



BIT



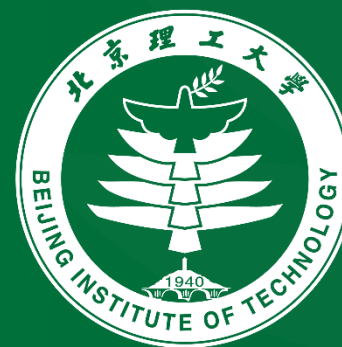
请在此输入副标题

BERT

答辩人：朱晓光 王海 陈立围

时间：2019/11/6

德以明理 学以精工



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



目录

CONTENTS

- 1 预训练模型的发展
- 2 BERT原理
- 3 After Bert
- 4 应用



1

预训练模型的发展

- Word2Vec
- ELMO
- GPT

离散表示 (one-hot)

apple = [1 0 0 0 0 0 0 0]

dog = [0 1 0 0 0 0 0 0]

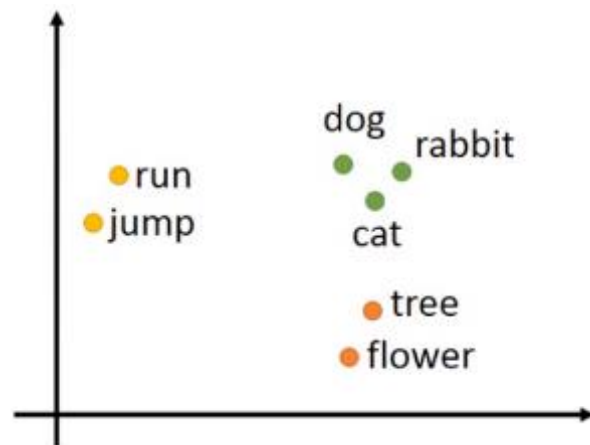
cat = [0 0 1 0 0 0 0 0]

⋮

优点： 简单、高维空间线性

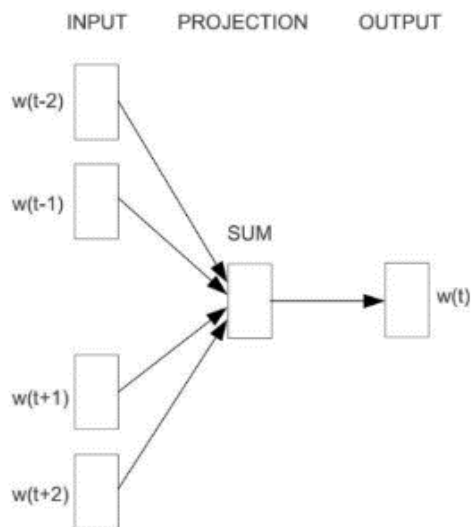
缺点： 稀疏、无语义信息

分布式表示 (word embedding)

