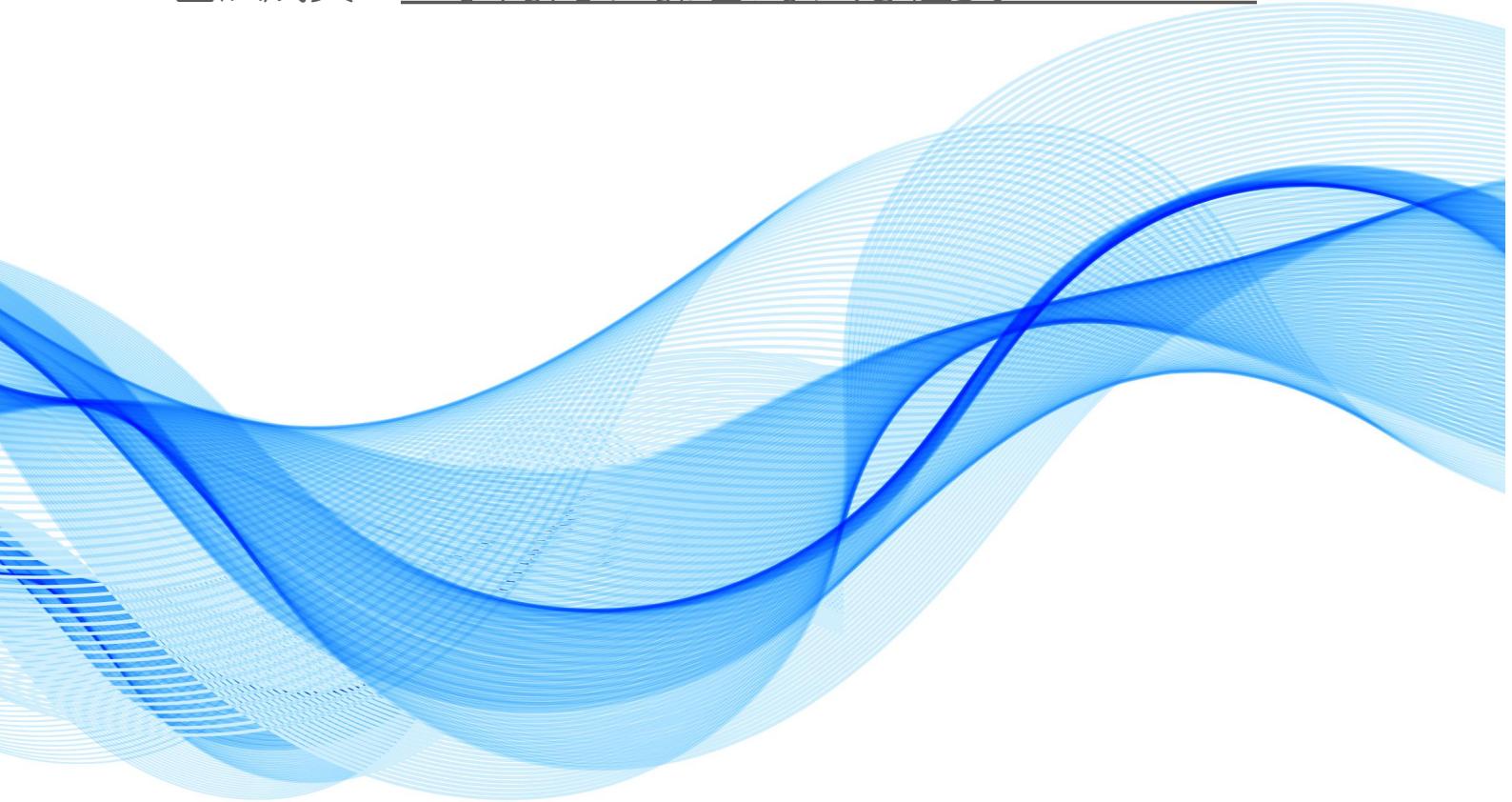


# 社会计算创新大赛决赛

作品名称： 智慧信访：政府数字化的智能应用

团队成员： 余海涛、郭睿琦、方桂安



# 目 录

<b>1. 相关背景 .....</b>	<b>1</b>
1.1 问题重述 .....	1
1.2 信访的定义 .....	1
1.3 古今信访的对比 .....	1
1.4 信访的问题 .....	1
1.5 智慧信访 .....	2
<b>2. 作品分析 .....</b>	<b>2</b>
2.1 应用场景分析 .....	2
2.2 用户需求分析 .....	2
2.3 问题解决思路分析 .....	3
<b>3. 数据探索 .....</b>	<b>4</b>
3.1 数据清洗 .....	4
3.2 数据预处理 .....	4
3.3 信访文本停用词过滤 .....	4
3.4 信访文本分词 .....	5
3.5 信访文本数据挖掘 .....	6
3.6 信访数据融合 .....	7
<b>4. 信访诉求的概要提取 .....</b>	<b>7</b>
4.1 建模思路 .....	7
4.2 抽取模型 .....	8
4.2.1 语料转换 .....	8
4.2.2 指标问题 .....	8
4.2.3 模型结构 .....	8
4.3 生成模型 .....	9
4.3.1 基础架构 .....	9
4.3.2 BIO Copy .....	9
4.3.3 稀疏 Softmax .....	10
4.4 效果展示 .....	10
<b>5. 信访件信息的自动比对 .....</b>	<b>11</b>
5.1 Match-Ignition 相似度判别算法总体架构 .....	11

5.2 句子级噪声过滤 .....	12
5.2.1 PageRank 算法 .....	12
5.3 单词级噪声过滤 .....	13
5.3.1 图 Transformer 模型 .....	13
5.3.2 PageRank 嵌入 Transformer .....	14
5.4 效果展示 .....	14
<b>6. 信访件所属的智能分类 .....</b>	<b>15</b>
6.1 信访四级分类类别信息 .....	15
6.2 基于 Bert-BiGRU 的信访分类模型 .....	17
6.2.1 信访文本向量化 .....	18
6.2.2 信访文本特征提取 .....	19
6.2.3 信访文本特征优化 .....	19
