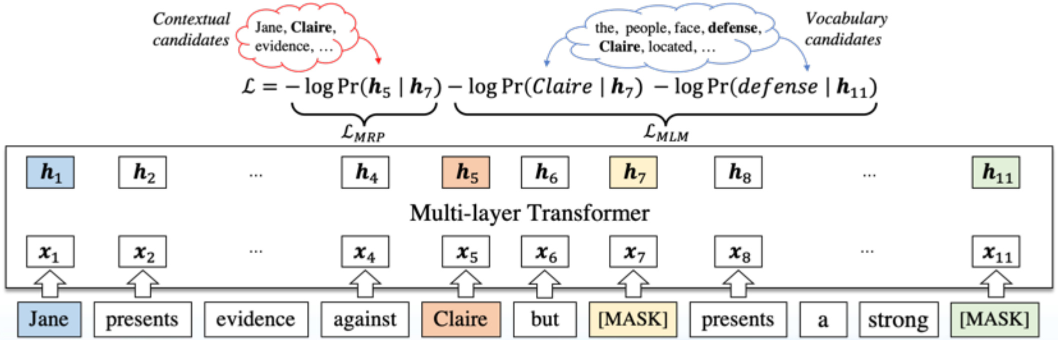
Wang et al. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. To appear at TACL.

Knowledgeable Tasks

- Coreference: Two or more expressions in a text refer to the same entity

  Antoine published ***The Little Prince*** in 1943. ***The book*** follows a young prince who visits various planets in space.

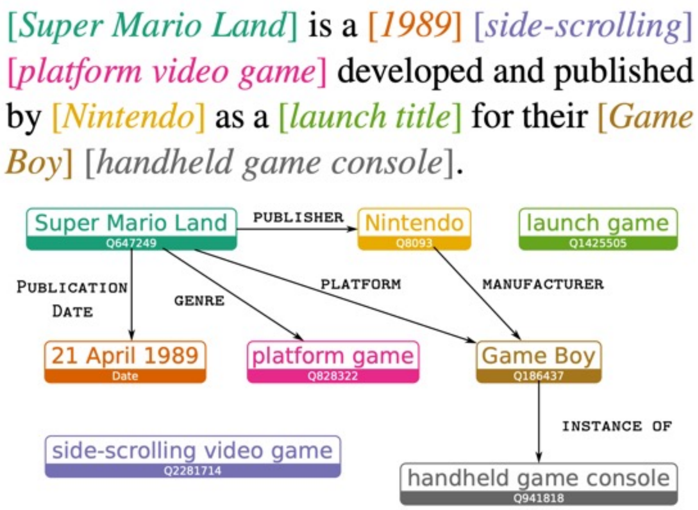
- CorefBERT: Learn coreferential reasoning ability from large-scale unlabeled corpus
  - Mask one or several mentions and requires model to predict the masked mention’s corresponding referents

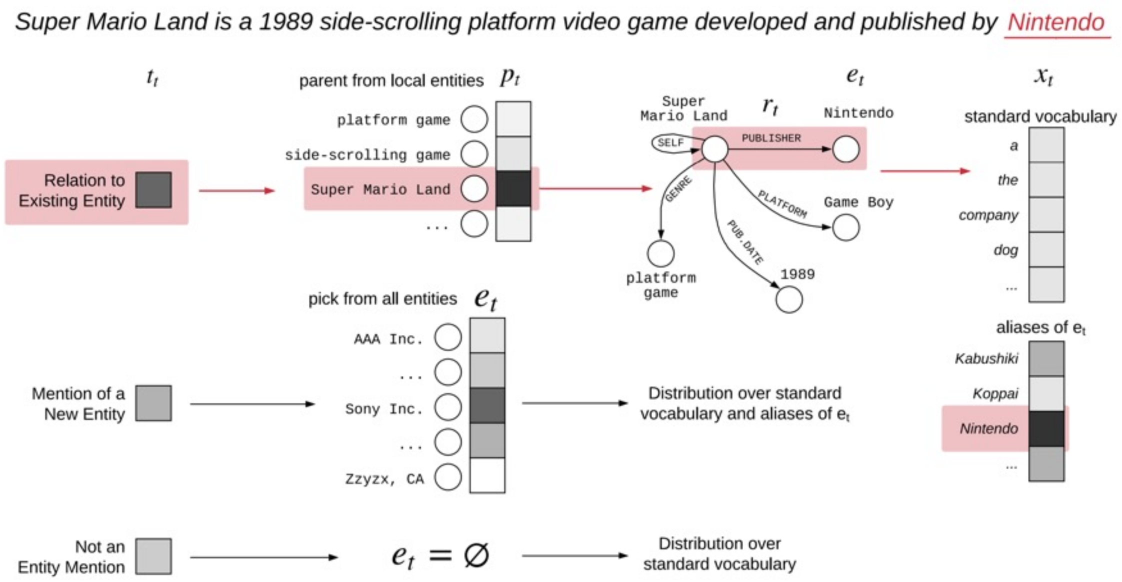


Ye et al. Coreferential Reasoning Learning for Language Representation. EMNLP2020.

Knowledgeable Framework

- LM with mechanisms for selecting and copying facts from KG

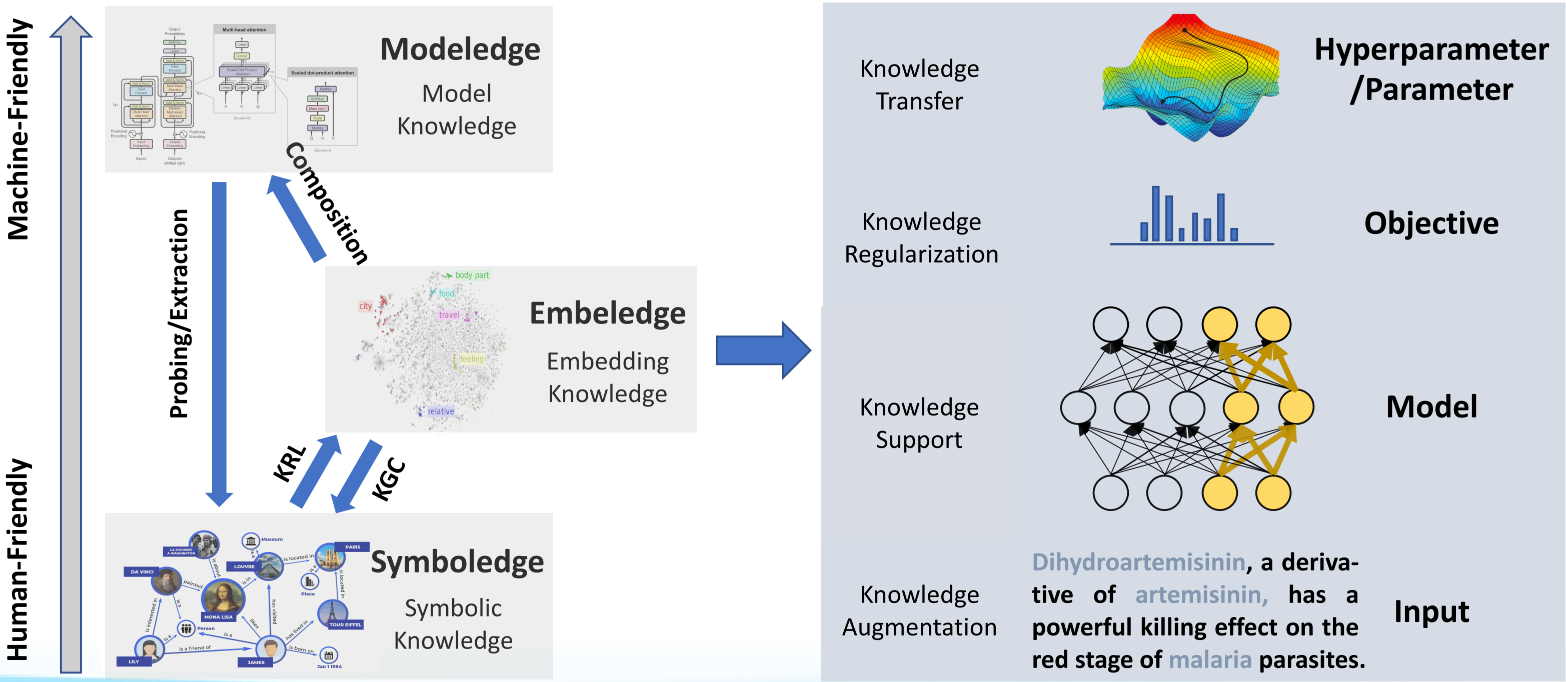




Robert L. Logan IV, Nelson F. Liu, Matthew E. Peters, Matt Gardner, Sameer Singh. Barack's Wife Hillary: Using Knowledge-Graphs for Fact-Aware Language Modeling. ACL 2019.

