基于增强语言模型和实体嵌入的实体链接技术

评测任务2：中文短文本实体链接 评测 第一名

报告人：潘春光

队伍：FREE

参数队员

潘春光
研二
文本分类、实体识别
实体链接
panchunguang@126.com

党金明
研二
关系抽取、实体链接
776039904@qq.com

杨智
研二
知识表示、实体识别
1780041410@qq.com

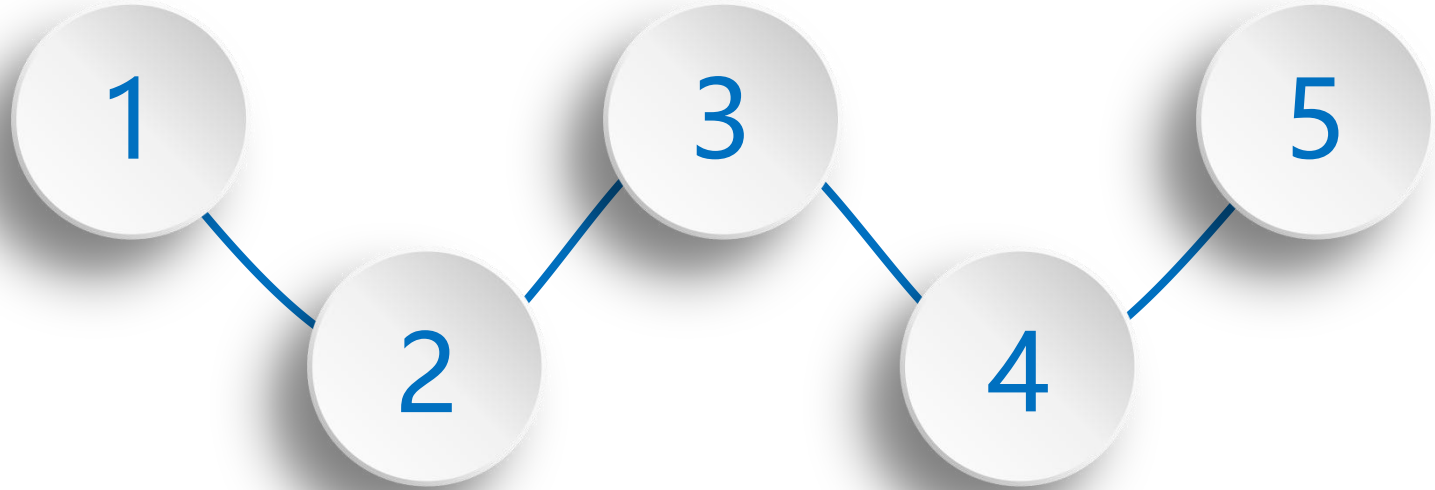
指导老师

张富
博士生导师
研究方向：知识图谱
zhangfu@cse.neu.edu.cn

程经纬
博士
研究方向：知识图谱
chengjingwei@cse.neu.edu.cn

东北大学-计算机学院-智能数据与知识工程实验室 *IDKE*

主目录



任务分析
Task Analysis

数据处理
Data Processing

实体识别
Entity Recognition

实体消歧
Entity Disambiguation

实验结果
Experimental Result

任务简介

实体链接

面向中文短文本的实体识别与链指（简称：ERL），即对于给定的一个中文短文本识别出其中的实体，并与给定知识库中的对应实体进行关联。ERL整个过程包括实体识别和实体链指两个子任务。

输入

```
{"text_id": "1",
 "text": "比特币吸粉无数，但央行的心另有所属|界面新闻 · jmedia"}
```

输出

```
"text_id":"1",
"text":"比特币吸粉无数，但央行的心另有所
        属|界面新闻 · jmedia"
"mention_data":[
        {
            "kb_id":"278410",
            "mention":"比特币",
            "offset":"0" },
        {
            "kb_id":"199602",
            "mention":"央行",
            "offset":"9" },
        {
            "kb_id":"215472",
            "mention":"界面新闻",
            "offset":"18" } ]
```

任务简介

训练数据

text 字段和 mention_data 字段，mention_data 里面包含连接的 mention 以及 kb_id。

知识库

subject_id，subject，alias，data 等字段，data 中包含多个 predicate、object。


```
"subject_id": "10001"
"subject": "胜利",
"alias": ["胜利"],
"type": ["Thing"],
"data": [
    {"predicate": "摘要", "object": "英雄联盟胜利系列皮肤是拳头公司制作的具有纪念意义限定系列皮肤之一。拳头公司制作的具有纪念意义限定系列皮肤还包括英雄联盟冠军系列皮肤。..."},
    {"predicate": "制作方", "object": "Riot Games"},
    {"predicate": "外文名", "object": "Victorious"},
    {"predicate": "义项描述", "object": "游戏《英雄联盟》胜利系列限定皮肤"}]
```

