全国知识图谱与语义计算大会，
CCKS-19，
2019年8月24-27，杭州

一种面向多需求的PDF文档信息抽取方法

余厚金 毛先领 黄河燕
北京理工大学计算机学院

I. 引言

目前，PDF已成为电子文档发行和数字化信息传播的一个标准，其广泛应用于学术界的交流以及各类公告的发行。如何从非结构化的PDF文档中抽取结构化数据是知识图谱领域所面临的一大挑战。本文利用Adobe公司开发的Acrobat DC SDK对PDF进行格式转换，从半结构化的中间文件进行信息抽取。相比已有的开源PDF解析方法，Acrobat导出的中间文件保存了更完整更准确的表格和文本段落信息，能应用于不同需求的信息抽取任务。

在CCKS 2019公众公司公告评测中，我们的方法获得总成绩第三名。在本次评测中，我们将公告文件（PDF格式）转换成XML。对于任务一，我们通过查找<Table>标签，获取PDF中所有的表格；然后根据表格的上下文，确定其名称，抽出符合条件的表格。对于任务二，我们首先抽出所有文本段落，使用Bi-LSTM-CRF进行命名实体识别，最后结合规则抽取信息点。

II.基于Acrobat DC SDK的PDF内容抽取系统

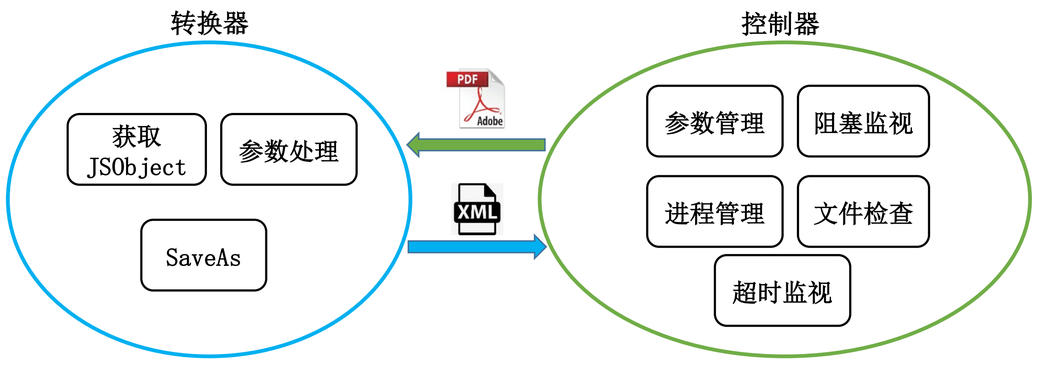

图 1. 基于Acrobat DC SDK的PDF内容抽取系统

左侧为转换器，包含3个模块，是PDF内容抽取的核心。右侧为控制器，包含5个模块，用于控制整个格式转换过程。

表 1. 转换格式之间的比较

格式	转换速度	能否直接提取表格	信息完整性	解析难度	解析速度
XML	快	**是**	好	**容易**	快
Word	慢	**是**	较好	一般	慢
Excel	较快	否	**很好**	较难	较快
TXT	**很快**	否	一般	难	**很快**
HTML	慢	**是**	很好	**容易**	较慢

抽取PDF文档中的表格可选择XML和Excel格式，中小规模文档集可选择Excel（召回率更高），XML更适合大规模文档集（效率和效果兼顾）。

III. 表格中的信息点抽取

XML文件中表格区域的树结构：

```
<Table>
  <TR>
    <TD>项目</TD>
    <TD>附注</TD>
    <TD>本期金额</TD>
    <TD>上期金额</TD>
  </TR>
</Table>
```

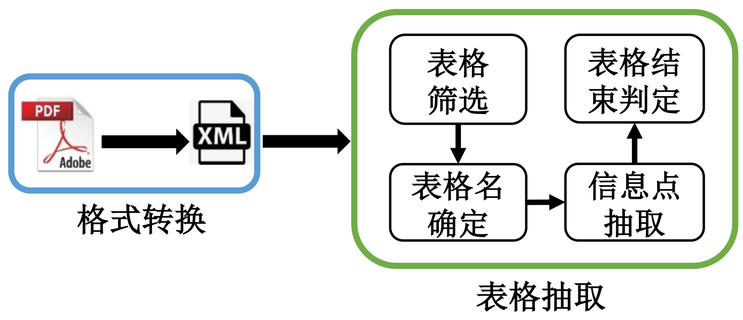

图 2. PDF文档表格抽取

查找<Table>标签获取所有表格，然后遍历表格，根据其上下文确定表格的名称，完成表格抽取。

表格结束判定：PDF文档中存在跨页的表格，会导致同一个表格在XML中被拆分。我们通过下一个表格的表头来判断其是否属于当前表格。为了增强鲁棒性，我们对每种表格统计其平均长度，以此设置最大表格长度。另外，构建项目名称字典，当抽取的表格项目名称80%出现在字典中，即断定该表格是需要抽取的表格。