Framework of Knowledgeable Learning

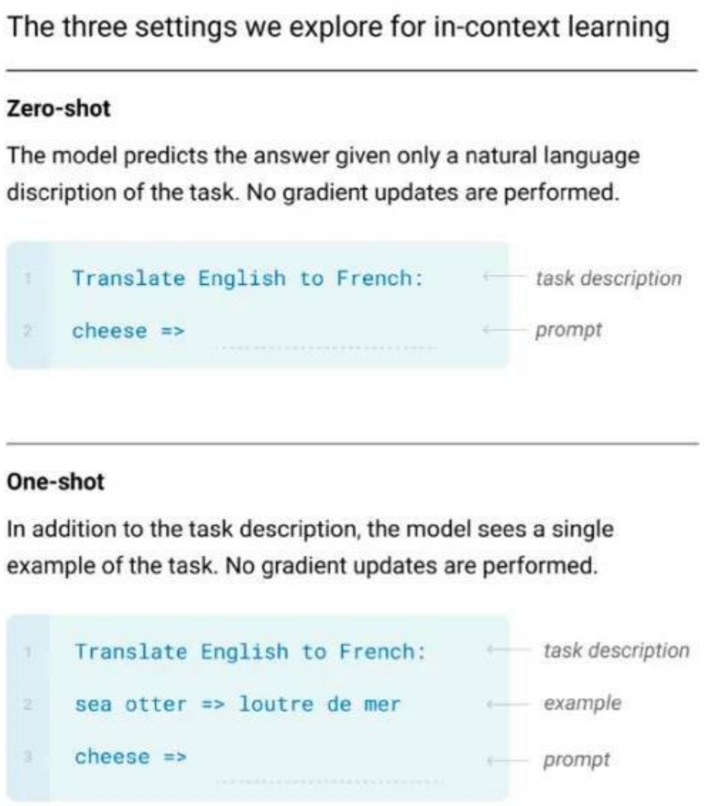
- More methods to incorporate multiple knowledge into deep learning



Model Knowledge Stimulation with Prompts

GPT-3 and Prompts

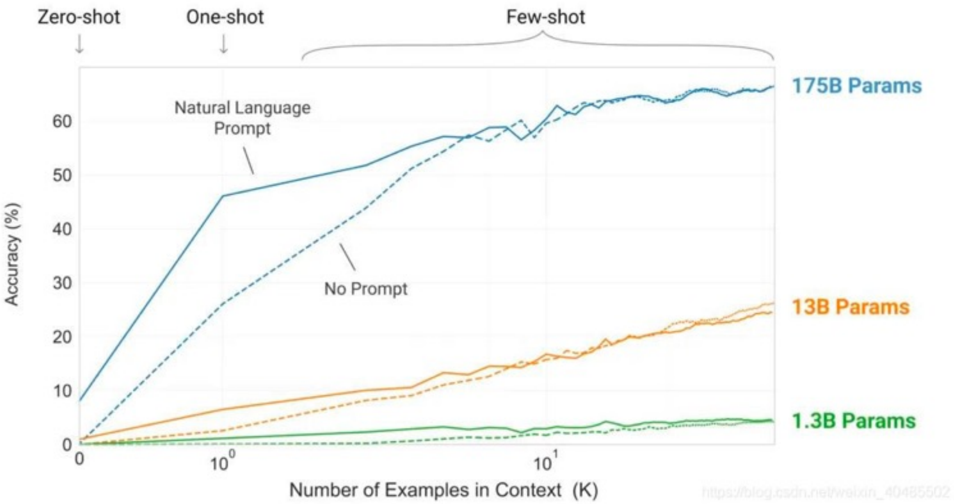
- GPT-3 has 175 billion parameters, almost impossible to fine-tune
- GPT-3 introduces prompts to stimulate knowledge in PLMs
- Prompts are typically task descriptions and language triggers to give models hints to generate words
- By adding prompts, downstream tasks are formalized as language modeling problems



Brown et al. GPT-3: Language Models are Few-Shot Learner. OpenAI 2020.

GPT-3 and Prompts

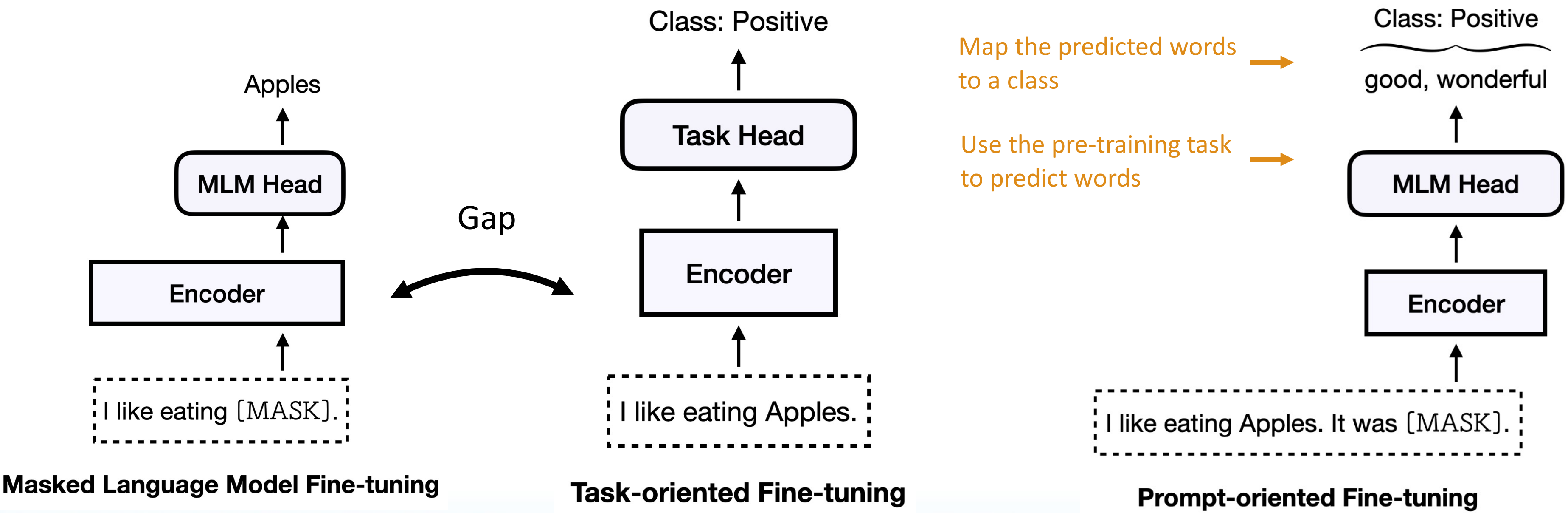
- Prompts stay untuned
- Prompts have great performance on few-shot and zero-shot tasks
- Big models contain more knowledge from large unlabeled corpora, and have better performance



Brown et al. GPT-3: Language Models are Few-Shot Learner. OpenAI 2020.