预处理

引入新的别名

原因

- 部分实体无法匹配：
  - 安妮‘海瑟薇：文本中间有特殊符号
  - 新浪微薄：文本中实体名错误
  - 国家质检总局：别名不在知识库中

方案

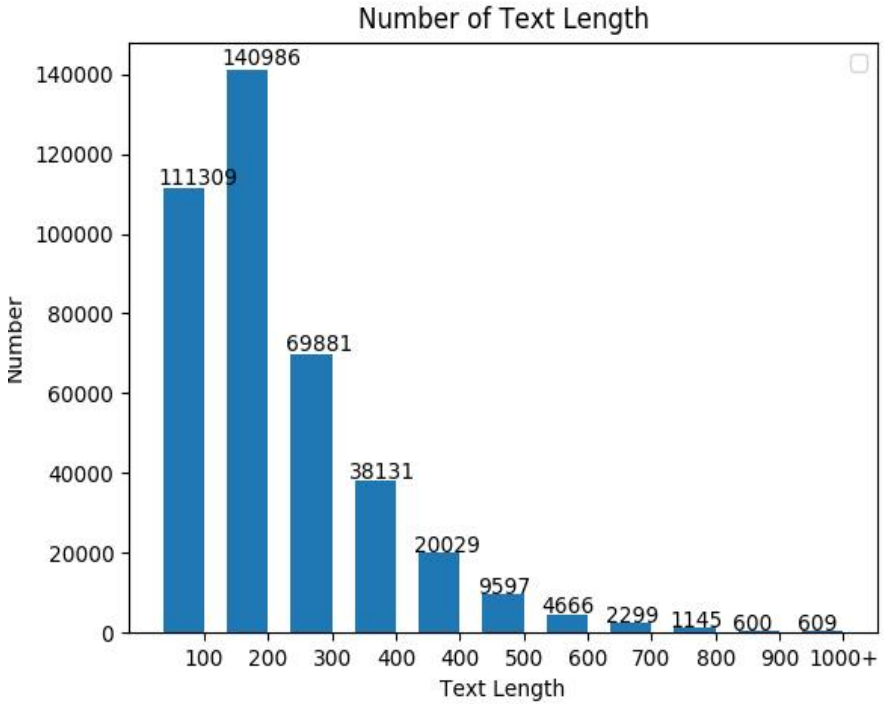
- 小写转换
- 特殊符号进行归一化处理
- 统计出现次数过多错误实体

```
'bilibili':{'总次数': 108, '哔哩哔哩': 94, '哔哩哔哩弹幕视频网': 5, '异地恋': 1, 'b站': 8}
```

预处理

实体描述文本

- 将所有的 predicate和object 相连可以得到实体描述文本。
- 文本过长
- 截断规则如下：
  - predicate项+object项的长度小于30 不截断
  - predicate项+object项的长度大于30按比例截断



