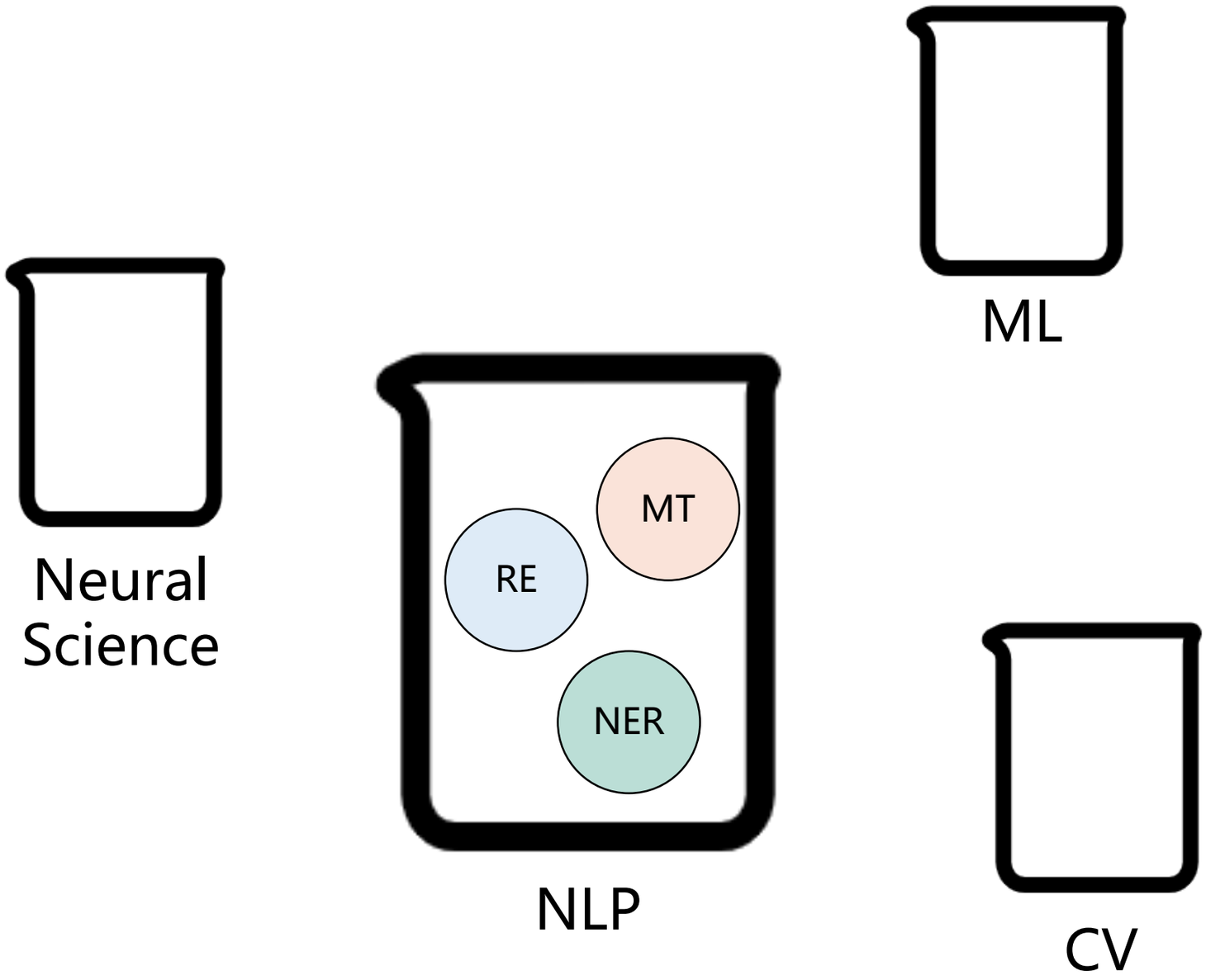
NLP研究杂谈

孙天祥

2022/4/19

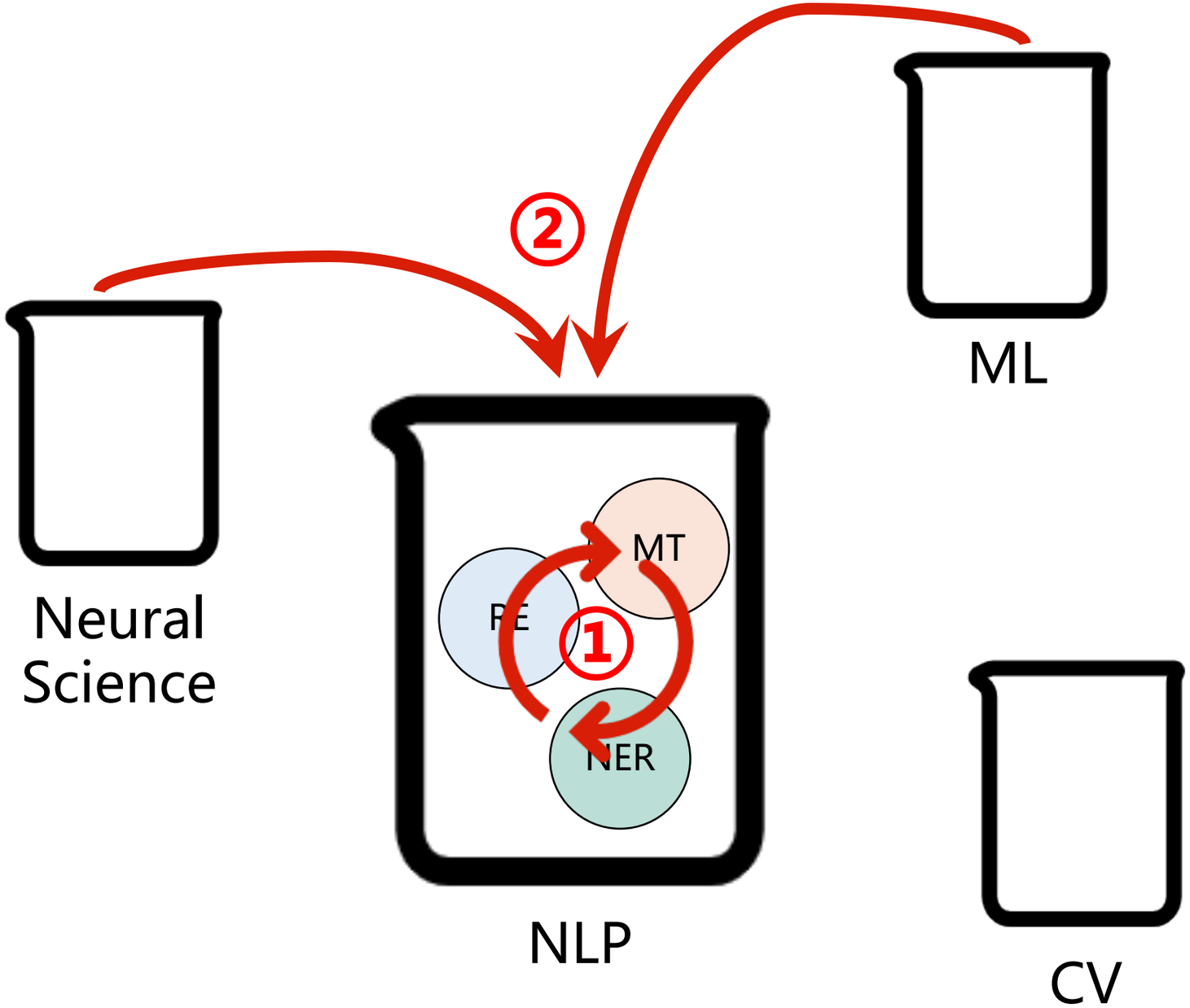
NLP研究划分



NLP研究划分

从哪来？（别人的研究）

- ① 内卷型
- ② 外卷型



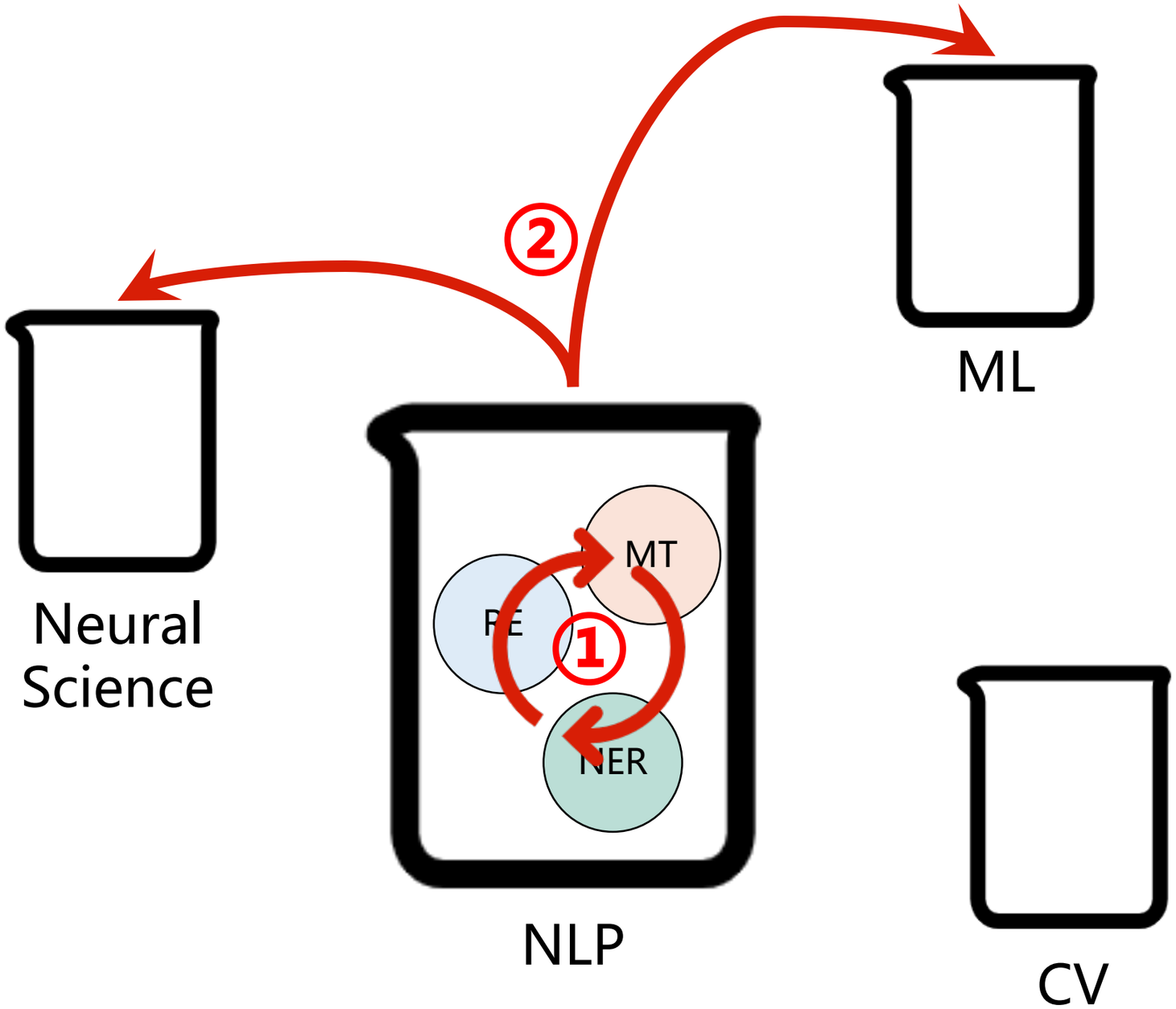
NLP研究划分

从哪来？（别人的研究）

- ① 内卷型
- ② 外卷型

到哪去？（别人的研究）

- ① NLP其他任务
- ② 其他领域



NLP研究划分

从哪来？（别人的研究） ➔ **技巧性**

- ① 内卷型
- ② 外卷型

到哪去？（别人的研究） ➔ **影响力**

- ① NLP其他任务
- ② 其他领域

技巧性 v.s. 影响力

技巧性工作（巧妙地引入了某种方法）

- XLNet, ELECTRA, Unsupervised MT/QA…

影响力工作（简单实用，被广泛使用）

- PtrNet, ELMo, BERT, …

技巧性 v.s. 影响力

一个技巧性工作的例子（Trick-Intensive Paper）

Unsupervised Question Decomposition for Question Answering

Ethan Perez[1 2] **Patrick Lewis**[1 3]
Wen-tau Yih[1] **Kyunghyun Cho**[2 4*] **Douwe Kiela**[1]
[1]Facebook AI Research, [2]New York University,
[3]University College London, [4]CIFAR Azrieli Global Scholar
perez@nyu.edu

技巧性 v.s. 影响力

技巧性工作（巧妙地引入了某种方法）

- XLNet, ELECTRA, Unsupervised MT/QA…

影响力工作（简单实用，被广泛使用）

- PtrNet, ELMo, BERT, …

NOTE: Audience (same task, general NLP, ML, AI, CS, Science, Normal people) is a key factor in the decision process of choosing best paper for conferences

一个例子

- EMNLP 2020 best paper award

Best Paper

Digital voicing of Silent Speech
David Gaddy and Dan Klein

Honourable Mention Papers

If beam search is the answer, what was the question?
Clara Meister, Ryan Cotterell and Tim Vieira

GLUCOSE: GeneraLized and COntextualized Story Explanations
Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran and Jennifer Chu-Carroll

Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems
Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre and Mark Cieliebak