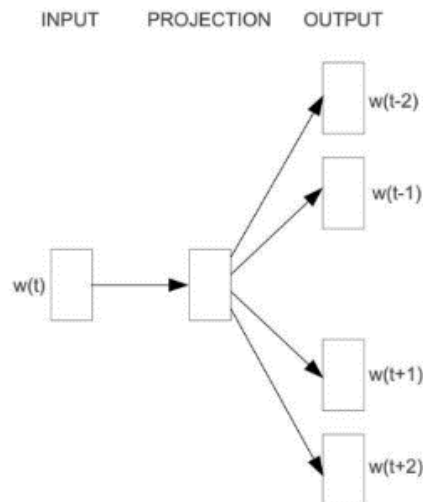


可计算词间关系、低维

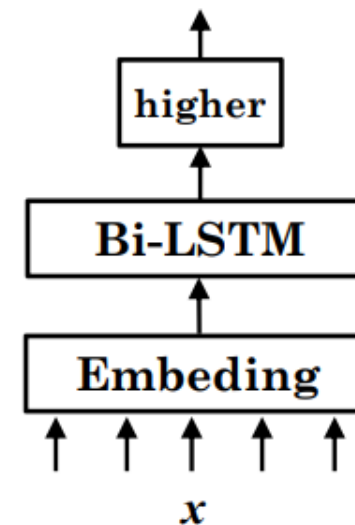
训练复杂



CBOW



Skip-gram

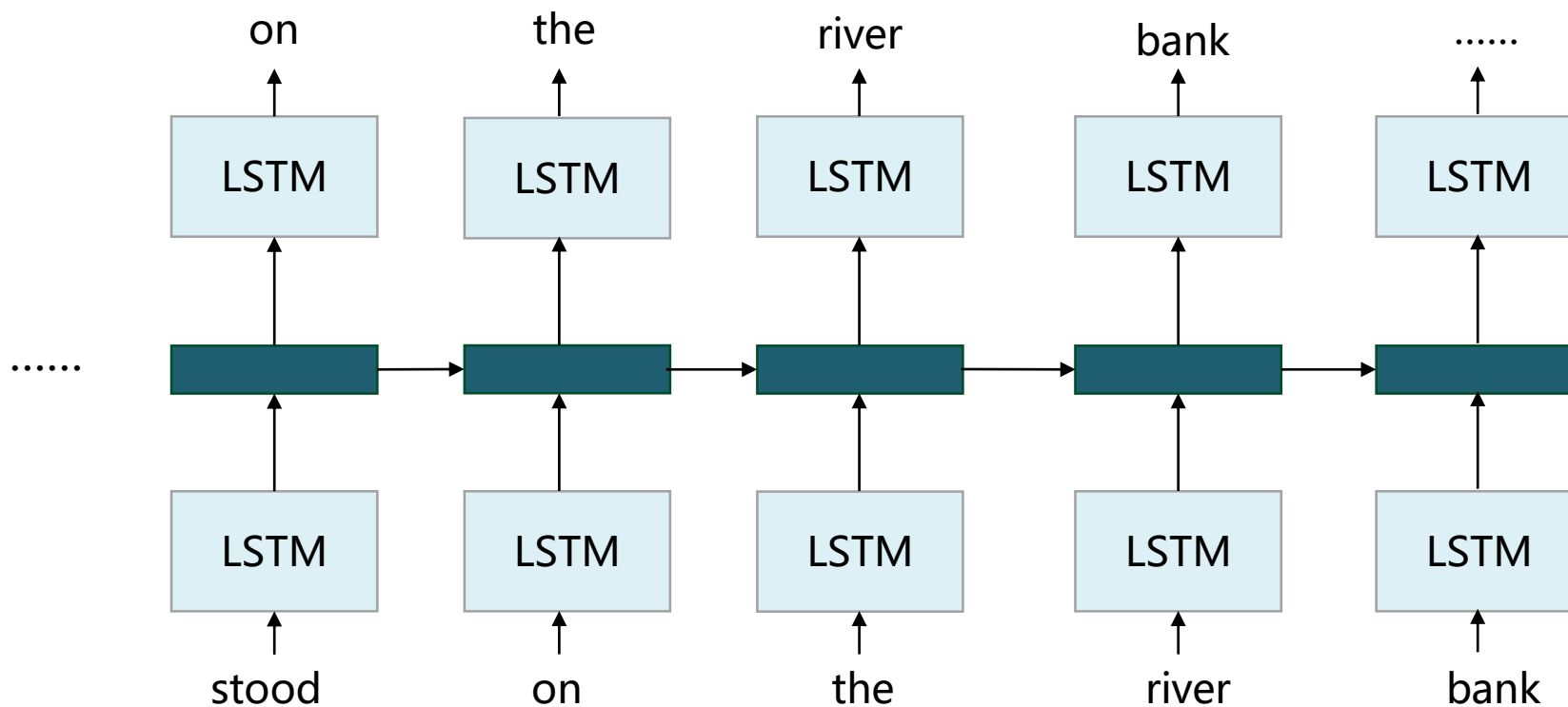


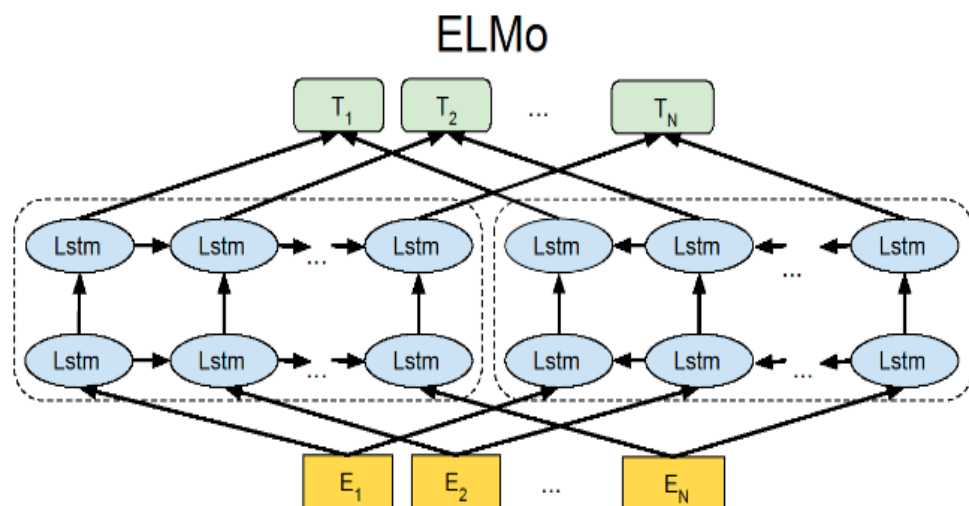
下游任务使用

缺陷：多义词问题

- Have you paid that **money** to the bank yet?
- They stood on the **river** bank to fish.

