

# 2025.5.23

---

2025.5.23

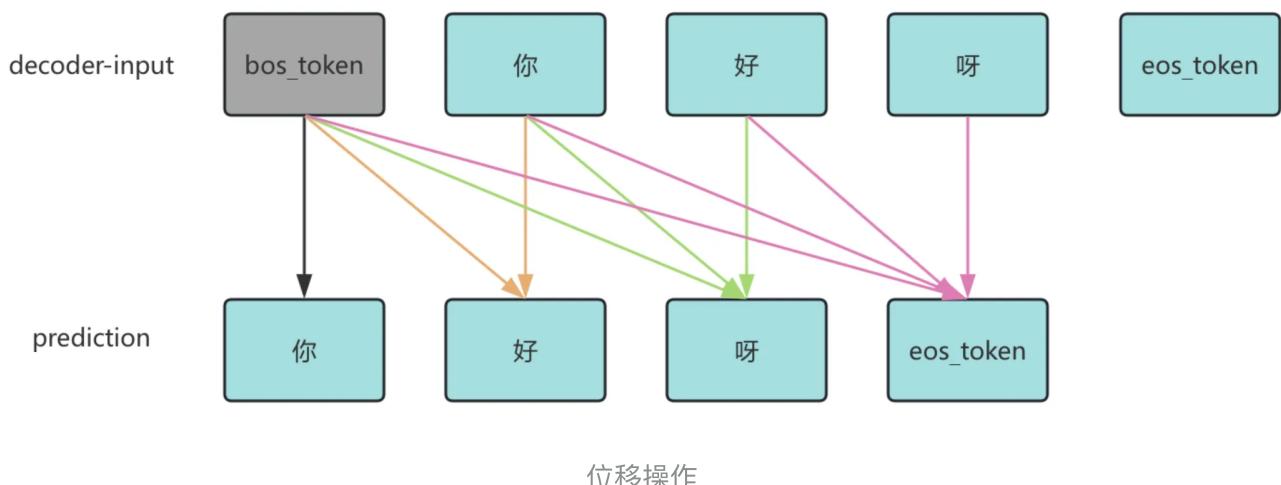
1. 机器翻译bug修复
2. 通过主题建模衡量开放式创新实践 论文

# 2025.5.23

## 1. 机器翻译bug修复

1. tokenizer本身没问题，可能是网络下载错误，重新下载就解决了
2. 预测没有采用自回归机制，而是一次性预测
3. decoder–input没有采用位移机制，初始化bos–token作为decoder–input是预测第一个词的关键。

encoder–decoder做seq2seq的时候，对于decoder–input和output有个关键点，就是decoder–input需要在第一个位置加上一个bos\_token，这个就是位移操作，因为是采用自回归预测的，所以在预测的时候，decoder–input初始化时bos\_token，这样可以预测第一个词，然后将预测的新token拼接到当前decoder–input，重复上述步骤，直到预测的token为eos\_token时停止。



训练epoch较少，模型不能够按照实际场景的语义来翻译，不过模型能够理解基本的单词对应的意思。

预测结果如图：

```

1 {
2   "english": "For greater sharpness, but with a slight increase in graininess, you can use a 1:1 dilution",
3   "chinese_reference": "为了更好的锐度,但是附带的会多一些颗粒度,可以使用这个显影剂的1:1稀释液。",
4   "chinese_prediction": "对于更锐利,但随着粒度的增加,您可以使用1:1 的开发者。"
5 }
6 {
7   "english": "He calls the Green Book, his book of teachings, \"the new gospel.\"", ->
8   "chinese_reference": "他还宣扬自己思想的所谓《绿皮书》称作“新福音书”。",
9   "chinese_prediction": "他叫绿书,他的书《新福音》。"
10 }
11 {
12   "english": "And the light breeze moves me to caress her long ear", ->
13   "chinese_reference": "微风推着我爱抚她的长耳朵",
14   "chinese_prediction": "光泽的光芒让我心心地照顾她长耳朵。"
15 }
16 {
17   "english": "They have the blood of martyrs is the White to flow ...",
18   "chinese_reference": "它们的先烈们的鲜血是白流了...",
19   "chinese_prediction": "他们有烈士的血是白的..."
20 }
21 {
22   "english": "Finally, the Lakers head to the Motor City to take on a Pistons team that currently owns the
23   "chinese_reference": "最后,在1月31日,湖人将前往汽车城底特律挑战活塞队,活塞近来在东部排名第二。",
24   "chinese_prediction": "最后,湖人头到莫斯科,带领一个巴克斯球队,现在拥有东会议第二纪录(1/331)。"
25 }
26 {
27   "english": "\\"The perfect match--my father loves names and Jackie loves money, \' sneered Alexander at th
28   "chinese_reference": "真是天造地设的一对—我父亲喜欢结交名人,杰姬爱金钱,”亚历山大在婚礼上讥讽道。他和克里斯蒂娜从头到
29   "chinese_prediction": “完美的匹配—我的父亲爱名字和杰克·爱金钱。”他结婚的亚历山大·亚历山大在婚礼上说,“无论他还是基督和
30 }
31 {
32   "english": "In 2006, Walmart was charged with racism when its recommendation engine paired Planet of the
33   "chinese_reference": "2006年,沃尔玛的推荐引擎将《人猿星球》与马丁·路德·金的纪录片配成了一对,为此沃尔玛遭到了种族歧视
34   "chinese_prediction": "2006年,沃尔玛被指控在接受引擎的飞机上对马丁·卢瑟福的文档进行称赞时被指控种族主义。"
35 }
36 {
37   "english": "The matte as main copper phase in the cleaning, slag was deter- mined by electron probe mic
38   "chinese_reference": "通过电子探针显微分析确定氧化渣中主要铜相为冰铜相。",
39   "chinese_prediction": "以电探针为主要铜相,采用微探针法测定了渣中铅的含量。"
40 }