Visually Grounded Compound PCFGs
Yanpeng Zhao and Ivan Titov

启发

从哪来？（别人的研究） → 技巧性 → **可遇不可求**

- ① 内卷型
- ② 外卷型

到哪去？（别人的研究） → 影响力 → **做有意义的方向**

- ① NLP其他任务
- ② 其他领域

什么是有意义的方向？

任务角度：能体现整个NLP领域的发展水平

- 正例：Parsing ➔ MT/QA ➔ PTM
- 负例：只输入不输出的领域

问题角度：包含NLP领域的本质问题

- 正例：降低语言模型困惑度
- 负例：重复生成问题（随着语言模型困惑度下降逐渐消失的问题）

不要做只输入不输出的任务！

NLP研究粗分

I类工作

- 贯穿NLP领域发展的主流脉络

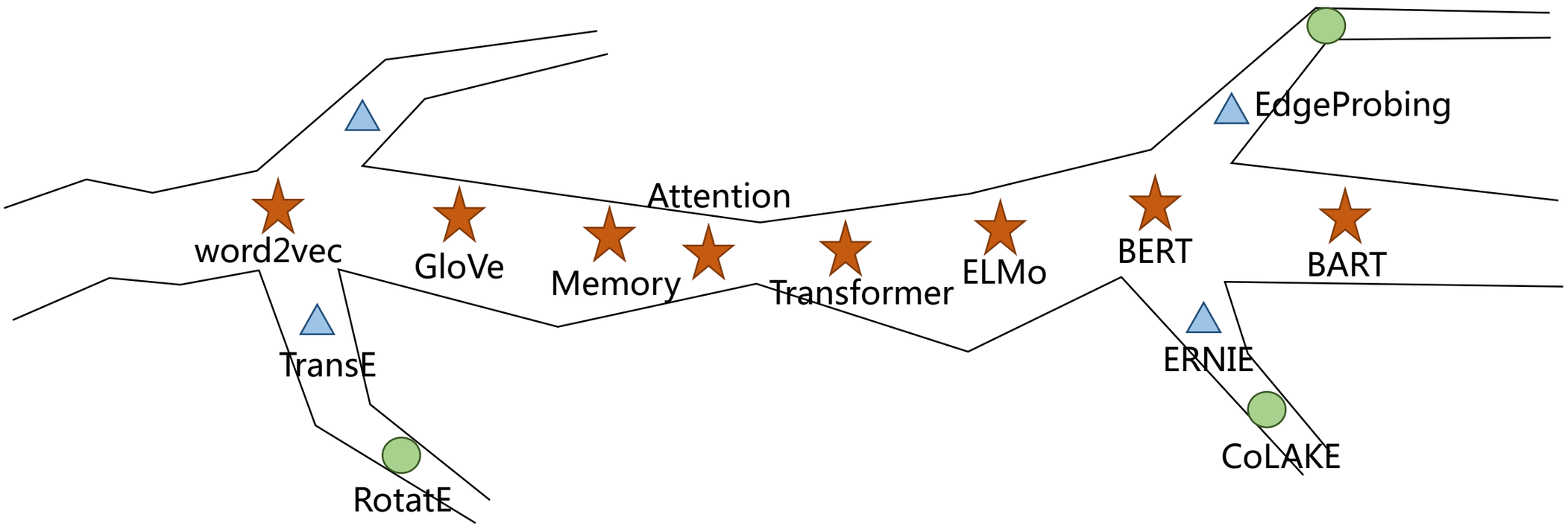
II类工作

- NLP发展脉络支流上的初创工作

III类工作

- 支流中下游工作

NLP研究粗分


EdgeProbing
Attention
word2vec
GloVe
Memory
Transformer
ELMo
BERT
BART
TransE
ERNIE
RotatE
CoLAKE

各类研究都是有迹可循的

III类工作

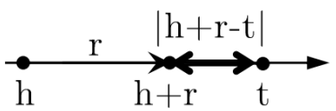
- 规律：寻找已有方法的缺陷完善之
- 特点：方法更完善，甚至很巧妙，性能也更好，但引用不如其前文。
- 例子：RotatE vs. TransE; CoLAKE vs. ERNIE

各类研究都是有迹可循的

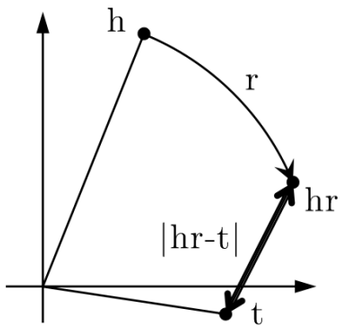
III类工作

- 举例说明

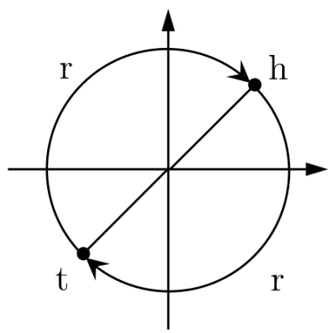
TransE ➔ RotatE: 欧氏空间推广至复空间来解决对称性问题



(a) TransE models $\mathbf{r}$ as translation in real line.



(b) RotatE models $\mathbf{r}$ as rotation in complex plane.



(c) RotatE: an example of modeling symmetric relations $\mathbf{r}$ with $r_i = -1$

各类研究都是有迹可循的

II类工作

- 规律：靠手速抓住历史的车轮
- 特点：功底扎实，嗅觉敏锐，手速快，推动历史在新时代的重演。
- 例子：ERNIE, EdgeProbing...

CoNLL 2016: Joint learning of the embedding of words and entities for named entity disambiguation.

ACL 2019: Enhanced Language Representation with Informative Entities

ACL 2015: How well do distributional models capture different types of semantic knowledge?

ICLR 2019: What do you learn from context? Probing for sentence structure in contextualized word representations

各类研究都是有迹可循的

II类工作

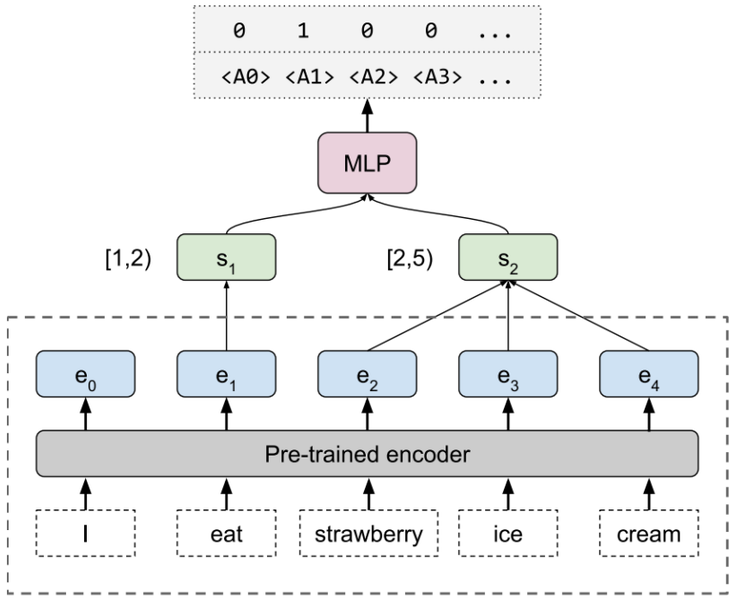
- 举例说明

  前人分析word2vec我分析BERT

2 Learning Semantic Properties of Concepts

The goal of this paper is to gain better understanding of the type of information DMs encode. We do so by evaluating the performance of a predictor trained on a DM-based representation to learn a semantic property. In this section, we describe the proposed learning task, the dataset and the DMs which serve as feature representations.

