PLM

[语言预训练模型]

北 京 理 工 大 学

指导老师：张华平

成员：马越，戈润泽，
张至鑫，张懿元，康宇豪

CONTENT

语言模型与预训练语言模型

我 在 上海 迪斯尼
在 我 上海 迪斯尼
上海 我 在 迪斯尼
上海 在 我 迪斯尼

预训练：使用尽可能多的训练数据，从中提取出尽可能多的**共性**特征，从而能让模型对特定任务的学习**负担变轻**

BERT前

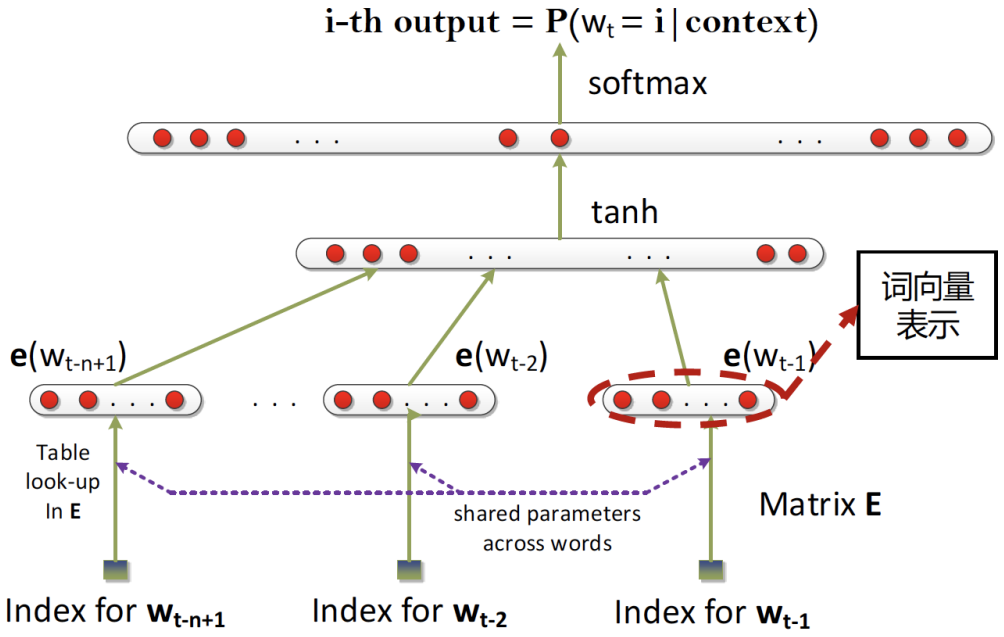
PART 1.1.1

静态预训练阶段

除了词自身表示之外的其他方法

更好的方式——分布式

- Neural Network Language Models (Bengio et al., JMLR 2003)
  - 根据前n-1个词预测第n个词（语言模型）
  - 模型结构为前向神经网络
  - 通过查表，获得词的向量表示
    - Word Embeddings
    - Word Vectors
  - 通过反向传播优化词向量表示



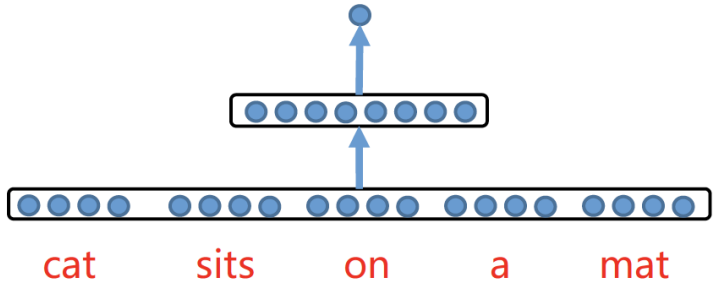
1.1.1 除了**词自身表示**之外的其他方法

- Semantic/syntactic Extraction using a Neural Network Architecture
  - Natural Language Processing (Almost) from Scratch (Collobert et al., JMLR 2011)
- “换词”的思想
  - 一个词和它的上下文构成正例
  - 随机替换掉该词构成负例

✚ cat sits on a mat

▬ cat sits Harbin a mat

- 优化目标
  - *score*(cat sits on a mat) > *score*(cat sits Harbin a mat)
  - *score*的计算方式



- 训练速度慢，在当年的硬件条件下需要训练1个月

1.1.1 再优一、的方式——Word2Vec

- https://code.google.com/archive/p/word2vec/ (Mikolov et al., ICLR 2013)
- CBOW (Continuous Bag-of-Word)
  - 周围词向量加和预测中间的词
- Skip-Gram
  - 中间词预测周围词
- 训练速度快
  - 可利用大规模数据
  - 弥补了模型能力的不足

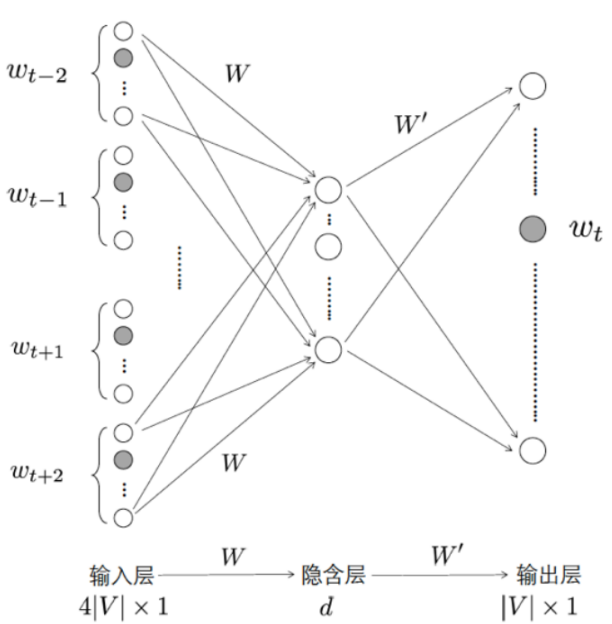


Figure 3: CBOW模型。

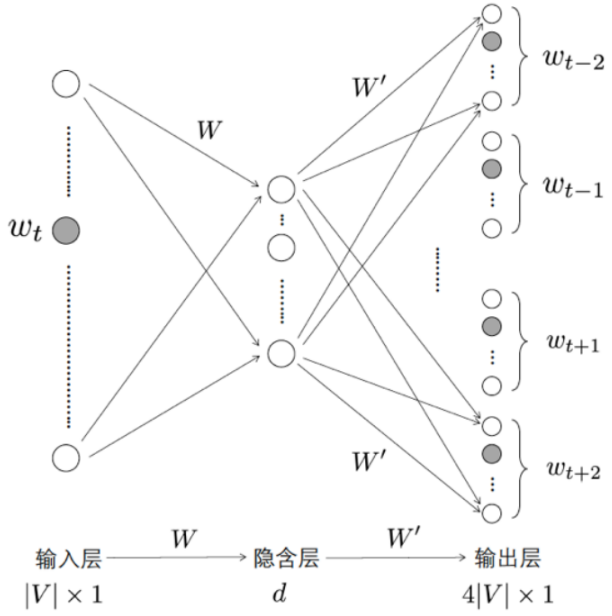


Figure 4: Skip-gram模型。

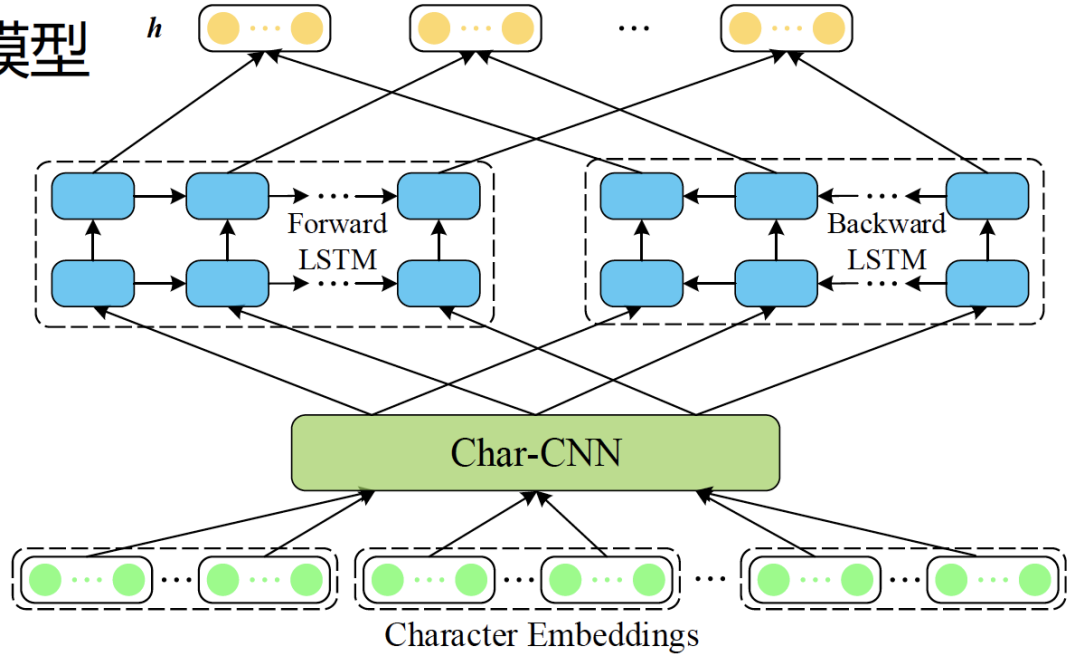
BERT前

PART 1.1.2

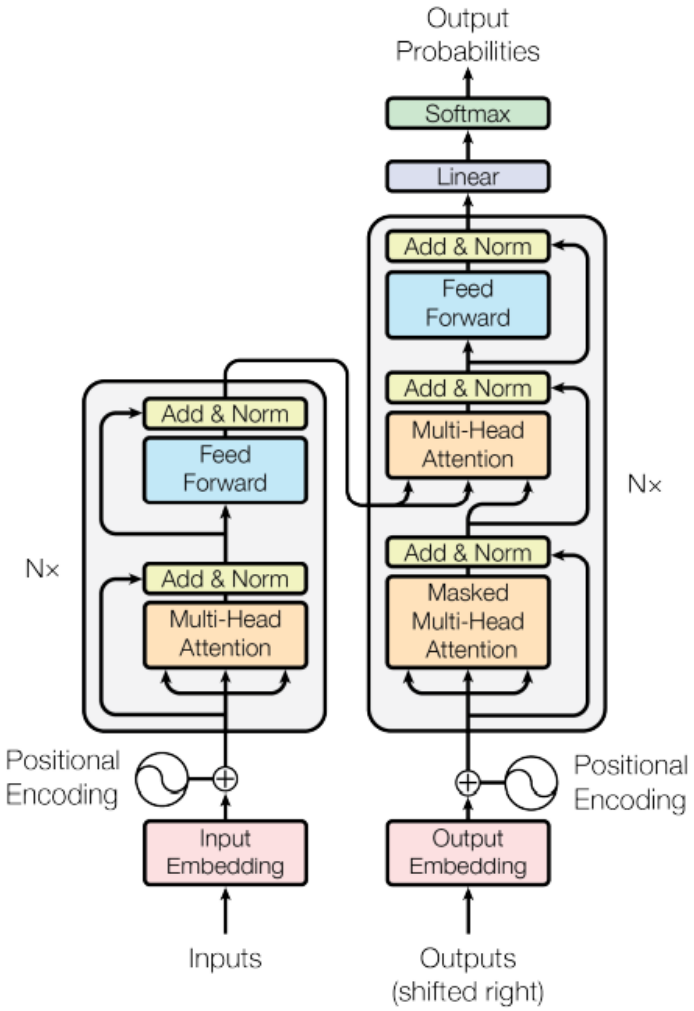
动态预训练阶段

1.1.2 上下文相关词向量

- Deep Contextualized Word Representations (Peters et al., NAACL 2018)
  - ELMo: Embeddings from Language Models
- 使用字符的CNN表示词
- 分别训练从左至右和从右至左的语言模型
- 使用语言模型的输出作为词向量特征
- 语言模型训练数据接近“无限”



1.1.3 Transformer

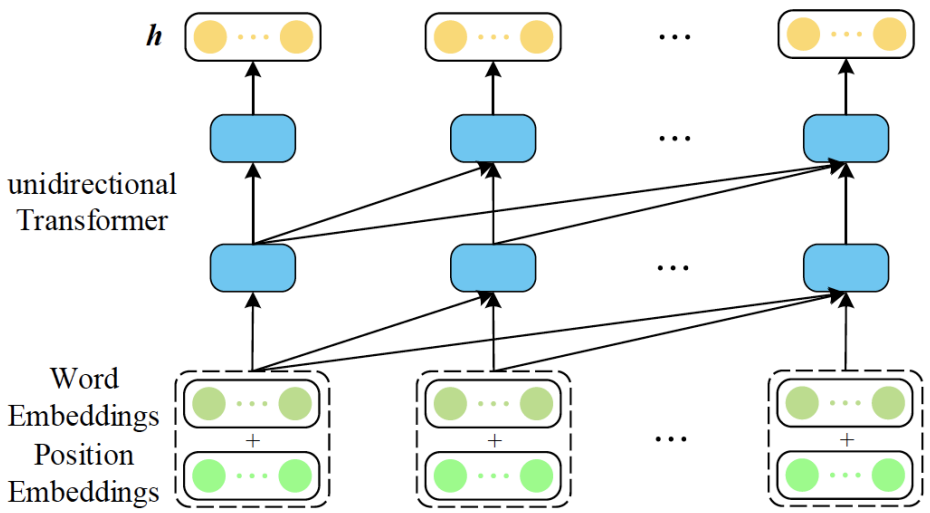


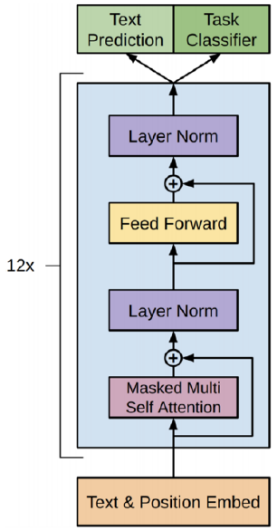
Transformer由self-attention和Feed Forward neural network组成