根据上下文动态调整词向量



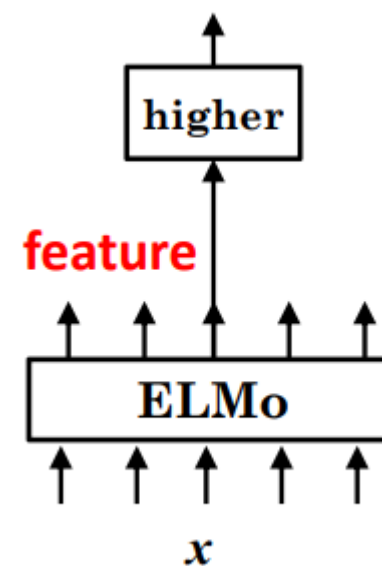


语义特征

句法特征

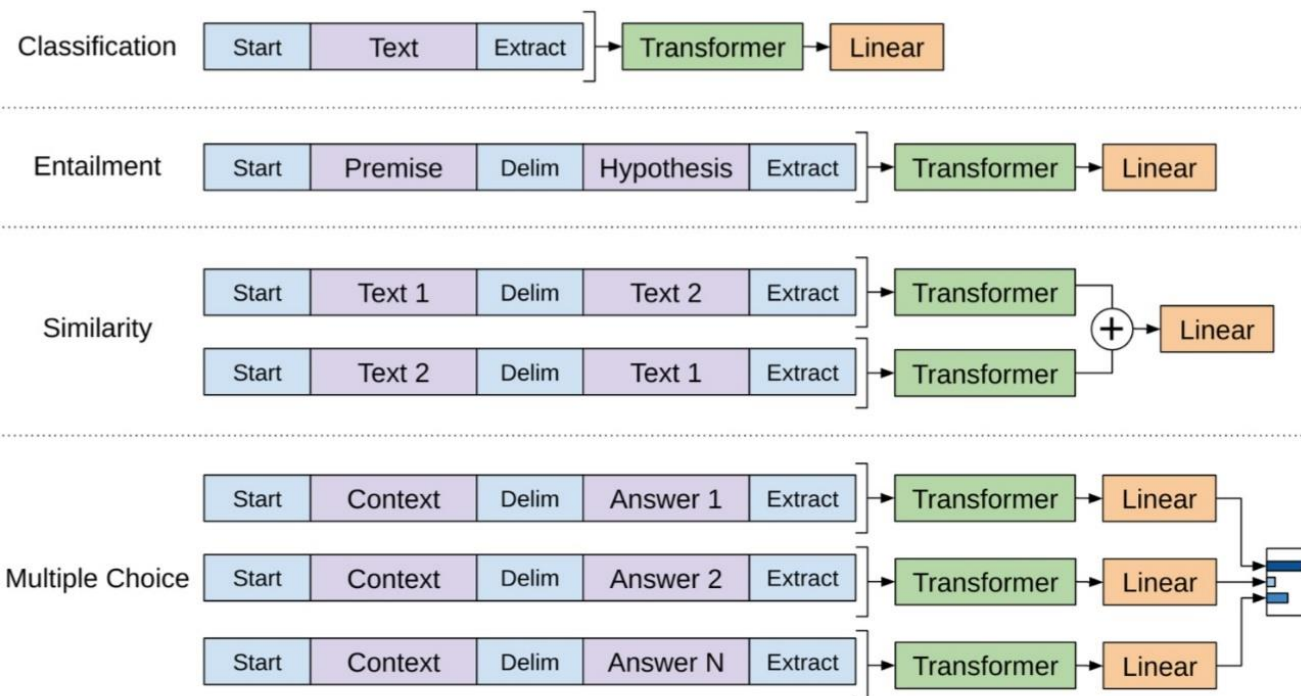
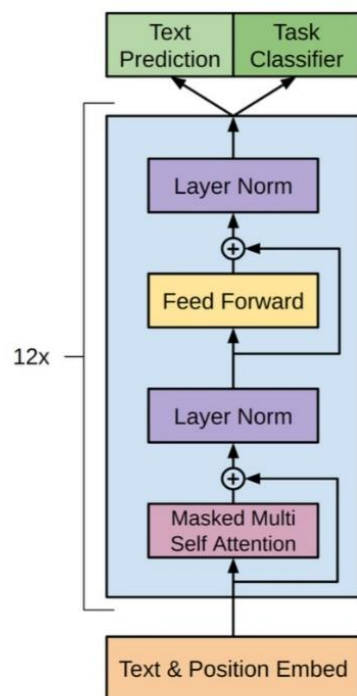
