CRF项目总结报告

1. 项目介绍

1.1 项目目标：从交通肇事罪案件文书中获取每个案件发生的原因。

1.2 相关技术：crf条件随机场

1.3 相关工具和语言：工具：crf++工具包；语言：python

2 项目过程总结

2.1 项目实施步骤：

2.1.1 人工抽取事发原因作为训练集和验证集：

一共从928个案件中人工截取了对应案件的事发原因，其中80%作为训练集，20%验证集。

2.1.2 整理训练集和验证集格式：

把事发原因所在段落的句子整理成crf标准的训练集格式和验证集格式，训练文件由若干个句子组成（可以理解为若干个训练样例），不同案件的句子之间通过换行符分隔。每个句子可以有若干组标签，最后一组标签是标注。项目中用到四列，即第一列、第二列和第三列都是已知的数据，第四列是要预测的标注，根据第一列的词语、第二列的词性和第三列的关键词特征，预测第四列的标注。

2.1.3使用crf++工具训练模型：

使用crf++ 工具包对训练集进行训练得到模型，再使用该模型对验证集进行标注，然后用模型评价指标对标注结果进行验证和评估，通过修改特征模板和参数来改进模型。

2.1.4 整理测试集格式：

把未抽截取过事发原因的案件作为测试案件，从中取出每个案件基本信息的句子，对句子进行分词和标注，整理成crf标准测试集格式。测试文件同训练文件类似，每个句子可以有若干组标签，但最后一组标签设为固定值。

2.1.5 提取测试集中案件的事发原因：

用crf++生成的模型提取测试集中案件的事发原因

2.1.6 与基于规则的模型对比

用基于CRF模型的性能与基于规则匹配模型的性能进行对比。

2.2 crf++训练方案

2.2.1特征选择：

在CRF标注算法中，模型的性能取决于特征的选择和特征函数的定义。 CRF模型的特征一般分为三类：原子特征，复合特征和全局变量特征。 这个项目中选择的特征包括：单词，词性和事件原因的关键词。具体特征类别如表1所示：

表1 特征类别

|  |  |  |
| --- | --- | --- |
| 特征 | 特征内容 | 样例 |
| 词 | 词语本身 | 如：超速、超载等 |
| 词性 | 词性标签（结巴分词的标准） | 如：v、n、f 等 |
| 关键词 | 原因开始的关键词 | 如：因、由于等 |
| 原因结束的关键词 | 如：时、处、中等 |
| 涉及驾车行驶的动词 | 如：起步、倒车、刹车、侧滑、操作等 |
| 涉及交通法规的名词 | 如：安全、文明、规范、原则、规定等 |
| 涉及事故发生状态的关键词 | 如：不周、不慎、疏于、大意、失控等 |
| 涉及案发环境的关键词 | 如：施工、狭窄、冰雪、湿滑、夜间等 |
| 非关键词 |  |

2.2.2 标注符号

关键词的特征标签与结果的词位标注的符号如表2所示：

表2 序列标注符号含义

|  |  |  |  |
| --- | --- | --- | --- |
| 关键词符号 | 含义 | 词位符号 | 含义 |
| Start | 原因开始的关键词 | B | 原因短语开始词 |
| End | 原因结束的关键词 | M | 原因短语中间词 |
| Drive | 涉及驾车行驶的动词 | E | 原因短语结尾词 |
| Rule | 涉及交通法规的名词 | O | 单个词的原因短语 |
| Status | 涉及事故发生状态的关键词 | N | 非原因词语 |
| Ement | 涉及案发环境的关键词 |  |  |
| N | 非关键词 |  |  |

2.2.3 数据集准备

从案件文书中获取包含事发原因的固定位置的段落，给段落分词，得到词和词性，再添加关键词的特征标签和结果的词位标注，构造crf++ 的训练集和验证集的标准数据格式。生成的部分训练文件如图1所示：