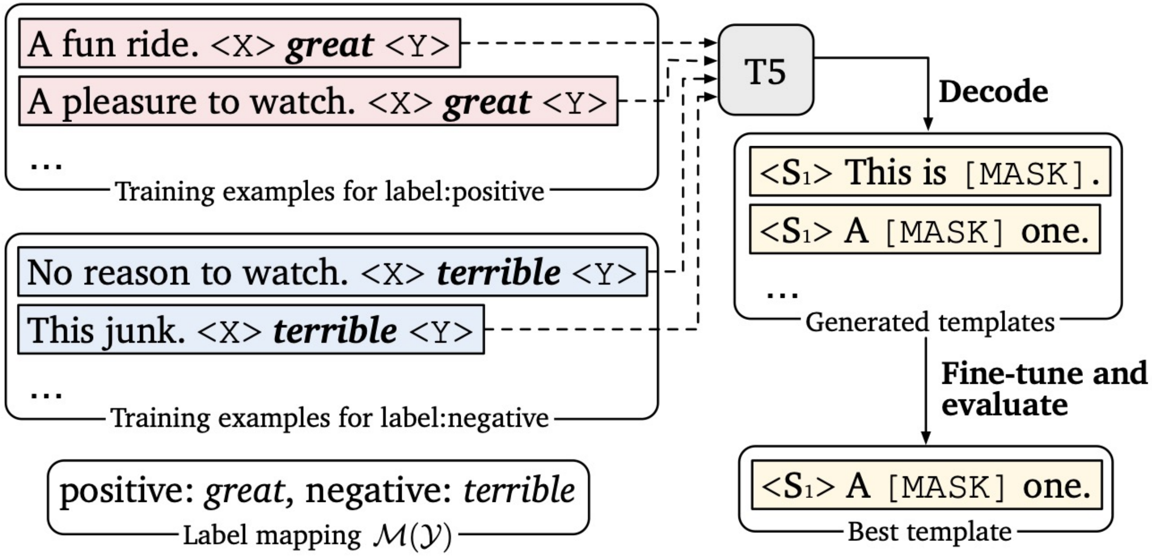
Prompt-Oriented Fine-Tuning

- Prompts can be tuned together with PLMs for downstream tasks



Prompt-Oriented Fine-Tuning

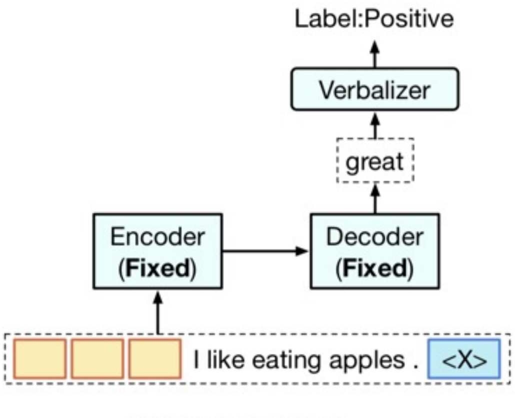
- Auto generated prompts
- Use encoder-decoder model to generate templates



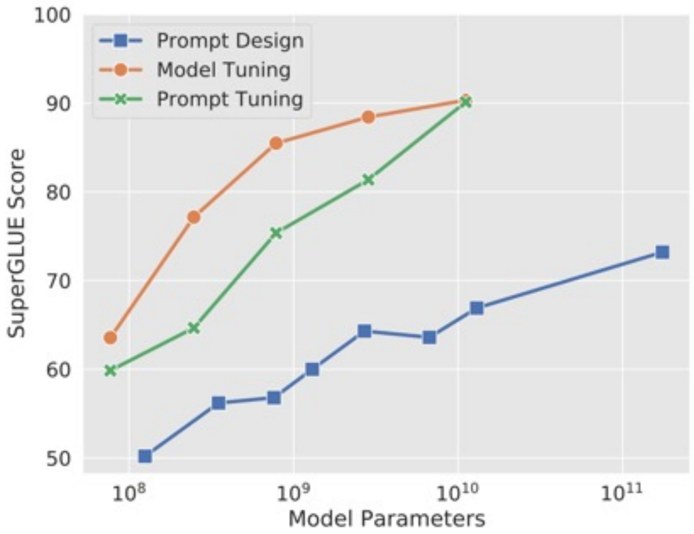
Making Pre-trained Language Models Better Few-shot Learners. 2021

Prompt Tuning

- Keep PLM fixed and tune soft prompts
- Achieve comparable performance with tuning all model parameters



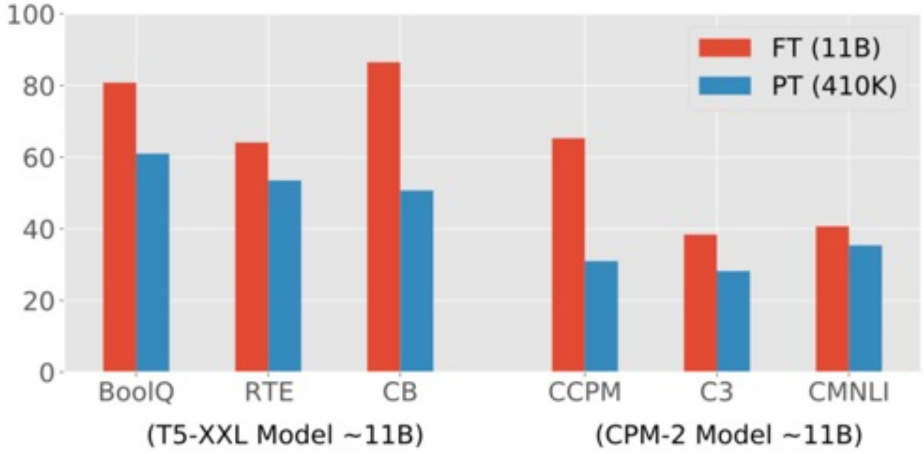
(d) Prompt Tuning

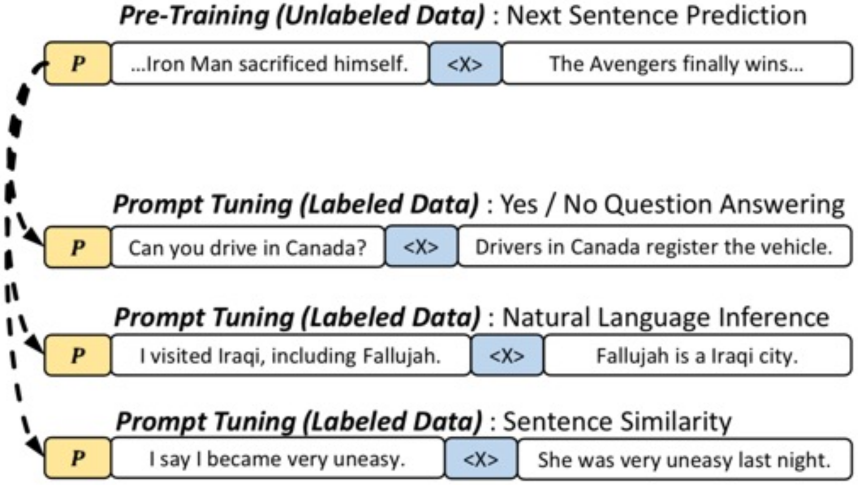


Lester et al. The Power of Scale for Parameter-Efficient Prompt Tuning. Google 2021.

Pre-trained Prompt Tuning

- Tuning soft prompts under few-shot setting is not easy
- Pre-train general soft prompts, keep PLMs fixed and tune pre-trained soft prompts for downstream tasks





Gu et al. PPT: Pre-trained Prompt Tuning for Few-shot Learning. 2021.