6.2.4 信访文本分类算法 .....	20
6.2.5 信访文本分类评价指标 .....	21
6.3 信访文本四级分类 .....	21
<b>7. 信访件单位的智能分派 .....</b>	<b>23</b>
7.1 事权单位实体识别 .....	23
7.2 双向长短时记忆网络 .....	24
7.3 BiLSTM 与 CRF 算法 .....	25
7.4 信访件转送到事权单位 .....	25
<b>8. 基于信访大数据的缠访和群体事件预警模型 .....</b>	<b>26</b>
<b>9. 作品的意义与价值 .....</b>	<b>28</b>
9.1 已有成效 .....	28
9.2 预期成效 .....	28

## 1. 相关背景

### 1.1 问题重述

在信访投诉举报场景下，对群众诉求概要提取，自动比对、智能分类和智能分派。

(1) 群众诉求概要提取：对群众的网上、来信、来电、来访等 4 类信访形式提出的具体诉求内容进行概要提取，概括的内容简洁明了、语义通顺并能保持原意，让业务员快速理解信访件内容。

(2) 信访件信息自动比对：使用合理的算法对诉求内容等信息进行相似度判别，辅助业务员对重复信访件的判别。

(3) 智能分类：通过对信访件诉求的意图分析，按照目前信访的 4 级分类标准，智能判断信访件所属分类。

(4) 智能分派：根据智能分类结果及问题属地等信息自动为信访件选择转送相应的事权单位。

(5) (拓展) 缠访和群体事件预警：构建缠访和群体事件预警模型，提高政府对社会风险的感知、预测和防范能力。

### 1.2 信访的定义

信访（即“人民来信来访”的简称），俗称上访，是中国大陆特有的政治表达、请愿及申诉方式。按照官方定义，信访指中华人民共和国公民、法人或者其他组织采用书信、电子邮件、传真、电话、走访等形式，向各级政府、或者县级以上政府工作部门反映冤情、民意，或官方（警方）的不足之处，提出建议、意见或者投诉请求等等。

### 1.3 古今信访的对比

在中国古代，魏晋以后历朝中央政府到各级地方政府设登闻鼓，并设专职机构或人员，遇有击鼓者需立即受理或上报。“人有穷冤则挝鼓，公车上表其奏”，以作“用下达而施于朝”之用。这也是“击鼓鸣冤”的由来。北宋专设登闻鼓院（鼓院）和登闻检院（检院），两院均受理吏民申诉之状。

而当今，为处理信访事宜，中华人民共和国国务院办公厅专门设立有国家信访局，各级政府、人大及政协也设有信访办公室。通常，规模以上各国有企事业单位及人民团体等均设有信访办公室。