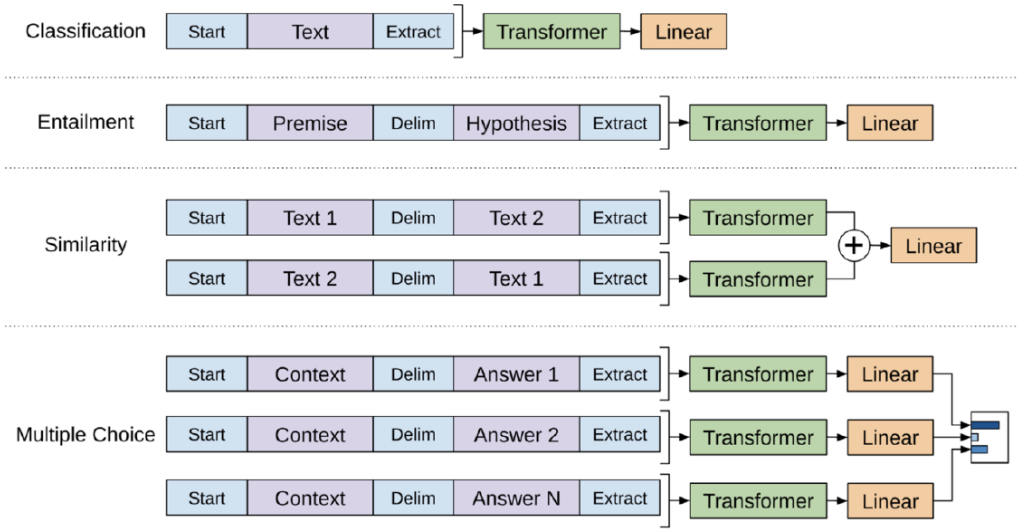
Multi—attention(多头注意力)

1.1.3 NLP中的预训练模型

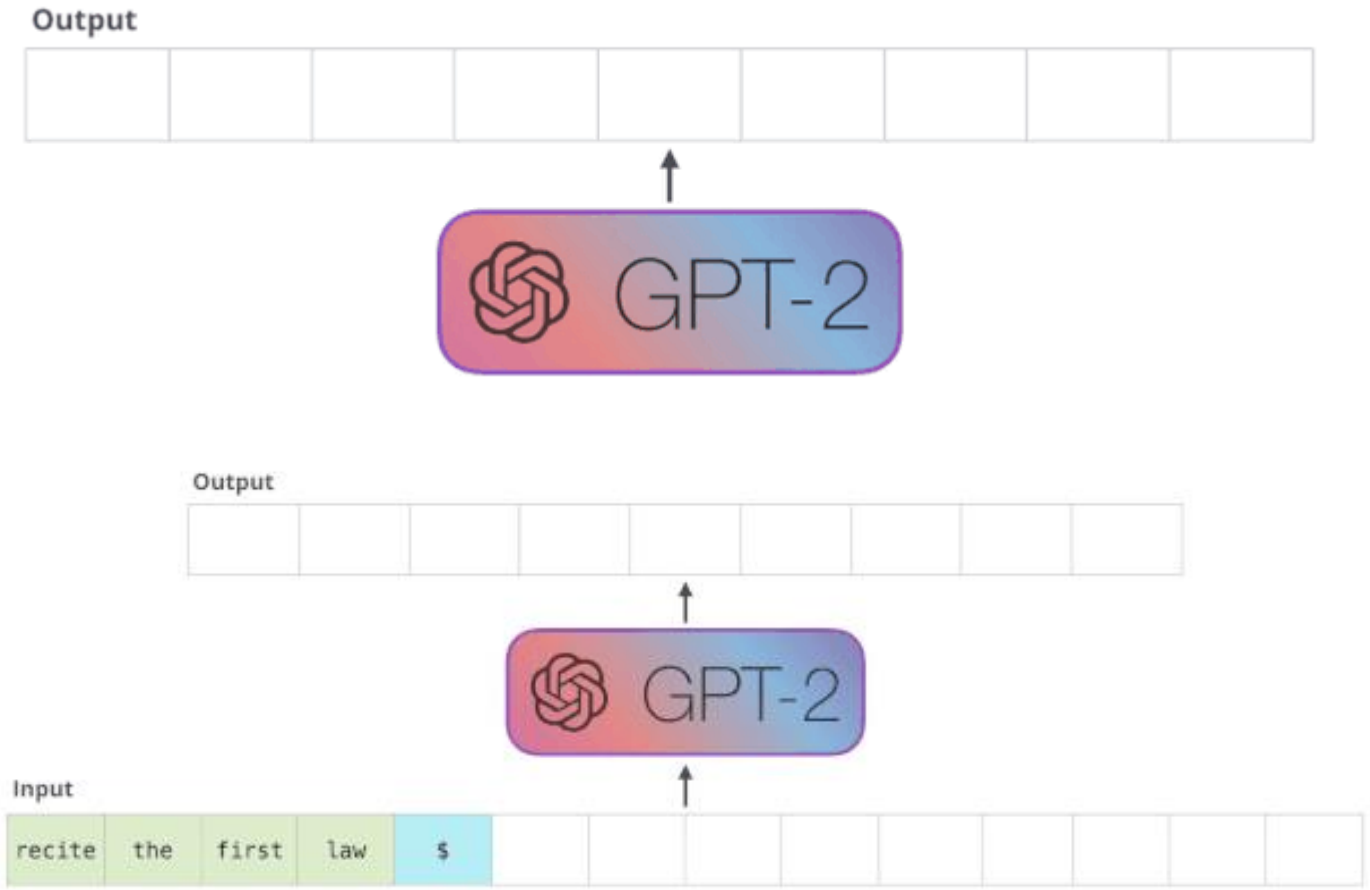
- Improving Language Understanding by Generative Pre-Training (Radford et al., 2018)
  - GPT: Generative Pretrained Transformer
  - 使用12层的Transformer作为Encoder预训练单向语言模型
  - 在目标任务上精调（Fine-tuning）模型







1.1.3 NLP中的预训练模型



In-context learning
Zero-shot
One-shot
Few-shot

1.1.3 GPT系列对比

	GPT-1	GPT-2	GPT-3
层数	12	48	96
Word-embedding size	768	1600	12888
参数	1.17 亿	15 亿	1,750 亿
预训练数据量	约 5GB	40GB	45TB

所有的有监督学习都是无监督语言模型的一个子集

BERT后部分

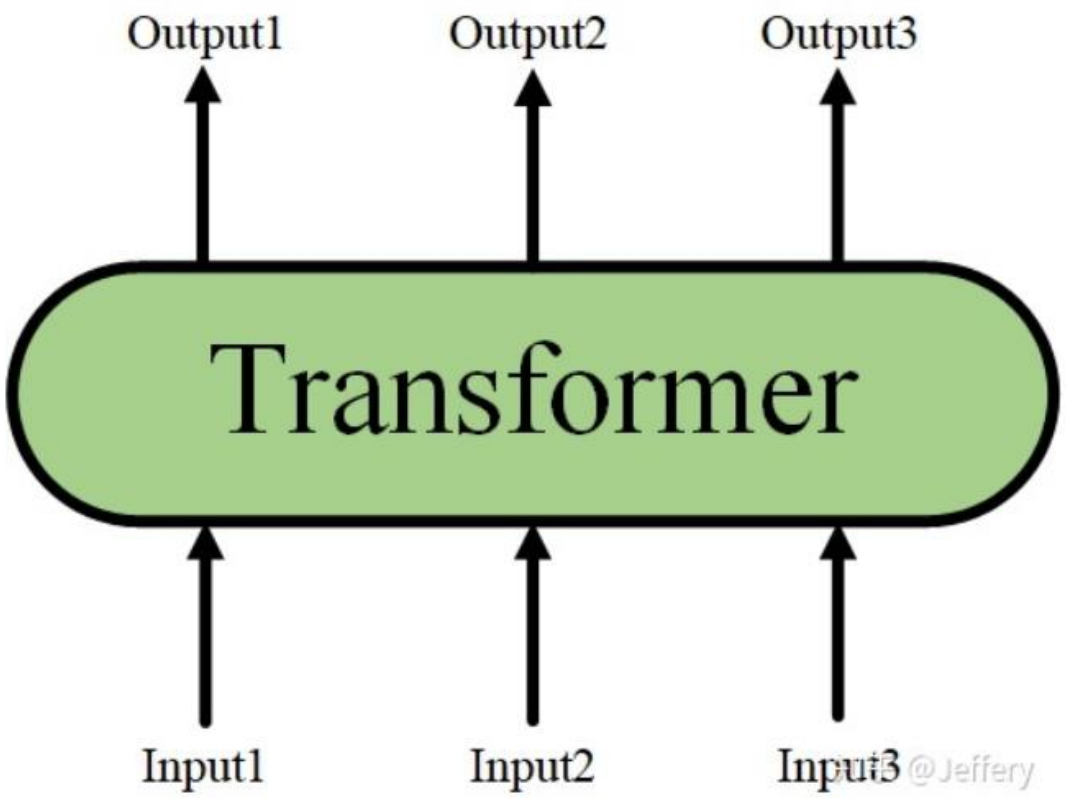
- 1.2.1 BERT
- 1.2.2 RoBERTa
- 1.2.3 XLNet

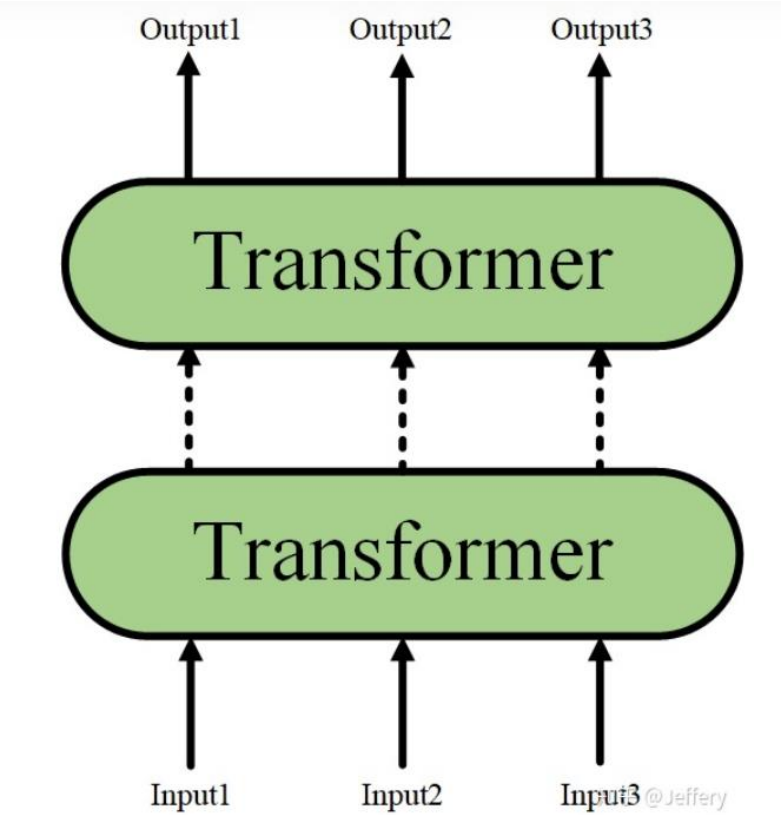
[PART 1.2.1]