Pre-trained Prompt Tuning

- As compared with prompt tuning, pre-trained prompt tuning works better under few-shot settings

			English Tasks						
	Model	Method	SST-2 Acc.	SST-5 Acc.	RACE-m Acc.	RACE-h Acc.	BoolQ Acc.	RTE Acc.	CB F1
FT (11B)	T5-Small	-	72.8$_{3.1}$	31.1$_{0.4}$	26.4$_{0.6}$	26.3$_{0.5}$	59.2$_{0.6}$	54.0$_{1.7}$	70.1$_{4.6}$
	T5-Base	-	74.6$_{2.7}$	28.8$_{1.8}$	27.2$_{0.5}$	26.7$_{0.2}$	61.9$_{2.1}$	56.1$_{2.3}$	70.4$_{2.6}$
	T5-Large	-	89.1$_{2.2}$	42.4$_{1.2}$	48.2$_{1.6}$	43.2$_{1.7}$	74.6$_{0.9}$	64.4$_{3.4}$	82.3$_{2.2}$
	T5-XL	-	89.6$_{3.2}$	38.4$_{5.1}$	55.0$_{2.8}$	50.9$_{2.6}$	77.2$_{2.1}$	62.3$_{6.8}$	81.9$_{9.0}$
	T5-XXL	-	91.4$_{0.8}$	40.6$_{2.0}$	**62.9$_{3.9}$**	**54.8$_{3.0}$**	80.8$_{2.4}$	64.1$_{2.0}$	**86.5$_{5.3}$**
PT (410K)	T5-XXL	Vanilla PT	70.5$_{15.5}$	32.3$_{8.3}$	34.7$_{8.2}$	31.6$_{3.5}$	61.0$_{5.3}$	53.5$_{3.5}$	50.7$_{4.1}$
		Hybrid PT	87.6$_{6.6}$	40.9$_{2.7}$	53.5$_{8.2}$	44.2$_{6.4}$	79.8$_{1.5}$	56.8$_{2.6}$	66.5$_{7.2}$
		LM Adaption	77.6$_{7.5}$	36.2$_{3.6}$	27.3$_{0.2}$	26.5$_{0.4}$	62.0$_{0.3}$	55.3$_{1.0}$	61.2$_{1.7}$
		PPT	93.5$_{0.3}$	**50.2$_{0.7}$**	60.0$_{1.2}$	53.0$_{0.4}$	66.43$_{5.7}$	58.9$_{1.6}$	71.2$_{6.2}$
		Hybrid PPT	93.8$_{0.1}$	50.1$_{0.5}$	62.5$_{0.9}$	52.2$_{0.7}$	**82.0$_{1.0}$**	59.8$_{3.2}$	73.2$_{7.0}$
		Unified PPT	**94.4$_{0.3}$**	46.0$_{1.3}$	58.0$_{0.9}$	49.9$_{1.3}$	76.0$_{2.7}$	**65.8$_{2.1}$**	82.2$_{5.4}$

Gu et al. PPT: Pre-trained Prompt Tuning for Few-shot Learning. 2021.

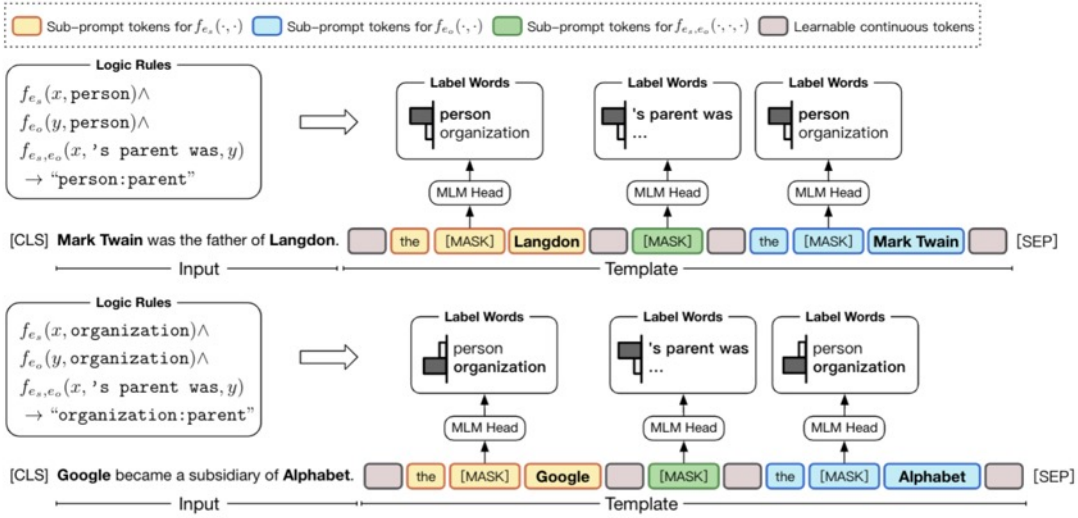
A Brief Comparison

- Performance comparisons among different strategies

Model Parameters	Tuning	Full data	Few-shot Data
Tune	Classifier	—	—
Tune	Prompts	≈	↑
Fix	Classifier	↓	↓
Fix	Prompts	≈ (Big Model)	≈ (Big Model with PPT)

Knowledgeable Prompt Tuning

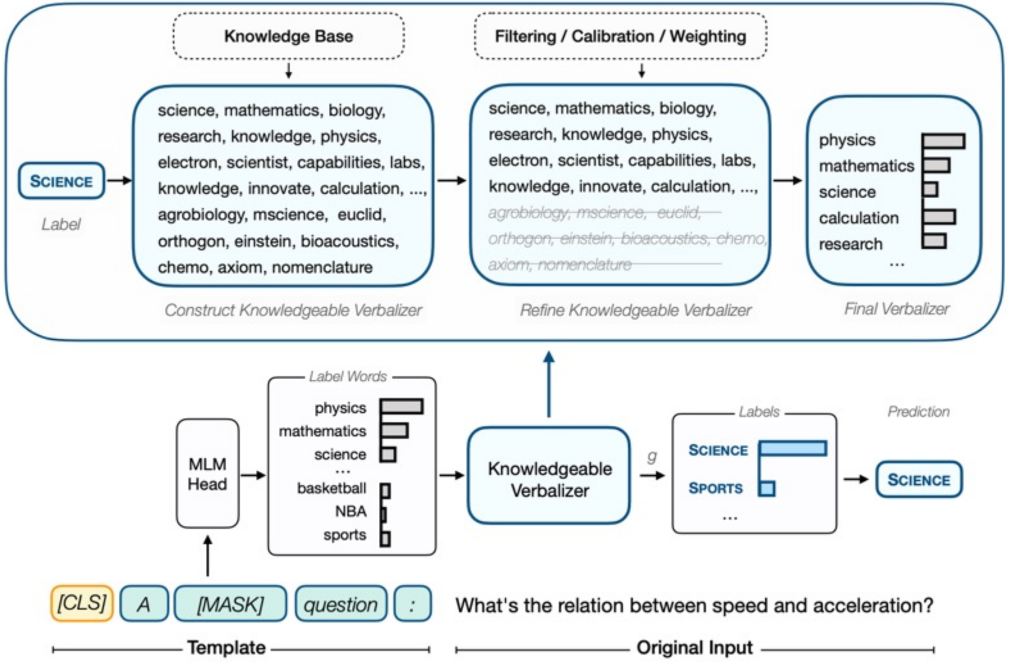
- Combine prompts (model knowledge) with human prior knowledge
- Use logic rules to enhance prompt tuning to downstream classification tasks



Han et al. PTR: Prompt Tuning with Rules for Text Classification. 2021.