### 1.4 信访的问题

#### 1.4.1 非中心化和权力分散问题严重

中央政府最多只能管理到县级的事务，地方官员拥有非常多的权力。这样便增加了官员在“天高皇帝远”的情况下渎职的机会。很多学者认为，人治色彩过分浓厚，是信访制度的弊病。由于信访没有一定的途径，也没有特定的制度管理，所以很容易出现官僚作风，甚至恐吓信访人的情况。

#### 1.4.2 政府级别多，反映困难

一个普通百姓，上面有乡（镇、街道）、县、市、省四级政府，才到中央政府。对于一些在乡郊地区的国民，还有村民委员会（村委会）、党支部。上访的最佳手段就是到北京去反映情况，但这样对于一个普通百姓来说是异常艰难的事。第一，是路程遥远；第二，是经费问题；第三，是上述的渎职或压迫信访人问题。而且法律上不容许越级上访，因此实际上很多这样的例子都给打回票，信访人的诉求往往被打回到信访对象的地方政府处理，从而使信访变得无效。

## 1.5 智慧信访

在信访大数据背景下，运用成熟的大数据技术深入挖掘数据背后所隐藏的规律或问题、及时发现矛盾风险点是新时期信访工作改革的重点。

在政府数字化改革背景下，我们希望通过深度学习技术，让群众在信访投诉举报时，诉求能够自动概要、自动比对、智能分类和智能分派，提高解决问题的效率。

## 2. 作品分析

### 2.1 应用场景分析

信访部门除了要处理信访平台本身的数据，还要与多部门对接，收集市长信箱等其他端口的信访数据，然后汇总统计、分类转办。

有些信访件是同一件事通过多种渠道提交，属于重复事项，然而信访平台未完全实现数据整合，面临着重复统计、重复上报的情况，需要采取人工方式对重复和无效的信访数据进行比对筛查，严重影响了办理效率。

长期以来，缠访在信访服务工作中屡见不鲜。由于缠访行为具有持续时间长、产生根源复杂、影响社会秩序等特点。如何调解缠访也逐渐成为信访部门和社会管理部门正在面临的重点和难点问题。

群体事件对社会有严重的负面影响，我国群体事件的增长趋势虽然得到控制，然而数量仍居高不下。信访数据能够适时且灵敏地反映社会各方面的问题，通过信访案例能够及时捕捉社会的内部矛盾和问题。

### 2.2 用户需求分析

- (1) 对于信访业务员，需要 AI 辅助剔除重复和无效的信访件以减轻负担。
- (2) 对于信访业务员，通过对群众诉求进行概要提取，概括的内容简洁明了、语义通顺并能保持原意，让业务员快速理解信访件内容。
- (3) 对于信访业务员，通过对信访件诉求分析，按照信访的 4 级分类标准，智能判断信访件所属分类，减少业务员工作压力。
- (4) 对于信访业务员，根据智能分类结果及问题属地，自动为信访件选择转送相应的事权单位，辅助业务员的转送工作。
- (5) 对于信访部门，需要精准识别缠访者并探究引起缠访行为的深层次原因，以建立解决缠访问题的信息化长效机制。

(6) 对于政府部门，需要通过对信访问题的分析，及时发现社会矛盾和冲突，以便有效解决这些问题，预防群体事件的发生。

### 2.3 问题解决思路分析

针对信访的相关数据，首先进行数据处理操作。主要包括：数据清洗、数据预处理、信访文本停用词过滤、信访文本分词、信访文本数据挖掘和信访数据融合等步骤。数据清洗的目的是去除“脏”数据、信访无关数据等，并设计规则用于判断数据是否为正确逻辑。数据预处理可以选出有价值的数据。通过过滤停用词，可以提高对信访文本分析的准确率。在数据预处理后要对文本分词，作为自然语言处理的数据源。通过信访文本数据挖掘，建立一定数量的信访领域新词表，并生成信访文本中出现的高频词汇库。此外，使用公开数据集与信访数据进行融合。

针对群众诉求概要问题，我们设计了信访数据处理流程，建立了 SPACES 文本摘要模型，通过“抽取+生成”的方法，端到端的对信访内容进行摘要，作用是为了简洁有效地对群众信访内容进行概要提取，使摘要简洁明了、语义通顺并能保持原意，让业务员快速理解信访件内容并同时便于后续自动比对、智能分类和智能分派。

