# 一种面向公安警情领域的事件抽取方法

邓秋严 谢松县 郑菲 程琛 彭立宏 曾道建

湖南数定智能科技有限公司,长沙

广州市公安局指挥中心情报信息处,广州

湖南师范大学,长沙

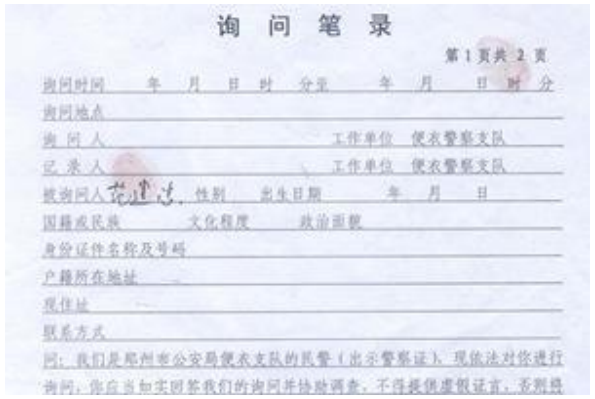
# 任务背景

## 警情文本

110接警

接警文本

警方询问

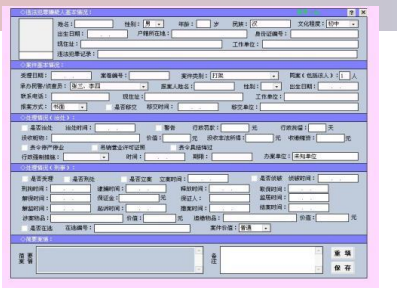



笔录文本

# 任务背景

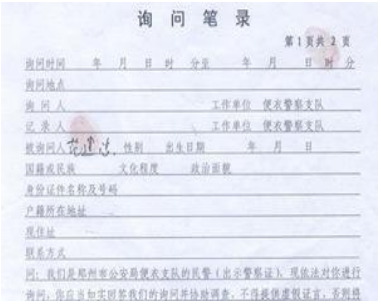
## 警情文本

110接警



接警文本

警方询问




笔录文本

目标：从接警文本、笔录文本提取结构化的事件信息

例句：李华在家吸食海洛因

| | |
|---|---|
| 事件类型 | 吸毒事件 |
| 吸食者 | 李华 |
| 吸毒地点 | 家里 |
| 毒品类型 | 海洛因 |

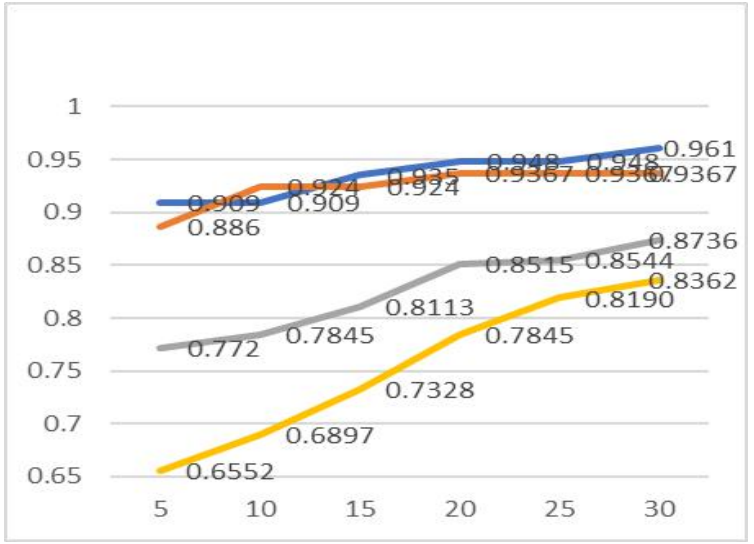
# 任务背景

## 警情文本特点

| 例句 | 现象 |
| --- | --- |
| 例1：我吸的是冰毒，在家里吸的，大约吸了0.1g。 | 触发词冗余 |
| 例2：我就问他一只“货”（冰毒）多少钱，他说600元，我说拿一只给我，他就放了两小包在老地方，我就从老地方将毒品拿回来了。 | 触发词不明显 |