Rubinstein et al. (2015)



EdgeProbing (2019)

各类研究都是有迹可循的

II类工作

- 举例说明

  前人改word2vec我改BERT

$$P(w_o|e_i) = \frac{\exp({\mathbf{V}_{e_i}}^\top \mathbf{U}_{w_o})}{\sum_{w \in W} \exp({\mathbf{V}_{e_i}}^\top \mathbf{U}_w)}$$

Wikipedia2vec (2016)

➔

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

ERNIE (2019)

各类研究都是有迹可循的

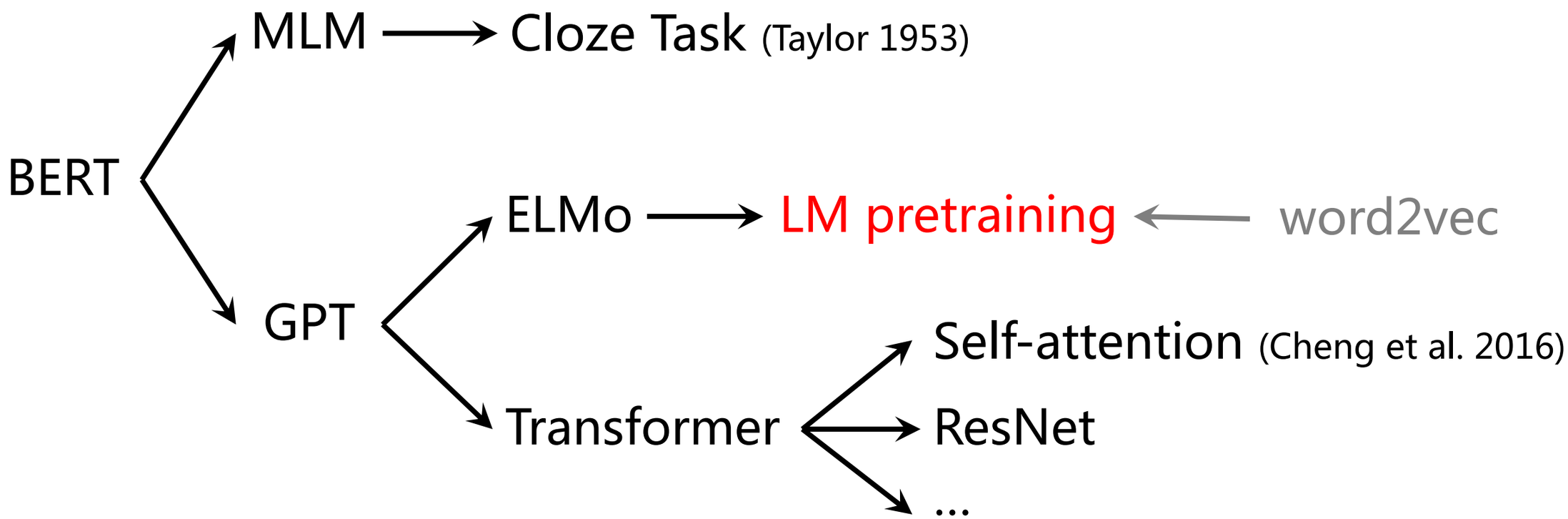
I类工作

- 规律：找到真正work的东西scale之
- 特点：善于发现巨人肩膀的骨质增生处。
- 例子：Transformer, ELMo, BERT...
- 冷知识：I类工作大多是内卷型

各类研究都是有迹可循的

I类工作

- 举例说明



各类研究都是有迹可循的

I类工作

- 举例说明

Language Model We consider a *language model* based on a simple fixed window of text of size ksz using our NN architecture, given in Figure 2. We trained our language model to discriminate a two-class classification task: if the word in the middle of the input window is related to its context or not. We construct a dataset for this task by considering all possible ksz windows of text from the entire of English Wikipedia (`http://en.wikipedia.org`). Positive examples are windows from Wikipedia, negative examples are the same windows but where the middle word has been replaced by a random word.

Results: Language Model Because the language model was trained on a huge database we first trained it *alone*. It takes about a week to train on one computer. The embedding obtained in the word lookup-table was extremely good, even for uncommon words, as shown in Table 1. The embedding obtained by training on labeled data from WordNet "synonyms" is also good (results not shown) however the coverage is not as good as using unlabeled data, e.g. "Dreamcast" is not in the database.

The resulting word lookup-table from the language model was used as an initializer of the lookup-table used in MTL experiments with a language model.

Collobert and Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. ICML 2008

各类研究都是有迹可循的

I类工作

- 举例说明

Pre-training of the shared layer with neural language model For model-III, the shared layer can be initialized by an unsupervised pre-training phase. Here, for the shared LSTM layer in Model-III, we initialize it by a language model [Bengio *et al.*, 2007], which is trained on all the four task dataset.

Model	SST-1	SST-2	SUBJ	IMDB	AvgΔ
Single Task	45.9	85.8	91.6	88.5	-
Joint Learning	47.1	87.0	92.5	90.7	+1.4
+ LM	47.9	86.8	93.6	91.0	+1.9
+ Fine Tuning	**49.6**	**87.9**	**94.1**	**91.3**	+2.8

Liu et al. Recurrent Neural Network for Text Classification with Multi-Task Learning. IJCAI 2016.

各类研究都是有迹可循的

I类工作

- 总结：Mikolov et al.将LM从多任务学习拿出来做了word2vec来做迁移学习，此后word2vec➔ELMo➔GPT就是LM模型的scale之路，最终BERT将LM替换为MLM以适应双向编码

- 讨论：word2vec和BERT的出现是一条非常清晰且自然的发展脉络而不是机械降神，但回头看从2008年LM在多任务中被发现很有用，到2013年Mikolov et al.提出word2vec间隔了5年，再到2018年Peters et al.提出ELMo也间隔了5年，是什么阻碍了这一主线的发展？