针对信访件信息自动比对问题，我们参考目前的 SOTA 模型 Match-Ignition，将 PageRank 算法嵌入 Transformer 模型，对信访件进行相似度判别，对于怀疑相似的信访件不仅返回相似度，同时还将相似处高亮展示，辅助业务员进行判别。

针对信访件所属的智能分类，首先依次分析信访四级分类中的类别信息，针对分类标准中的一级分类共计为 20 类。建立基于 Bert-BiGRU 的信访分类模型，使用 Bert 模型进行文本向量化，并设计使用 BiGRU 完成文本特征提取，同时经过 Attention 的筛选后可以获得最优特征。最后，使用 TF-IDF 算法将信访数据按照四级标准，从上至下依次分类。针对信访数据样本的分布不平衡问题，设计通过 2 方面进行改进，改变数据分布，或使用优化算法。

针对信访件单位的智能分派，首先进行事权单位实体识别操作，并根据信访的事权单位，建立相应的事权单位名称实体。使用双向长短时记忆网络，来进行命名实体识别，得到事权单位名称。并设计模型的一些超参数，定义为全局变量。同时，结合条件随机场 CRF，使模型在预测时也能考虑上下文之间的关系。最后，将信访件转送到事权单位时，采用多级匹配方式，先根据行政区划代码信息进行初步的直接匹配，若匹配不唯一时再依据关键词、相似度一级识别出的事权单位实体进行进一步的匹配，最后转送到对应的事权单位。

在赛题前 4 问结果的基础上，我们拓展了作品的内容。在已有的缠访和群体事件预警模型的基础上，我们利用信访件的相似度判别和意图分析有效解决了现有预警模型重複信访判定不清的问题。

### 3. 数据探索

#### 3.1 数据清洗

数据清洗是必不可少的环节，其结果直接影响模型效果。数据清洗包括将缺失的数据补充完整，将重复、多余的数据筛选清除，将错误的数据纠正或删除，最后整理成为可以进一步加工、使用的有效数据。

数据清洗的目的是去除“脏”数据、重复数据及与信访无关的数据。

由于信访数据不是标准化数据，而数据库中存储的数据都是规范化、形式统一的数据，在处理该问题时，程序判断数据是否为正确的逻辑数据时，设计主要判断的规则如表所示。

表 3-1 数据判断规则

类型	规则
数据缺失	a.从系统再次导入； b.手工补录； c.根据逻辑补填； d.放弃。
数据重复	a.完全重复则去除重复； b.根据时间去除； c.人工去除； d.根据业务逻辑去除。
数据错误	a. 对于异常值，通过区间限定去除； b.格式错误，通过规则回复； c.人工干预； d.历史数据近似值。
数据不可用	a.按规则适配； b.关键字匹配； c.枚举转换。

#### 3.2 数据预处理

数据处理对数据挖掘的结果质量有关键影响。它包括数据预处理和信息预处理两部分。数据预处理可以选择有价值的数据，过滤无价值的数据。

另外，由于计算机不能直接处理自然语言。因此需要对文本数据进行信息预处理。

我们设计进行的信访诉求数据的预处理，主要步骤如下：

- (1) 剔除无效数据，如数据集中的一些空文档；
- (2) 删除空行，空行数据毫无意义且占用资源；
- (3) 过滤所有乱码、标点符号和特殊字符，保留具有语义价值信息的中英文文本；
- (4) 使用 jieba 库中的 cut 方法的精准模式对文本进行分词。

#### 3.3 信访文本停用词过滤

去除停用词，停用词是指句子中起连接作用的词，它可使句子通顺，但对语义理解几乎没有帮助。

中文语句中存在着介词、副词、语气词和连词等不具有实际意义的词汇和标点符号，停用词中还有些属于高频词汇，例如“的”、“了”、“我”和“及”，这些高频词汇在某些模型中会对结果产生影响。

通过需要删除这种无用词，可以提高后续信访文本分析的准确率。

### 3.4 信访文本分词

在进行文本分类或数据挖掘等自然语言处理任务时，数据来源是词汇，并不是一段文本或一个句子，将词汇作为数据源是自然语言处理最方便的操作，在数据预处理后要对文本分词。

输入数据一般是数值类型，然而信访诉求中的数据类型主要是：中文汉字、英文词汇和拉丁文字母等，因此需要转换成模型能够计算的数值型数据，或者说是将文字嵌入到数学模型中，也就是词嵌入方式。