某些事件类型触发词覆盖率较低

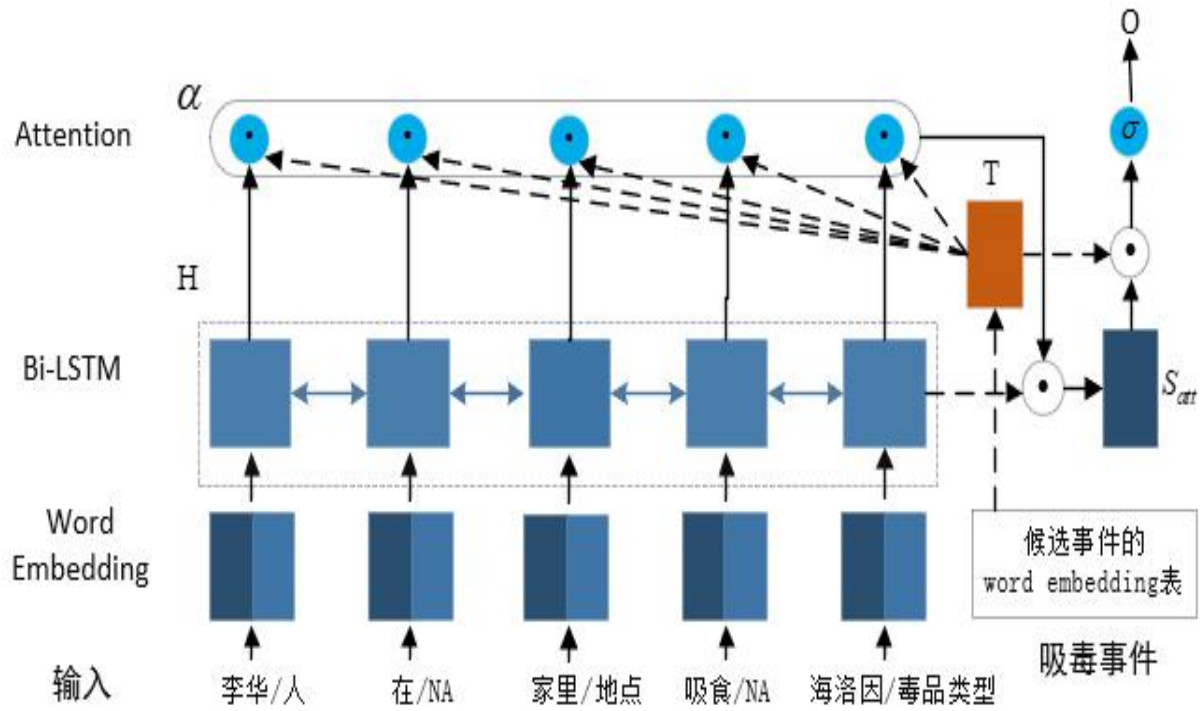
| 核心词个数 | 吸毒事件 | 贩毒事件 | 资金流入事件 | 资金流出事件 |
| --- | --- | --- | --- | --- |
| 5 | 0.9090 | 0.8860 | 0.7720 | 0.6552 |
| 10 | 0.9090 | 0.9240 | 0.7845 | 0.6897 |
| 15 | 0.9350 | 0.9240 | 0.8113 | 0.7328 |
| 20 | 0.9480 | 0.9367 | 0.8515 | 0.7845 |
| 25 | 0.9480 | 0.9367 | 0.8544 | 0.8190 |
| 30 | 0.9610 | 0.9367 | 0.8736 | 0.8362 |




基于无触发词的事件识别方法

# 无触发词的事件识别方法

网络结构




输入：基于警情数据自训练的命名实体识别

# 数据集

来源

- 5694个涉毒案件、诈骗案件的笔录文本
- 事件类型及角色

| 事件类型 | 子事件类型 | 论元角色 |
| --- | --- | --- |
| 涉毒事件 | 吸毒事件 | 吸毒时间、吸毒地点、吸食者、吸毒类型、吸毒剂量、吸毒方式 |
| | 贩毒事件 | 贩毒时间、贩毒地点、贩毒者、购买毒品者、贩毒类型、贩毒数量、贩毒金额 |
| | 购买毒品事件 | 购买毒品时间、购买毒品地点、购买毒品者、贩毒者、购买毒品类型、购买毒品数量、购买毒品金额 |
| | 持毒事件 | 持毒时间、持毒地点、持毒者、持毒类型、持毒数量 |
| 诈骗事件 | 资金流入事件 | 时间、地点、流入资金额、流入资金账号、流出资金方式 |
| | 资金流出事件 | 时间、地点、流出资金额、流出资金账号、流出资金方式 |

