BERT在商品实体检索中的应用

DeepBlueAI 队伍

深兰科技（上海）有限公司

报告人：潘春光

团队介绍

- 团队介绍

深兰北京AI研发中心 隶属于深兰科技（上海）有限公司，核心成员来自清华、北大、上交、浙大、北邮、北航等知名院校，致力于计算机视觉、自然语言处理和数据挖掘等领域的研究与开发。团队依靠其自身的技术优势，主要负责公司AI平台的研发工作，并在CVPR、ICCV、ECCV、NeurIPS、KDD、SIGIR、PAKDD、ACM MM、IEEE ISI 及 ICPR等众多世界计算机科学及人工智能领域顶级赛事上获得**二十多项冠军**。

- CCKS 2020 获奖

1. CCKS 2020 基于标题的大规模商品实体检索 (第 1 名 & 技术创新奖)
2. CCKS 2020 面向金融领域的篇章级事件主题与要素抽取:task1 事件主体抽取 (第 1 名)
3. CCKS 2020 新冠知识图谱构建与问答：task1:新冠百科知识图谱类型推断 (第 1 名)
4. CCKS 2020 面向中文短文本的实体链指 (第 2 名)

团队介绍

- 团队奖项

ICPR 2020 Large-scale t Product n Recognition (第 1 名)

ACM MM 2020 Video Object Detection (第 1 名)

ECCV 2020 GigaVision Task1:Object Detection (第 1 名)

ECCV 2020 GigaVision Task2:Multi-Object Tracking (第 1 名)

CVPR 2020 NightOwls Detetection Challenge:Task1 (第 1 名)

CVPR 2020 NightOwls Detetection Challenge:Task2 (第 1 名)

CVPR 2020 UG2+ PRIZE CHALLENGE Track 1-task 1 (第 1 名)

CVPR 2020 NTIRE Perceptual Extreme Super-Resolution (PSNR 第 1 名)

IEEE FG 2020 Compound Emotion challenge (第 1 名)

ICCV 2019 COCO & Mapillary (第 1 名)

ICCV 2019 CVWC Challenge: Tiger Pose Detection (第 1 名)

ICCV 2019 VisDrone Challenge: Multi-Object Tracking (第 1 名)

ICCV 2019 VisDrone Challenge: Object Detection in Videos (第 1 名)

NeurIPS 2019 D-City BDD100K 目标检测挑战赛 (第 1 名)

NeurIPS 2019 AutoNLP (第 1 名)

ACM MM 2019 Relation understanding in videos (第 1 名)

KDD Cup 2019 AutoML Track (第 1 名)

SIGIR 2019 eBay Data Challenge (第 1 名)

CVPR 2019 Cassava Disease Classification (第 1 名)

IEEE ISI-World Cup 2019(Task 1) (第 1 名)

(AI 研习社) 安全帽佩戴检测赛 (第 1 名)

AIIA 2019 面向存量市场的 4G 用户消费预测 (第 1 名)

任务介绍

- 任务介绍

- CCKS 2020：基于标题的大规模商品实体检索，任务为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

  - **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：四盒粉，宝宝痱子粉

输出：硼酸氧化锌散

任务分析

- 难点

1. 商品标题一般较短
2. 输入文本中可能无法识别出实体指代词
3. 商品标题中存在很多变异指代，没有给定的指代映射表

- 信息检索

1. 与基于知识库的实体链接任务不同
2. 没有明确的的实体指称

数据处理

- 数据特点

1. 图书类别太多
2. 训练集中text_id 不唯一
3. 相同标题文本对应多个实体ID
4. 知识库存在部分相似实体

- 训练集格式


```
{
    "text_id": 81228,
    "text": "四盒粉，宝宝痱子粉",
    "implicit_entity": [{"subject": "硼酸氧化锌散", "subject_id": 23813}]
}
```

数据处理

- 知识库格式

1. 实体描述文本：Predicate 与 object 相连得到描述文本
2. 连接顺序：产地、功能、症状、主要成分、生产企业、规格


```
{
    "type": "Medical",
    "subject_id": 23813,
    "subject": "硼酸氧化锌散",
    "data": [
        {"predicate": "生产企业", "object": "中国医科大学附属盛京医院"},
        {"predicate": "主要成分", "object": "本品为复方制剂。其组分为：每盒含氧化锌12.5g、硼酸
12.5g"},
        {"predicate": "症状", "object": "本品具有收敛、止痒、吸湿、杀菌作用。用于预防和治疗成人和婴
幼 儿各种原因引起的痱子。"},
        {"predicate": "规格", "object": "50g"},
        {"predicate": "产地", "object": "中国"}]
}
```