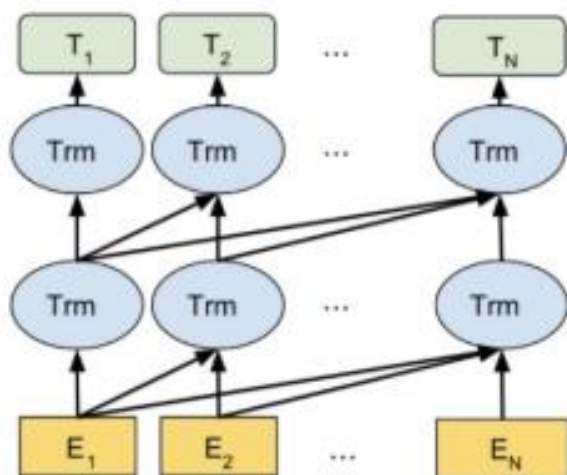
单词特征

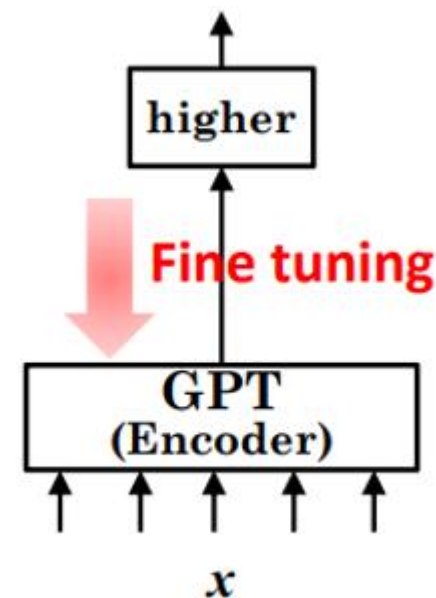
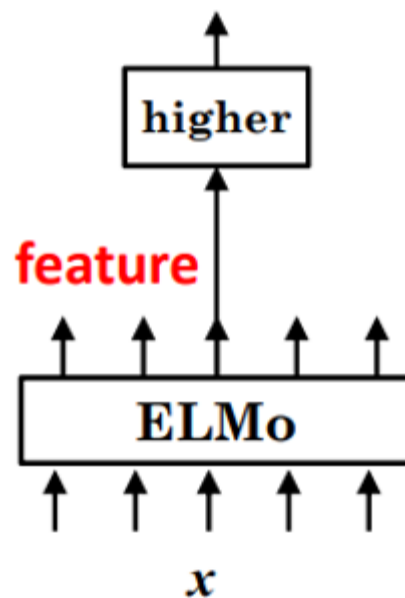
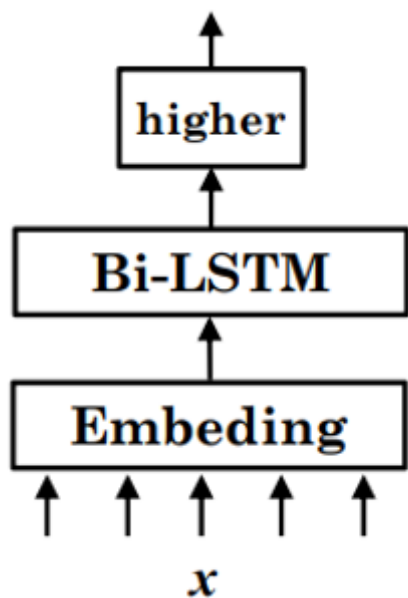
+