- 今天，LM之路已基本见顶，之后是什么？

各类研究都是有迹可循的

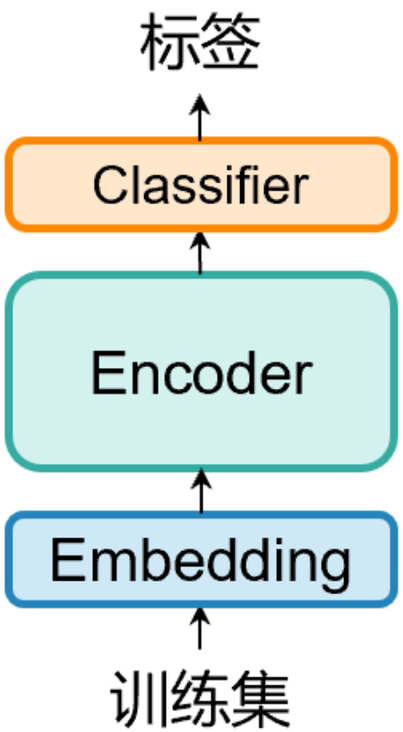
高目标

- 求其上者得其中，求其中者得其下。目标是Best paper才能做出I类工作，目标是I类工作只能做出II类工作，目标是灌水一般中不了。

脚踏实地

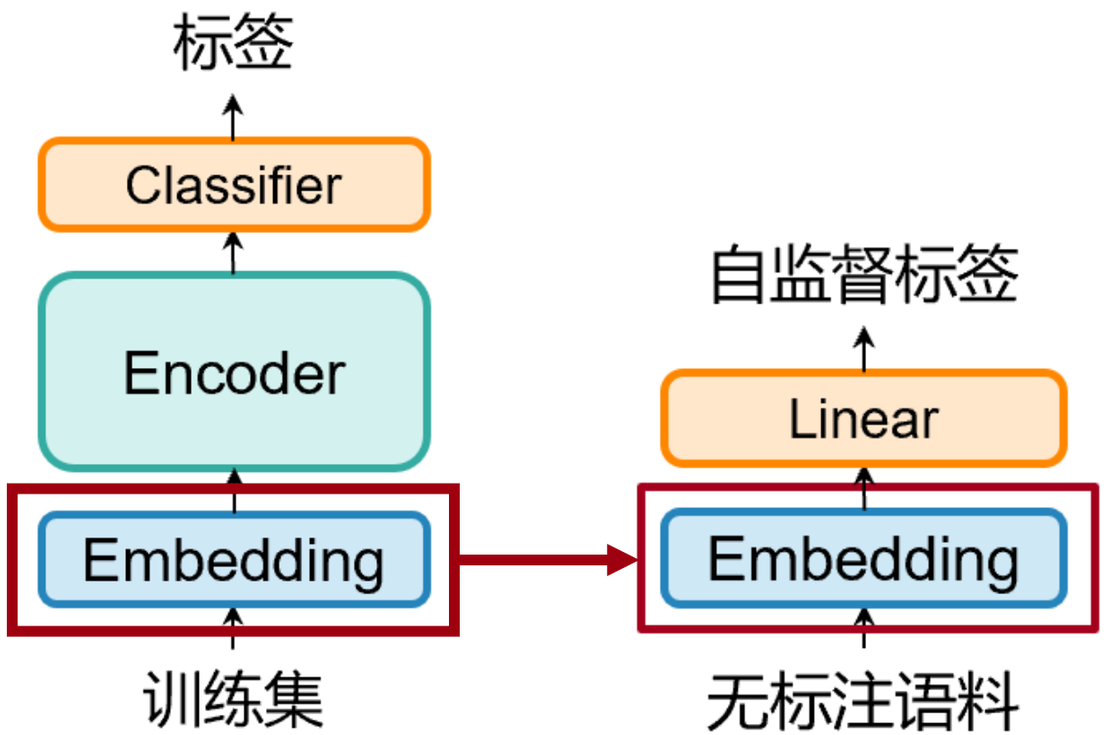
- 不要轻视照搬方法的好工作（比如别人分析word2vec的方法我拿来分析BERT），所有经得起考验的工作都是建立在扎实的写作和实验功底上的。

