基于阅读理解的文档级人物属性抽取

CCKS2022

刘资蕴， 张世奇，陈文亮

主要贡献

本文继马进的工作，发现人物属性抽取的序列级别标注数据会产生标注位置问题和属性值重合的问题，进而提出将任务转化为阅读理解任务，该任务的形式化描述见右图1，并基于此任务构造了一份基于阅读理解的文档级人物属性抽取数据集。

方法

1. 构造文档级人物属性抽取数据，步骤包括数据源获取、扩充人物属性、属性标准化、属性值过滤和介绍文本过滤等，流程见右图2。
2. 采用了BERT-MRC和BERT-CRF-MRC两种阅读理解基线模型对该数据进行评测。
3. 在模型输出的结果后添加归一化规则来输出最终的属性值。

结果

构建约44万条文档级任务属性抽取数据，数据基本信息见下表

数据量	446309
平均句长	7315
短文本平均句长	829
平均属性数量	4.6

模型抽取结果如下表

	BERT-MRC	BERT-CRF-MRC
短文本	88.63%	91.95%
长文本	72.36%	75.15%

讨论

本文并没有考虑比较复杂的属性，例如职业，作品等。并且属性的归一化采用简单的规则方式。在未来工作中可以考虑采用限制属性值的长度，或者引入词典，将这些属性也加入进来。并且考虑提高属性的归一化质量，来提升系统准确率。另外本文只采用了两种模型进行评测，对于该数据模式采用什么样的方法来抽取最有效仍需要大量的实验。

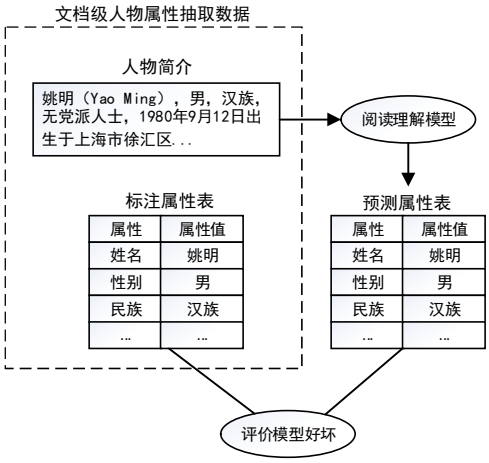

图1 文档级人物属性抽取数据形式样例

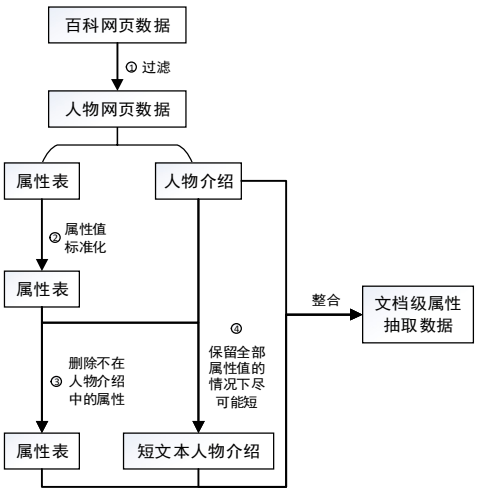

图2 数据构建框架

联系方式：
email zyliu129@stu.suda.edu.cn

2022全国知识图谱与语义计算大会
China Conference on Knowledge Graph and Semantic Computing
秦皇岛 8.24-8.27