Schema约束的知识抽取系统架构

梁家卿
上海数眼科技发展有限公司
复旦大学知识工场

Outline

- 问题描述与数据集分析
- 抽取框架
  - 框架选型
  - 第一步：存在关系识别
  - 第二步：关系元素抽取
- 后检验
  - 先验不确定知识库：从训练集中构建的知识库
  - 知识一致性检验

问题描述与数据集分析

问题描述

- 给定Schema
  - 50个属性
  - 包含subject和object的概念约束

{"object_type": "地点",	"predicate": "祖籍",	"subject_type": "人物"}
{"object_type": "人物",	"predicate": "父亲",	"subject_type": "人物"}
{"object_type": "地点",	"predicate": "总部地点",	"subject_type": "企业"}
{"object_type": "地点",	"predicate": "出生地",	"subject_type": "人物"}
{"object_type": "目",	"predicate": "目",	"subject_type": "生物"}
{"object_type": "Number",	"predicate": "面积",	"subject_type": "行政区"}
{"object_type": "Text",	"predicate": "简称",	"subject_type": "机构"}
{"object_type": "Date",	"predicate": "上映时间",	"subject_type": "影视作品"}

- 输入
  - 自然语言句子
- 输出
  - 句子中包含的所有给定Schema相关的三元组

《李烈钧自述》是2011年11月1日人民日报出版社出版的图书，作者是李烈钧

o {"object": "人民日报出版社", "object_type": "出版社", "predicate": "出版社", "subject": "李烈钧自述", "subject_type": "书籍"}

o {"object": "李烈钧", "object_type": "人物", "predicate": "作者", "subject": "李烈钧自述", "subject_type": "图书作品"}

问题与数据集分析

- 问题：多分类多输出
  - 句子中可能包含多个关系
  - 对每个关系，可能有多个三元组
    - 这些三元组可能共享s或o
  - 对某个关系，s和o可能相同或者有重叠

- 数据集
  - 训练集中有大量缺失

抽取框架

框架选型

- End-to-End ?
  - 输出非常高维度，极度稀疏，难训练
    - E.g. 50 BIO encoding
- 分步抽取
  - 1. 实体识别->实体对关系分类
    - NER准确率 85%-90%
    - 大量无意义实体对

      《冰山上的来客》是戴冰执导的军事悬疑谍战片，由王洛勇、于荣光、努尔比亚等主演↓
      - => (戴冰，于荣光) ?
  - 2. 存在关系识别->抽取Sub/Obj
    - 效果估计：96% x 90%+
    - 类似事件抽取：先抽事件类型，再进行槽填充

第一步：存在关系识别

- 句子中存在哪些关系？

马志舟，1907年出生，陕西三原人，汉族，中国共产党，任红四团第一连连长，1907年逝世↓

=> 出生地，出生日期，民族，国籍

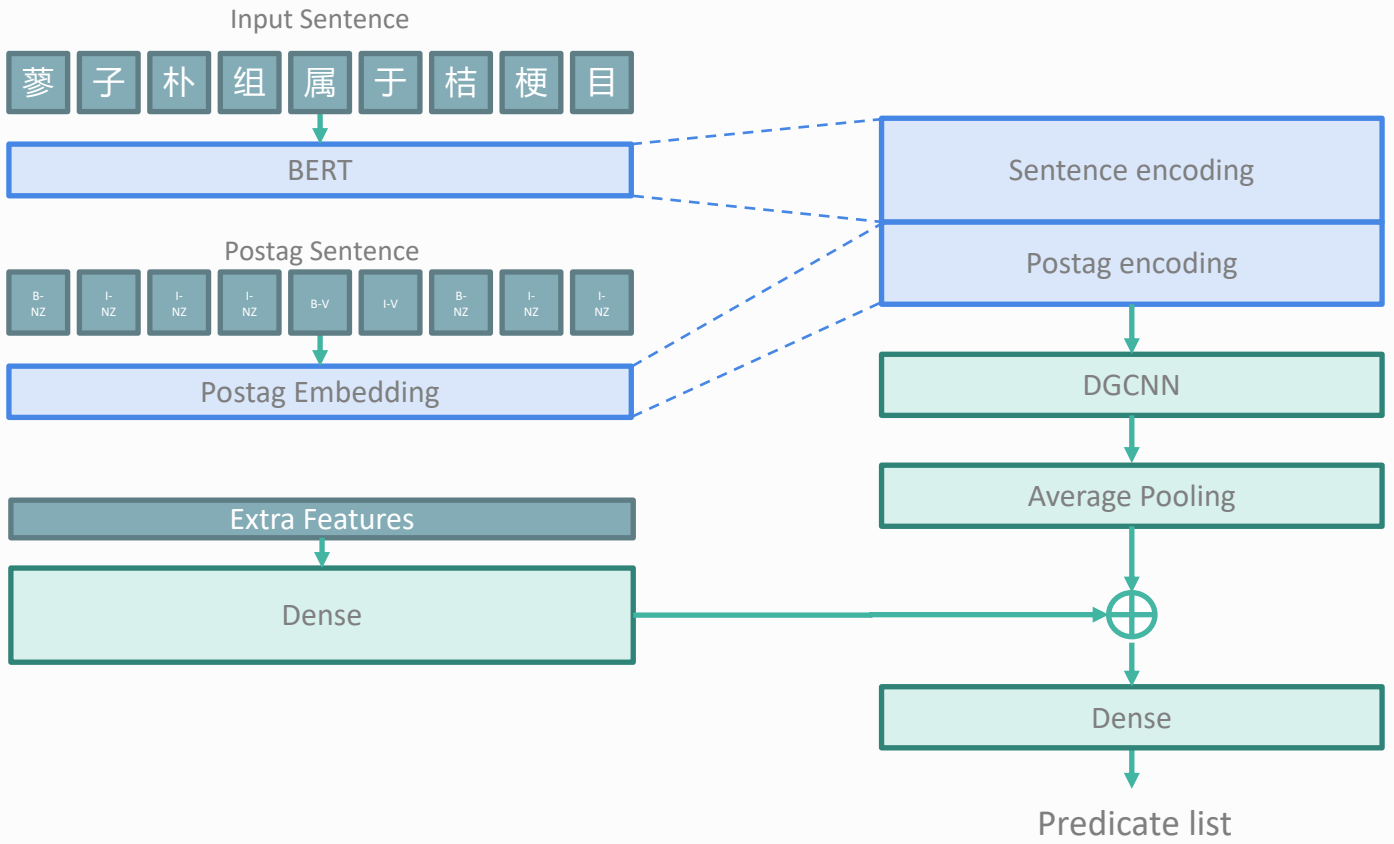
- 输入:
  - 句子：Token列表{ t_1 ,…,t_M }
- 输出:
  - 50类多分类输出 (50 relations)
    - 50个概率，互不冲突（ sigmoid输出，非softmax ）