将预处理后的原文信访诉求数据进行分词，原数据及分词结果，基于不同方法时结果分别如下：

原句部分语句	其是拱墅区祥符街道塘萍路 155 号上尚庭 9 棚 1002 室住户，反映 9 棚 1003 住户于 2021 年 5 月开始装潢，装潢中占用约 6 平米公共走廊、安装铝合金门搭建小房，在小房内摆设洗衣机等家电，且将空调外机安装在其家设备阳台空调外机上方，已向属地城管、消防反映，消防告知需在安全隐患，要求整改，请相关部门帮助核实处理。此件曾两次交办拱墅区，第一次反馈：2021 年 7 月 16 日 9 时 58 分工作人员陆清漪通过电话 88176110 联系信访人，经向祥符中队询问，目前已调取来图纸，还在进行比对核实中，后续根据调查情况做进一步处理。
分词法 1	其是   拱墅区   祥符   街道   塘萍   路   155   号   上   尚庭   9   棚   1002   室   住户   ,   反映   9   棚   1003   住户   于   2021   年   5   月   开始   装潢   ,   装潢   中   占用   约   6   平米   公共   走廊   、   安装   铝合金   门   搭建   小房   ,   在   小房   内   摆设   洗衣机   等   家电   ,   且   将   空调   外机   安装   在   其家   设备   阳台   空调   外机   上方   ,   已   向   属地   城管   、   消防   反映   ,   消防   告知   需在   安全隐患   ,   要求   整改   ,   请   相关   部门   帮助   核实   处理   。   此件   曾   两   次   交   办   拱   墅   区   ,   第   一   次   反   馈   :   2021   年   7   月   16   日   9   时   58   分   工   作   人   陆   清   漪   通   过   电   话   88176110   联   系   信   访   人   ,   经   向   祥   符   中   队   询   问   ,   目   前   已   调   取   来   图   纸   ,   还   在   进   行   比   对   核   察   中   ,   后   续   根   据   调   查   情   况   做   进   一   步   处   理   。
分词法 2	其/ 是/ 拱墅区/ 祥符/ 街道/ 塘/ 萍/ 路/ 155/ 号/ 上/ 尚/ 庭/ 9/ 棚/ 1002/ 室/ 住户/ , / 反映/ 9/ 棚/ 1003/ 住户/ 于/ 2021/ 年/ 5/ 月/ 开始/ 装潢/ , / 装潢/ 中/ 占用/ 约/ 6/ 平米/ 公共/ 走廊/ 、 / 安装/ 铝合金/ 合金/ 金门/ 搭建/ 小房/ , / 在/ 小房/ 房内/ 摆设/ 洗衣/ 洗衣机/ 等/ 家电/ , / 且/ 将/ 空调/ 外/ 机/ 安装/ 在/ 其家/ 设备/ 阳台/ 空调/ 外/ 机上/ 上方/ , / 已向/ 属地/ 城管/ 、 / 消防/ 防反/ 反映/ , / 消防/ 告知/ 需/ 在/ 安全/ 安全隐患/ 隐患/ , / 要求/ 整改/ , / 请/ 相关/ 部门/ 帮助/ 核实/ 实处/ 处理/ 。 / 此/ 件/ 曾/ 两/ 次/ 交/ 办/ 拱/ 墅/ 区/ , / 第/ 一/ 第/ 一/ 次/ 反/ 馈/ : / 2021/ 年/ 7/ 月/ 16/ 日/ 9/ 时/ 58/ 分工/ 工作/ 工作人员/ 作人/ 人员/ 陆/ 清漪/ 通/ 过/ 电/ 话/ 88176110/ 联/ 系/ 信/ 访/ 人/ , / 经/ 向/ 祥/ 符/

	中队/ 询问/ , / 目前/ 已/ 调取/ 取来/ 图纸/ , / 还/ 在/ 进行/ 比对/ 核实/ 中/ , / 后续/ 根据/ 据调查/ 调查/ 情况/ 做/ 进一步/ 一步/ 处理/ 。
--	--

信访数据经过预处理后，得到格式化文本数据。且均为 UTF-8 纯文本格式。可以收集近几年信访数据，在预处理后的信访数据中提取出 80%的数据作为训练集，10%的数据作为验证集，10%的数据作为测试集。数据提取完毕之后，训练集、验证集、测试集分别作为一个文档，便于后续实验使用。