实体识别

BERT-CRF 模型

- 存在问题
  - 实体边界错误
  - 识别实体不全
  - 没有利用到知识库的信息

BERT-EntityNameEmbedding 模型

- 充分利用知识库信息
- 弥补上述所有缺点

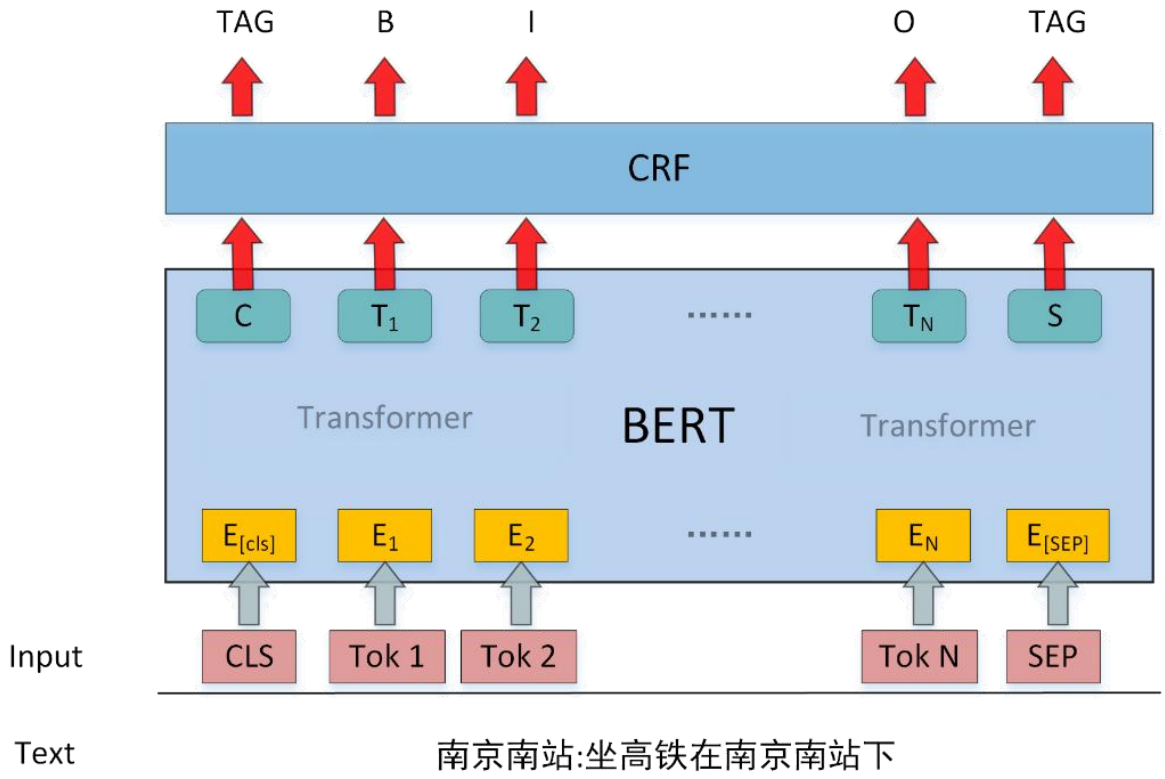
两个模型融合

实体识别

BERT-CRF

模型细节

- BIO 标记
- BERT 的[CLS]、[SEP] 位置用标签 TAG 表示
- 9折交叉验证



BERT-EntityNameEmbedding

BERT-ENE

基本思路

- 构建实体名称字典
- 实体名称嵌入
- 正向最大匹配
- BERT-ENE 模型对匹配的结果进行筛选

BERT-EntityNameEmbedding

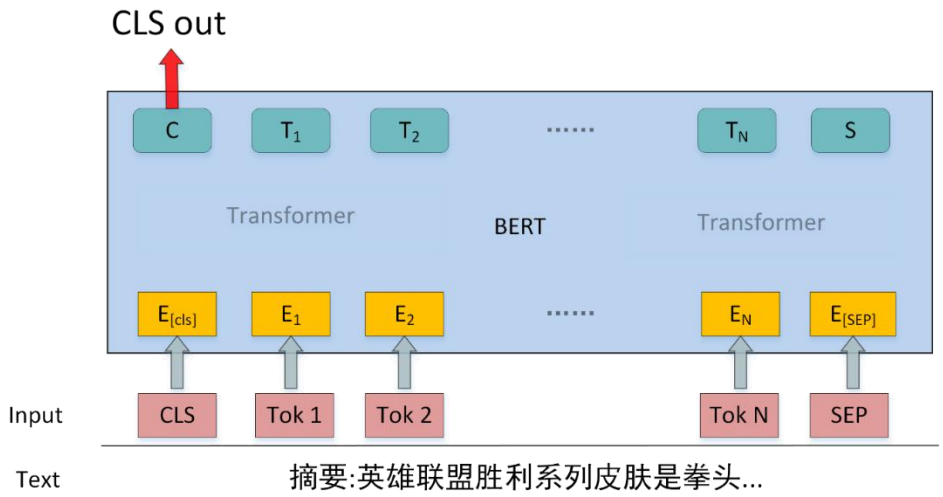
实体名称字典

```
{'胜利': ['10001', '19044', '37234', '38870', '40008', '85426', '86532', '140750']}
```

- 构建：实体名称、别名
- 形式：
  - key：实体名字
  - value：kb_id 列表

实体名称嵌入

- [CLS] 位置的输出向量
- 实体名称只对应单个实体，直接作为实体名称的嵌入
- 实体名称只对应多个实体，多个向量求平均



BERT-EntityNameEmbedding

正向最大匹配

- 全部匹配

```
text: 《大话英雄·联盟》-原创-高清视频
result: [('大话英雄·联盟', 1), ('联盟', 6), ('原创', 10), ('高清视频', 13), ('视频', 15)]
```