- 8105条标注数据，包含10117个标注事件

# 无触发词的事件识别性能分析

**GRU+CRF:**通过TF-IDF方法获得触发词，将任务转化成序列标注
**DMCNN:** 动态多池化神经网络方法

事件识别结果

| 模型 | 吸毒事件 | 贩毒事件 | 购买毒品事件 | 持毒事件 | 资金流入事件 | 资金流出事件 | 微平均值 | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | F | F | F | F | F | P | R | F |
| GRU+CRF | 0.662 | 0.723 | 0.613 | 0.794 | 0.585 | 0.758 | 0.831 | 0.735 | 0.780 |
| DMCNN | 0.954 | 0.892 | 0.742 | 0.892 | 0.667 | 0.901 | **0.962** | 0.809 | 0.880 |
| **ours/字** | **0.966** | 0.902 | **0.968** | 0.931 | 0.823 | 0.989 | 0.949 | 0.935 | **0.942** |
| **ours/词** | 0.964 | 0.840 | 0.941 | 0.928 | 0.740 | **0.991** | 0.938 | 0.932 | 0.935 |

# 无触发词的事件识别性能分析

**GRU+CRF:**通过TF-IDF方法获得触发词，将任务转化成序列标注
**DMCNN:** 动态多池化神经网络方法

事件识别结果

| 模型 | 吸毒事件 | 贩毒事件 | 购买毒品事件 | 持毒事件 | 资金流入事件 | 资金流出事件 | 微平均值 | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | F | F | F | F | F | P | R | F |
| GRU+CRF | 0.662 | 0.723 | 0.613 | 0.794 | 0.585 | 0.758 | 0.831 | 0.735 | 0.780 |
| DMCNN | 0.954 | 0.892 | 0.742 | 0.892 | 0.667 | 0.901 | **0.962** | 0.809 | 0.880 |
| **ours/字** | **0.966** | 0.902 | **0.968** | 0.931 | 0.823 | 0.989 | 0.949 | 0.935 | **0.942** |
| **ours/词** | 0.964 | 0.840 | 0.941 | 0.928 | 0.740 | **0.991** | 0.938 | 0.932 | 0.935 |

| 核心词个数 | 吸毒事件 | 贩毒事件 | 资金流入事件 | 资金流出事件 |
|---|---|---|---|---|
| 5 | 0.9090 | 0.8860 | 0.7720 | 0.6552 |
| 10 | 0.9090 | 0.9240 | 0.7845 | 0.6897 |
| 15 | 0.9350 | 0.9240 | 0.8113 | 0.7328 |
| 20 | 0.9480 | 0.9367 | 0.8515 | 0.7845 |
| 25 | 0.9480 | 0.9367 | 0.8544 | 0.8190 |
| 30 | 0.9610 | 0.9367 | 0.8736 | 0.8362 |

触发词覆盖率高的事件识别性能影响较小，两种方式的性能相差较小

# 无触发词的事件识别性能分析

**GRU+CRF:**通过TF-IDF方法获得触发词，将任务转化成序列标注
**DMCNN:** 动态多池化神经网络方法

事件识别结果

| 模型 | 吸毒事件 | 贩毒事件 | 购买毒品事件 | 持毒事件 | 资金流入事件 | 资金流出事件 | 微平均值 | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | F | F | F | F | F | P | R | F |
| GRU+CRF | 0.662 | 0.723 | 0.613 | 0.794 | 0.585 | 0.758 | 0.831 | 0.735 | 0.780 |
| DMCNN | 0.954 | 0.892 | 0.742 | 0.892 | 0.667 | 0.901 | **0.962** | 0.809 | 0.880 |
| **ours/字** | **0.966** | 0.902 | **0.968** | 0.931 | 0.823 | 0.989 | 0.949 | 0.935 | **0.942** |
| **ours/词** | 0.964 | 0.840 | 0.941 | 0.928 | 0.740 | **0.991** | 0.938 | 0.932 | 0.935 |