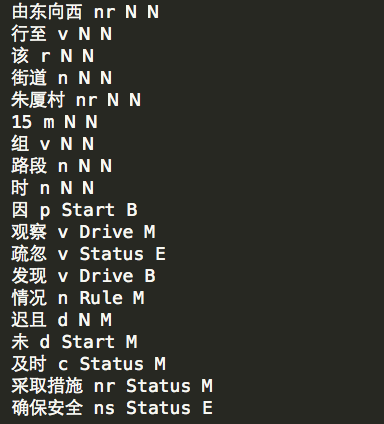


图1 部分训练文件

2.2.4 模板文件：

模版文件中的每一行是一个模板。每个模板都是由%x[row,col]来指定输入数据中的一个token。row指定到当前token的行偏移，col指定列位置。在构建特征模板时，使用原子特征和复合特征，上下文特征窗口为3。部分模版文件如图2所示：

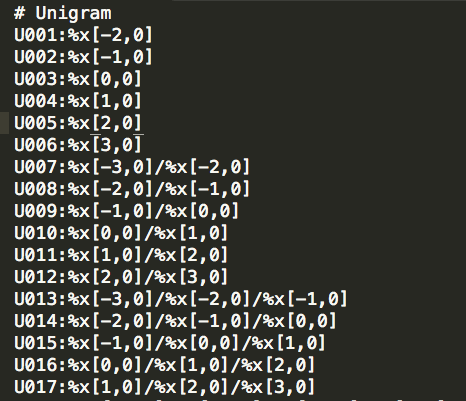
****

图2 部分模版文件

3 项目结果

3.1 生成模型的评价指标

准确率（P）= 正确标注的词数 /  总共标注的词数 \* 100%

召回率（R）= 正确标注的词数 /  标准结果中标注的词数 \* 100%

F 值  = 正确率 \* 召回率 \* 2 / (正确率 + 召回率)

（F 值即为正确率和召回率的调和平均值）

3.2 基于CRF的模型性能评估

基于CRF的试验分为三组，第一组：只考虑词本身；第二组：基于第一组，加入词性特征；第三组：基于第三组，加入关键词特征。三组试验模型的评估结果如表3所示：

表3 基于CRF的试验结果 ％

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 结果 | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  | | --- | | 第1组（词） | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | P | R | F |  |  | | | |  | | --- | | 第2组（词+词性） | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | P | R | F |  |  |  | | | |  | | --- | | 第3组（词+词性+关键词） | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | P | R | F |  |  | | | |
| B | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 75.07 | 77.13 | 76.09 | 79.30 | 75.91 | 77.57 | 82.70 | 80.18 | 81.42 | | 87.39 | 83.97 | 85.64 | 85.89 | 90.42 | 88.10 | 91.45 | 88.86 | 90.13 | | 78.99 | 81.4024 | 80.18 | 84.08 | 80.49 | 82.24 | 86.88 | 84.76 | 85.80 | | 85.71 | 50.00 | 63.16 | 83.33 | 41.67 | 55.56 | 58.33 | 58.33 | 58.33 | |
| M |
| E |
| O |
| 总体值 | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 83.18 | 81.90 | 82.54 | 84.34 | 85.33 | 84.83 | 88.63 | 86.16 | 87.38 | |

4 与基于规则的模型对比

用基于CRF模型的性能与基于规则匹配模型的性能进行对比。取两种基于规则匹配的模型。

4.1模型一：根据开始关键词和结束关键词匹配。

从开始关键词到结束关键词视为一个原因短语，即遇到开始关键词则标注 B ，B之后的词标注M，直到遇到结束关键词标注 E。对比结果如表4所示：

表4 CRF与基于规则（匹配开始与结束关键词）的试验比较 ％

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 结果 | CRF（词 + 词性 +关键词特征）   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | P | R | F |  |  |  | | 基于规则（匹配开始与结束关键词）   |  |  |  | | --- | --- | --- | | P | R | F | |
| B | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 82.70 | 80.18 | 81.42 | 75.38 | 44.82 | 56.21 | | 91.45 | 88.86 | 90.13 | 49.57 | 78.20 | 60.68 | | 86.88 | 84.76 | 85.80 | 13.04 | 01.83 | 03.20 | | 58.33 | 58.33 | 58.33 | 0 | 0 | 0 | | |
| M |
| E |
| O |
| 总体值（B，M，E，O） | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 88.63 | 86.16 | 87.38 | 51.37 | 56.36 | 53.75 | | |

4.2 模型二：根据开始关键词和固定词数匹配。

从开始关键词取词数为n的短语作为一个原因短语，即遇到开始关键词则标注 B ，从B开始的第 n 个词标注为E，中间的词标注M。分别取词数 n 为3，4，5，6，7 。对比结果如表5和表6所示：