基础模型

- BERT+DGCNN

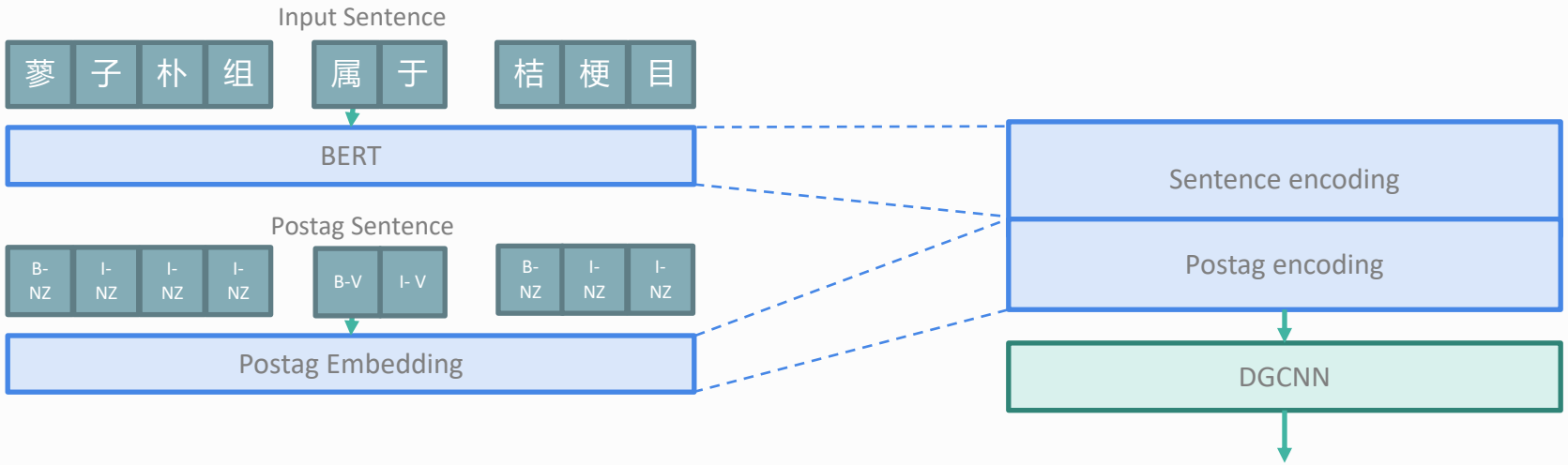
DGCNN(dilated gated CNN):

$$\boldsymbol{h} = \boldsymbol{x} \otimes \left(1 - \mathcal{C}_\sigma(\boldsymbol{x})\right) + \mathcal{C}(\boldsymbol{x}) \otimes \mathcal{C}_\sigma(\boldsymbol{x}),$$
$$\mathcal{C}_\sigma(\boldsymbol{x}) = \sigma\Big(\mathcal{C}(\boldsymbol{x}) \otimes (1 + \varepsilon)\Big),$$