- 正向最大匹配

```
text: 《大话英雄·联盟》-原创-高清视频
result: [('大话英雄·联盟', 1), ('原创', 10), ('高清视频', 13)]
```

- 正确匹配

```
text: 《大话英雄·联盟》-原创-高清视频
result: [('大话英雄·联盟', 1), ('视频', 15)]
```

经过实验分析，采用正向最大匹配效果最好

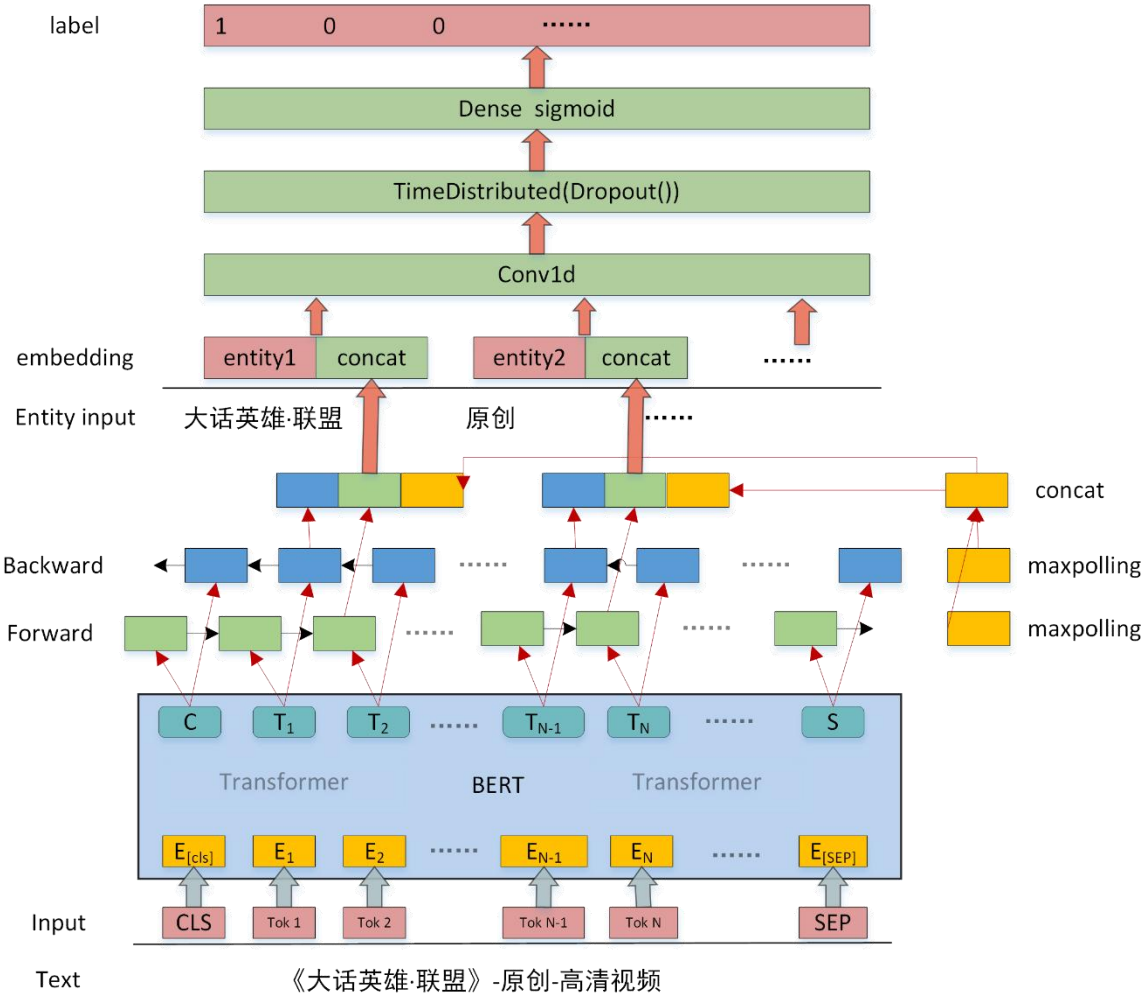
BERT-ENE

思路

- 短文本中实体名称的语义特征
- 知识库中实体的语义特征
- 通过计算两个语义特征的相似性进行分类

模型细节

- 文本语义特征
  - 实体名称右边界对应前向GRU向量
  - 实体名称左边界对应后向GRU向量
  - BIGRU 最大池化 向量
- 知识库特征
  - 实体名称对应嵌入