表5 基于规则（匹配开始关键词与固定词数）的试验结果 ％

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 结果 | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  | | --- | | 词数为3 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | P | R | F |  |  | | | |  | | --- | | 词数为4 | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | P | R | F |  |  |  | | | |  | | --- | | 词数为5 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | P | R | F |  |  | | | |
| B | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 69.80 | 54.27 | 61.06 | 70.17 | 50.91 | 59.01 | 70.086 | 50.00 | 58.36 | | 79.61 | 19.84 | 31.77 | 73.43 | 33.24 | 45.76 | 67.44 | 42.72 | 52.30 | | 18.83 | 13.72 | 15.87 | 22.84 | 13.72 | 17.14 | 13.10 | 06.71 | 08.87 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| M |
| E |
| O |
| 总体值 | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 56.88 | 25.19 | 34.92 | 61.47 | 32.64 | 42.64 | 59.33 | 36.84 | 45.46 | |

表6 基于规则（匹配开始关键词与固定词长）的试验结果 ％

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 结果 | 词数为6   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | P | R | F |  |  |  | | 词数为7   |  |  |  | | --- | --- | --- | | P | R | F | |
| B | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 71.68 | 49.39 | 58.48 | 70.78 | 47.26 | 56.67 | | 65.56 | 50.05 | 56.76 | 62.98 | 55.72 | 59.13 | | 23.49 | 10.67 | 14.68 | 17.97 | 07.01 | 10.09 | | 0 | 0 | 0 | 0 | 0 | 0 | | |
| M |
| E |
| O |
| 总体值（B，M，E，O） | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 61.33 | 41.93 | 49.81 | 59.74 | 44.23 | 50.83 | | |

5 项目结果分析

5.1 对比试验结果

所有试验模型的评估结果如图3所示，实验显示：基于CRF的模型性能优于基于规则匹配模型的性能。其中，词＋词性＋关键词的特征组合效果最佳。

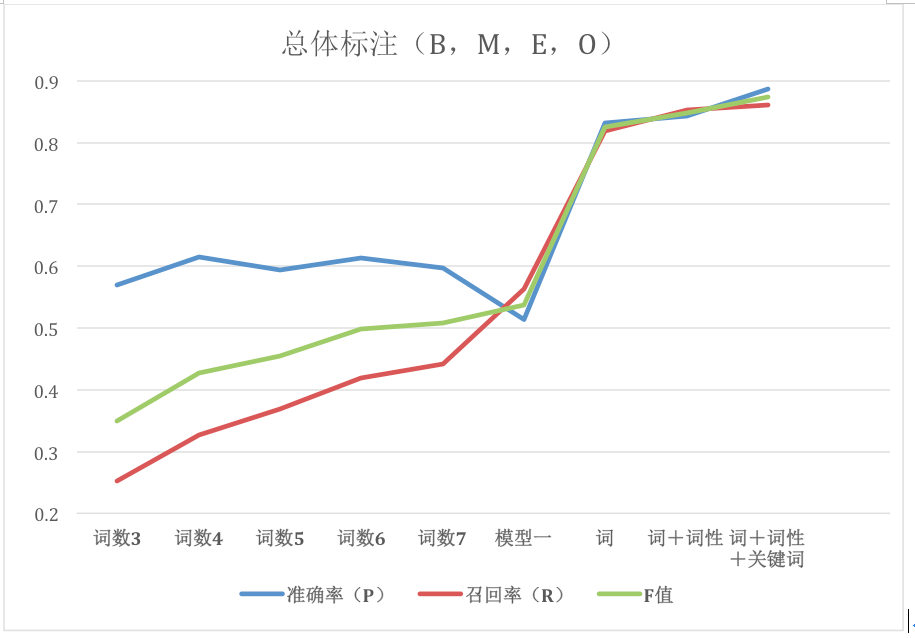


图3 试验结果

5.2 原因分析：

两个基于规则模型共同缺点：开始关键词不能涵盖所有起始位置；单独词的原因短语不能匹配到。此外，模型一中结束位置没有明显规律，标注B之后很难遇到结束关键词，导致B之后的大量非原因词被标注为M，造成准确率低下。模型二中原因短语包含的词数范围跨度大且词数不固定，导致标注的错误率较高。