```

40M模型  
11epoch效果

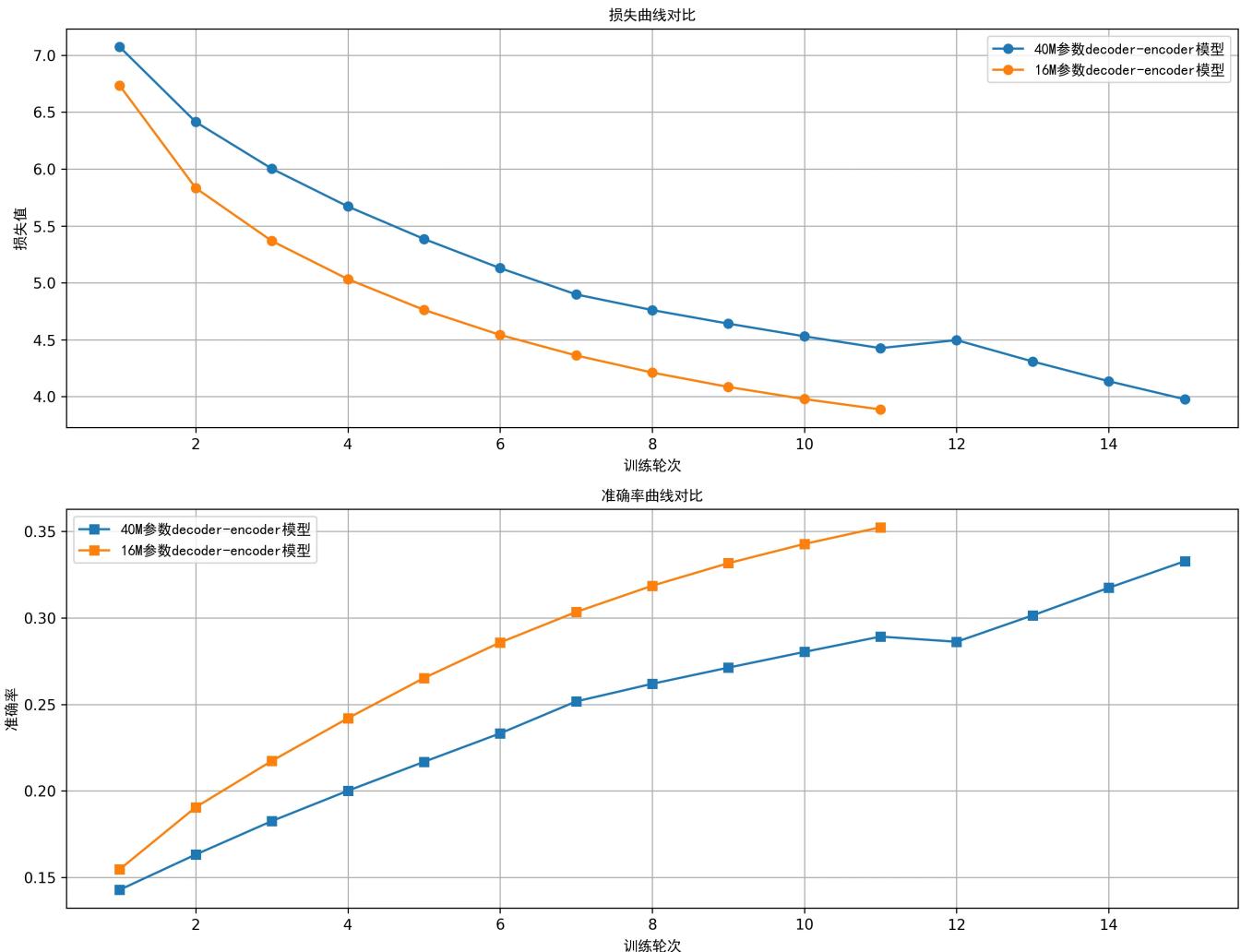
```

1 {
2   "english": "For greater sharpness, but with a slight increase in graininess, you can use a 1:1 dilution",
3   "chinese_reference": "为了更好的锐度,但是附带的会多一些颗粒度,可以使用这个显影剂的1:1稀释液。",
4   "chinese_prediction": "对于更大的伤害,但随着谷物的增加,你可以使用1/4的解决方案。"
5 }
6 {
7   "english": "He calls the Green Book, his book of teachings, \"the new gospel.\"", ->
8   "chinese_reference": "他还宣扬自己思想的所谓《绿皮书》称作“新福音书”。",
9   "chinese_prediction": "他叫格林布雷克,他的书《新福音》。"
10 }
11 {
12   "english": "And the light breeze moves me to caress her long ear", ->
13   "chinese_reference": "微风推着我爱抚它的长耳朵",
14   "chinese_prediction": "灯光让我想起她长长的耳朵。"
15 }
16 {
17   "english": "They have the blood of martyrs is the White to flow ...",
18   "chinese_reference": "它们的先烈们的鲜血是白流了...",
19   "chinese_prediction": "他们有谁的血流,是白云飞..."
20 }
21 {
22   "english": "Finally, the Lakers head to the Motor City to take on a Pistons team that currently owns the
23   "chinese_reference": "最后,在1月31日,湖人将前往汽车城底特律挑战活塞队,活塞近来在东部排名第二。",
24   "chinese_prediction": "最后,湖人头到莫斯科,带领一个巴克斯球队,目前只有一场东区队的第二场比赛。"
25 }
26 {