LM发展轨迹的另一角度



Before 2013
NLP from scratch

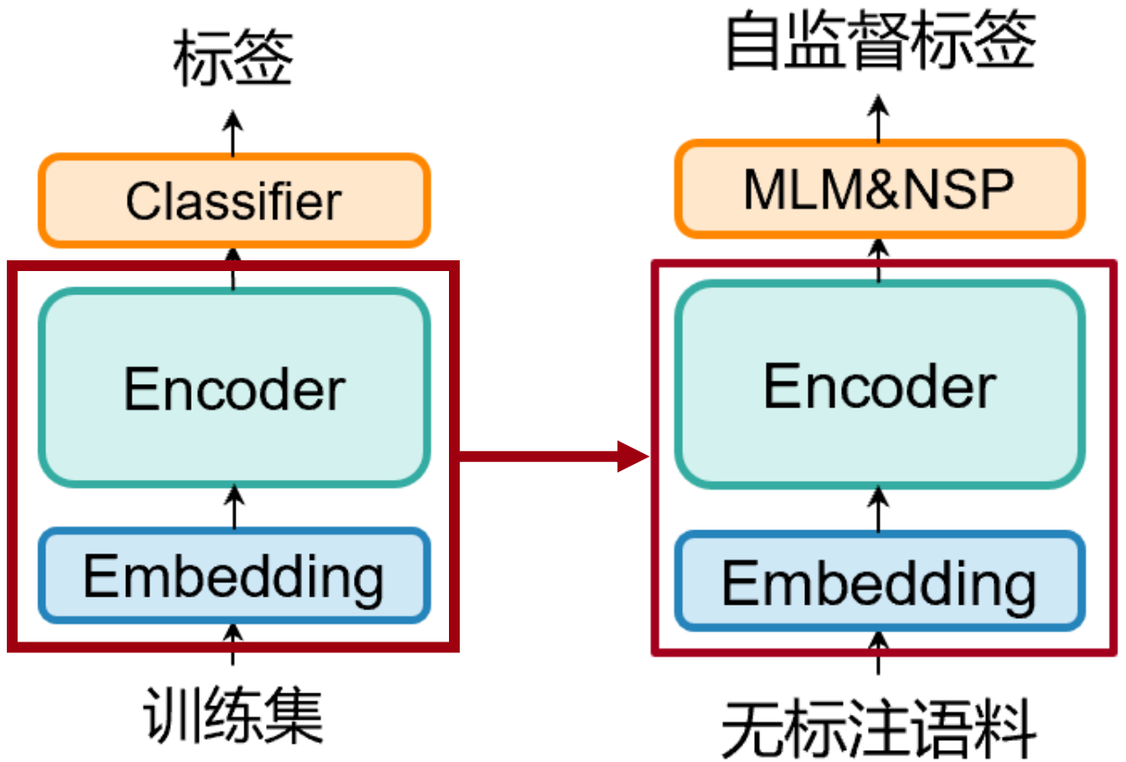
LM发展轨迹的另一角度



2013 - 2018

Pretrained word embeddings: word2vec, GloVe, …

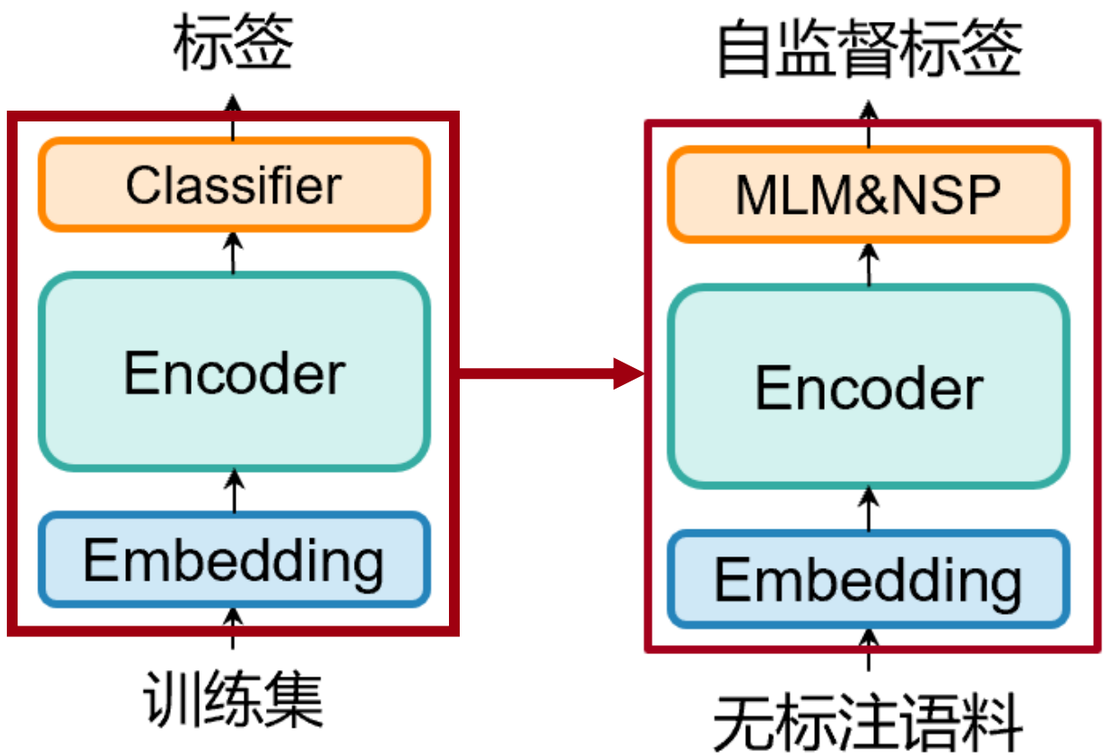
LM发展轨迹的另一角度



2018 - 2019

Pretrained LM (almost): ELMo, GPT, BERT, …

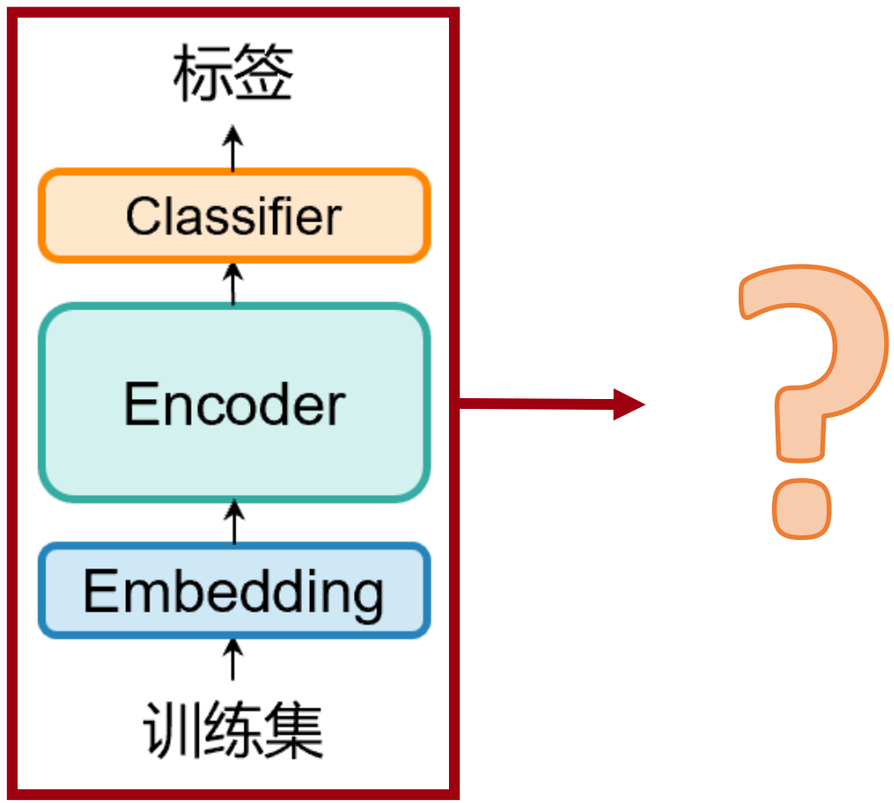
LM发展轨迹的另一角度



2019 - Now

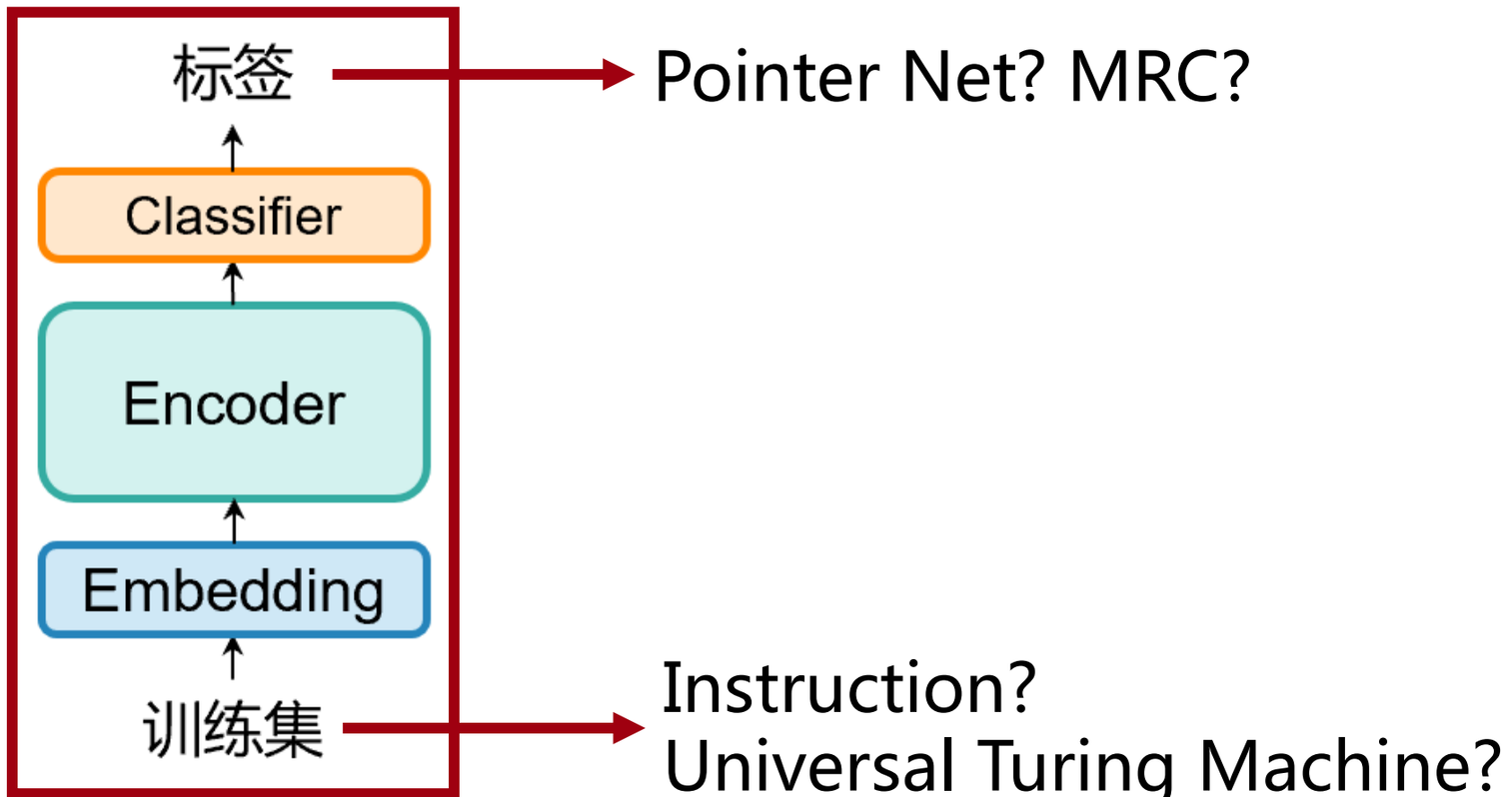
Pretrained LM: GPT-3, Prompt Learning, …