BERT

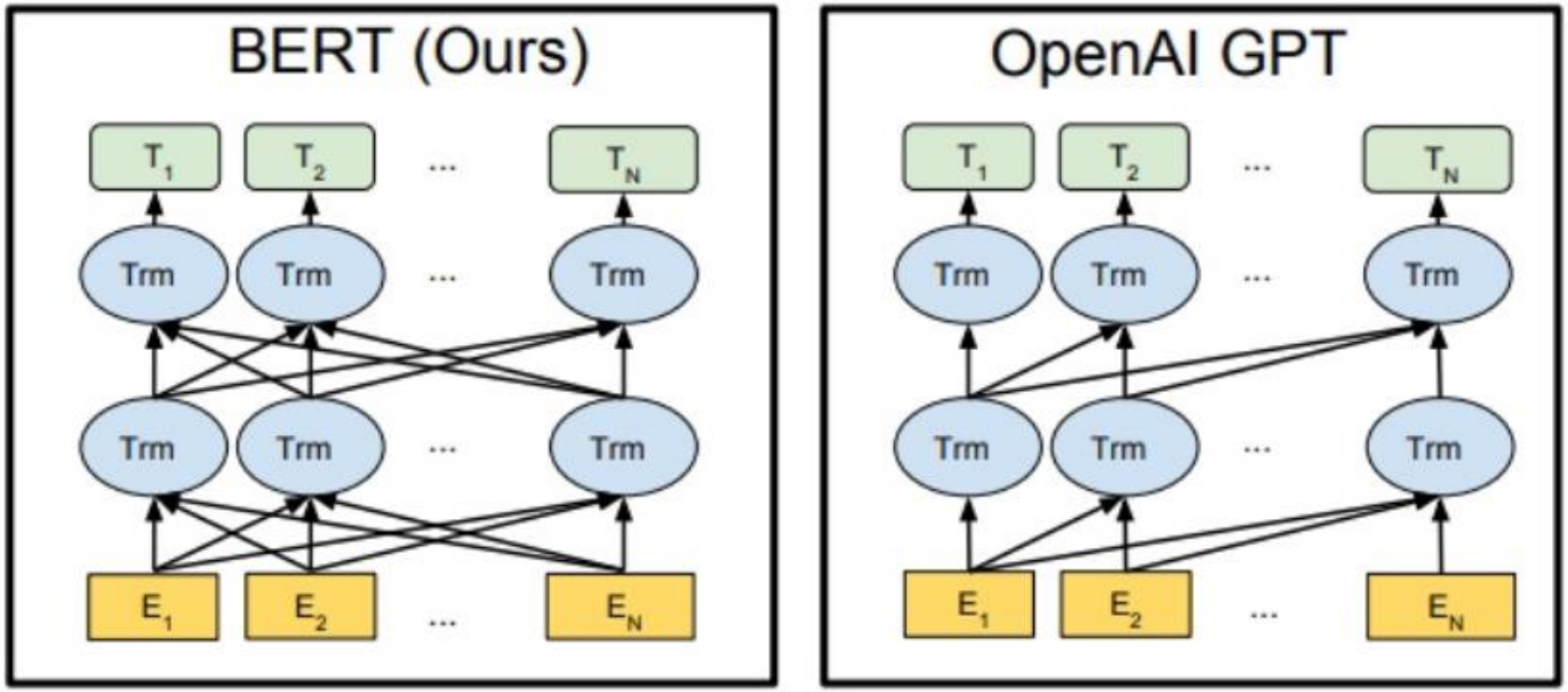
1.2.1 BERT

基本组成部分





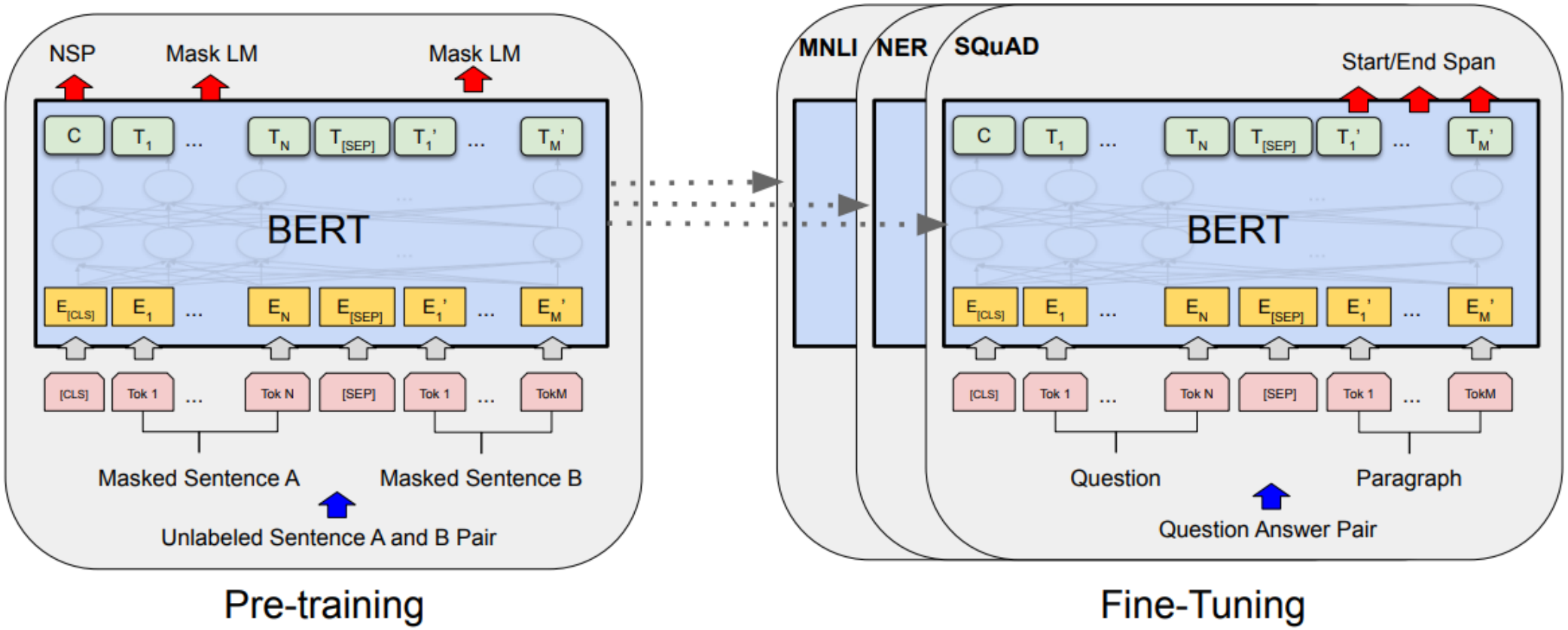
1.2.1 BERT



BERT和GPT一样均是采用的transformer的结构，与GPT相比，BERT是双向结构的，而GPT是单向的

1.2.1 BERT

预训练与微调



1.2.1 BERT

BERT的输入

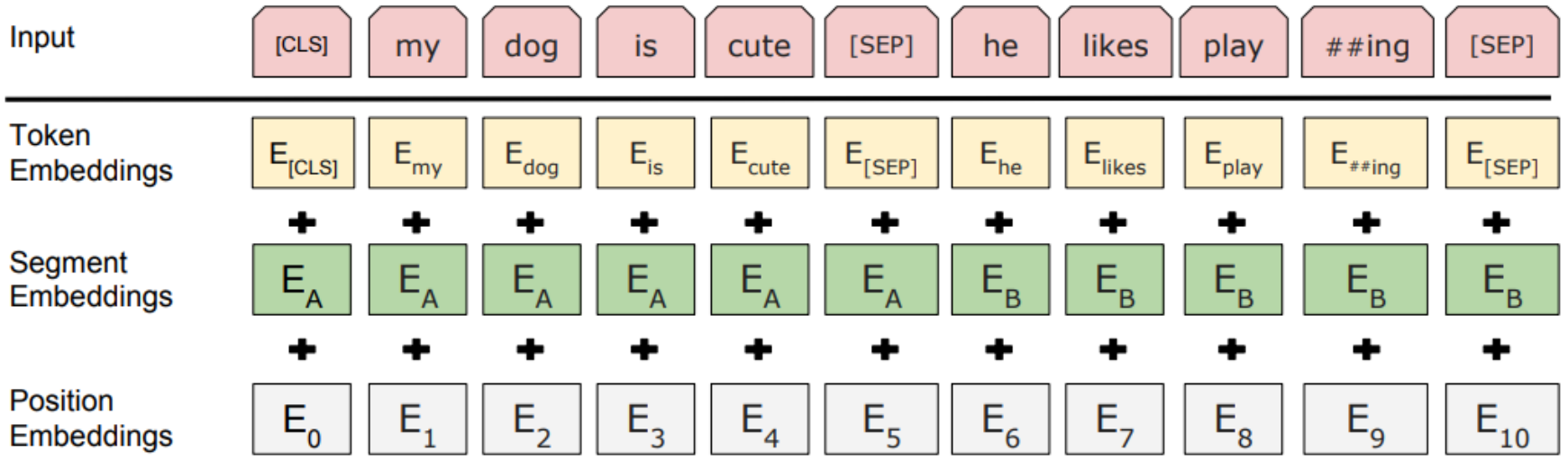


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

token embeddings:查询字向量表将文本中的每个字转换为一维向量

segmentation embeddings:描述句子对之间的关系

position embeddings:Transformer 模型不能记住时序，所以人为加入表示位置的向量

1.2.1 BERT的预训练阶段包括两个任务

1.Masked Language Model

随机mask每一个句子中15%的词，用其上下文来做预测，例如：

my dog is hairy → my dog is [MASK]

此处将hairy进行了mask处理，然后预测mask位置的词是什么，但是该方法有一个问题，因为是mask 15%的词，其数量已经很高了，这样就会导致某些词在fine-tuning阶段从未见过，为了解决这个问题，进行如下的处理：

- 80%的时间是采用[mask]，my dog is hairy → my dog is [MASK]
- 10%的时间是随机取一个词来代替mask的词，my dog is hairy -> my dog is apple
- 10%的时间保持不变，my dog is hairy -> my dog is hairy

1.2.1 BERT的预训练阶段包括两个任务

2. Next Sentence Prediction

选择一些句子对(A,B)，其中50%的B是A的下一条句子，剩余50%的B是从语料库中随机选择的，学习(A,B)的相关性，添加这样的预训练的目的是目前很多NLP的任务比如QA和NLI都需要理解两个句子之间的关系，从而能让预训练的模型更好的适应这样的任务。

[PART 1.2.2 RoBERTa]

1.2.2 RoBERTa

在BERT基础上做了几点调整

1）训练时间更长，batch size更大，训练数据更多；

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

2）移除了next predict loss

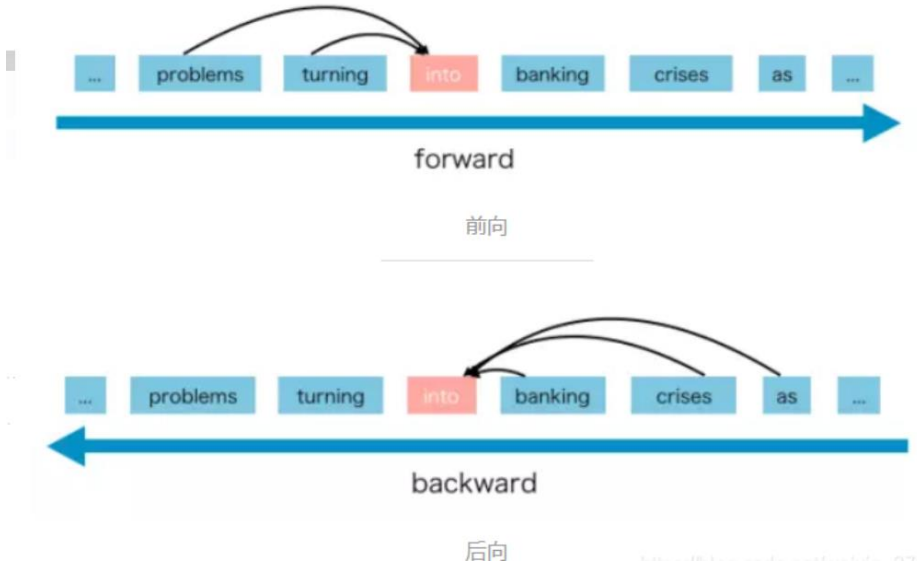
3）将BERT中的static masking调整为dynamic masking

[PART 1.2.3]

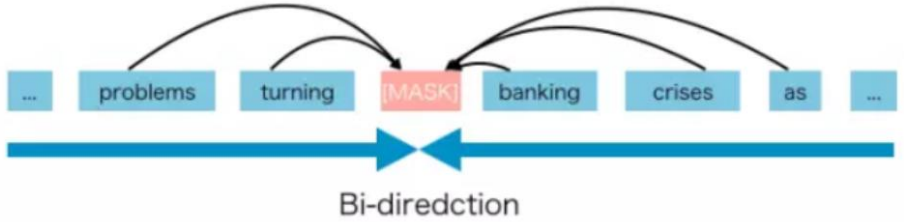
XLNet

1.2.3 XLNet

XLNet 论文中将当前预训练模型分为了两类 AR (Auto Regression，自回归) 和 AE (Auto Encoder，自编码器)。



前面介绍的GPT 就是一种 AR 方法，不断地使用当前得到的信息预测下一个输出 (自回归)。而 BERT 是一种 AE 方法，将输入句子的某些单词 mask 掉，然后再通过 BERT 还原数据。

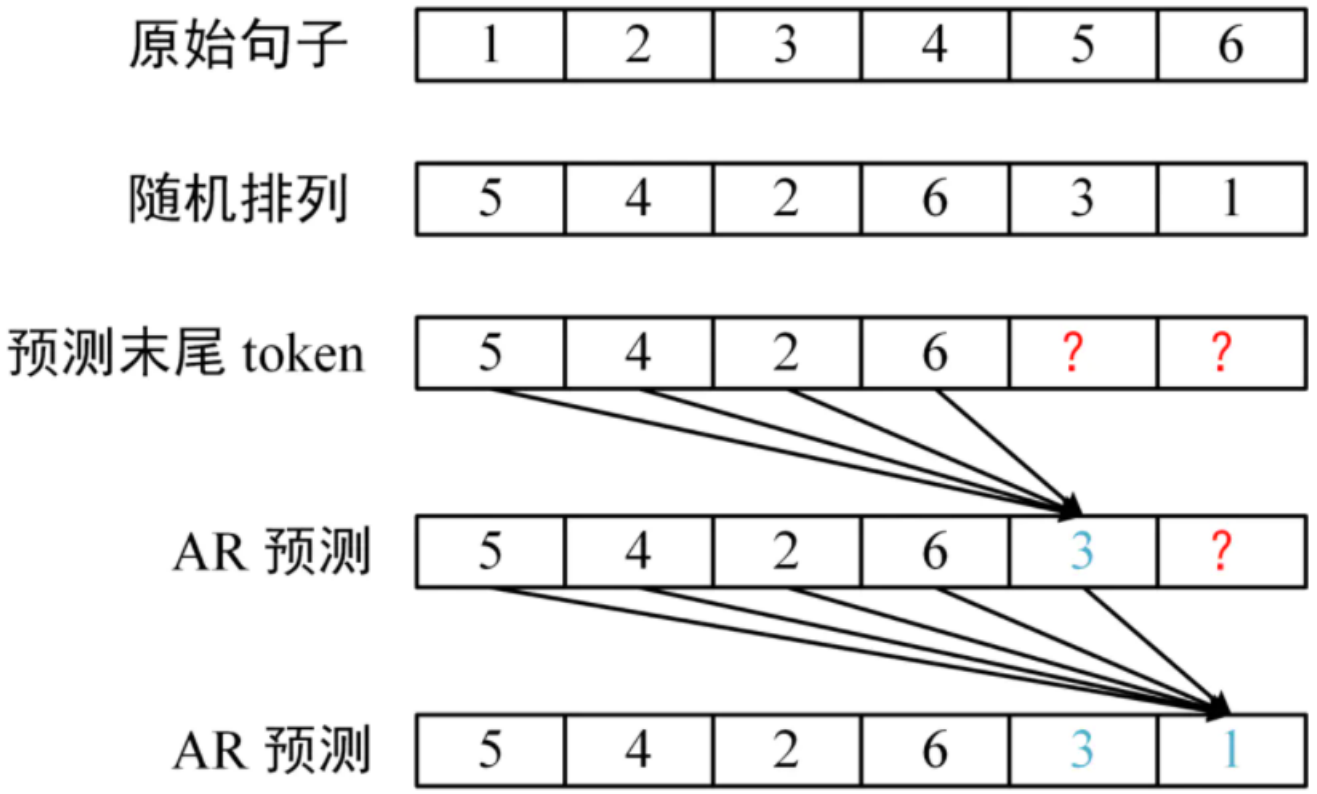


AR 的方法可以更好地学习 token 之间的依赖关系，而 AE 的方法可以更好地利用深层的双向信息。因此 XLNet 希望将 AR 和 AE 两种方法的优点结合起来，XLNet 使用了 **Permutation Language Model (PLM)** 实现这一目的。