下游任务使用

OpenAI GPT





	优点	缺点
ELMO	1. 考虑了上下文信息 2. 适配下游任务比较简单	1. LSTM效果较差 2. 不适用于生成任务
GPT	1. Transformer可捕捉更长范围的信息 2. 可并行化，训练速度快 3. 适用于语言生成任务	1. 只利用了上文的信息，没有利用下文的信息 2. 下游任务需要重新调整数据结构



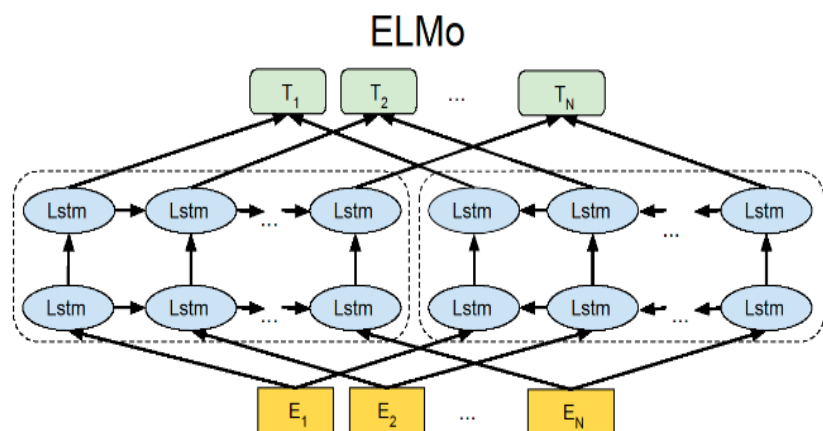
2

Bert原理

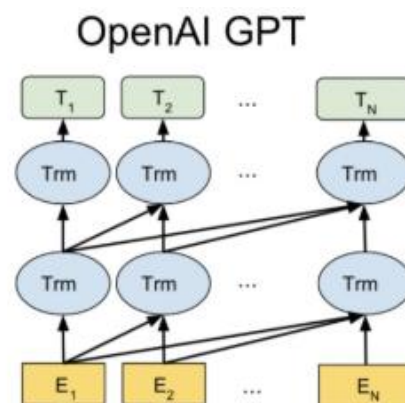
主要特征：

1. 利用了真正双向的Transformer
2. Mask-LM
3. Next Sentence Prediction
4. 通用任务框架

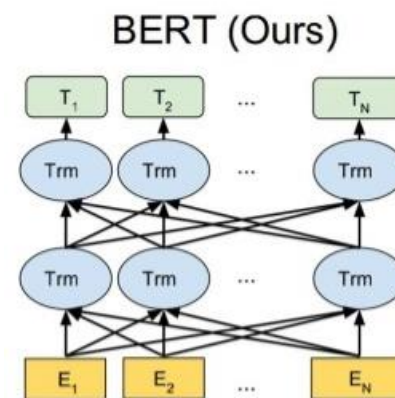
对比:



ELMo两个单向模型信息的生硬连接



单向：从左到右



双向的Transformer
“你中有我，我中有你”