实体识别

模型融合

BERT-CRF & BERT-ENE

- 融合规则
  - 两个模型预测结果存在交叉，则选取 BERT-ENE 的结果
  - 单字实体选取 BERT-CRF 模型的结果

结果

Model	Precision	Recall	F1
BERT-CRF	0.8316	0.8121	0.8218
BERT-ENE	0.8224	0.8157	0.8191
BERT-CRF & BERT-ENE	0.8268	0.8534	0.8398

实体消歧

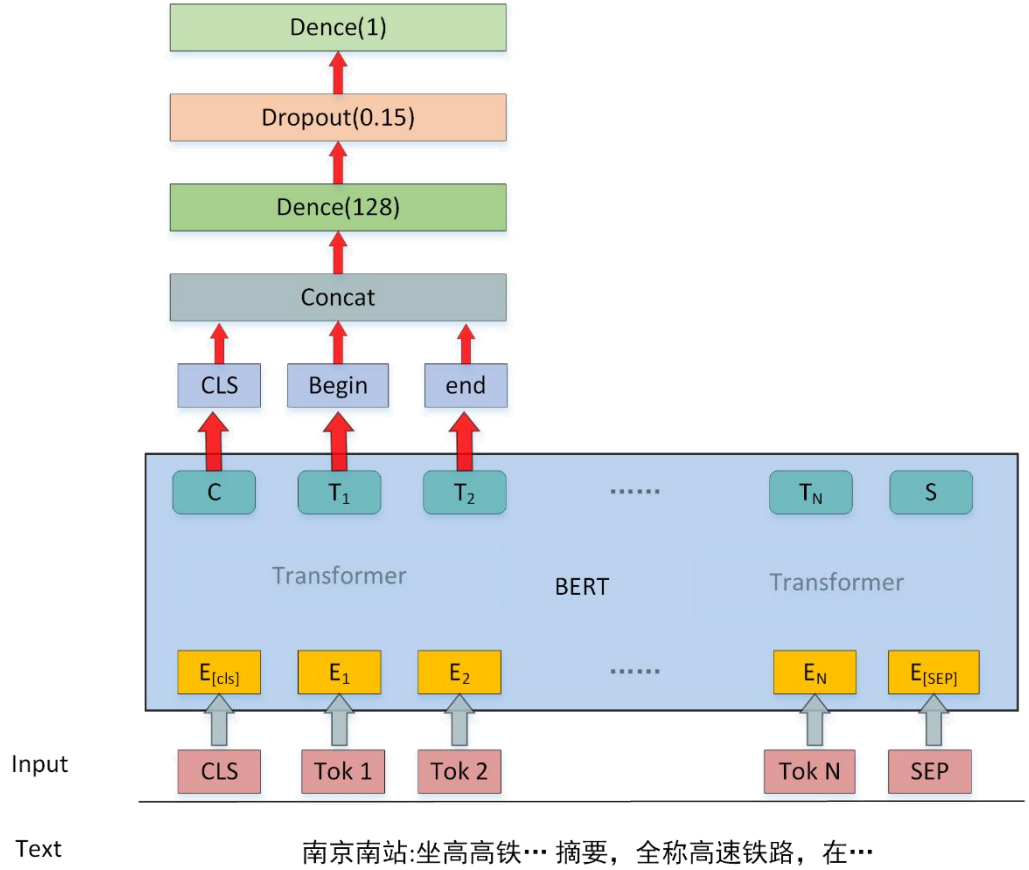
基于BERT二分类

思路

- 候选实体打分
- 打分排序

模型细节

- 输入：
  - 短文本
  - 待消歧实体的描述文本
- 特征：
  - CLS 位置向量
  - 实体开始位置向量
  - 实体结束位置向量



结果

Github：https://github.com/panchunguang/ccks_baidu_entity_link

初赛结果

#	Δ	队伍名	分数
1	—	FREE	0.78586
2	—	Team KG	0.77793
3	—	烟雾弹大师法棍诺	0.77730

复赛结果

#	队伍名	分数
1	FREE	0.80143
2	Team KG	0.79965
3	观	0.79654

感谢

Thanks!

恳请各位评委批评指正！