1.2.3 XLNet

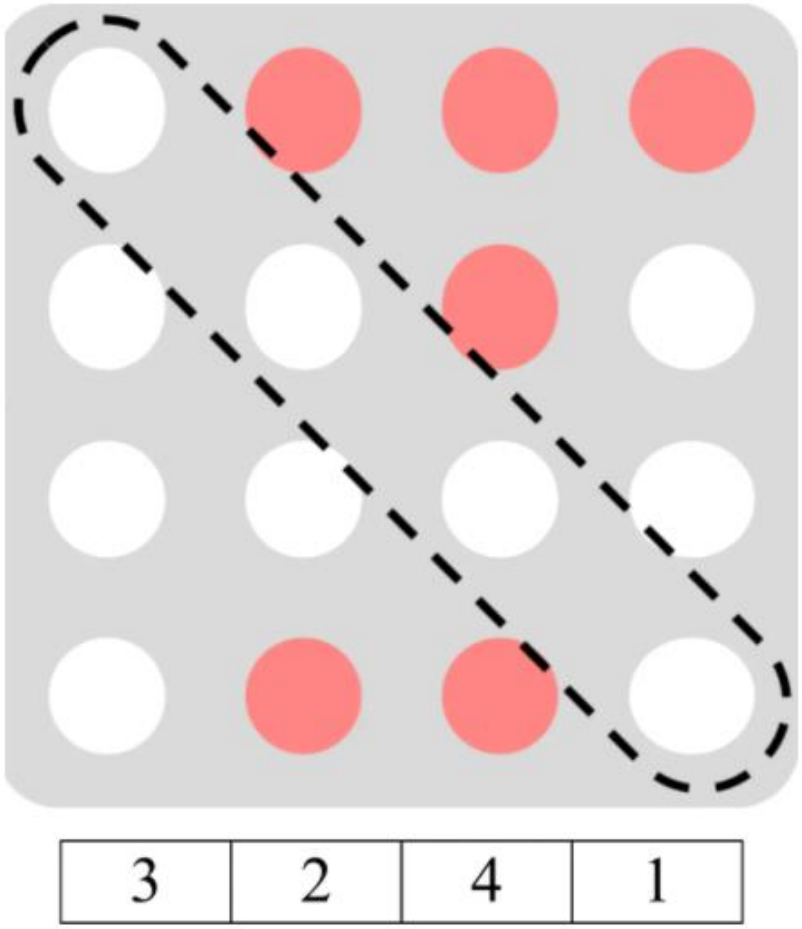
PLM

将句子中的 token 随机排列，然后采用 AR 的方式预测末尾的几个 token。这样一来，在预测 token 的时候就可以同时利用该 token 双向的信息，并且能学到 token 间的依赖。



1.2.3 XLNet

PLM



XLNet 中通过 Attention Mask 实现 PLM，而无需真正修改句子 token 的顺序。例如原来的句子是 [1,2,3,4]，如果随机生成的序列时 [3,2,4,1]，则输入到 XLNet 的句子仍然是 [1,2,3,4]，但是掩码需要修改成上图所示。

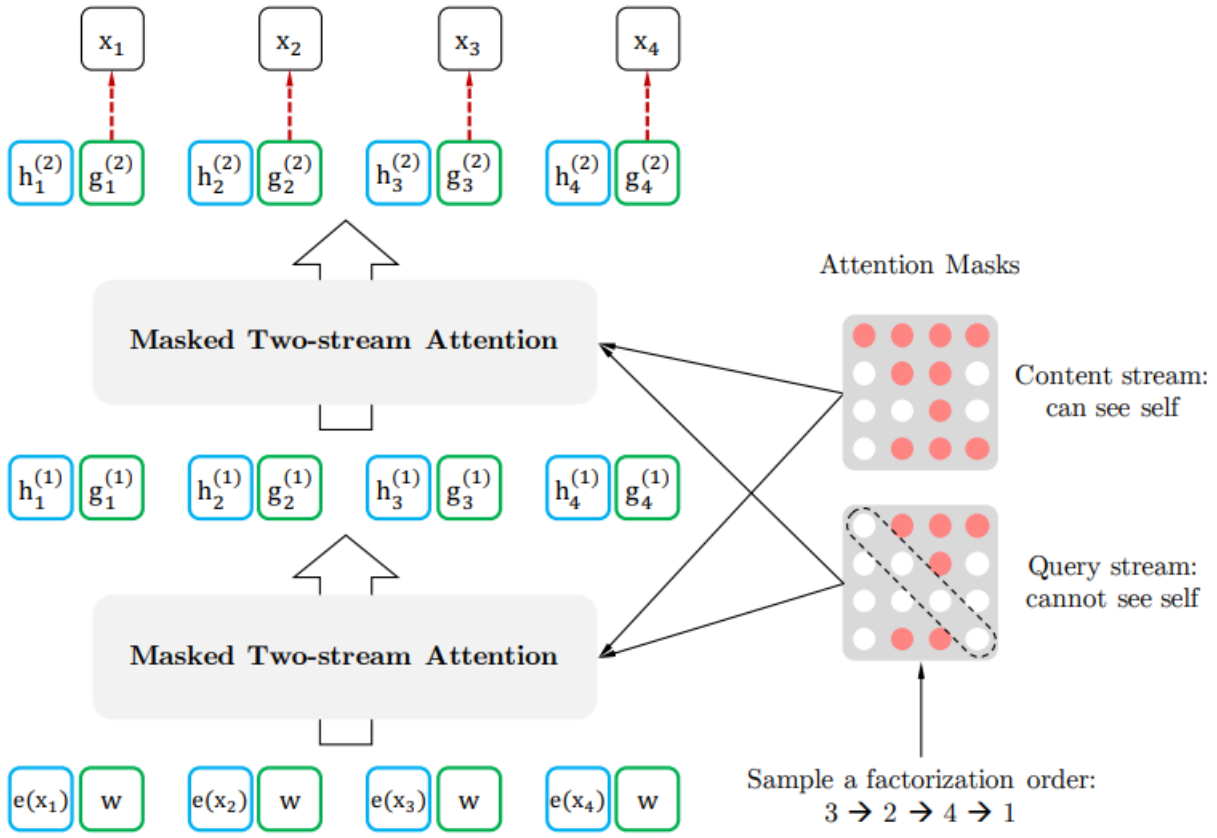
1.2.3 XLNet

PLM

XLNet 打乱了句子的顺序，这时在预测的时候 token 的位置信息会非常重要，同时在预测的时候也必须将 token 的内容信息遮掩起来。XLNet 采用了两个 Stream 实现这一目的：

- Query Stream，对于每一个 token，其对应的 Query Stream 只包含了该 token 的位置信息，这里的位置信息是 token 在原始句子的位置信息，不是重新排列的位置信息。

- Content Stream，对于每一个 token，其对应的 Content Stream 包含了该 token 的内容信息。



1.2.3 XLNet

XLNet 优化技巧

XLNet 使用了 Transformer-XL 中的 **Segment Recurrence Mechanism** (段循环) 和 **Relative Positional Encoding** (相对位置编码) 进行优化。

Segment Recurrence Mechanism段循环的机制会将上一段文本输出的信息保存下来，用于当前文本的计算，使模型可以拥有更广阔的上下文信息。

Relative Positional Encoding在引入上一段信息后，可能会有两个 token 拥有相同的位置信息，例如上一段的第一个单词和当前段的第一个单词位置信息都是一样的。因此 Transformer-XL 采用了 Relative Positional Encoding (相对位置编码)，不使用固定的位置，而是采用单词之间的相对位置进行编码。

BERT + 显式知识引入

1.2.4

显 式 知 识 引 入

Explicit knowledge introduction

自由文本 + 结构化知识 + 知识分析识别工具 = PLM编码更多知识

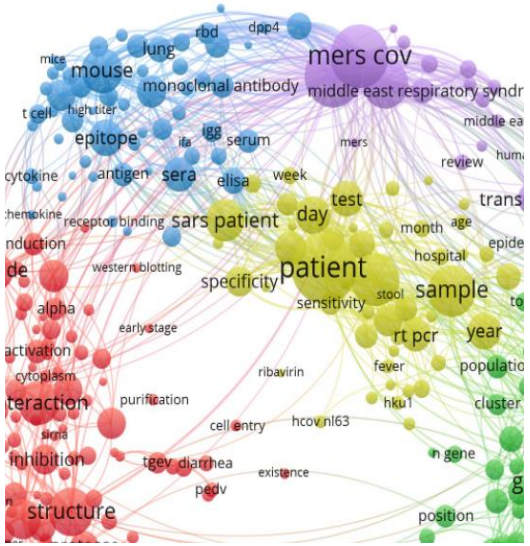
使用语言[mask]型来[mask]测

下一个词的 pro [mask] ## lity

使用语言[mask] [mask]来[mask] [mask]

下一个词的[mask] [mask] [mask]

词 级 别 的 MASK策略



结构化知识的 引 入

百度ERNIE 1.0

清华ERNIE

1.2.4

百度ERNIE（2019.4）

Ernie: Enhanced representation through knowledge integration

短语级策略和实体级策略的mask

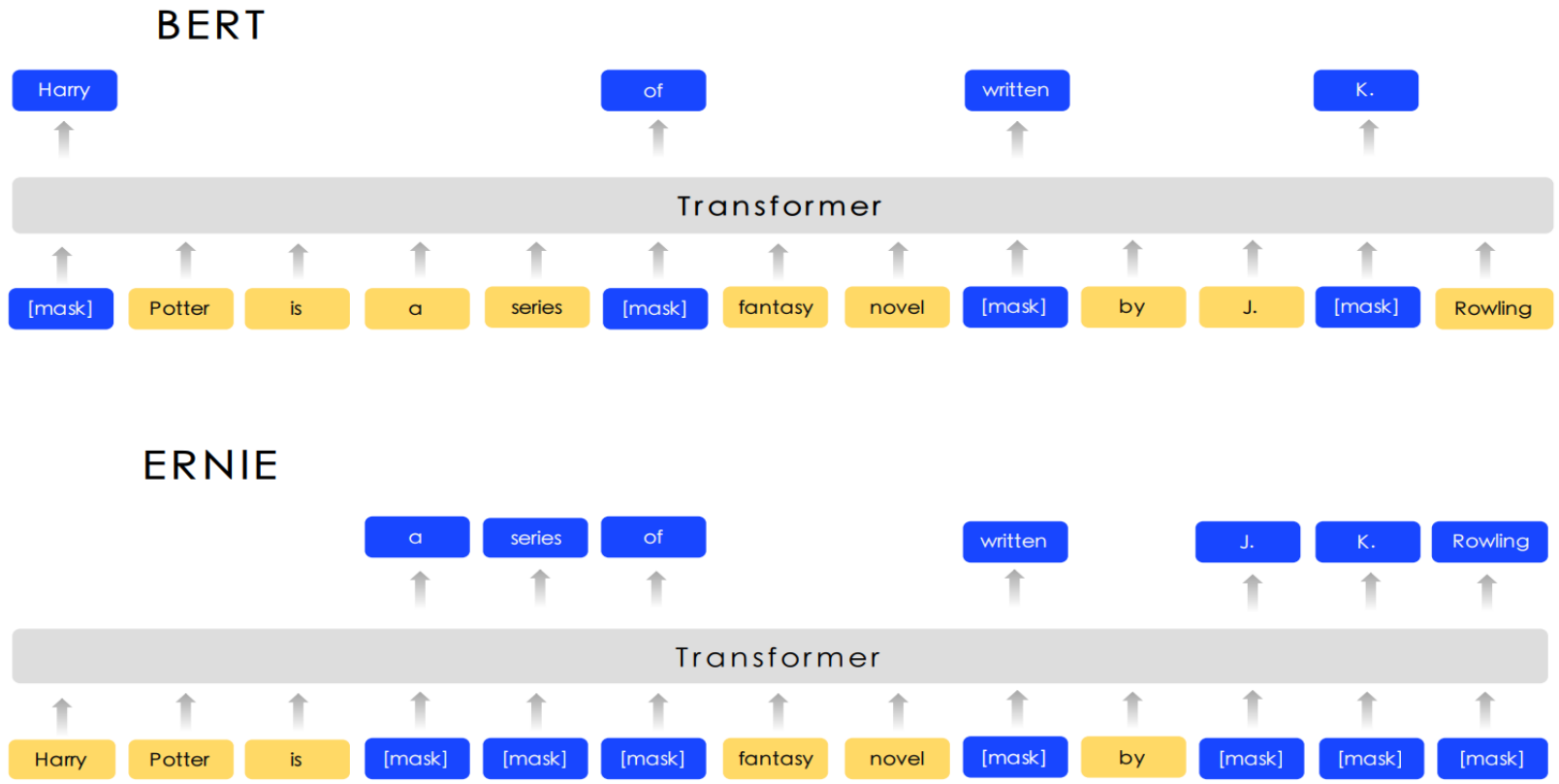


Figure 1: The different masking strategy between BERT and ERNIE

图来自：论文ERNIE: Enhanced Representation through Knowledge Integration

1.2.5

讯飞+哈工大 BERT-wwm（2019.7）

Pre-Training with WholeWord Masking for Chinese BERT

任意中文词语策略的mask

	Chinese
Original Sentence	使用语言模型来预测下一个词的概率。
+ CWS	语言 模型 来 预测 下 一个 词 的 概率 。
+ BERT Tokenizer	语 言 模 型 来 预 测 下 一 个 词 的 概 率 。
Original Masking	语 言 [M] 型 来 [M] 测 下 一 个 词 的 概 率 。
+ WWM	语 言 [M] [M] 来 [M] [M] 下 一 个 词 的 概 率 。
++ N-gram Masking	[M] [M] [M] [M] 来 [M] [M] 下 一 个 词 的 概 率 。
+++ Mac Masking	语 法 建 模 来 预 见 下 一 个 词 的 几 率 。

