基于BERT的事件主体抽取

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

任务旨在从文本中抽取事件类型和对应的事件主体。即给定文本 T，抽取 T 中所有的事件类型集合 S，对于 S 中的每个事件类型 S ，从文本T 中抽取 S 的事件主体。其中各事件类型的主体实体类型为公司名称或人名或机构名称。

- **输入：** 一段文本T。
- **输出：**事件类型和事件主体。

示例

输入：“法院裁定公司A需向公司B一次性赔付500万” 。

输出：事件类型：“重大赔付”。 事件主体：“公司A”。

任务介绍

- 难点

1. 类别分布极度不均衡
2. 存在大量NAN数据
3. 句子杂乱，多个句子连在一起

- Pipeline方案

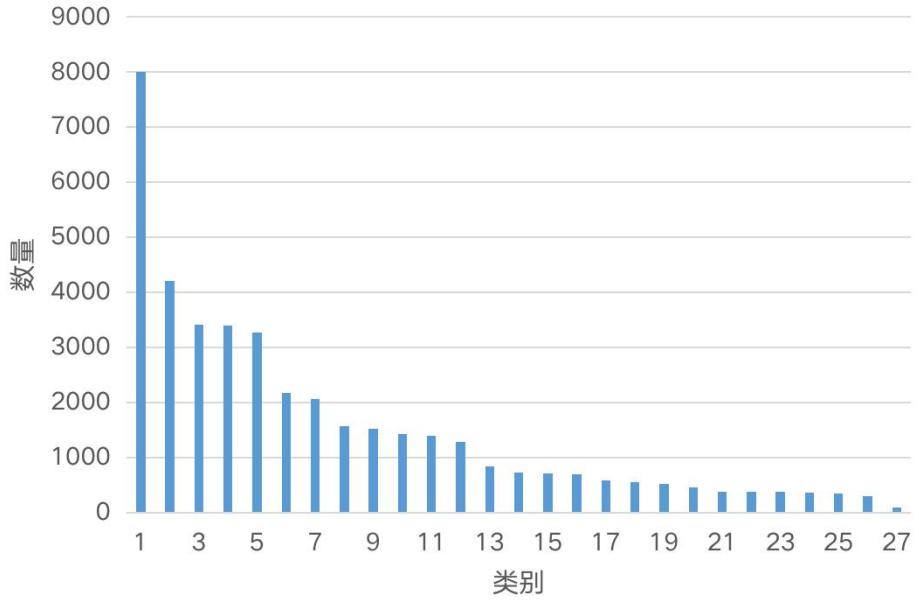
1. 事件类型识别

   多标签分类

2. 事件主体识别

   阅读理解

   序列标注



数据处理

- 数据增强

1. 事件主体随机替换
2. 每个Batch 随机替换20%事件主体
3. 提升模型的泛化性，使类型识别不依据特定实体

例1：

聚祥公司法人代表黄某、刘某等三人被深圳市人民检察院批准延长羁押期限一个月