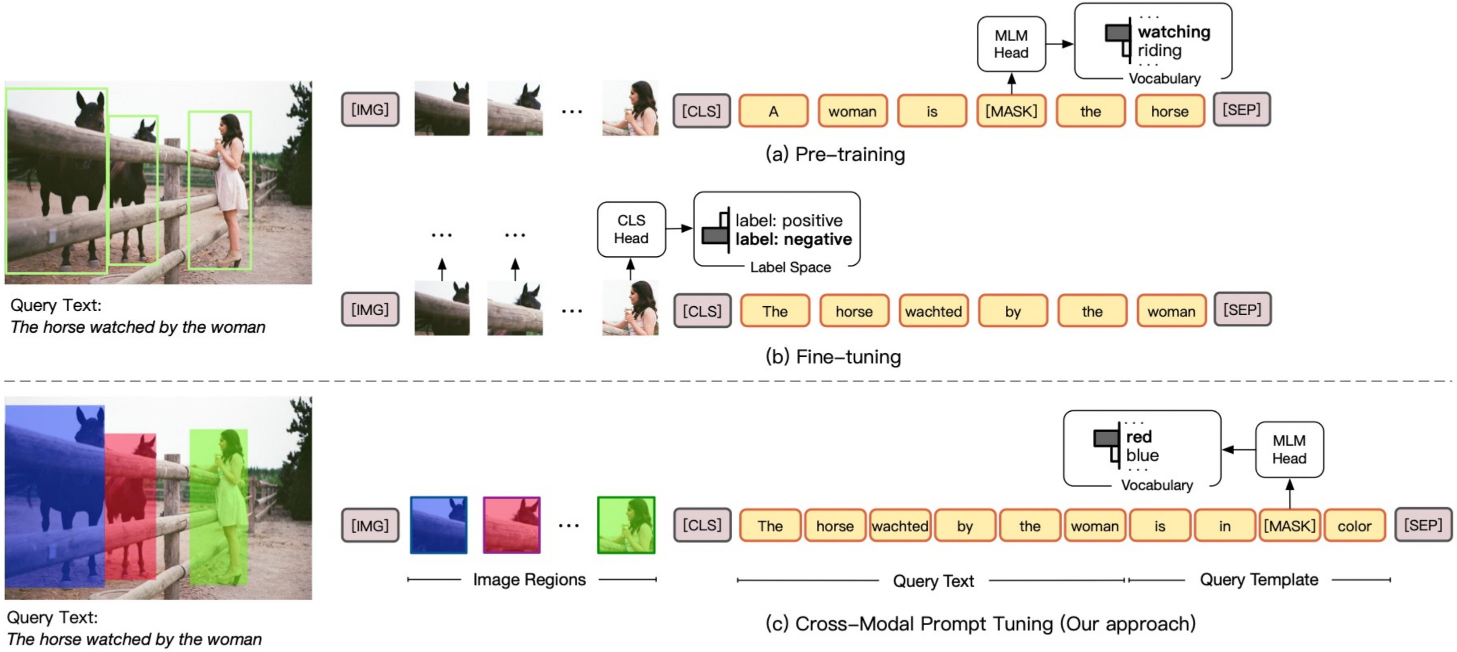
Knowledgeable Prompt Tuning

- Incorporate knowledge base into verbalizer design in prompt tuning



Cross-modal Prompt Tuning

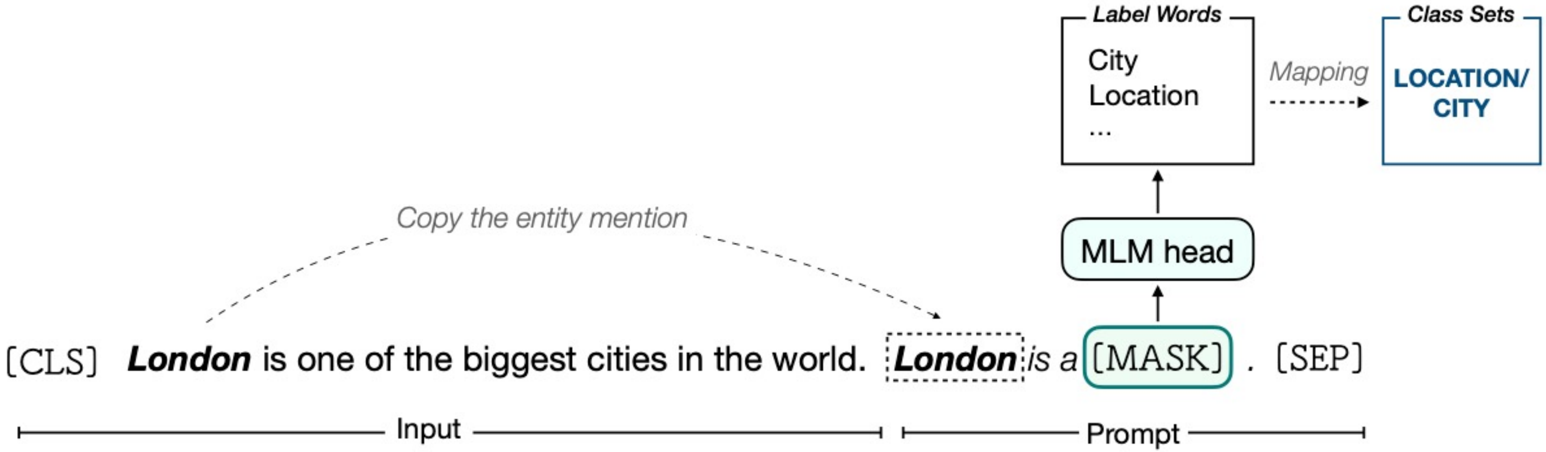
- Cross-model prompts: use prompt-learning in computer vision



CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models. 2021

Application: Information Extraction

- 60~80-way classification for fine-grained entity typing



Prompt-learning for Fine-grained Entity Typing. 2021

Application: Information Extraction

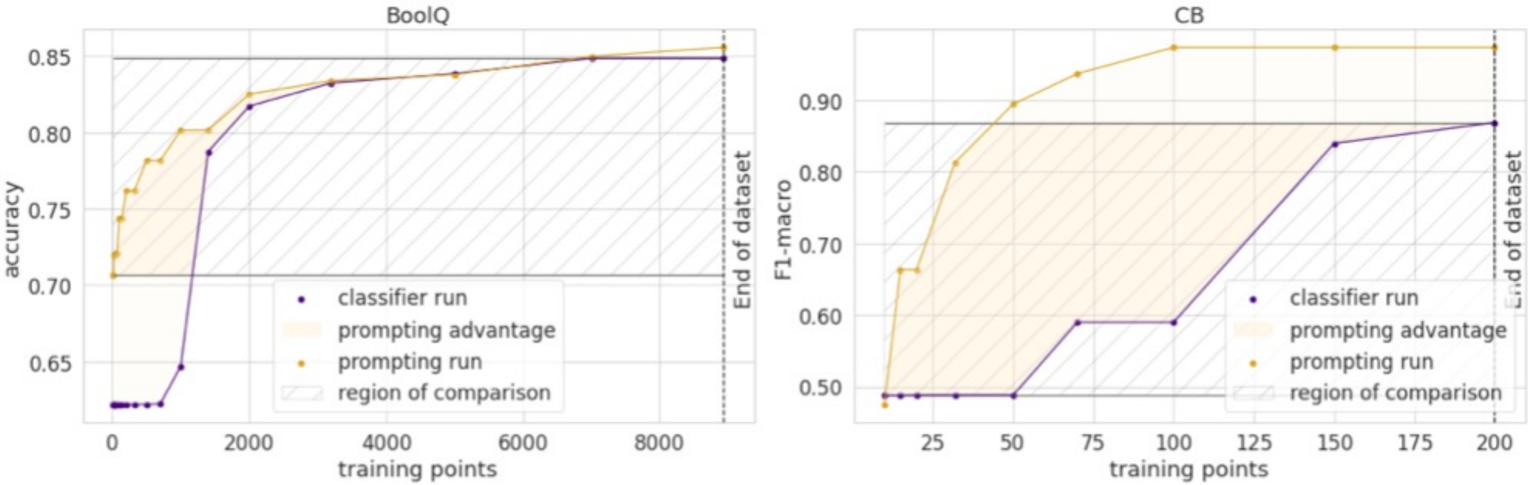
- 60~80-way classification for fine-grained entity typing