数据处理

- 图书类别实体

1. 图书类别实体太多
2. 知识库实体：**27.7w**
3. 图书类: **27.3w**
4. 医药类: **4.4k**
5. 训练集：8.3w 图书类 44 个

- 训练集中text_id 不唯一

1. 在多数情况下大家会默认 text_id 是唯一的
2. 利用text_id唯一性，会导致标注错误

数据处理

- 相同标题文本对应多个实体ID

1. 标题文本不包含实体信息


```
{"text_id": 22473, "text": "药品", "implicit_entity": [{"subject": "丁苯羟酸乳膏",
"subject_id": 268655}]}
{"text_id": 105526, "text": "药品", "implicit_entity": [{"subject": "肿节风软胶囊",
"subject_id": 53176}]}
```


2. 标题文本对应的两个实体都具有关系


```
{"text_id": 134542, "text": "正品米菲司同片铜片", "implicit_entity": [{"subject": "米菲司酮
片", "subject_id": 140181}]}
{"text_id": 21246, "text": "正品米菲司同片铜片", "implicit_entity": [{"subject": "司米安米非司
酮片", "subject_id": 134662}]}
```

数据处理

- 相同标题文本对应多个实体ID

  3. 标题文本对应的两个实体一个为正确标注每一个为错误标注


```
{"text_id": 132115, "text": "阿达帕林", "implicit_entity": [{"subject": "福牌阿胶阿胶片",
"subject_id": 216530}]}
{"text_id": 45692, "text": "阿达帕林", "implicit_entity": [{"subject": "维A酸乳膏",
"subject_id": 230257}]}
```


- 处理方式

  直接删除

数据处理

- 相似实体


```
{"type": "Medical", "subject_id": 172360, "subject": "肾石通颗粒", "data": [{"predicate": "生产企业", "object": "河北万岁药业有限公司"}, {"predicate": "主要成分", "object": "金钱草、王不留行（炒）、萹蓄、延胡索（醋制）、鸡内金(烫)、丹参、木香、瞿麦、牛膝、海金沙。"}, {"predicate": "症状", "object": null}, {"predicate": "规格", "object": ["15g*10袋(万岁)"]}, {"predicate": "功能", "object": null}]}

{"type": "Medical", "subject_id": 31946, "subject": "肾石通颗粒", "data": [{"predicate": "生产企业", "object": "修正药业集团股份有限公司"}, {"predicate": "主要成分", "object": "金钱草、王不留行（炒）、萹蓄、延胡索（醋制）、鸡内金（烫）、丹参、木香、瞿麦、牛膝、海金沙。"}, {"predicate": "症状", "object": null}, {"predicate": "规格", "object": ["15g*10袋(修正)"]}, {"predicate": "功能", "object": null}]}
```