| 核心词个数 | 吸毒事件 | 贩毒事件 | 资金流入事件 | 资金流出事件 |
|---|---|---|---|---|
| 5 | 0.9090 | 0.8860 | 0.7720 | 0.6552 |
| 10 | 0.9090 | 0.9240 | 0.7845 | 0.6897 |
| 15 | 0.9350 | 0.9240 | 0.8113 | 0.7328 |
| 20 | 0.9480 | 0.9367 | 0.8515 | 0.7845 |
| 25 | 0.9480 | 0.9367 | 0.8544 | 0.8190 |
| 30 | 0.9610 | 0.9367 | 0.8736 | 0.8362 |

触发词覆盖率低的事件识别性能影响较大，而基于无触发词方式此时有较大的优势

# 无触发词的事件识别性能分析

**GRU+CRF:**通过TF-IDF方法获得触发词，将任务转化成序列标注
**DMCNN:** 动态多池化神经网络方法

事件识别结果

| 模型 | 吸毒事件 | 贩毒事件 | 购买毒品事件 | 持毒事件 | 资金流入事件 | 资金流出事件 | 微平均值 | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | F | F | F | F | F | P | R | F |
| GRU+CRF | 0.662 | 0.723 | 0.613 | 0.794 | 0.585 | 0.758 | 0.831 | 0.735 | 0.780 |
| DMCNN | 0.954 | 0.892 | 0.742 | 0.892 | 0.667 | 0.901 | **0.962** | 0.809 | 0.880 |
| **ours/字** | **0.966** | 0.902 | **0.968** | 0.931 | 0.823 | 0.989 | 0.949 | 0.935 | **0.942** |
| **ours/词** | 0.964 | 0.840 | 0.941 | 0.928 | 0.740 | **0.991** | 0.938 | 0.932 | 0.935 |