Shot	Metric	Few-NERD		OntoNotes		BBN	
		Fine-tuning	PLET	Fine-tuning	PLET	Fine-tuning	PLET
1	Acc	8.94	43.87 (+34.93)	3.70	38.97 (+35.27)	0.80	40.70 (+39.90)
	MiF	19.85	60.60 (+45.75)	18.98	59.91 (+40.93)	5.79	49.25 (+43.46)
	MaF	19.85	60.60 (+40.75)	19.43	61.42 (+41.99)	4.42	48.48 (+43.06)
2	Acc	20.83	47.78 (+26.95)	7.27	39.19 (+31.92)	6.68	41.33 (+34.65)
	MiF	32.67	62.09 (+29.42)	24.89	61.09 (+36.20)	13.70	54.00 (+40.30)
	MaF	32.67	62.09 (+29.42)	25.64	62.68 (+37.04)	13.23	51.97 (+38.74)
4	Acc	33.09	57.00 (+23.91)	11.15	38.39 (+27.24)	19.34	52.21 (+32.87)
	MiF	44.14	68.61 (+24.47)	27.69	59.81 (+32.12)	27.03	61.13 (+34.10)
	MaF	44.14	68.61 (+24.47)	28.26	60.89 (+32.63)	24.69	58.91 (+34.22)
8	Acc	46.44	55.75 (+9.31)	18.37	39.37 (+21.00)	27.01	44.30 (+17.29)
	MiF	57.76	68.74 (+10.98)	38.16	57.97 (+19.81)	40.19	56.21 (+16.02)
	MaF	57.76	68.74 (+10.98)	37.77	58.32 (+20.55)	39.50	55.15 (+15.65)
16	Acc	60.98	61.58 (+0.60)	32.26	42.29 (+10.03)	39.67	55.00 (+15.33)
	MiF	71.59	72.39 (+0.80)	51.40	60.79 (+9.39)	49.01	62.84 (+13.83)
	MaF	71.59	72.39 (+0.80)	51.45	61.80 (+10.35)	47.09	62.38 (+15.29)

Prompt-learning for Fine-grained Entity Typing. 2021

Analysis: Effectiveness in few-shot learning

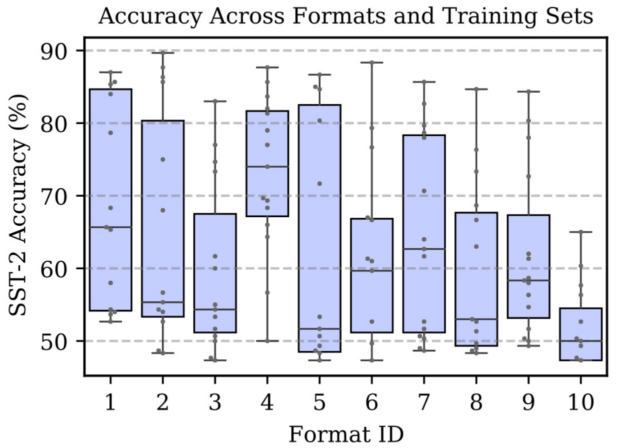
- How many data points is a prompt worth?
- Using 50 examples with prompts is comparable with 200 data points



How Many Data Points is a Prompt Worth? 2021.

Analysis: Stability

- Templates have huge impact, and different templates means different context for [MASK]
- Human-defined, automatically generated, randomly initialized…



Prompt-learning could be unstable for different templates

Dataset	Metric	Method: PLET	Method: PLET (S)
Few-NERD	Acc	17.55	23.99 (+6.44)
	MiF	28.39	47.98 (+19.59)
	MaF	28.39	47.98 (+19.59)
OntoNotes‡	Acc	25.10	28.27 (+3.17)
	MiF	33.61	49.79 (+16.18)
	MaF	37.91	49.95 (+12.04)
BBN	Acc	55.82	57.79 (+1.97)
	MiF	60.64	63.24 (+2.60)
	MaF	59.99	64.00 (+4.01)

Zero-shot entity typing. With appropriate templates, the performance is promising

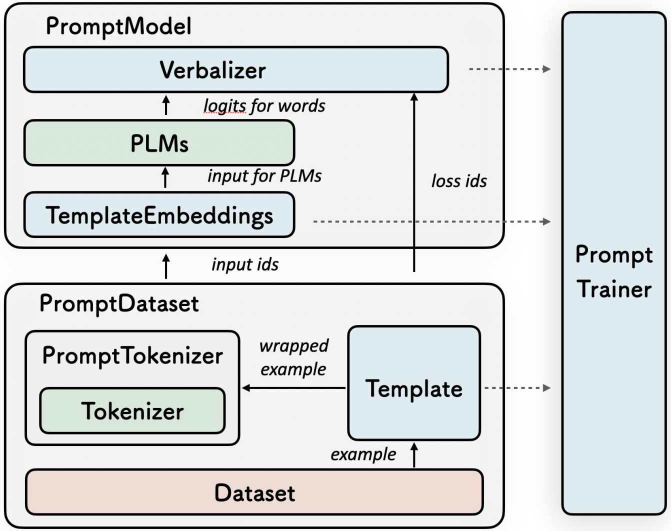
Calibrate Before Use: Improving Few-Shot Performance of Language Models.

Implementation Issues for Prompt-learning

- Prompt-learning is a synthesis of pre-trained tasks, deep models, human prior knowledge and current tasks
- The implementation may face problems
  - What ? What model? What template? Hard or soft? What verbalizer?
  - When? When to insert the template?
  - Where ? Where to insert the template?
  - How ? How to generate templates and verbalizers?

OpenPrompt: A Prompt-learning Programming Framework

https://github.com/thunlp/OpenPrompt

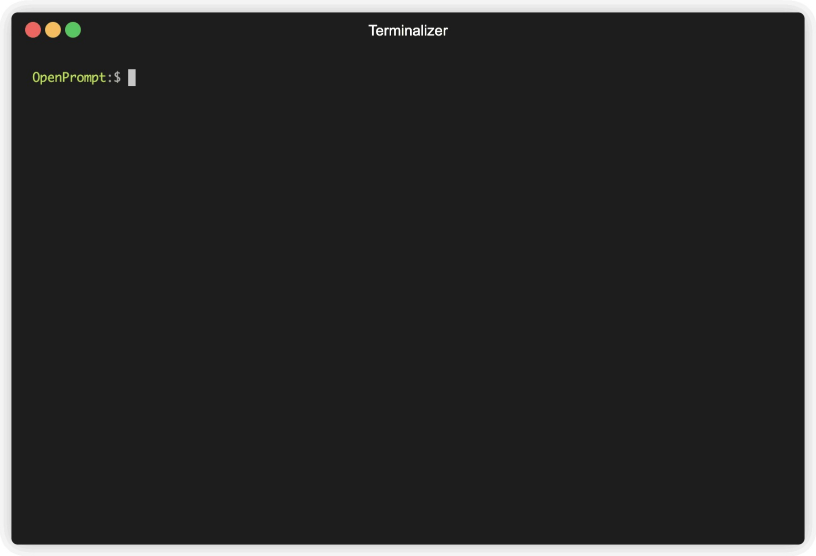


Wrapper Class: *User doesn't need to modify. In order to make prompt tuning resemble traditional PyTorch pipeline*