处理方式：保留训练集出现的实体

召回模块

- 传统召回方式

1. TF-IDF、BM25
2. DSSM,CLSM

- 缺陷

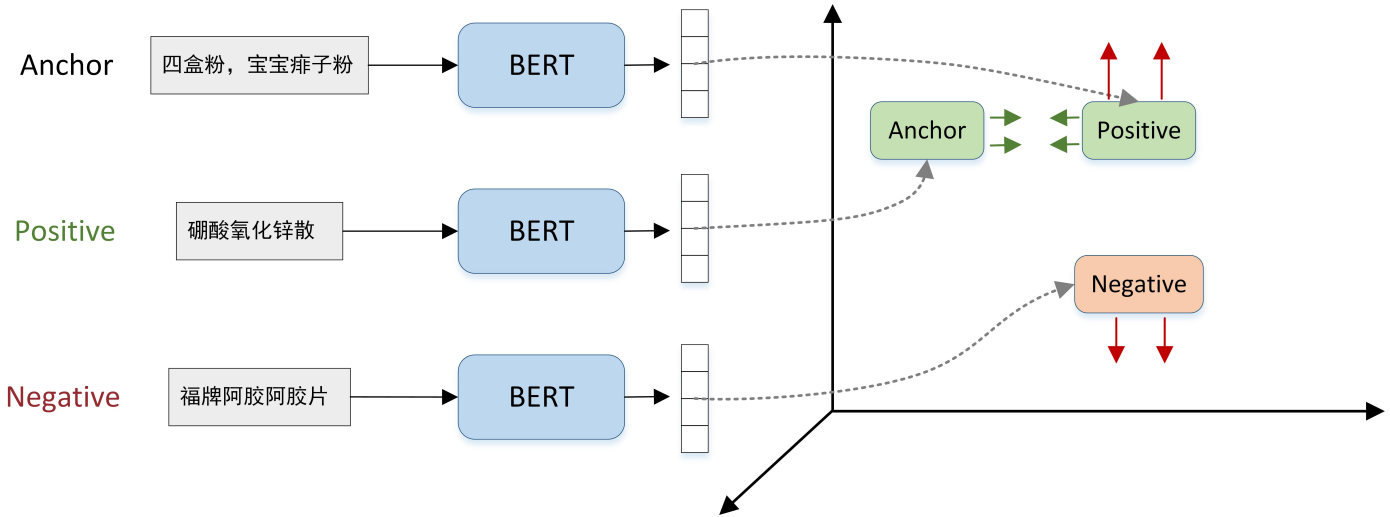
1. 传统的词特征不适用与当前数据集
2. 仅仅使用了静态词向量

- 方案

1. Triplet network
2. loss： Triplet loss
3. 网络：BERT

召回模块

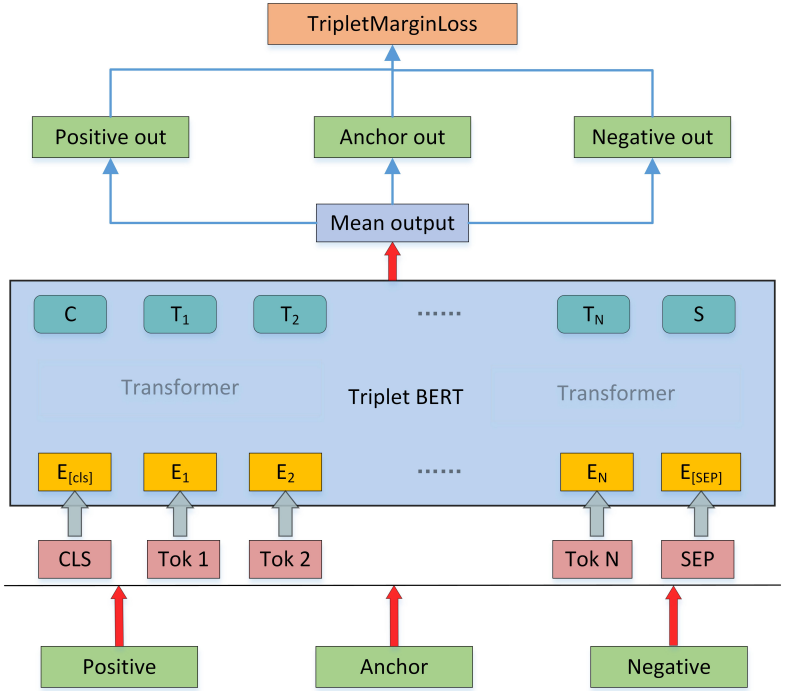
- Triplet loss



$$L\left(r_a, r_p, r_n\right) = \max\left(0, m + \mathrm{d}\left(r_a, r_p\right) - \mathrm{d}\left(r_a, r_n\right)\right)$$