Mask-LM

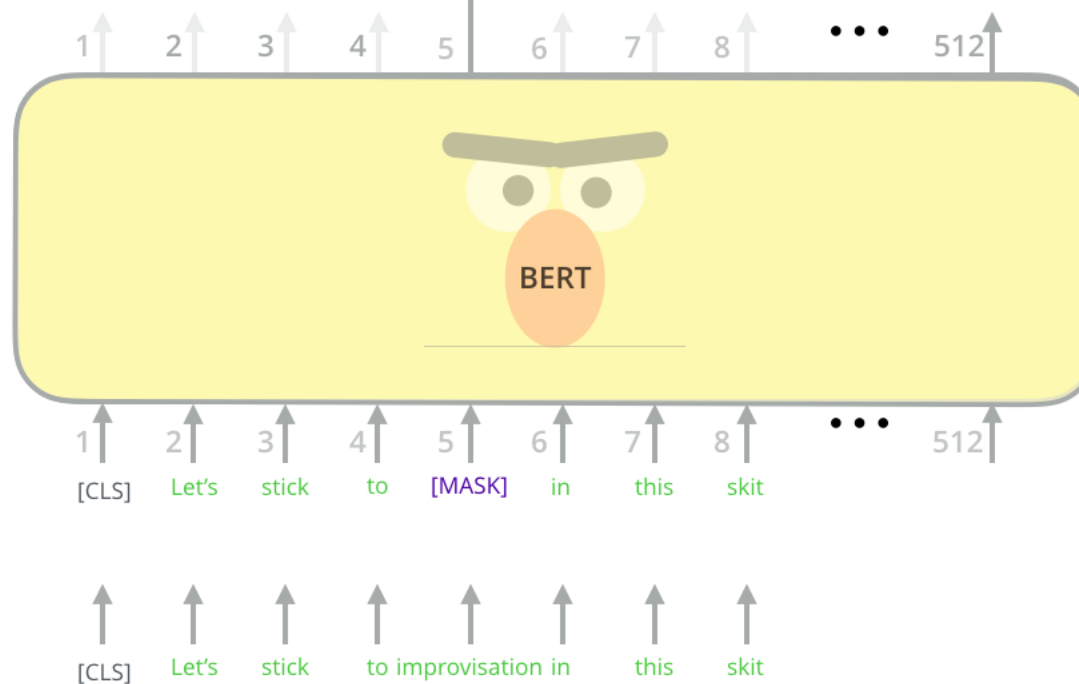


Use the output of the masked word's position to predict the masked word

Possible classes:
All English words

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zyzzzyva

FFNN + Softmax



Randomly mask
15% of tokens

Input