### 3.5 信访文本数据挖掘

信访文本中蕴含了大量的特殊词汇，仅使用公开的词汇库进行分词可能不能达到较好效果，比如“有关职能部门”一词的分词结果为：“有关”、“职能”、“部门”。这样的分词显然达不到要求。故我们设计，先运用新词挖掘工具对信访文本数据进行探索分析，得到一定数量的信访领域新词表。

另外，还可以整理信访领域停用词库。其中包括“年”、“月”、“日”、“号”、“字”、“人”等 82 个词。

最后，将新词库和停用词表中的单词导入了分词工具，结合噪声处理和中文分词技术，以得到规范化的信访数据库。有助于后续实验任务的使用。

依据信访文本中，出现生成的高频词汇，以及整理社会语义网络中出现的共现高频率词汇。我们设计将近几年的信访文本中的高频词汇，按照信访分类标准等整理为如下形式：

表 3-2 高频与共现词库

信访分析类目	高频词汇	高频共现词汇
信访主题	农村农业	土地、农村、农民
	国土资源	土地、工程、资源、
	城乡建设	建设、发展、住房、土地、交通、工程、环境、公路、道路、小区物业、经济、资金
	劳动和社会保障	工资、保障、住房、劳动、补贴、人力资源、保险、改革、合同、职工、公务员、教师
	卫生计生	健康、安全
	教育文体	教育、收费、考试、教师、学生、教育局
	民政	无
	政法	交通、法律、人力、保险、公安、户口
	经济管理	收费、企业、市场、经济、改革、经营、合同、安全
	交通运输	交通、公路、道路、车辆、派出所、交警、安全
	商贸旅游	企业、市场、改革、经营、合同
	科技与信息产业	无
	环境保护	环境、小区、健康、安全

	党务政务	机关、干部	无
	组织人事	机关、考试、规范、改革、岗位、公务员	无
	纪检监察	机关、服务、收费、考试、规范、车辆、干部、派出所、交警、公安	无

### 3.6 信访数据融合

收集到的信访数据没有提供摘要，人工摘要再进行训练标注成本较高，故我们的模型使用了由哈工大深圳研究生院智能计算研究中心收集、整理的大规模中文短文摘要数据集 LCSTS 进行预训练。

**Short Text:**水利部水资源司司长陈明忠今日在新闻发布会上透露，根据刚刚完成的水资源管理制度的考核，有部分省接近了红线的指标，有部分省超过红线的指标。在一些超过红线的地方，将对一些取用水项目进行区域的限批，严格地进行水资源论证和取水许可的批准。

Mingzhong Chen, the Chief Secretary of the Water Devision of the Ministry of Water Resources, revealed today at a press conference, according to the just-completed assessment of water resources management system, some provinces are closed to the red line indicator, some provinces are over the red line indicator. In some places over the red line, It will enforce regional approval restrictions on some water projects, implement strictly water resources assessment and the approval of water licensing.

**Summarization:**部分省超过年度用水红线指标 取水项目将被限批

Some provinces exceeds the red line indicator of annual water using, some water project will be limited approved

图 3-1 LCSTS 数据对示例

该语料库包含超过 200 万条真实的中文短文本，每个文本的作者都给出了简短的摘要。内容均来自于中国微博网站新浪微博，更贴近于民生民情。从而确保在对群众诉求概要提取时生成的摘要简洁明了、语义通顺并能保持原意。

## 4. 信访诉求的概要提取

对通过预处理的群众网上、来信、来电、来访等 4 类信访形式的具体诉求内容进行概要提取，使概括的内容简洁明了、语义通顺并能保持原意，让业务员快速理解信访件内容并便于后续自动比对、智能分类和智能分派。

### 4.1 建模思路

综合信访数据特性，我们不难想到应该采取“抽取+生成”相结合的方式进行摘要，并配合一些新方法来保证摘要的忠实程度与提升最终的效果。最终使用到的模型被命名为 SPACES：

- (1) S: Sparse Softmax (新设计的 Softmax 替代品);
- (2) P: Pretrained Language Model (预训练模型);
- (3) A: Abstractive (抽象式，即生成式);
- (4) C: Copy Mechanism (新设计的 Copy 机制);
- (5) E: Extractive (抽取式);