注：在训练集上，我们的方法F1值达到了0.95，理论最佳F1值达到了0.99（忽略附注中的空格，以及不区分0和0.00）。

IV. 文本段落中的信息抽取

我们把这个任务建模为序列标注问题，首先进行命名实体识别，需要识别3类实体：人名，原因，职位；然后结合规则抽取信息点。首先获得BIO训练数据，然后训练Bi-LSTM-CRF，选择最好的模型进行预测。

训练数据生成：评测数据只提供JSON格式的信息点，因此需要生成序列标注的训练数据。由于JSON中的信息点几乎都是从文本中原样抽取（除了合并项），我们使用下面过程获得BIO标注数据：

（1）抽取XML中的文本段落，除去空白符，分句。

（2）抽取JSON中的信息点，得到所有子串。

（3）遍历所有句子，每个句子所有字初始标记为O；对每个子串，查找其在句子中的所有位置，分别标注B和I。

这种方式不能标注包含合并项的信息点，因为无法匹配到该子串。

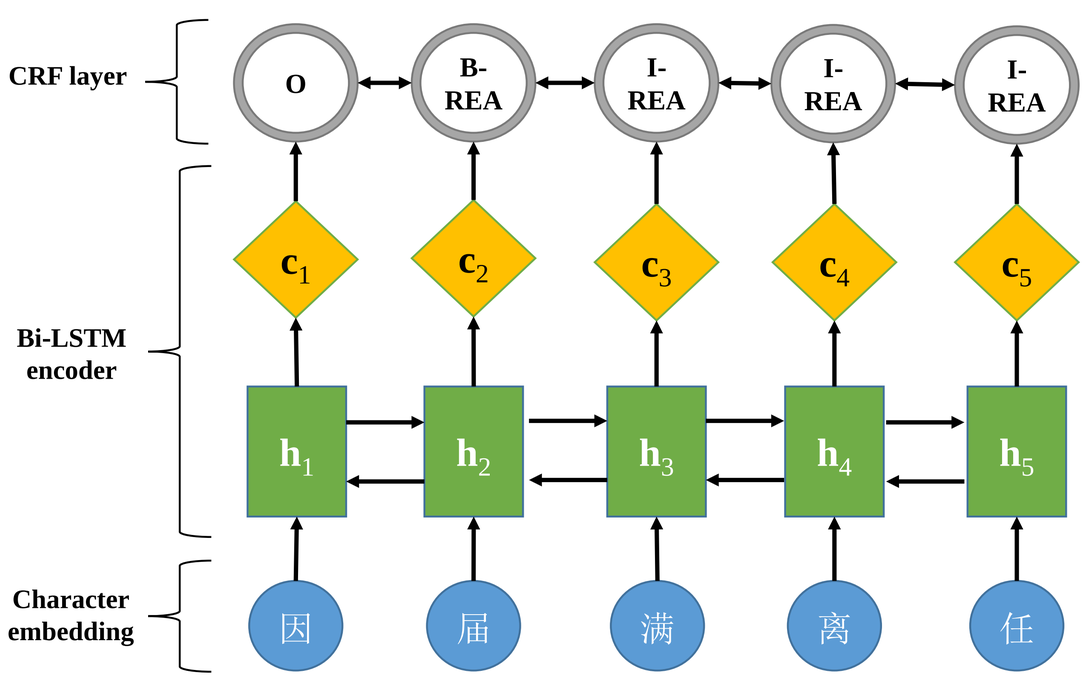

图 3. Bi-LSTM-CRF模型

第一层是字嵌入层，我们使用金融新闻预训练的词向量。第二层Bi-LSTM层可以有效地使用过去和未来的输入信息并自动提取特征。第三层CRF层给每个句子中的字打上BIO标签。

注：由于时间原因，我们只在句子级别进行信息点抽取，训练数据中确实存在跨句的信息点。