基于字和词的模型总体偏差不大，数据偏口语等因素，基于字模型总体均衡性要好于基于词的模型

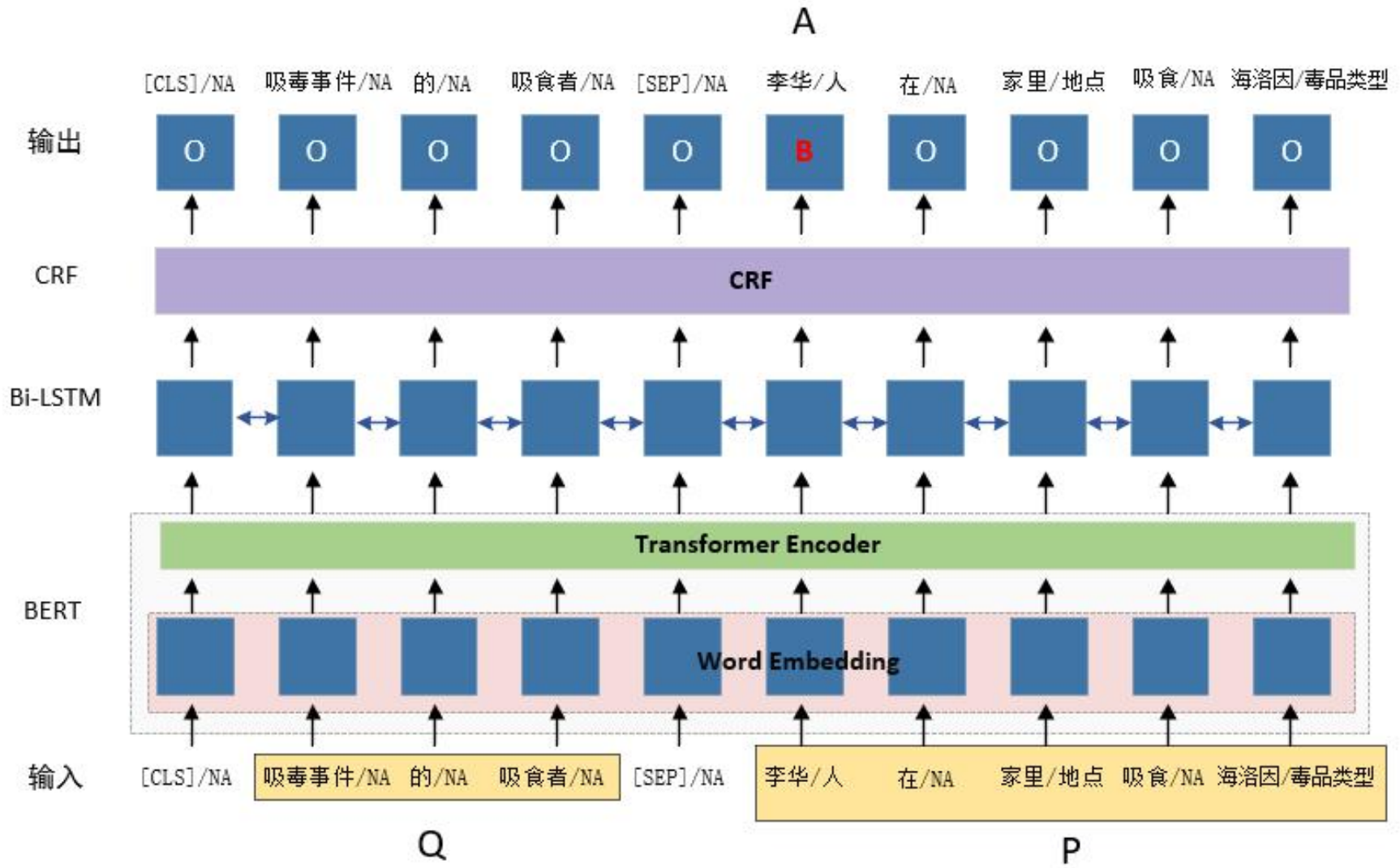
# 无触发词的事件识别性能分析

利用TF-IDF方法和attention机制获取排名前五的词的比较

| 事件类型 | TF-IDF | | | | | attention | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 吸毒事件 | 吸食 | 冰毒 | 方式 | 毒品 | 注射 | 吸食 | 吸食毒品 | 方式 | 毒品 | 吸毒 |
| 贩毒事件 | 毒品 | 贩卖 | 冰毒 | 白色 | 包装 | 卖 | 毒品 | 冰毒 | 贩卖 | 共计 |
| 购买毒品事件 | 购买 | 毒品 | 冰毒 | 微信 | 人民币 | 毒品 | 购买 | 买 | 海洛因 | 吸食 |
| 持毒事件 | 毒品 | 冰毒 | 白色 | 塑料袋 | 包装 | 毒品 | 贩卖 | 吸食 | 有 | 搜出 |
| 资金流入事件 | 对方 | 人民币 | 返还 | 微信 | 转入 | 人民币 | 转入 | 返还 | 提现 | 收到 |
| 资金流出事件 | 对方 | 转账 | 人民币 | 微信 | 诈骗 | 转账 | 二维码 | 刷 | 金额 | 转给 |

