

任务7：基于标题的大规模商品实体检索 评测总体

阿里安全

黄龙涛 高级算法专家

任务背景

- 电商场景中，由于商品标题存在堆砌关键词、虚假宣传、违规隐藏等问题，导致通过标题不一定能准确地匹配到目标商品。
- 该评测任务旨在对于给定的一个商品标题，参赛系统需要匹配到该标题在给定商品库中的对应商品实体。
- 存在的挑战：
 - 输入文本中可能无法识别出实体指代词
 - 商品标题中存在很多变异指代，没有给定的指代映射表
 - 商品标题一般较短，上下文语境不丰富，须对上下文语境进行精准理解

任务描述

- **输入：**
输入文件包括若干行商品标题。
- **输出：**
输出文本每一行包括此标题对应的商品实体，即给定知识库中商品ID，只返回最相关的1个结果。
- **示例：**

输入：

[转卖]现货初中学生巨文库 武志红2016年新作 浙江出版社传
俄罗斯代购印度睫毛增长液3ml,断货王! 如果有需要请咨询客服
泰国正品DC套装yh产品定居曼谷提供视频支持直发

输出：

.....
1
231
467
.....

数据集描述

● 商品库

该任务的商品知识库包括的约**30万**个商品实体。知识库中的每个实体都包含一个subject-ID，一个字符串名称，及与此实体相关的一系列三元组 <subject, predicate, object>信息形式。知识库中每行代表知识库的一条记录，每条记录的格式为一个json格式。predicate-id和object-id的值都为0，subject-id的值为一个正整数，说明subject总是对应知识库中的一个实体。

● 标注数据集

标注数据集由训练集、开发集和评估集组成。其中训练集中包括**9万**条商品标题，开发集和评估集分别包括**1万**条数据。

```
{  
  "subject_id": "1",  
  "subject": "巨婴国",  
  "type": ["Publication"],  
  "data": [  
    {  
      "predicate": "出版社",  
      "object": "浙江人民出版社"  
    },  
    {  
      "predicate": "书名",  
      "object": "巨婴国"  
    },  
    {  
      "predicate": "作者",  
      "object": "武志红"  
    },  
    {  
      "predicate": "出版时间",  
      "object": "2016年12月"  
    },  
    {  
      "predicate": "装帧",  
      "object": "平装"  
    }  
  ]  
}
```

```
{  
  "text_id": "1",  
  "text": "[转卖]现货初中生巨文库 武志红 2016 年新作 浙江出版社传",  
  "implicit_entity": [  
    {  
      "subject": "巨婴国",  
      "subject_id": "1"  
    }  
  ]  
}
```

评测方法

- 对于给定的商品标题，参赛系统输出的结果为商品知识库中的商品ID。我们通过将输出结果与人工标注集合S进行比较得到**准确率(Precision)**作为评价指标。
- 给定M个商品标题输入 $T = \{t_1, t_2, t_3, \dots, t_M\}$ ，系统输出结果中与标注结果相同的数量为N，则准确率为：

$$Precision = \frac{N}{M}$$

评测结果

#	队伍名	分数
1	DeepBlueAI	0.88489
2	喵喵喵 ☰	0.87947
3	tonyxu	0.87744
4	HairLossKnight ☰	0.87721
5	qianrenjian	0.86898
6	这次还会是分母吗	0.85295
7	zhdun ☰	0.84641
8	ECNU-NLP ☰	0.83049
9	nlpca ☰	0.81977
10	阿巴阿巴阿巴 ☰	0.81797

共**331**支队伍参赛

冠军:

深兰科技（上海）有限公司

亚军:

平安集团，阅文集团

季军:

南京希音电子商务有限公司

方法分析

- 由于暴力检索时间复杂度会非常高，参赛队伍基本都使用了召回+排序的策略
- 召回：多分类、相似度计算、
- 排序：粗排/精排、句子对匹配、loss设计、特征工程
- 数据处理：
 - 实体描述文本生成
 - 小样本、混淆样本处理
- 预训练模型
 - Roberta、ERNIE使用最多
 - electra效果不佳

THANKS