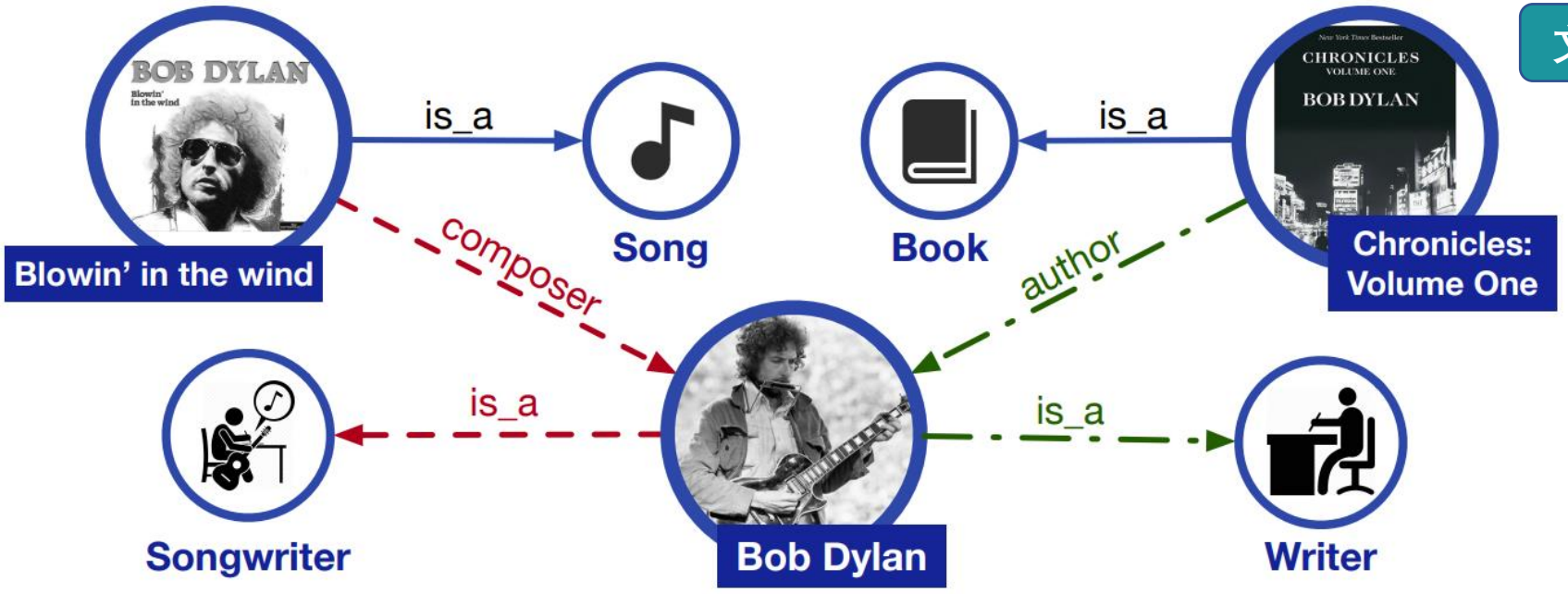
图来自：论文Pre-Training with Whole Word Masking for Chinese BERT

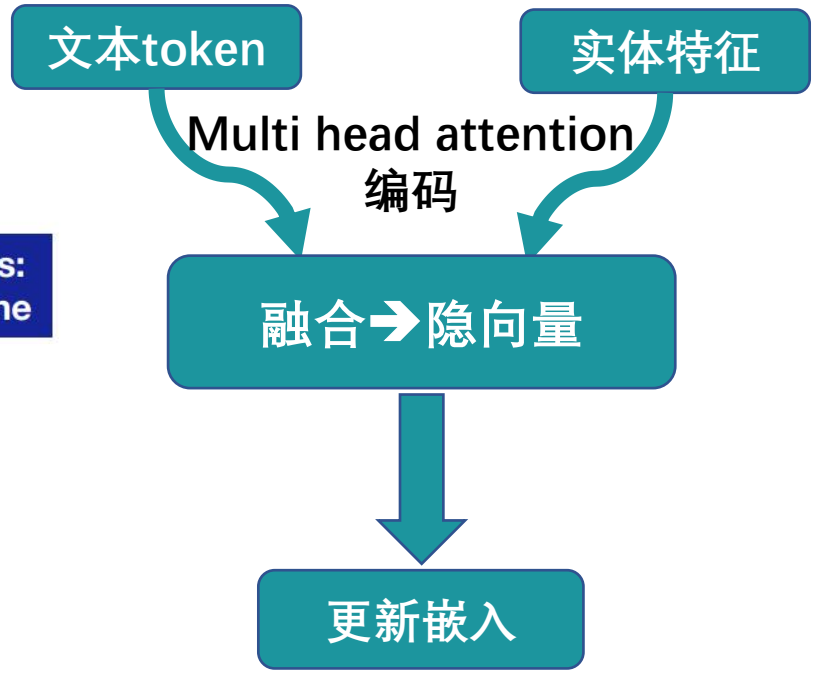
1.2.6

清华ERNIE（2019.5）

ERNIE: Enhanced language representation with informative entities

外接神经网络模型融合文本以外的信息





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

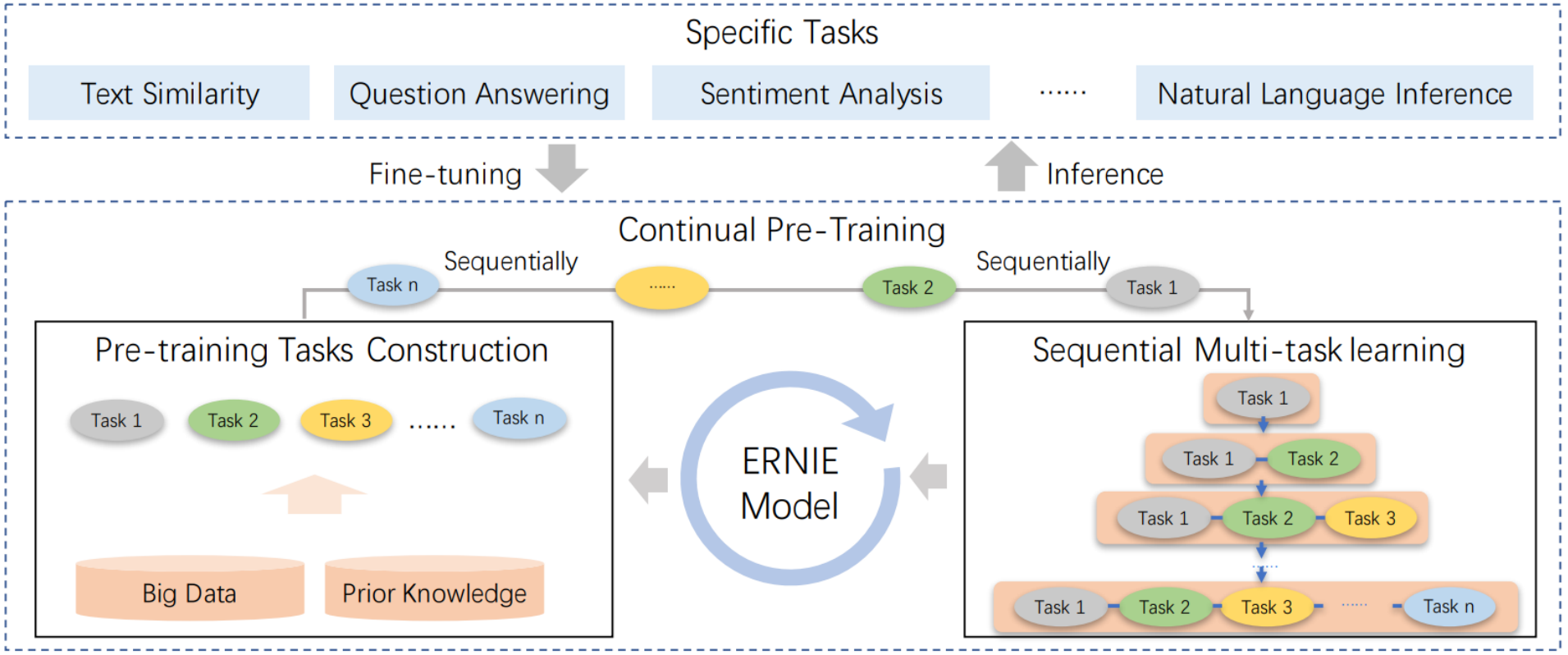
图来自：论文ERNIE: Enhanced Language Representation with Informative Entities

1.2.7

百度ERNIE 2.0（2019.7）

A continual pre-training framework for language understanding

持续多任务学习学习预训练模型



共现信息，句法语义

任务

增量构建

增量训练分布式表示

词汇、句法编码的跨任务能力up

图来自：论文ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding

BERT后改进方面总结

其他预训练目标

百度 ERNIE 1.0
百度 ERNIE 2.0
脸书 Span
BERT
……

融入知识图谱

清华ERNIE 1.0
KnowBERT
K-BERT
……

更加精细的调参

脸书 RoBERTa

跨语言，跨模态

M-BERT
脸书 XLM

VideoBERT
……

模型压缩与加速

模型蒸馏：
DistilBERT
华为 TinyBERT

词表的优化：
ALBERT
……

端到端(编解码)