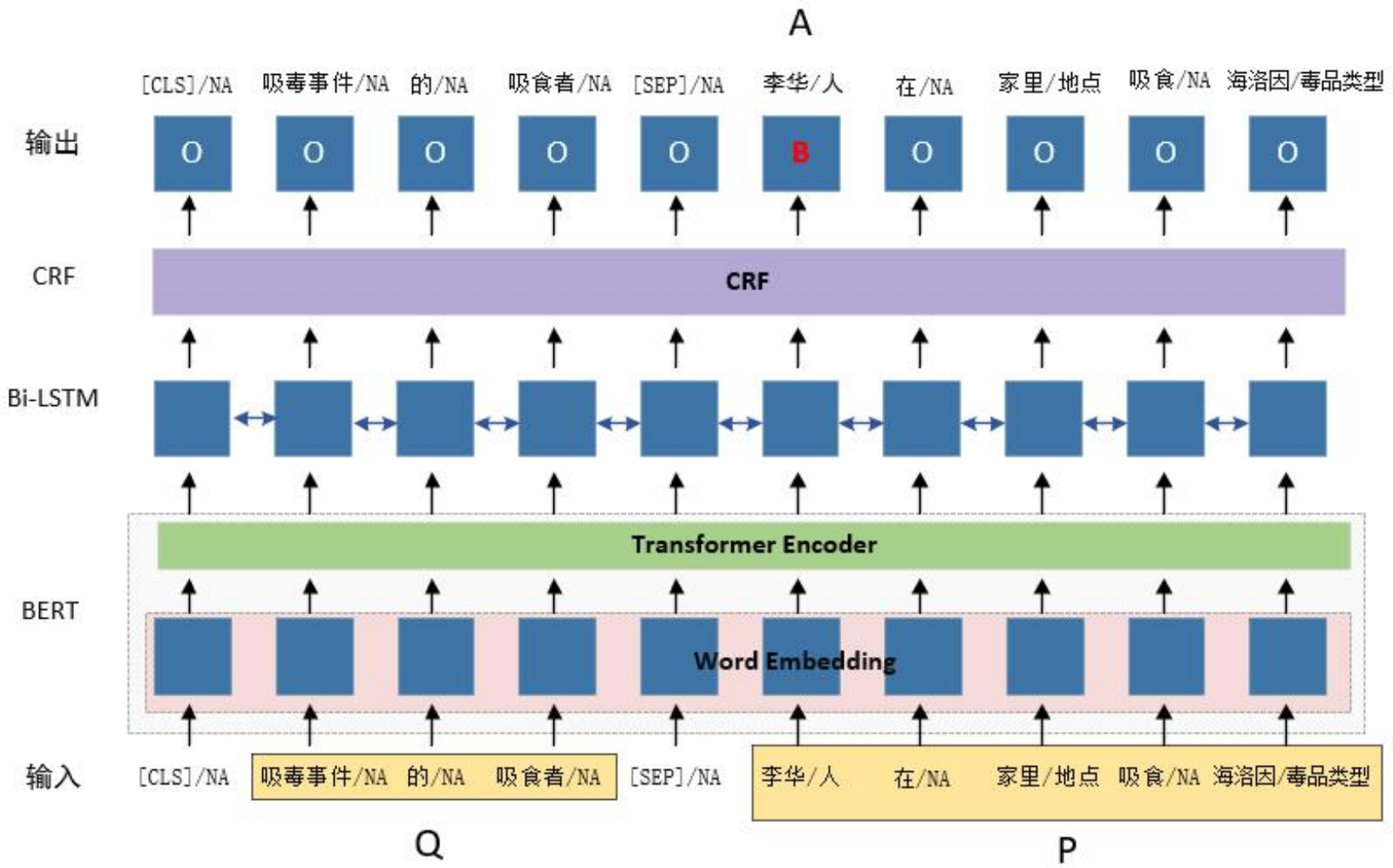
# 阅读理解的论元角色分类方法

网络结构



# 阅读理解的论元角色分类方法

网络结构




**问题Q的构造**:某某事件(事件类型)的某某性质(事件论元角色)

| 论元 | 针对论元的问题 |
| --- | --- |
| 时间 | 吸毒事件的时间? |
| 地点 | 吸毒事件的地点? |
| 吸食者 | 吸毒事件的吸食者? |
| 类型 | 吸毒事件的类型? |
| 剂量 | 吸毒事件的剂量? |
| 方式 | 吸毒事件的方式? |

# 阅读理解的论元角色分类性能分析

论元角色分类结果

| 模型 | drug taking | drug selling | drug buying | drug having | money -in | money -out | micro-average | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | F | F | F | F | F | P | R | F |
| GRU+CRF | 0.801 | 0.717 | 0.500 | 0.393 | 0.674 | 0.823 | 0.877 | 0.679 | 0.765 |
| DMCNN | **0.861** | 0.793 | 0.784 | 0.621 | 0.826 | 0.883 | 0.821 | 0.872 | 0.846 |
| **MRC** | 0.856 | 0.878 | **0.806** | **0.626** | **0.866** | **0.898** | 0.831 | **0.895** | **0.862** |

# 结论

- 事件识别的微平均F1达到0.942
- 事件论元角色分类的微平均F1达到0.862
- 方向：多个同种类型事件的事件抽取

# THANKS
# Q&A