27   "english": "\\"The perfect match--my father loves names and Jackie loves money, \' sneered Alexander at th
28   "chinese_reference": "真是天造地设的一对—我父亲喜欢结交名人,杰姬爱金钱,”亚历山大在婚礼上讥讽道。他和克里斯蒂娜从未
29   "chinese_prediction": “完美匹配—我的父亲名字叫和杰克逊,“她坐在婚礼上,珍妮斯·亚历山大,北方北方,北方,他还是1717岁。”
30 }
31 {
32   "english": "In 2006, Walmart was charged with racism when its recommendation engine paired Planet of the
33   "chinese_reference": "2006年,沃尔玛的推荐引擎将《人猿星球》与马丁·路德·金的纪录片配成了一对,为此沃尔玛遭到了种族歧视
34   "chinese_prediction": "2006年,沃尔玛被评为《和平》杂志,《和平》杂志《路》杂志《路》杂志上传说。"
35 }
36 {
37   "english": "The matte as main copper phase in the cleaning, slag was deter- mined by electron probe mic
38   "chinese_reference": "通过电子探针显微分析确定氧化渣中主要铜相为冰铜相。",
39   "chinese_prediction": "电探针显微分析仪测定了该细胞的主要铜相色谱。"
40 }

```

110M模型  
15epoch效果

## Transformer模型训练历史对比



## 2. 通过主题建模衡量开放式创新实践 论文

先前的开放式创新（OIPs）对公司表现的影响的研究得到的结果不一致，本文提出的方法基于无监督学习技术，利用NLP和LDA，利用NLP从公司的OIPs抽取信息，创建初始关键词篮。

- 本文重新审视OIPs和公司财务表现的关系，结果表明一个公司的整体开放水平和提升的财务表现相关。
- 在部分的特定OIPs中发现了倒U型的关系，这一点部分的支撑了先前综述中出现的开放性悖论。
- 内部R&D与个体OIPs的互补性也随着实践不同而不同。
- 这些OIPs的影响也因为行业不同而不同。

本文的结论是：开放式创新对业务的影响是复杂的，不存在普遍适用的最佳实践来开展开放式创新。