LM发展轨迹的另一角度



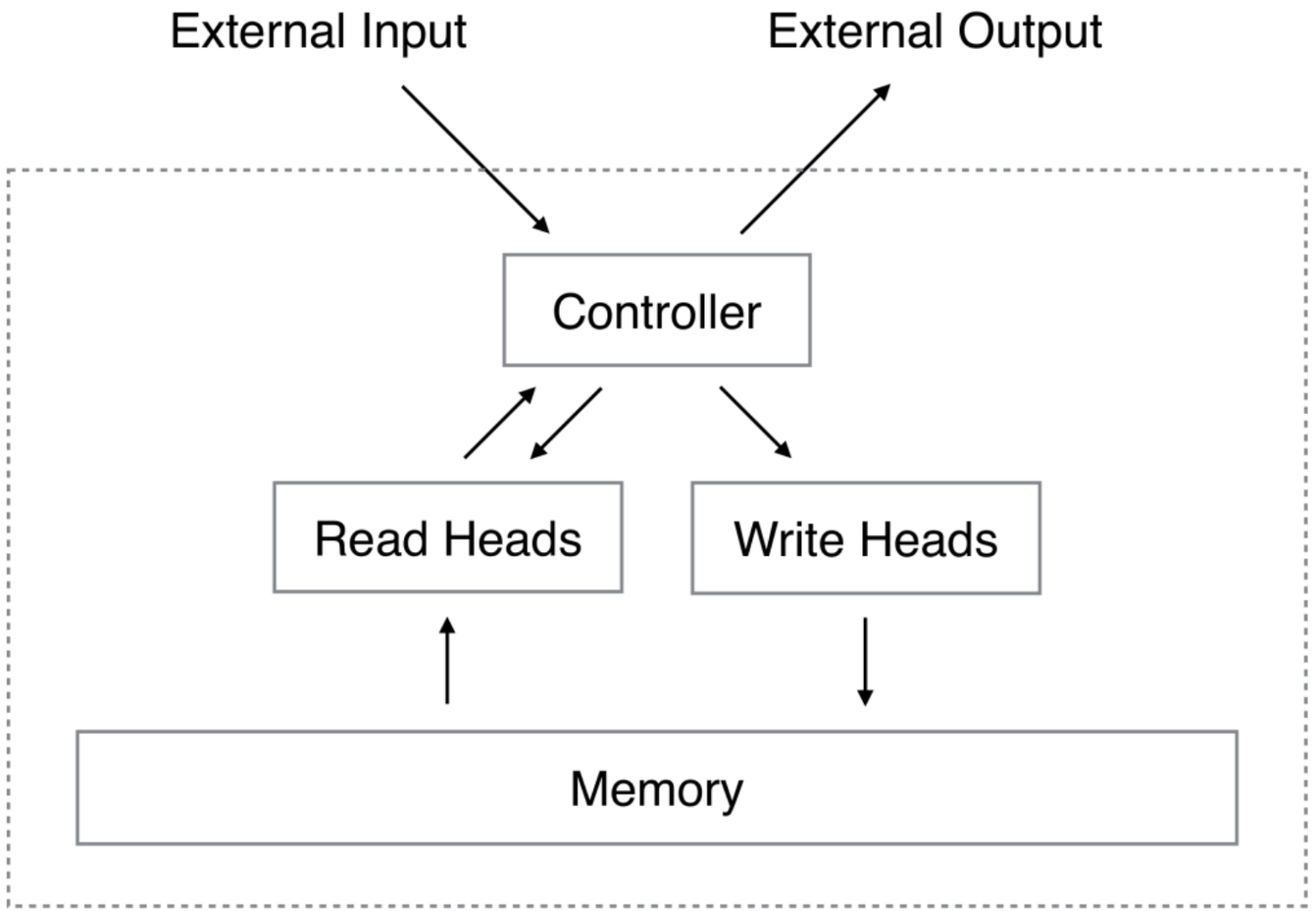
After 2022
What is the next?

LM发展轨迹的另一角度



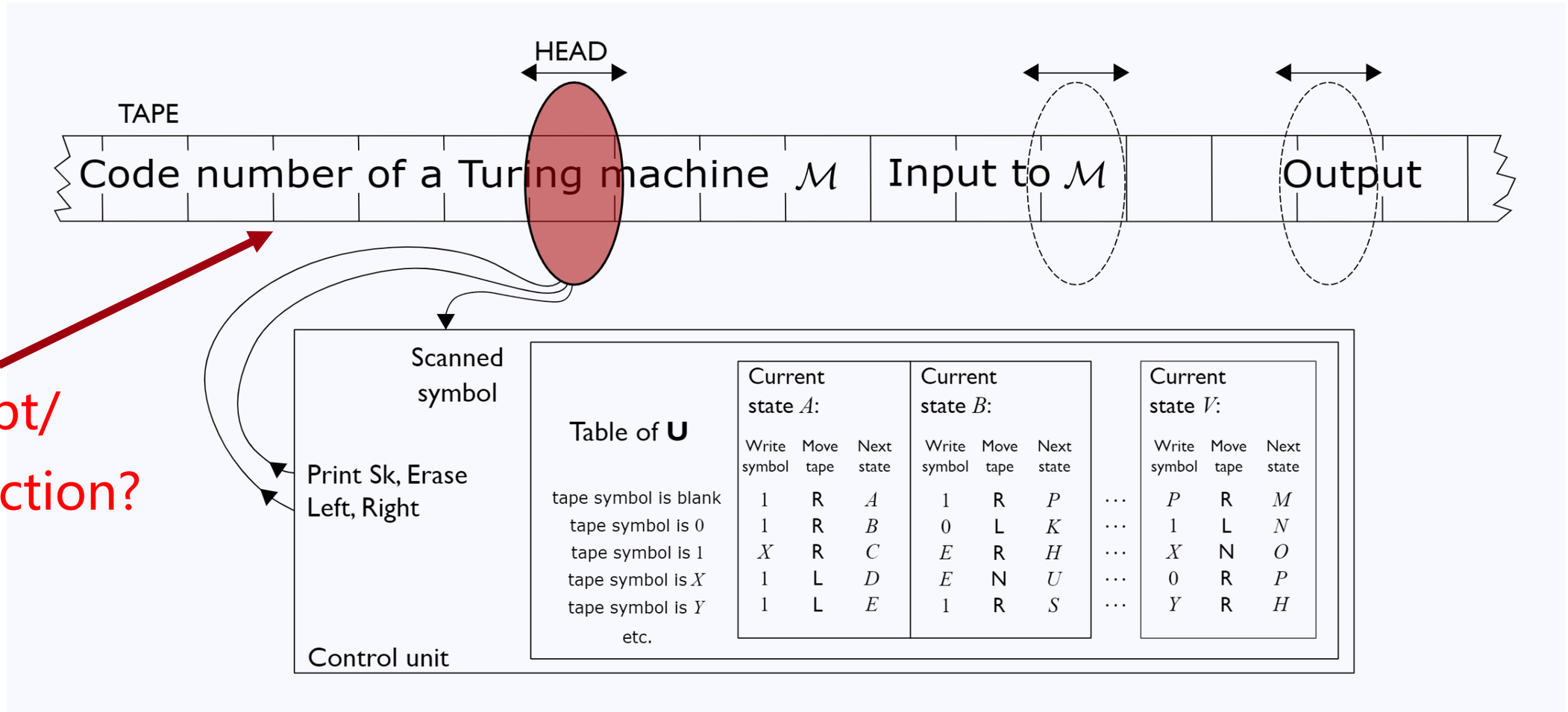
After 2022
What is the next?

LM发展轨迹的另一角度



Neural Turing Machine

LM发展轨迹的另一角度



Prompt/
Instruction?