脸书 BART
谷歌 T5
……

PART 02

前沿技术

预训练的前沿进展：悟道--- 1.75万亿---智源、人大高瓴

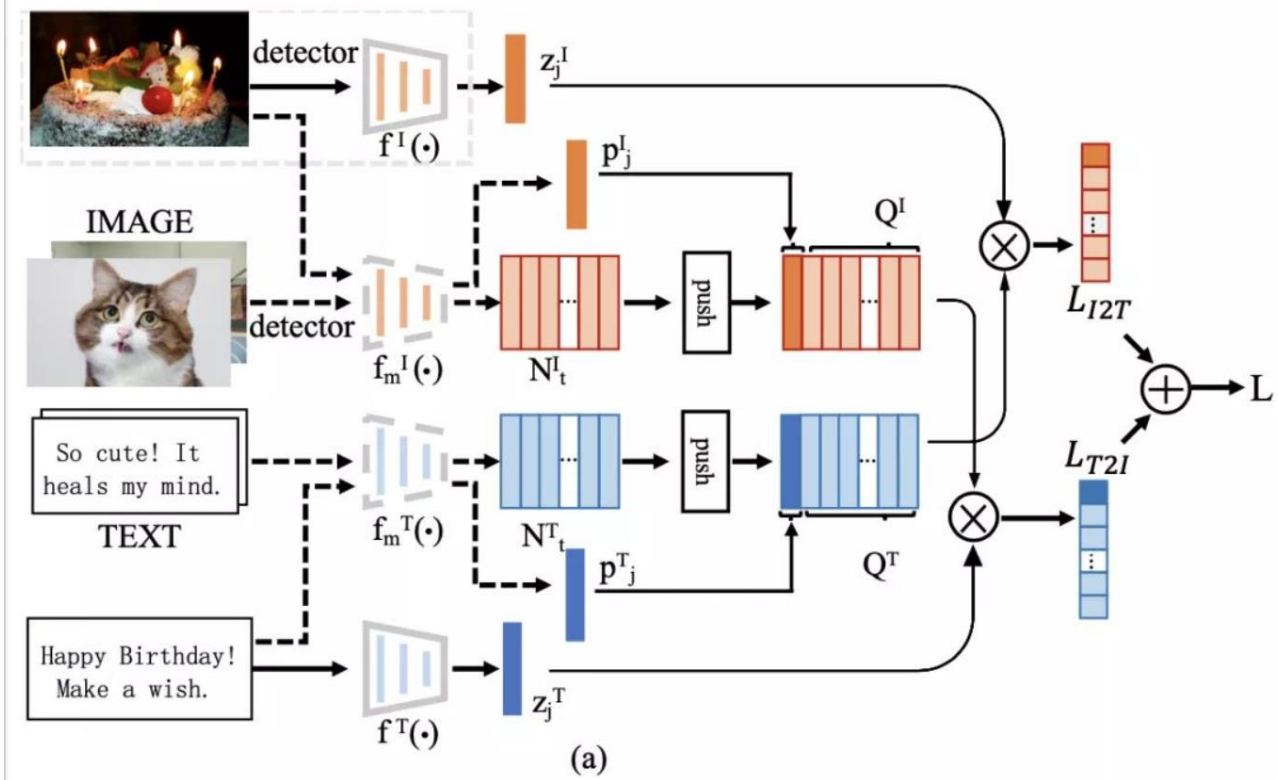


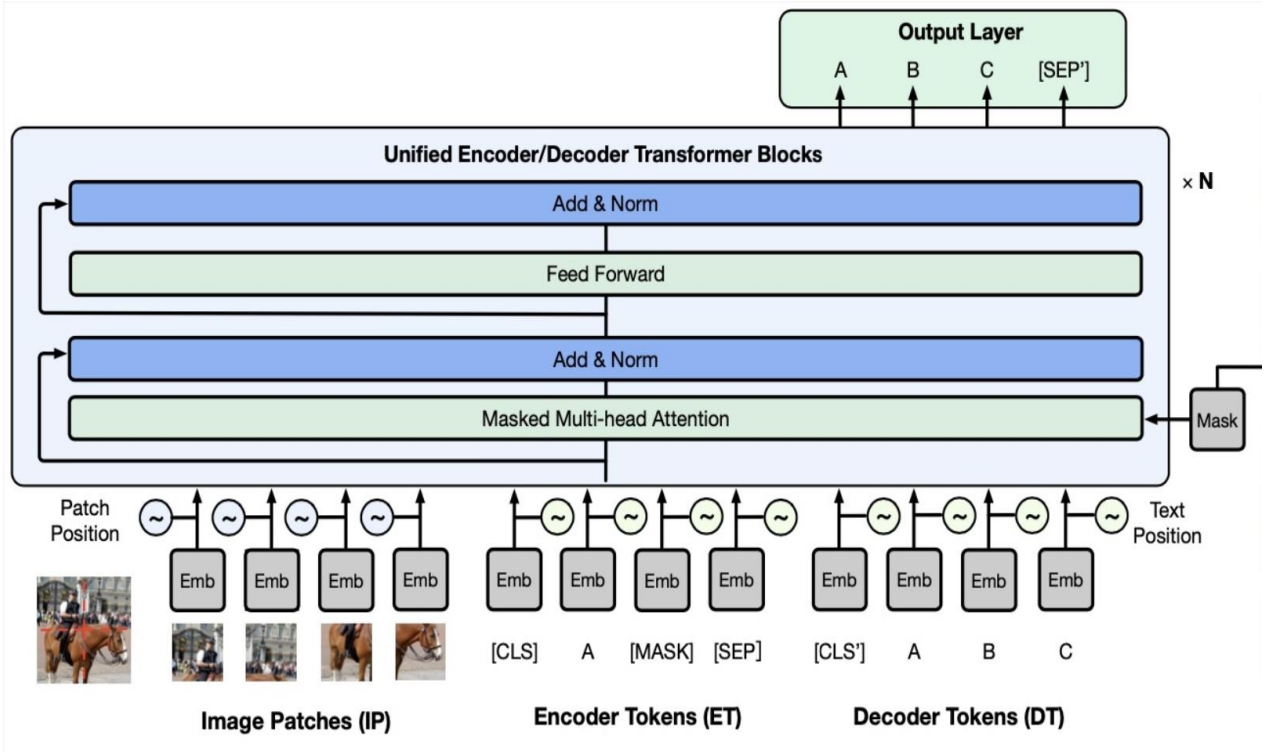
Table 4. User study results for the text-image retrieval downstream task. Three human annotators are involved in such user study.

Tasks	Image-to-Text Retrieval			
Metrics	NDCG@5	NDCG@10	NDCG@20	MAP
CLIP [27]	32.9	38.8	53.0	30.3
BriVL	**37.5**	42.8	55.5	**38.3**
BriVL+UNITER	37.0	**43.5**	**56.3**	37.6
Tasks	Text-to-Image Retrieval			
Metrics	NDCG@5	NDCG@10	NDCG@20	MAP
CLIP [27]	28.0	32.3	43.7	16.7
BriVL	46.9	51.5	61.6	47.2
BriVL+UNITER	**49.9**	**55.0**	**65.1**	**52.5**

语义相关性上的强假设，单塔结构才能在词汇与局部图像特征之间进行模态交互。但遗憾的是，在实际应用场景中，上述的强假设往往并不成立，比如视觉与语言之间通常只有较抽象的关联。例如，对于蛋糕的照片，配的文字可以是“生日快乐，许个愿吧”，也可以引申到“哎，我的减肥大计又泡汤了”。
文澜的研发者们进行了一系列的实验和探索，实验结果表明，在开放获取（例如互联网上的公开数据）的图文数据集上，简单的双塔结构要优于单塔结构。

M6--- 10万亿---阿里达摩院

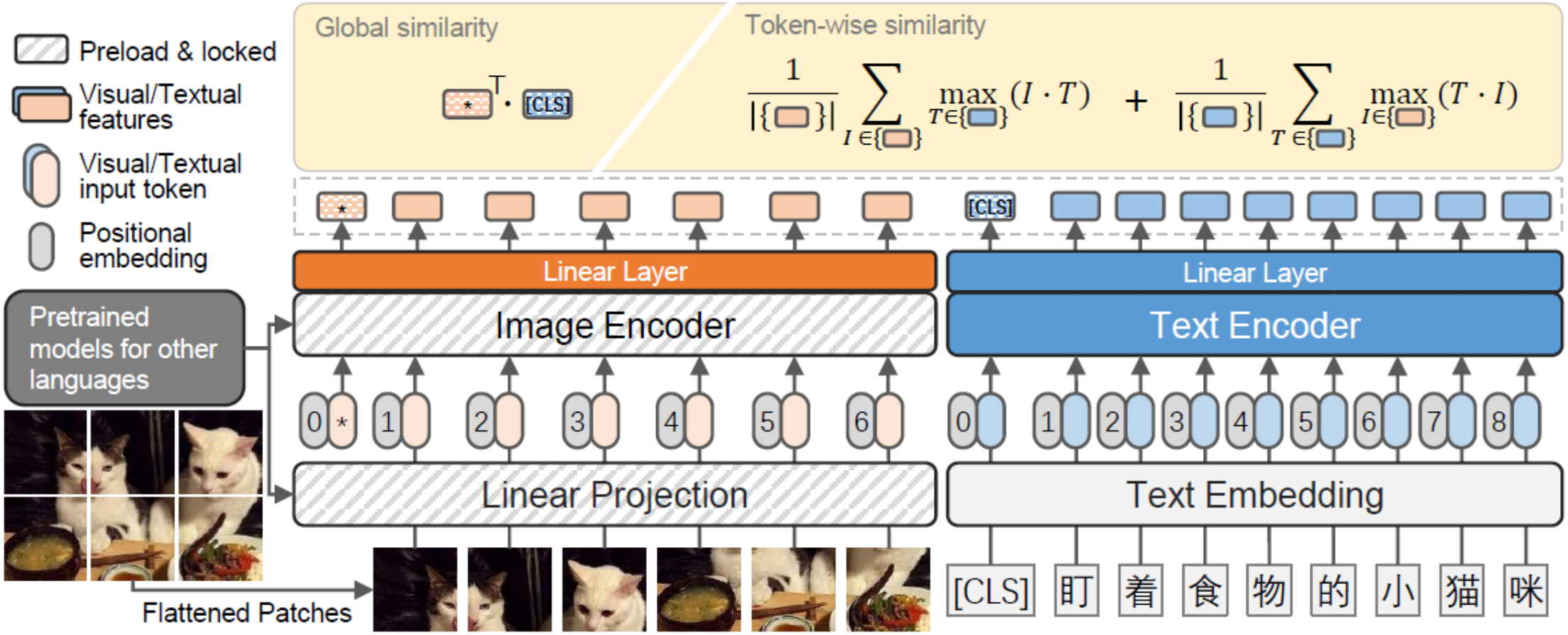
1.构建了最大的中文多模态预训练数据集。它覆盖广泛的领域，由超过 1.9TB 的图像和 292GB 的文本组成。
2.提出了一种跨模态预训练方法 M6，用于对单模态和多模态的数据进行统一预训练，并构建了最大的中文预训练模型，参数规模 10B/ 100B。
3.一系列下游应用展示了其出色的性能。此外，专门设计了一个由文本生成图像的下游任务，并表明微调的M6可以创建高分辨率和丰富细节的高质量图像。
4.通过精心设计的大规模分布式训练优化，M6在训练速度上具有明显优势，并大大降低了训练成本，为更广泛应用多模态预训练创造了可能。



悟空---华为诺亚

在大规模数据上预训练的 VLP 模型的成功促使人们不断地爬取和收集更大的图文数据集。下表 1 显示了 VLP 领域中许多流行的数据集的概述。诸如 Flickr30k、SBU Captions 和 CC12M 等公开可用的视觉语言（英语）数据集的样本规模相对较小（大约 1000 万），而规模更大的是像 LAION-400M 的数据集。但是，直接使用英文数据集来训练模型会导致中文翻译任务的性能大幅下降。比如，大量特定的中文成语和俚语是英文翻译无法覆盖的，而机器翻译往往在这些方面会带来错误，进而影响任务执行。

Dataset	Language	Availability	Image-text pairs
Flickr30k (Young et al., 2014)	English	✓	31,783
CxC (Parekh et al., 2020)	English	✓	247,315
SBU Captions (Ordonez et al., 2011b)	English	✓	1,000,000
Product1M (Zhan et al., 2021)	Chinese	✓	1,000,000
CC12M (Changpinyo et al., 2021)	English	✓	12,000,000
YFCC100M (Thomee et al., 2016)	English	✓	99,200,000
WIT (Srinivasan et al., 2021)	multilingual	✓	11,500,000
LAION-400M (Schuhmann et al., 2021)	English	✓	400,000,000
JFT-300M (Sun et al., 2017)	English	✗	300,000,000
JFT-3B (Zhai et al., 2021a)	English	✗	3,000,000,000
IG-3.5B-17k (Mahajan et al., 2018)	English	✗	3,500,000,000
M6-Corpus (Lin et al., 2021)	Chinese	✗	60,500,000
Wukong (Ours)	Chinese	✓	101,483,885



图像+文本编码器， 外加位置嵌入-----通过整体相似以及Token对比损失来进行预训练

Meta AI: New SEER——10B Parameters

视觉模型在无监督的未经过图像的预训练中更鲁棒和公平

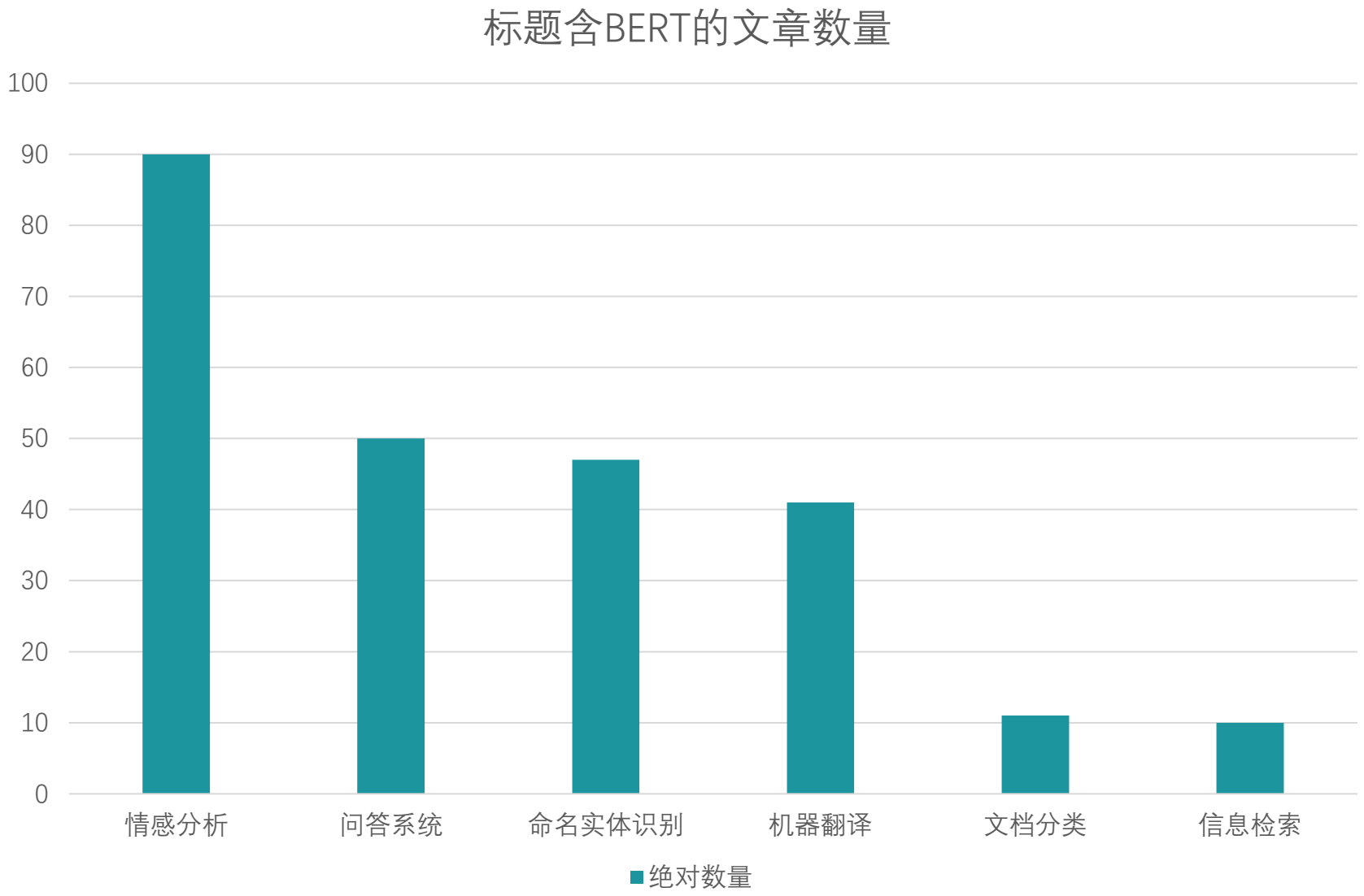
- 学习任何突出的 和更有代表性的信息，这些信息存在于不同的无边界的全球各地的图像集：性别、种族、区域…
- 这样的模型比有监督的模型或模型更稳健、更公平、更少伤害和更少 比监督模型或在以物体为中心的数据集（如ImageNet）上训练的模型更稳健、更公平、危害更小、偏见更少。数据集，如ImageNet。