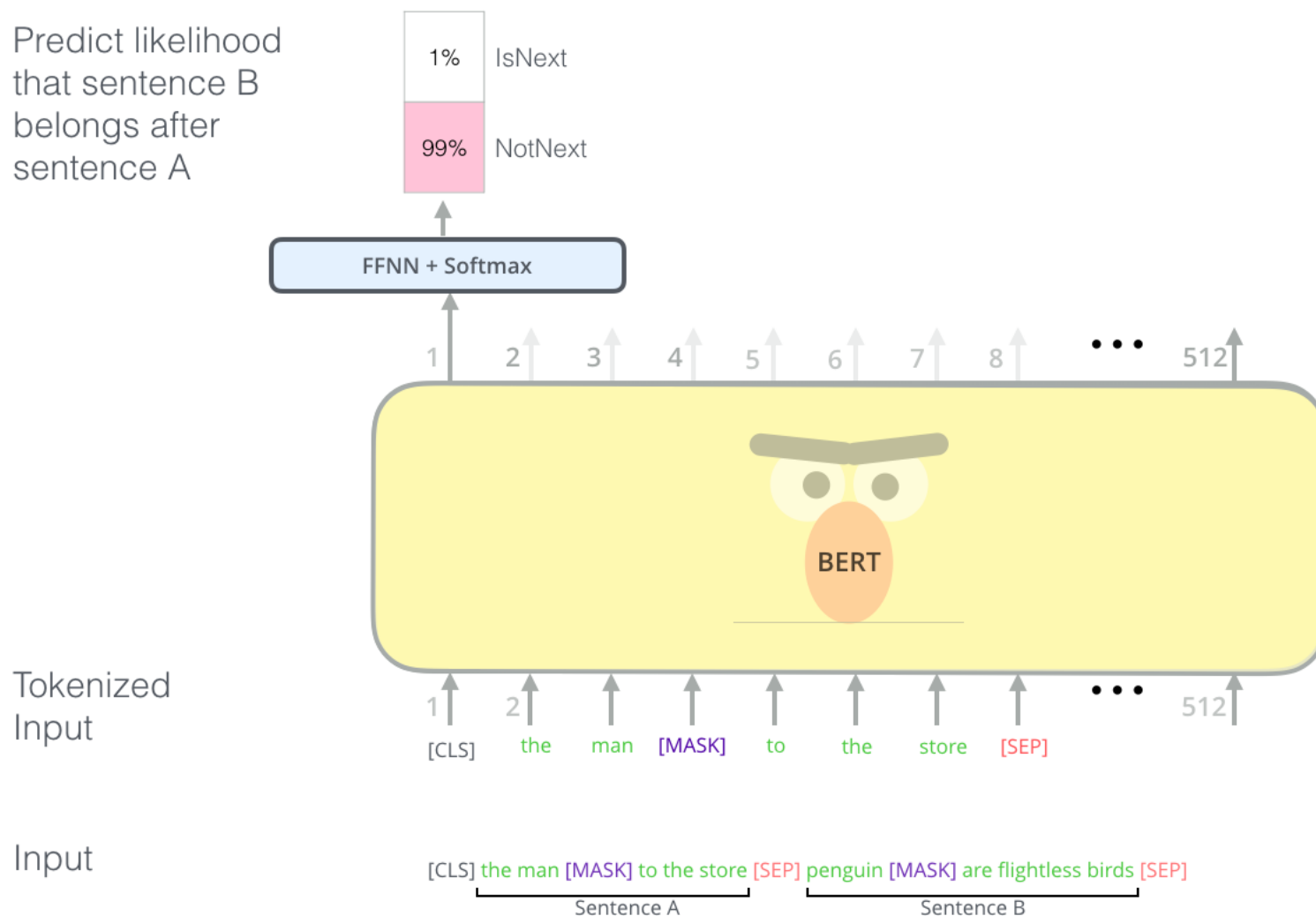
BERT针对如何做 “[MASK]” ， 做了一些更深入的研究， 它做了如下处理

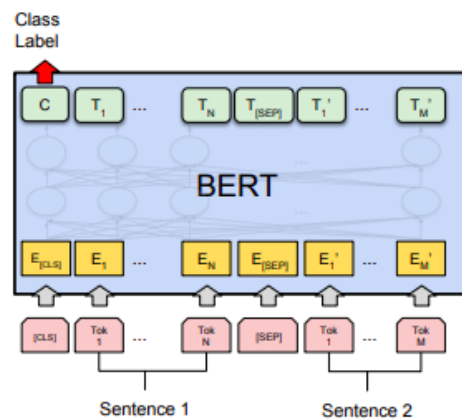
- 选取语料中所有词的15%进行随机mask
 - 选中的词在80%的概率下被真实mask
 - 选中的词在10%的概率下不做mask， 仍然保留原来真实的词
 - 让输出向量朝真实输入Embedding有一个偏移
 - 选中的词在10%的概率下不做mask， 而被随机替换成其他一个词
 - 增强鲁棒性。
 - “居安思危， 思则有备， 有备无患， 敢以此规”