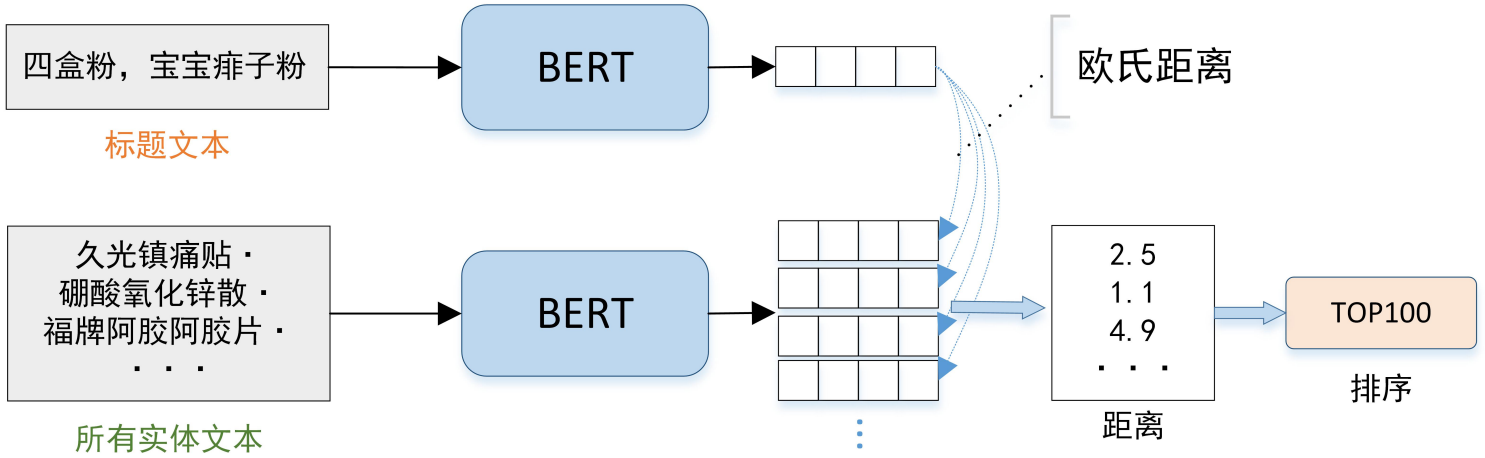
召回模块

- 训练
- 动态负采样：
  - 每个Batch随机选取负样本
- BERT模型
  1. ernie-1.0
  2. roberta-wwm
- 模型融合
  1. 训练集： 交叉验证
  2. 测试集：得分求平均



召回模块

- 预测



- 召回率：98.6%

排序模型

- 排序模型

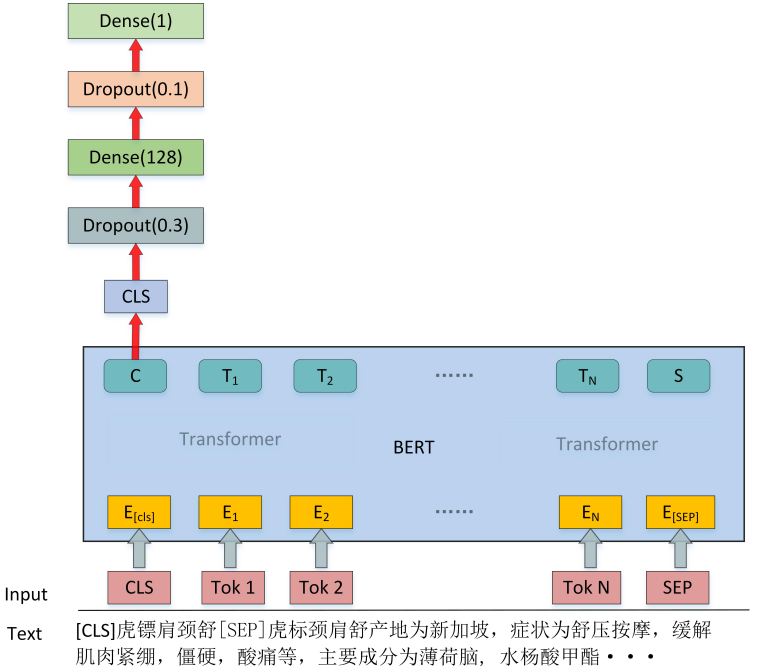
1. 二分类模型
2. 利用CLS向量特征

- 输入

标题文本：虎镖肩颈舒

实体描述文本：虎标颈肩舒产地为新加坡，症状为舒压按摩，缓解肌肉紧绷，僵硬，酸痛等，主要成分为薄荷脑, 水杨酸甲酯

连接方式：[CLS] 标题文本 [SEP] 实体描述文本 [SEP]



排序模型

- Top100 -> top10
- 动态负采样：
  - 每个Batch随机选取负样本
  - 在top100 中选取
  - 负样本个数：3个
- 预训练模型：

1. ernie-1.0
2. roberta-wwm

- 准确率：93%

排序模型

- Top10 -> top1
- 动态负采样：
  - 每个Batch随机选取负样本
  - 在top10 中选取
  - 负样本个数：2个
- 预训练模型

1. ernie-1.0
2. roberta-wwm
3. bert-wwt

- 准确率：83%

排序模型

- 最终成绩

最终得分 - CCKS 2020：基于标题的大规模商品实体检索

如果你发现有参赛者用多个账户参加比赛，请联系管理员.

#	队伍名	分数	最终提交次数
1	DeepBlueAI	0.88489	37
2	喵喵喵	0.87947	74
3	tonyxu	0.87744	70
4	HairLossKnight	0.87721	46
5	qianrenjian	0.86898	60

招聘

岗位： NLP、 CV、ML

bjhr@deepblueai.com

luozp@deepblueai.com

负责人微信

Q &A

THANKS