额外特征：分词信息

- BERT中文版本是按字级别划分Token
  - 浪费了数据集提供的 分词/词性标注/NER信息
  - 将分词和词性标注信息以BIO编码输入模型



损失函数设计

$$\theta_i = \mathrm{sgn}\Big((y_i - 0.5)(y_{ti} - 0.5)\Big)$$

$$\delta_i = \begin{cases} 1, & 0.5 - m < y_i < 0.5 + m \text{ or } \theta_i < 0 \\ 0, & \text{Otherwise.} \end{cases}$$

边界指示函数：只有惩罚模型还未充分学到的样本而忽略已可分样本，以解决样本噪声和类别不平衡问题

原始交叉熵

$$L_R^* = -\frac{\sum_{i=1}^{N} \delta_i y_{ti} \log y_i}{\sum_{i=1}^{N} y_{ti}} - \mu \log\left(1 - \frac{\delta_i (1 - y_{ti}) y_i}{\sum_{i=1}^{N} (1 - y_{ti})}\right)$$

缓解训练样本的漏标问题，我们将未标注视为负样本，而负样本可能有假阴性，因此减弱负样本的惩罚有助于使模型更接近真实值。

第二步：关系元素抽取

- 枚举第一步预测的每个属性，从句子中标注Sub和Obj

- **输入:**
  - 属性，属性信息，句子
- **输出:**
  - 句子标注Subject和Object

模型

- 除了输出部分，模型和关系识别完全相同
- 输入的属性和属性信息与句子一同连成长序列输入
  - E.g.

    图书作品，作者，人物。《东方奥斯维辛》是1999年中国青年出版社出版的图书，作者是谭元亨

输出设计

- 常规：BIO encoding，使用CRF输出
  - E.g.
    - 作 曲 | 《 离 开 》是 由 张 宇 谱 曲 ， 演 唱
    - X X X O B-S I-S O O O B-O I-O O O O O O
  - S和O不能重叠，难以获取候选项后使用其他规则排序
- 双指针输出
  - 类似SQuAD等阅读理解数据集上的输出
  - 每个token有4个输出
    - 作为S开头的概率、作为S结尾的概率、作为O开头的概率、作为O结尾的概率
    - 这些概率不互斥（使用Sigmoid输出而非Softmax）
    - => 可以计算句子中每个子串作为S的概率和作为O的概率
    - 便于加入先验分数、选取多个候选项、阈值筛选
    - 对于较长的实体效果较好

输出解码过程

1. 分别产生Sub候选集合和Obj候选集合
   - 枚举句子中字串，按预测的起始概率和结束概率之积作为分数排序
2. 排除覆盖和相交部分
   - 若候选项中有覆盖和相交部分，则按分数删去其中一些使候选项不相交
     - E.g. 设“张三”，“李四”，“张三和李四”都在候选项中，
     - 若score(张三)+score(李四)> score(张三和李四)，则保留“张三”和“李四”，反之则只保留“张三和李四”。
3. 匹配S和O形成三元组
   - 候选项少，全匹配
   - 候选项多，最近匹配

后检验

NLP技术不能解决所有问题

MV拍摄花絮周杰伦与搭档方文山再次连手写出创新的中国风歌曲《雨下一整晚》，MV找来浪花兄弟的Darren担任男主↓

- => (方文山，雨下一整晚) 作词 or 作曲？

片名 卧虎藏龙导演 李安主演 周润发、杨紫琼、章子怡片名 大红灯笼高高挂导演 张艺谋主演 巩俐、何赛飞片名

- => (卧虎藏龙 ，李安) 主演 or 导演？
  (大红灯笼高高挂，张艺谋) 主演 or 导演？

王雷和李小萌这对夫妻可谓是娱乐圈中有名的神仙眷侣，他们夫妻二人不仅在演艺事业上有着自己独到的特点，在现实生活中也拥有幸福的人生↓

- => 谁是丈夫谁是妻子？

- 在语法层面很难解决，但人类很容易做出正确选择
  - 人类拥有常识：方文山是词人，不作曲
- 不允许外部数据
  - => 用数据集内部数据挖掘常识

知识一致性检验

- 先验不确定知识库：从训练集中构建的知识库
  - 挖掘实体的细分概念：
    - 方文山：人物，作词家 or 作曲家？
  - 对高度共现的概念进行区分
    - 作词 vs 作曲，男人 vs 女人， 导演 vs 主演， 作家 vs 编剧，etc…

    - 人物A观测到 99次作词，0次作曲 => A只会作词不会作曲
    - 人物B观测到 29次作词，30次作曲 => B既会作曲也会作词
    - 人物C观测到 2次作词，0次作曲 => C不确定身份

- 预测的结果和先验知识库不一致的会被删除
  - E.g. 方文山属于观测到只作词不作曲的人物 预测其作曲的三元组会被删除

其他Tricks

- 连续的顿号实体会只在句中保留第一个，若预测到第一个其他的也同时加入结果
- 使用jieba分词做数据增强
- 使用BIO输出的模型做回退
- 加入FAQ中定义的规则
- 采用BERT相同的优化器和设置进行训练
- …

- 最终效果：89.3% F1 on testset
- 投入使用效果：87.1% F1，单模型无Ensemble

共同构建万物之网

公众号

kw.fudan.edu.cn

www.shuyantech.com