权健公司法人代表黄某、刘某等三人被深圳市人民检察院批准延长羁押期限一个月

例2：

宏达股份(600331)副总经理辞职**ST北生**(600556)实施退市警示 去年营收不足千万

宏达股份(600331)副总经理辞职南宁百货(600556)实施退市警示 去年营收不足千万

数据处理

- 特征词提取

1. 少数类型特征词提取
    1. 针对该类别提取该类别对应的关键词
2. 解决类别不均衡问题

例：

履行连带担保责任：连带、担保、清偿

- 提取168个特征词

1. One-hot 编码
2. 句子中出现该关键词，则标1

事件类型识别

- 模型

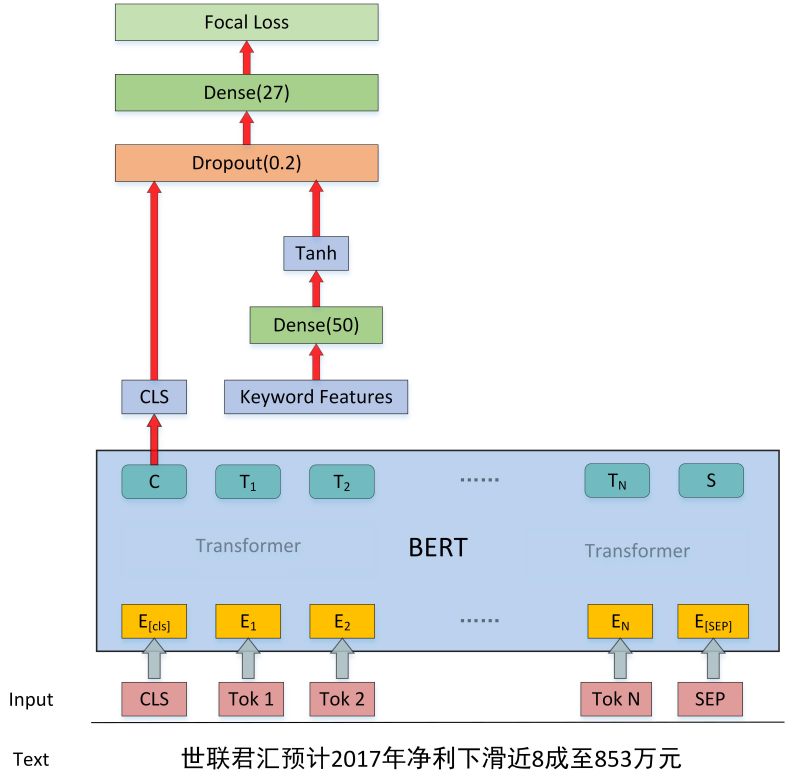
  1. CLS位置向量特征
  2. 提取特征词特征
  3. 多标签分类，NAN label 为0

- focal loss

  1. 解决类别不均衡问题

- 对抗训练

  1. FGM
  2. 提升模型鲁棒性



事件类型识别

- 训练细节

1. 6折交叉验证
2. chinese-roberta-wwm-ext
3. chinese-bert-wwm-ext
4. ernie-1.0
5. chinese-xlnet-base