Next Sentence Prediction

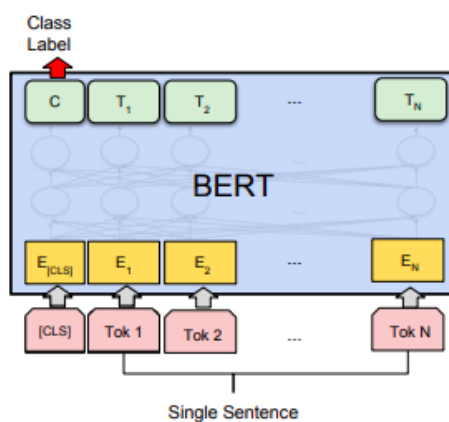


Predict likelihood
that sentence B
belongs after
sentence A

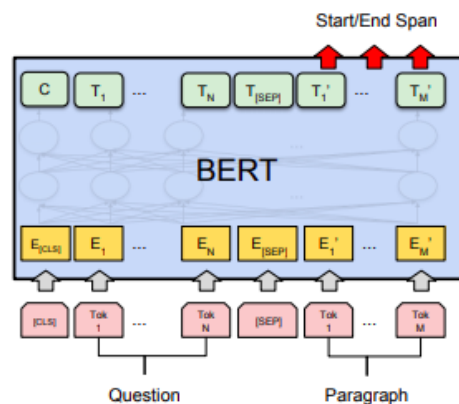




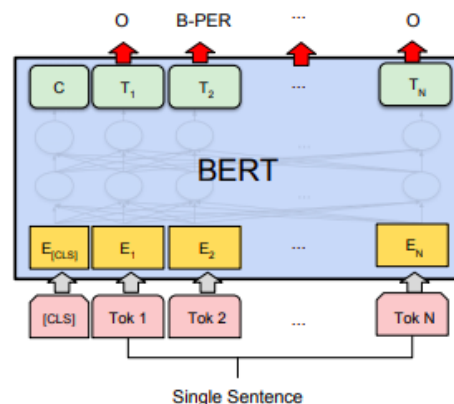
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA



(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER



3

After Bert

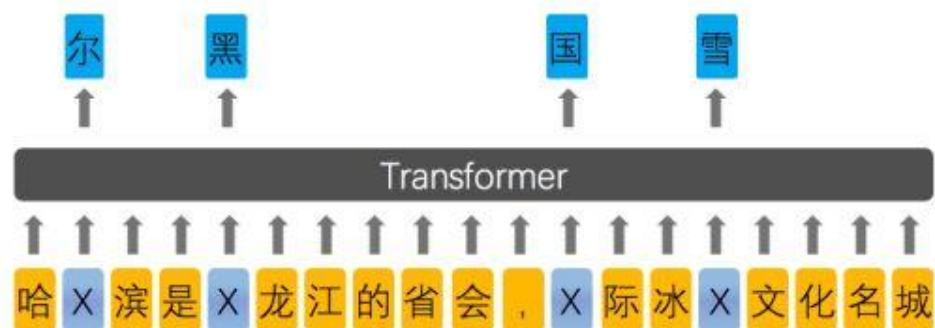


Ernie经常利用Bert的严肃捉弄他，搞得Bert毫无办法。他是因为喜欢Bert才笑话他，所以他的玩笑和笑话绝不是故意伤害。



Bert长期痛苦地忍受着 Ernie 的取笑，他对人特别认真，任何事都可以令他沉迷总是能原谅厄尼，永远做他的好朋友。

Learned by BERT



Learned by ERNIE



哈尔滨是黑龙江的省会，国际冰雪文化名城





4 应用

数据来源



数据处理

训练数据集: 38471条新闻

验证数据集: 4000条新闻

测试数据集: 4000条新闻

```
class DataProcessor(object):...
```



```
class NewsProcessor(DataProcessor):...
```



```
class InputExample(object):...
```



```
class InputFeatures(object):...
```

现在肉松包、不能吃原来都是棉花做滴！大家看看吧黑商！, 1

上海4000吨垃圾偷倒苏州太湖西山岛(图) (分享自@新浪新闻) ,0

匿名用户: 如何看待5月18日邵东县人民医院王俊医生被患者击打后脑勺致死? -来自读读日报,0



Dev result

Test result

success

```
def forward(self, input_ids, token_type_ids, attention_mask, labels=None):
    all_output, pooled_output = self.bert(input_ids, token_type_ids, attention_mask)
    logits = self.classifier(pooled_output)
    loss_fct = CrossEntropyLoss()

    if labels is not None:
        loss = loss_fct(logits, labels)
        return loss, logits
    return logits
```



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

感谢大家的聆听

德以明理
学以精工