PLM-related Class: *Support various PLM types.*

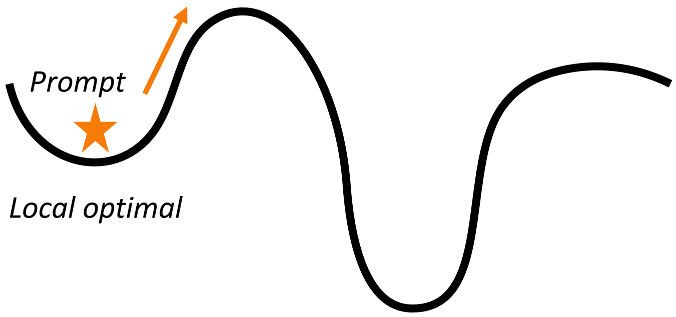
Prompt-related Class: *support many prompt-learning type. Also easy for reimplementation by user.*

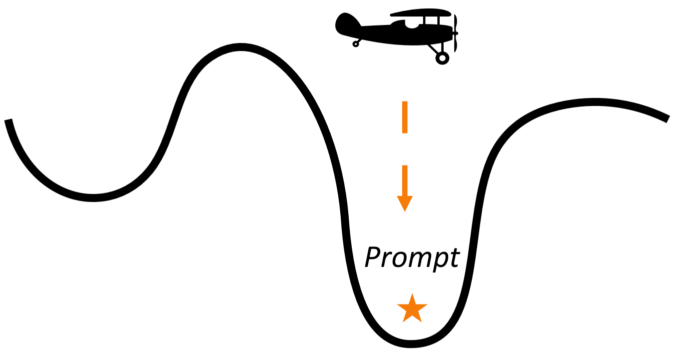
Dataset-related Class: *Support available dataset. Users can also use their own dataset once having the desired data format.*



Remaining Challenges

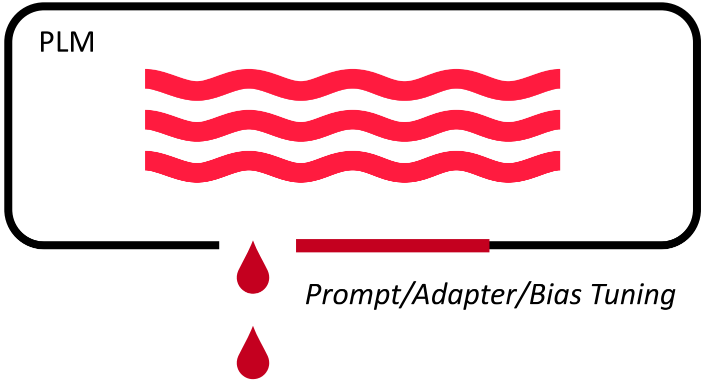
- Converge speed of prompt-tuning for super large models
- The convergence speed is still very slow
- Fast estimation for prompt-tuning





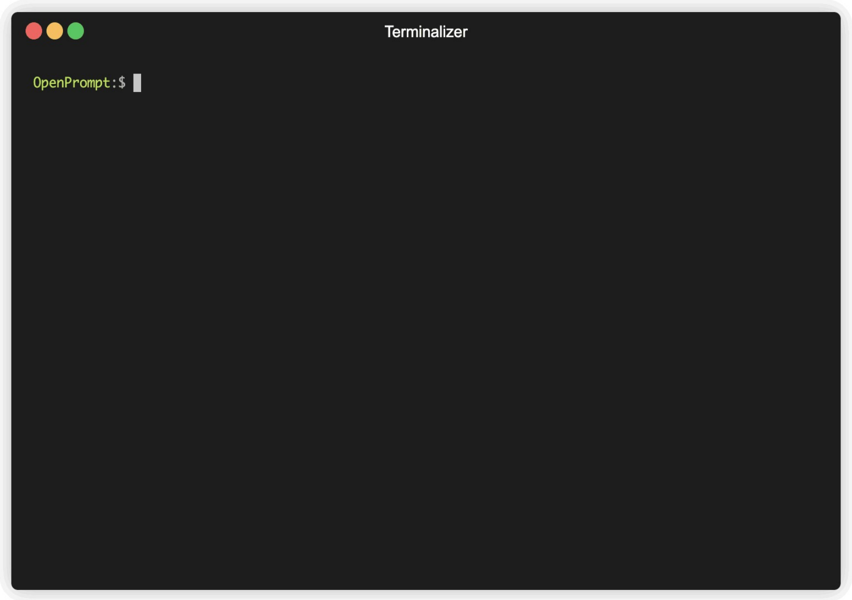
Remaining Challenges

- Only tune prompts, adapters, or biases. Are they all the same?
- Additional parameters in different pattern: contexts, MLPs, matrices…
- Assumption: They are just switches for knowledge distributed in PLMs



Remaining Challenges

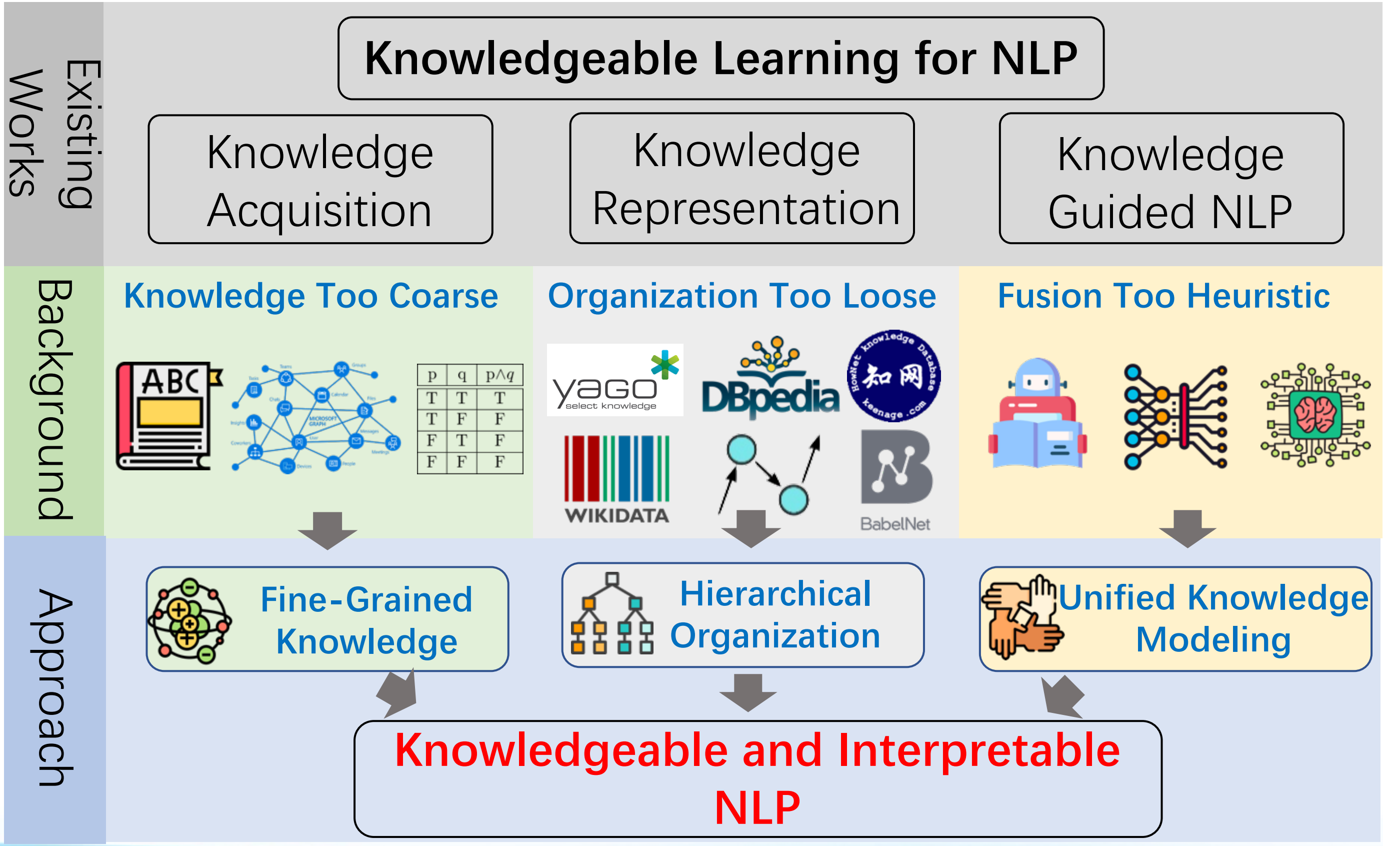
- Still vanilla pre-training?
- Pursue the grand unity for pre-training and model tuning
- A central model with toolkit can do all the things