- 模型融合

1. 概率求平均
2. 测试集30万文本，有类别的1000左右
3. 阈值调整：增大阈值使预测有类别的文本接近1000

事件主体识别

- 事件主体识别

  事件主体识别：即给定事件类型 S 和 文本 T ，从文本 T 中抽取 S 的事件主体。

  - **输入：**事件类型 和 一段文本T。
  - **输出：**该事件类型对应的事件主体。

  示例

  输入：“重大赔付”“法院裁定公司A需向公司B一次性赔付500万” 。
  输出：事件主体：“公司A”。

- 方案

  1. 基于BERT-LSTM-CRF的主体识别
  2. 基于MRC（阅读理解）的主体识别

事件主体识别

- BERT-LSTM-CRF

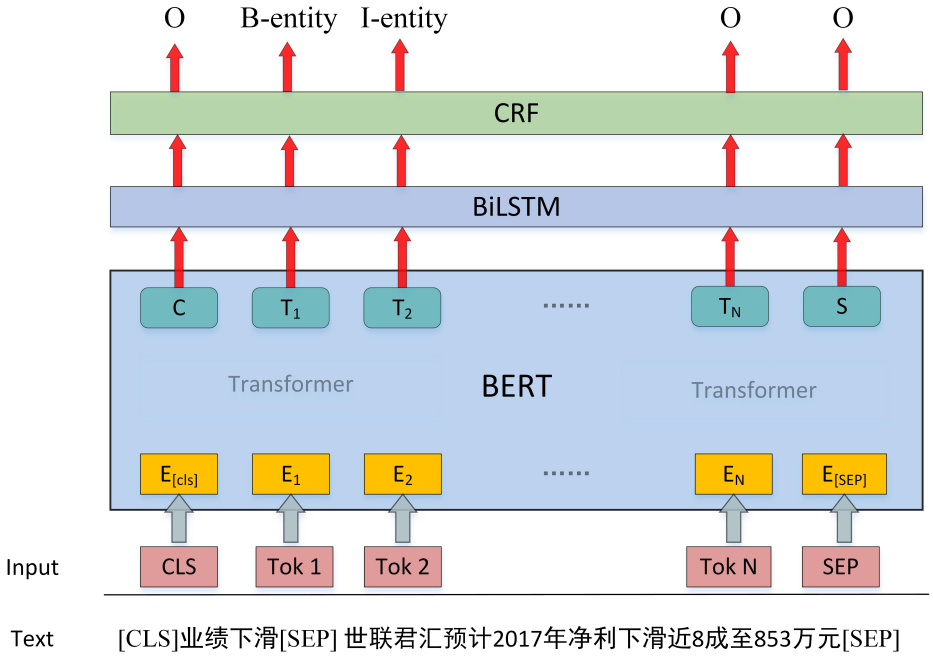
  1. 标注方式：BIO标注

  2. 输入：[CLS]事件类型[SEP]文本T[SEP]

- 训练细节

  1. FGM对抗训练

  2. 分层学习率：LSTM CRF 较大学习率



事件主体识别

- 基于MRC的主体识别

问题构建：

格式：

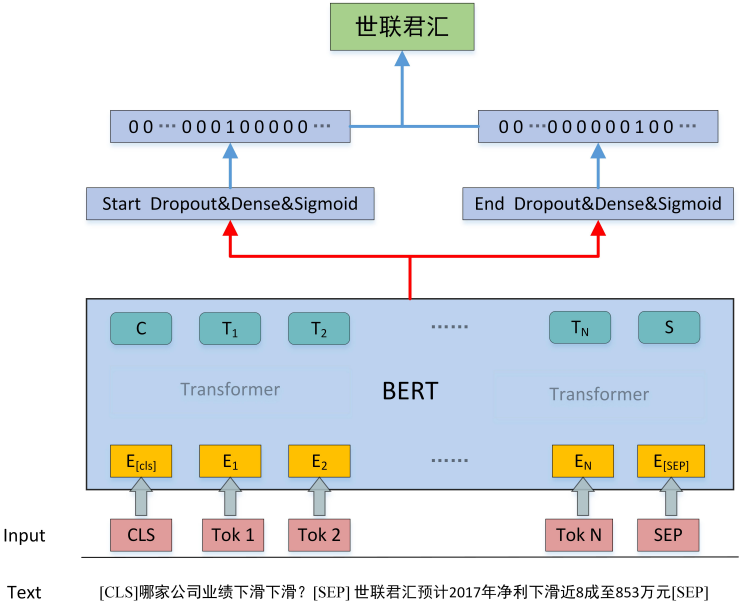
哪家公司 + 事件类型?

例：

哪家公司业务资产重组?

输入：问句与文本连在一起输入

[CLS]哪家公司业务资产重组? [SEP]当天，华菱钢铁股东大会审议通过包括关于重大资产置换在内的18个议案 [SEP]



事件主体识别

- 训练细节

模型	备注
BERT-CRF-chinese-roberta-wwm-ext	2组5折交叉验证
MRC-chinese-roberta-wwm-ext	1组5折交叉验证
MRC-chinese-bert-wwm-ext	1组5折交叉验证
MRC-ernie-1.0	1组5折交叉验证

- 模型融合

1. 模型投票
2. 共25个模型，票数13以上作为事件主体

结果分析

- 最终成绩

最终得分 - CCKS 2020：面向金融领域的篇章级事件主体与要素抽取（一）事件主体抽取

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	0.30781	9
2	人生苦短不会python	0.26632	8
3	Mark_4396	0.24624	5
4	DeepQ	0.24049	9
5	=★baseline★=	0.22077	6

招聘

岗位： NLP、 CV、ML

bjhr@deepblueai.com

luozp@deepblueai.com

负责人微信

Q &A

THANKS