Universal Turing Machine

LM发展轨迹的另一角度

Neural Turing Machine

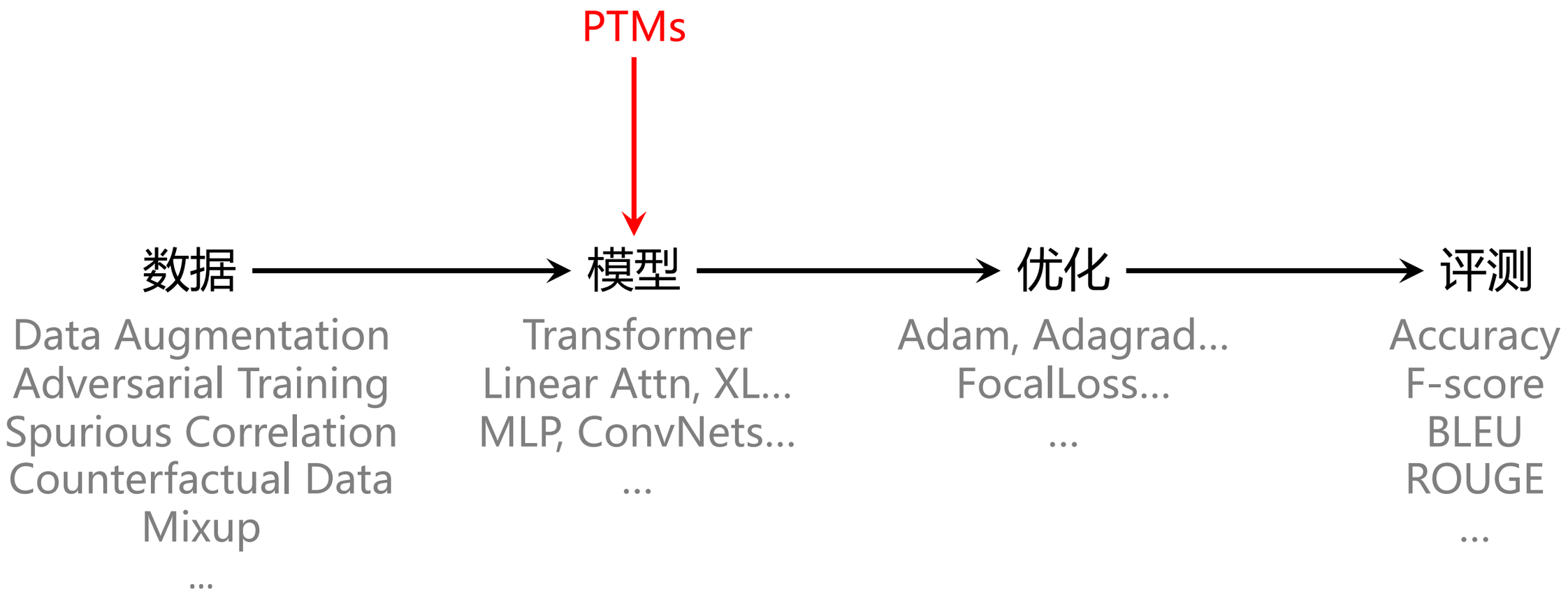
(Task-specific Model)

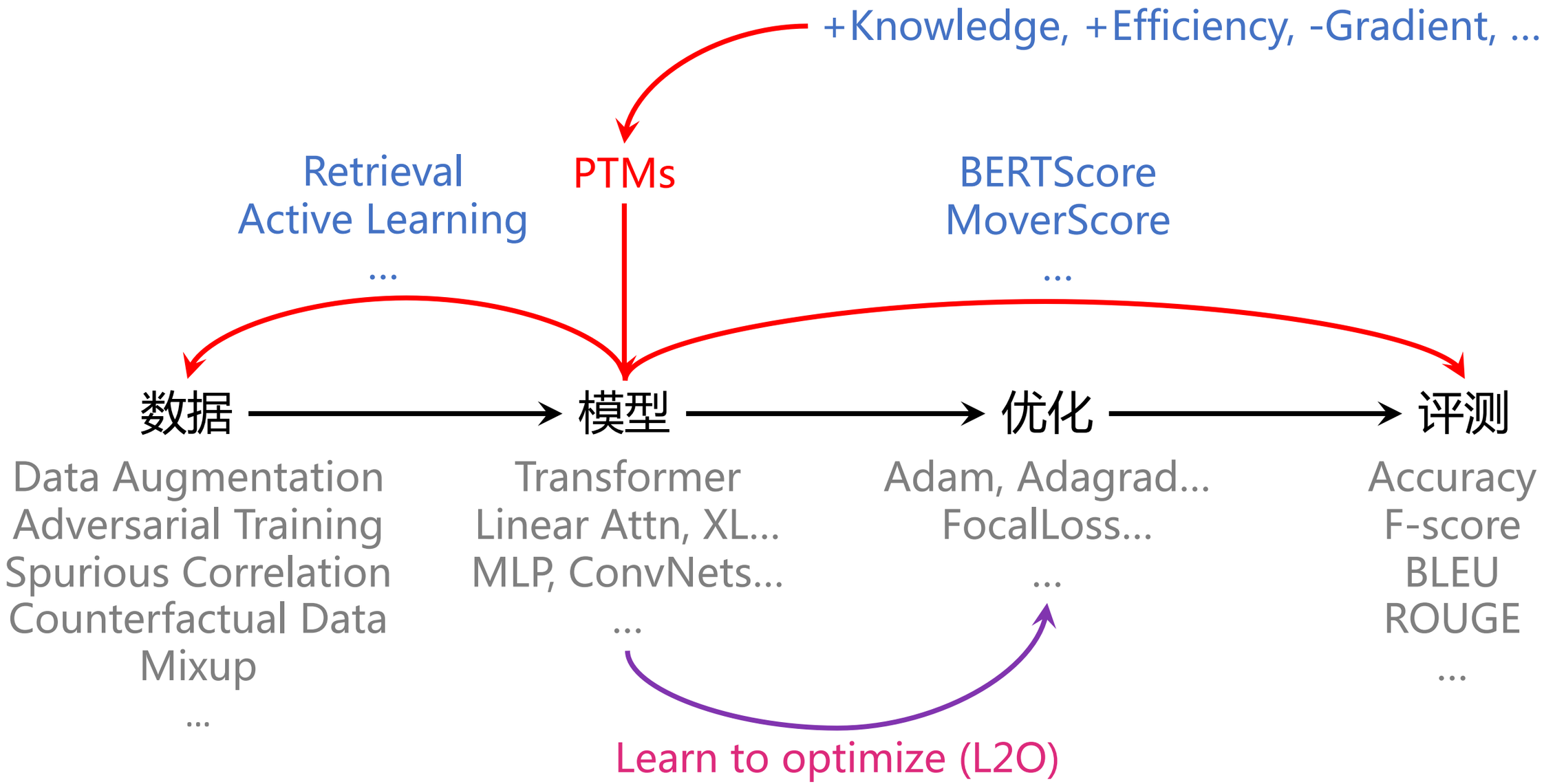
Instruction Tuning?

Neural **Universal** Turing Machine

(General Language Model)

A Perspective from the ML Pipeline


PTMs
数据
模型
优化
评测
Data Augmentation
Adversarial Training
Spurious Correlation
Counterfactual Data
Mixup
...
Transformer
Linear Attn, XL...
MLP, ConvNets...
...
Adam, Adagrad...
FocalLoss...
...
Accuracy
F-score
BLEU
ROUGE
...


+Knowledge, +Efficiency, -Gradient, ...
Retrieval
Active Learning
...
PTMs
BERTScore
MoverScore
...
数据
模型
优化
评测
Data Augmentation
Adversarial Training
Spurious Correlation
Counterfactual Data
Mixup
...
Transformer
Linear Attn, XL...
MLP, ConvNets...
...
Adam, Adagrad...
FocalLoss...
...
Accuracy
F-score
BLEU
ROUGE
...
Learn to optimize (L2O)

谢谢！