Summary

- Knowledge is the key to deep understanding of human languages
- Knowledge can be represented in appropriate ways: symbol vs. model
- Big PLMs are the most advanced approach to model knowledge
- Big PLMs do capture knowledge from plain text including commonsense
- The challenge is to stimulate and stabilize model knowledge in PLMs
- **Prompt Tuning** seems a promising approach to stimulate model knowledge for NLP
- Prompt Tuning is friendly to deploy big PLMs in applications, one PLM vs. thousands of prompts and applications

Future Work

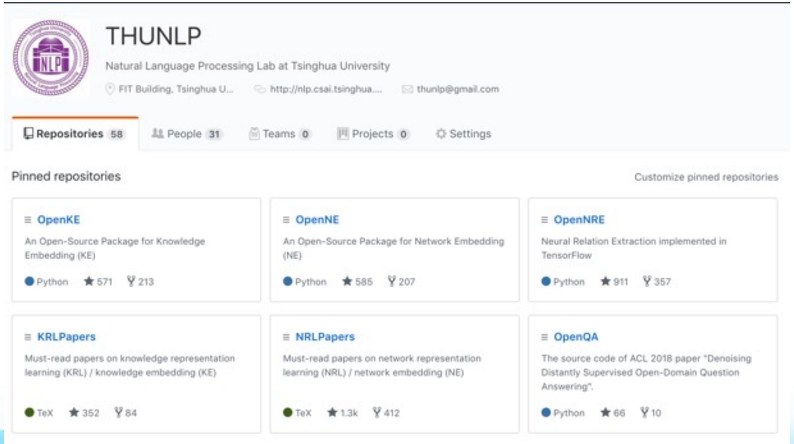


Existing Works
Knowledgeable Learning for NLP
Knowledge Acquisition
Knowledge Representation
Knowledge Guided NLP
Background
Knowledge Too Coarse
Organization Too Loose
Fusion Too Heuristic
yago select knowledge
DBpedia
WIKIDATA
BabelNet
Approach
Fine-Grained Knowledge
Hierarchical Organization
Unified Knowledge Modeling
Knowledgeable and Interpretable NLP

Open Source

- Packages for representation and acquisition of linguistic and world knowledge
- The projects obtain 40000+ stars on GitHub

https://github.com/thunlp



BMInf - https://github.com/OpenBMB

- Low-cost Inference Package for Big Pretrained Language Models (PLMs)

Implementation	GPU	Encoder Speed (tokens/s)	Decoder Speed (tokens/s)
BMInf	NVIDIA GeForce GTX 1060	533	1.6
BMInf	NVIDIA GeForce GTX 1080Ti	1200	12
BMInf	NVIDIA GeForce GTX 2080Ti	2275	19
BMInf	NVIDIA Tesla V100	2966	20
BMInf	NVIDIA Tesla A100	4365	26
PyTorch	NVIDIA Tesla V100	-	3
PyTorch	NVIDIA Tesla A100	-	7

Resource: **C**hinese **P**re-Trained **M**odels (CPM)

训练数据	任务
新闻	文本分类
百科	自然语言推理
对话	阅读理解
网页	完形填空
故事	对话生成
	实体生成

模型大小

参数量	109M	334M	2.6B
层数	12	24	32
隐向量维度	768	1,024	2,560
每层注意力数	12	16	32
注意力向量维度	64	64	80

arXiv:2012.00413 [pdf, other] cs.CL

CPM: A Large-scale Generative Chinese Pre-trained Language Model

Authors: Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, Maosong Sun

Abstract: ...as the training corpus of GPT-3 is primarily English, and the parameters are not publicly available. In this technical report, we release the Chinese Pre-trained Language Model (**CPM**) with generative pre-training on large-scale Chinese training data. To the best of our knowledge,... ▽ More

Submitted 1 December, 2020; **originally announced** December 2020.

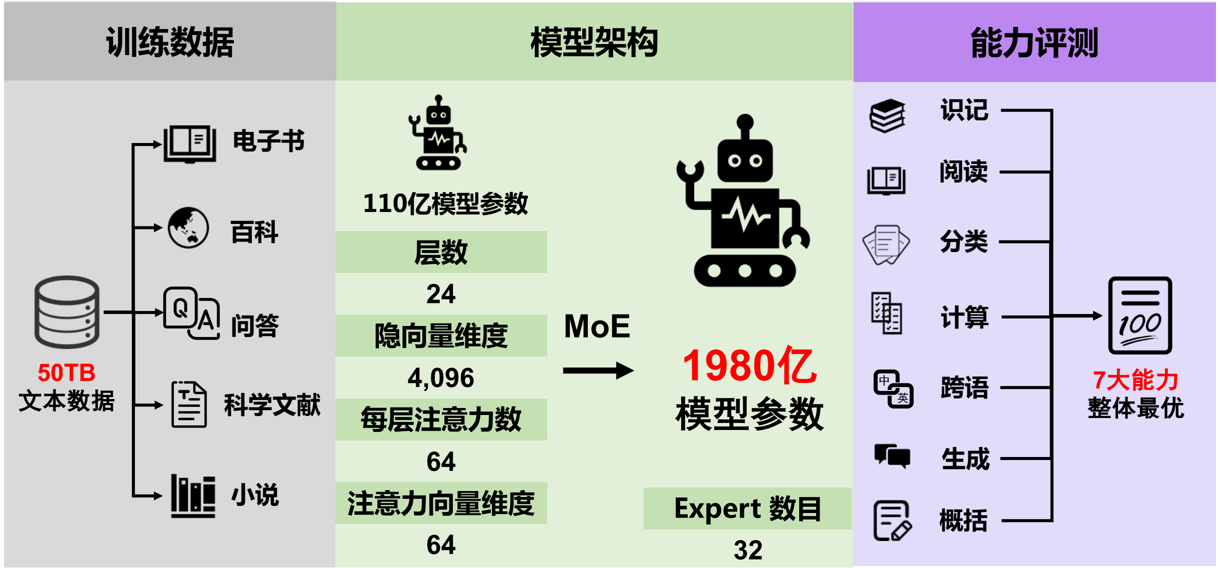
CPM-Generate

Chinese Pre-Trained Language Models (CPM-LM) Version-I

Python MIT 54 595 9 0 Updated 2 days ago

主页（含模型下载）　源码　技术报告

CPM-2: Large-scale **C**ost-effective **P**re-trained Language **M**odels



CPM-3: A Large-Scale **Continual** **P**re-Trained Language **M**odel

CPM-2: Large-Scale **Cost-Effective** **P**re-Trained Language **M**odels

CPM-1: A Large-Scale **Chinese** **P**re-Trained Language **M**odel

项目主页

开源代码

CPM-2技术报告

PLM综述论文

Thanks!

liuzy@tsinghua.edu.cn

http://nlp.csai.tsinghua.edu.cn/~lzy