(6) S: Special Words (将特殊词添加到预训练模型)。

## 4.2 抽取模型

### 4.2.1 语料转换

首先，我们需要记住的是，抽取模型只是过程而不是结果，我们还要把抽取的结果送入到 Seq2Seq 模型优化。因此，抽取模型的原则是“求全”，即尽量把最终摘要所需要的信息覆盖到。为此，我们按照如下规则将原始训练语料转换为抽取式语料：

1、自行构建分句函数，使得句子的颗粒度更细；

2、人工摘要的每个句子，都在原文中匹配与之相似度最高的那个句子（可以重复匹配）；

3、将所有匹配到的原文句子作为抽取句子标签。

### 4.2.2 指标问题

上述转换流程涉及到一个“相似度”的选择，根据前面的介绍，我们选择“以词为单位的加权 Rouge”作为评测指标。

以词为单位来算评测指标的做法，是为了使得信访内容中专有名词能够完全匹配上，比如本来是“中国人民共和国未成年人保护法”，模型预测成了“中国人民共和国文物保护法”，如果以字为单位的话，最长公共子序列为“中国人民共和国...保护法”，至少还是算对了大部分，但是如果以词为单位的话，两者就是不同的词，因此算全错。因此，以词为单位有利于专有名词匹配得更精准。

### 4.2.3 模型结构

我们使用的是以句为单位的序列标注模型作为抽取模型，句向量部分用“BERT+平均池化”来生成，并固定不变，标注模型主体方面则用 DGCNN 模型构建。

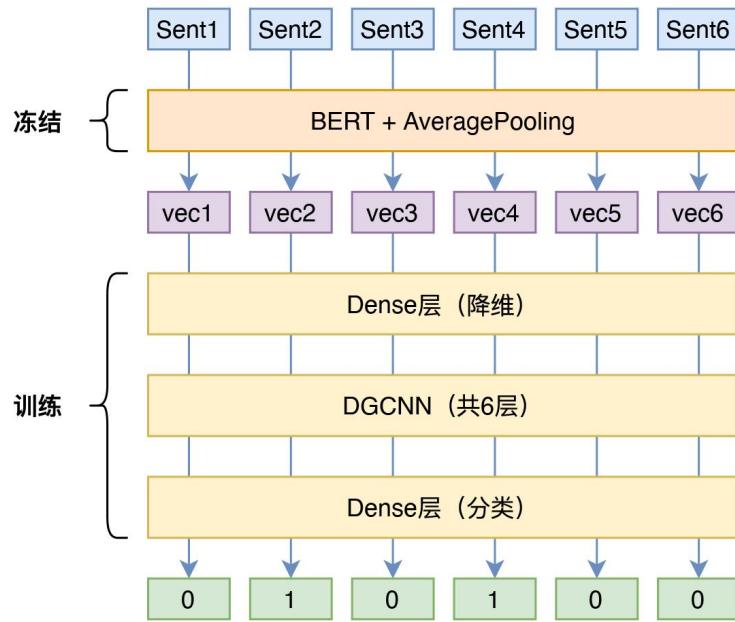


图 4-1 抽取模型架构

## 4.3 生成模型

### 4.3.1 基础架构

Seq2Seq 模型依然选择了经典的 UniLM，并且考虑到“输入+输出”的总长度基本上都超过 512 了，所以选择华为的 NEZHA 模型作为基础模型架构，因为 NEZHA 使用了相对位置编码，不限长度。

此外，在使用预训练模型方面，我们将部分信访四级分类词语加入到了 NEZHA 模型中，改变了中文预训练模型以字为单位的通用选择，这使得模型的效果和速度都有一定的提升。