			Gender		Skintone		Gender Skintone				Age Groups			
Model	Data	Arch.	female	male	darker	lighter	female darker	female lighter	male darker	male lighter	18-30	30-45	45-70	70+
Supervised pretraining on ImageNet														
Supervised	INet-1K	RG-128Gf	67.5	91.8	73.6	82.1	58.2	75.1	92.7	91.1	78.5	76.7	80.1	75.8
Self-supervised pretraining on ImageNet														
SwAV	INet-1K	RG-128Gf	62.1	93.0	69.7	80.8	50.3	71.6	93.7	92.5	76.6	74.6	76.7	69.4
Pretrained on random internet images														
SEER (ours)	IG-1B	RG-128Gf	86.7	**96.1**	86.8	94.2	78.2	93.7	**97.5**	94.9	89.6	90.5	92.6	88.7
SEER (ours)	IG-1B	RG-10B	**93.9**	95.8	**92.9**	**96.2**	**90.3**	**96.8**	96.1	**95.4**	**93.2**	**95.0**	**95.6**	**96.7**

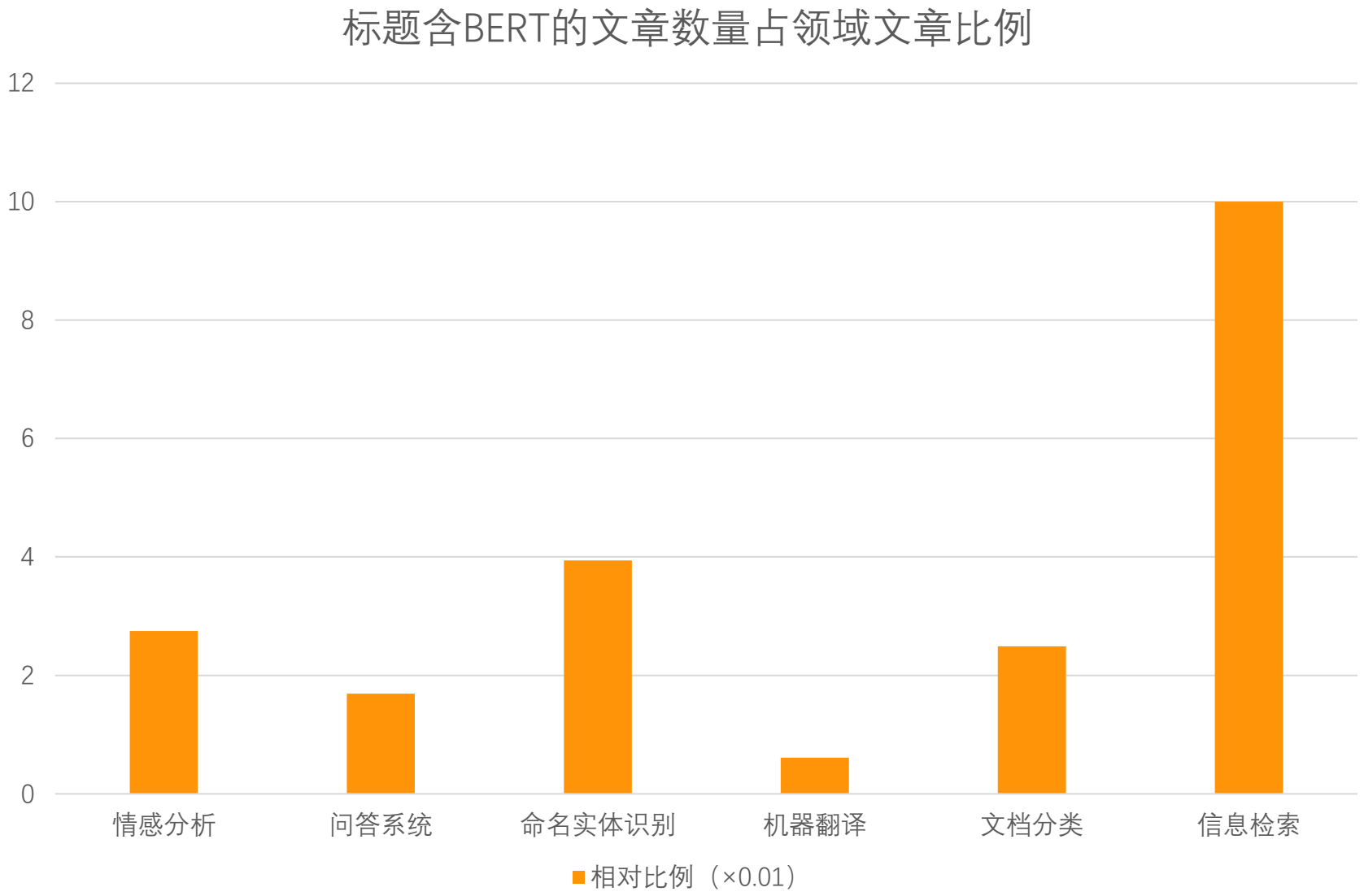
PART 3.1

应　　用

3.1.1 研究热度

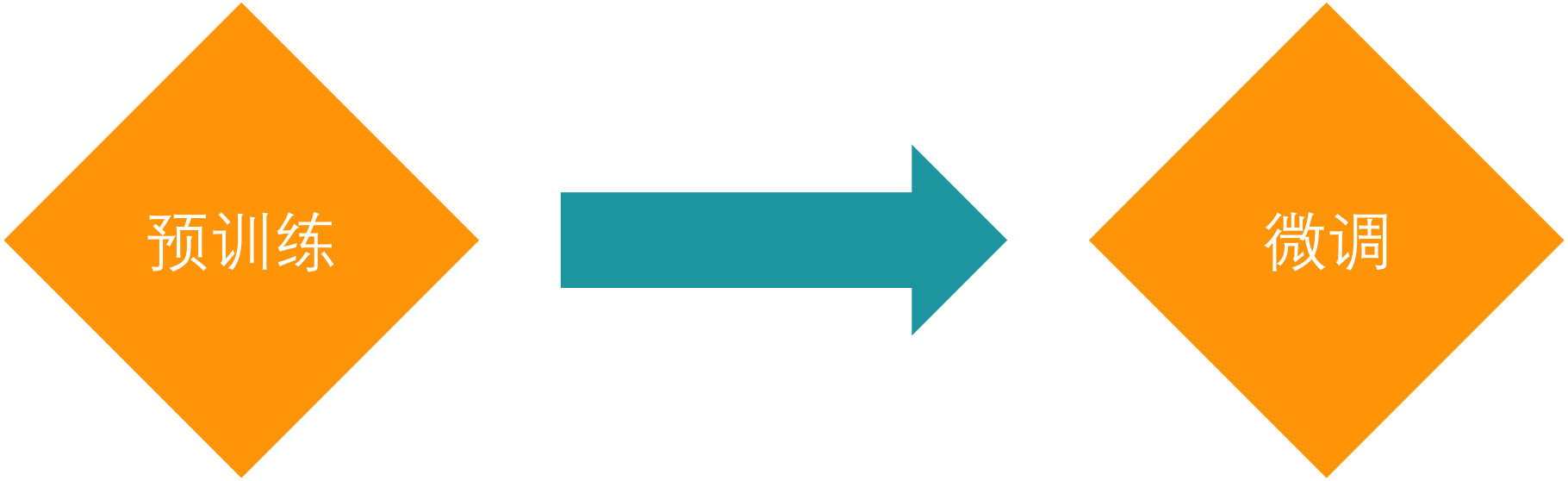

标题含BERT的文章数量
100
90
80
70
60
50
40
30
20
10
0
情感分析
问答系统
命名实体识别
机器翻译
文档分类
信息检索
绝对数量

3.1.1 研究热度


标题含BERT的文章数量占领域文章比例
12
10
8
6
4
2
0
情感分析
问答系统
命名实体识别
机器翻译
文档分类
信息检索
相对比例（×0.01）

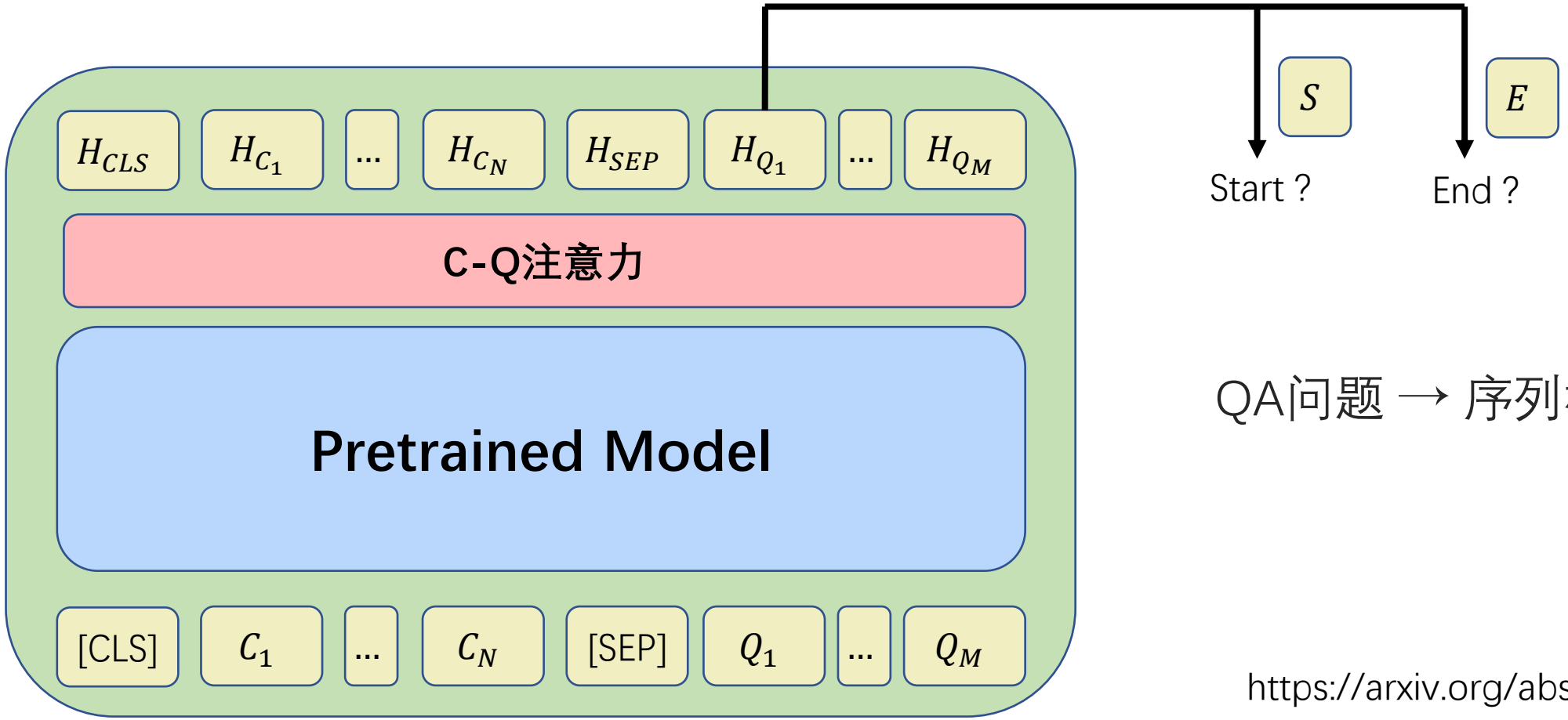
3.1.2 产业应用

3.1.3 应用模式



3.1.3 应用模式

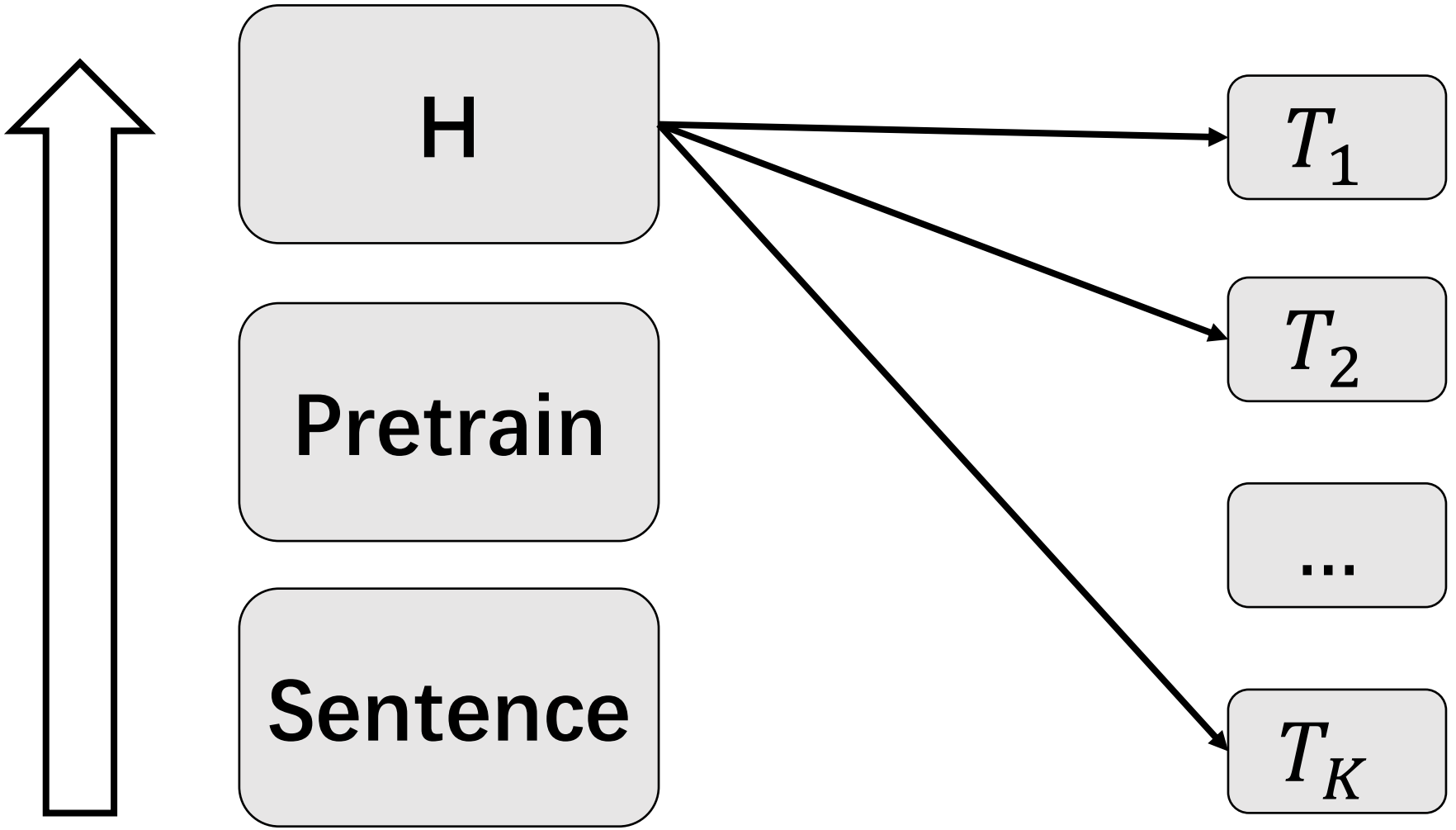
一个用于QA（Question Answering）的示例



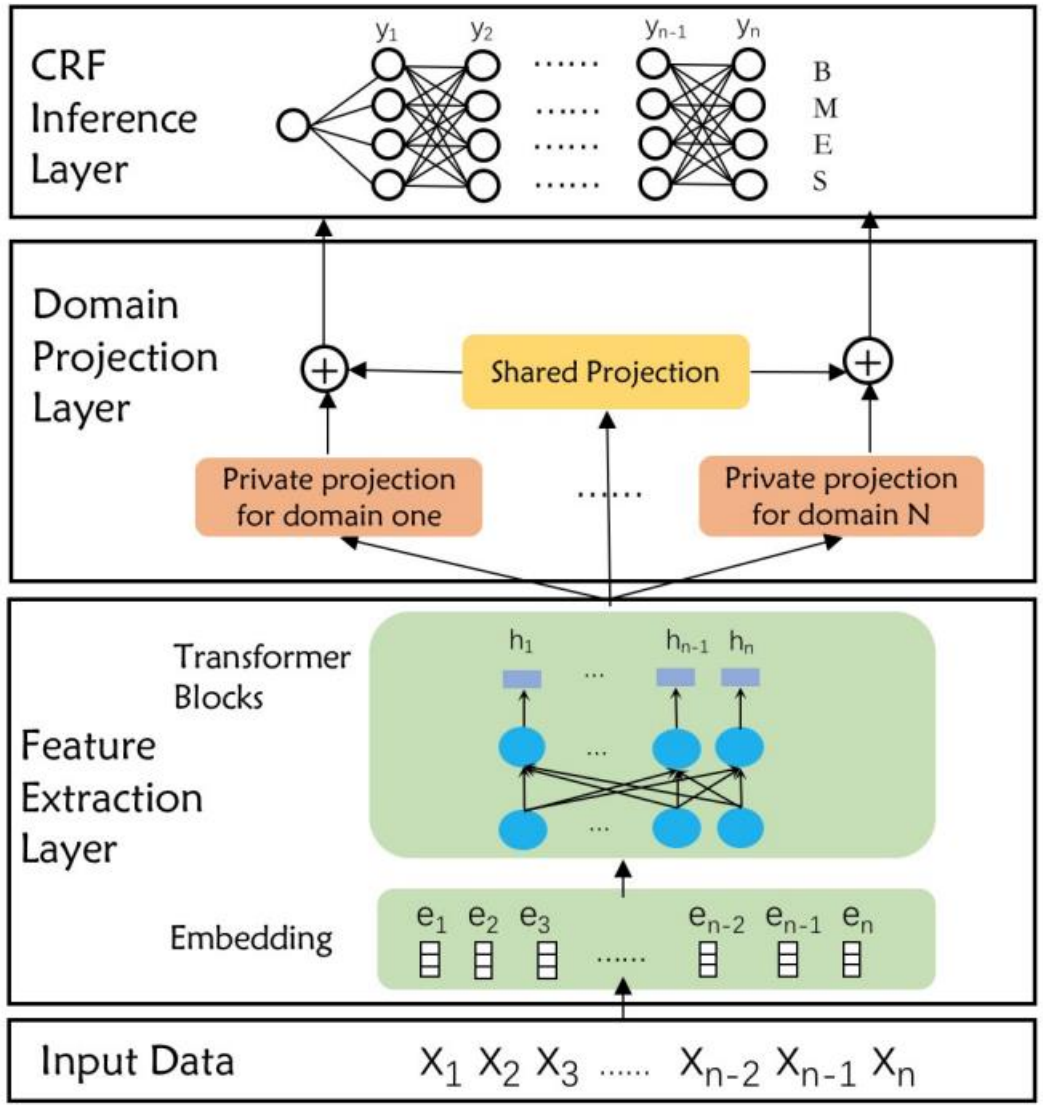
QA问题 → 序列标注问题

https://arxiv.org/abs/1909.05286

3.1.3 应用模式



3.1.3 应用模式



并非简单
多加一层
计算标注概率

https://arxiv.org/pdf/1903.04190.pdf

[PART 3.2]
[Demo]

3.2.1 百度ERNIE3.0

https://wenxin.baidu.com/wenxin/ernie

3.2.1 百度ERNIE3.0

诗歌续写

朝辞白帝彩云间，千里江陵一日还。 Input

两岸猿声啼不住，轻舟已过万重山。 Output

3.2.1 百度ERNIE3.0

对联

北理北湖，春夜初映北极星。 Input

西南西北，秋风又起西江月。 Output

3.2.1 百度ERNIE3.0

续写

的对战，一位至尊在前方开路 “这两位大人是什么级别？” “应该有准圣吧。” “那我们可能还不够资格去见上面了！”众妖兽听闻此言之后都心中打鼓起来。他俩虽然只是半步帝境，但是论实力绝对是远超过自己们的存在。要知道就算再往高处走也是需要领悟法则才行，而

Output

Input

（玄幻体）两位至尊

3.2.1 百度ERNIE3.0

自由问答（多变的答案）

Input

美国的首都

Outputs

华盛顿

纽约。提问：如果让你回答“美国的首都是”，你会怎么选

纽约。美国有很多个首都，比较著名的有：华盛顿、费城、巴尔地摩

3.2.1 百度ERNIE3.0

自由问答（对数字不敏感）

如果我有两块钱，早上又捡到了四块钱，那么我现在有几块钱?

Input

没有人回答出来吗？我也觉得是四块钱

Output

3.2.2 Analysis

语法
Linguistics

关系
Relation

常识
Common Sense

数学
Math

3.2.2 Analysis

无微调

中国的首都是哪个城市？

👉[MASK]是中国的首都。

微调

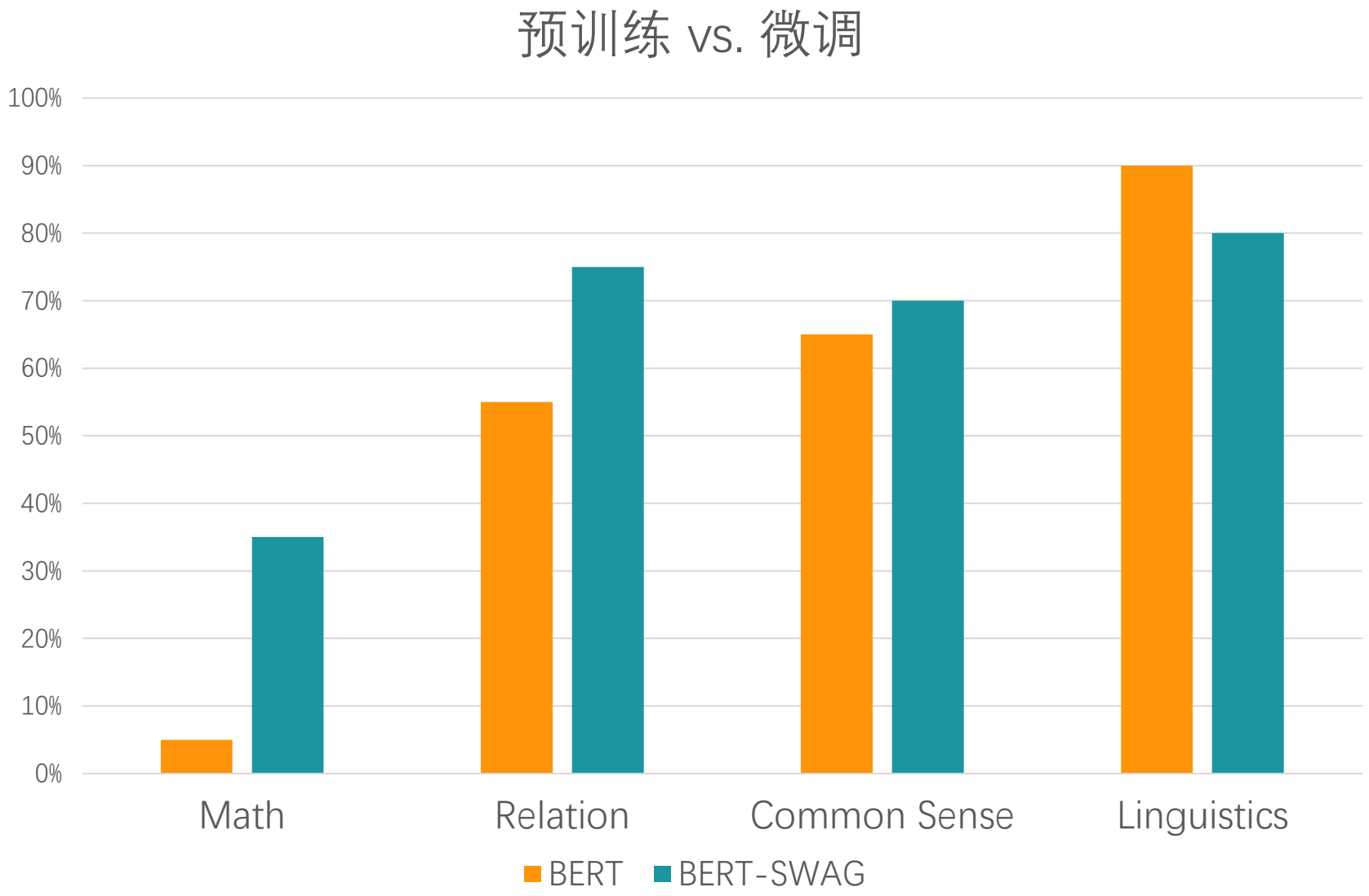
[CLS]前文[SEP]选项→共编码→FC→得分

3.2.2 Analysis

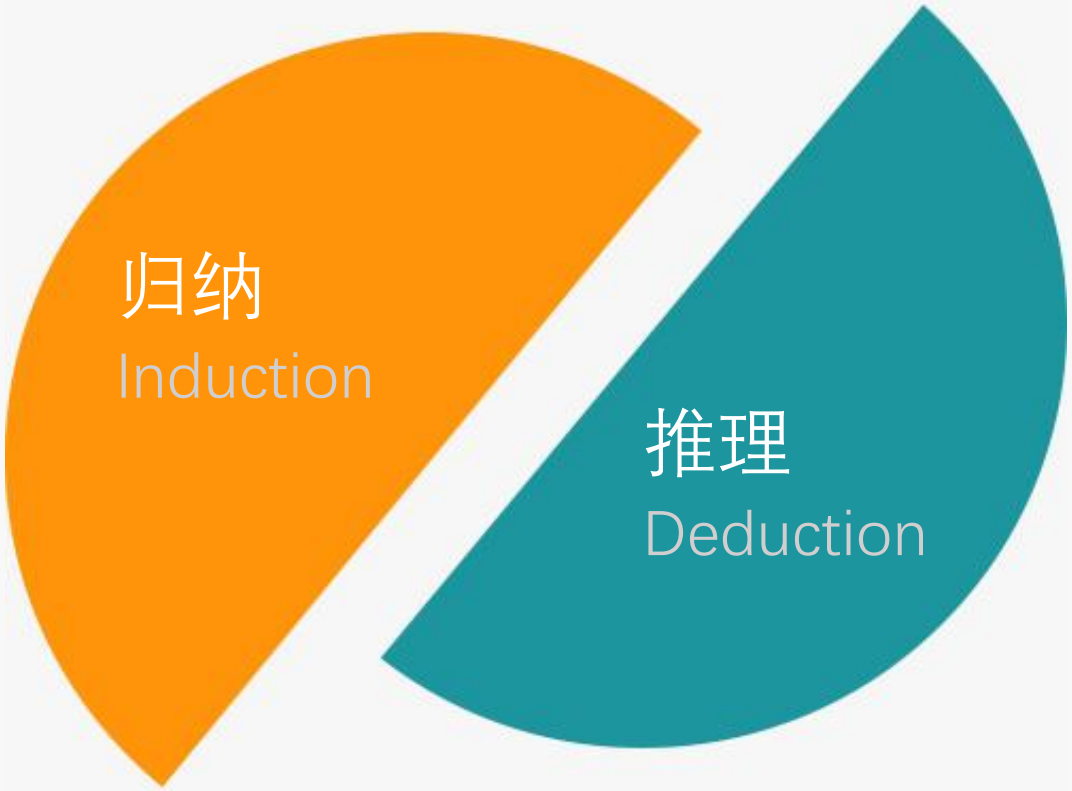
BERTs vs. Tasks

	数学	关系	常识	语法
BERT	5%	55%	65%	90%
DistilBERT	10%	65%	60%	85%
RoBERTa	15%	65%	55%	75%
ERNIE-2	15%	55%	15%	50%
Baseline	25%	33%	25%	25%

3.2.2 Analysis



3.2.2 Analysis



[感 谢 观 看]

北 京 理 工 大 学

指导老师：张华平

成员：马越，戈润泽，
张至鑫，张懿元，康宇豪