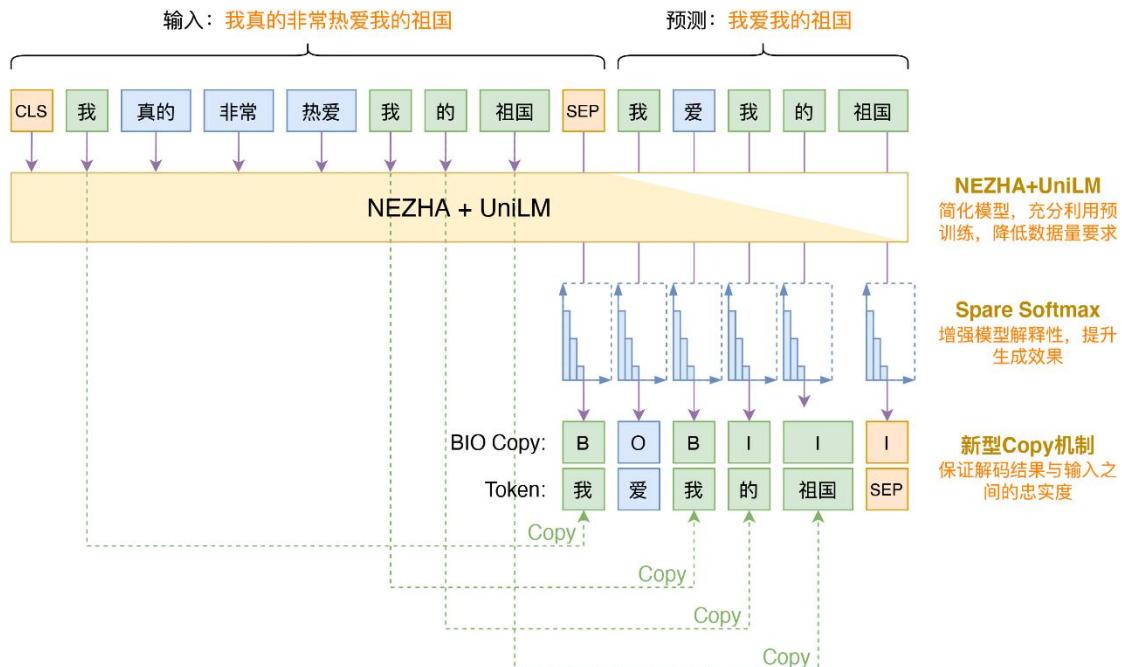


图 4-2 生成模型架构

### 4.3.2 BIO Copy

Copy 机制在摘要生成模型中并不新鲜，甚至可以说已经成为了生成式摘要的标配了。常规的 Copy 机制一般就是《Pointer Networks》的做法，但这种做法有两个不足之处：

- 1、每次只能 Copy 一个 token，不能保证 Copy 一个连续片段（n-gram）出来；
- 2、实现起来比较复杂，不够即插即用。

为此，我们使用了一种新型的 Copy 机制——BIO Copy，它实现起来非常简单，而且具有 Copy 连续片段的能力。

其实前面的图示已经展示了这种 Copy 机制，实际上它就是在 Decoder 部分多加一个序列预测任务，即原来 Decoder 建模的是每个 Token 的分布  $p(y_t|y < t, x)$ ，现在多预测一个标签分布，变为

$$p(y_t, z_t|y < t, x) = p(y_t|y < t, x)p(z_t|y < t, x)$$

其中  $zt \in \{B, I, O\}$ , 含义如下:

- (1) B: 表示该 token 复制而来;
- (2) I: 表示该 token 复制而来且跟前面 Token 组成连续片段;
- (3) O: 表示该 token 不是复制而来的。

那么, 训练时  $z$  的标签哪里来呢? 这里直接采用一种比较简单的方法: 算摘要与原文的“最长公共子序列”, 只要是出现在最长公共子序列的 token, 都算是 Copy 过来的, 根据 BIO 的具体含义设置不同的标签。比如前面图片中的例子, “我 真的 非常 热爱 我的 祖国”与“我 爱 我 的 祖国”的最长公共子序列“我 我 的 祖国”, 其中第一个“我”是单字, 标签为 B, 后面“我的 祖国”是一个连续片段, 标签为“B II”, 其他标签为 O, 所以总的标签为“B O B II”。

所以, 在训练阶段, 其实就是多了一个序列预测任务, 并且标签都是已知的, 实现起来很容易, 也不增加什么计算成本。至于预测阶段, 对于每一步, 我们先预测标签  $zt$ , 如果  $zt$  是 O, 那么不用改变, 如果  $zt$  是 B, 那么在 token 的分布中 mask 掉所有不在原文中的 token, 如果  $zt$  是 I, 那么在 token 的分布中 mask 掉所有不能组成原文中对应的 n-gram 的 token。也就是说, 解码的时候还是一步步解码, 并不是一次性生成一个片段, 但可以通过 mask 的方式, 保证 BI 部分位置对应的 token 是原文中的一个片段。

需要指出的是, Copy 机制的引入未必能明显提高分数, 但是 Copy 机制可以保证摘要与原始文本的忠实程度, 避免出现专业性错误, 这在实际使用中是相当必要的。

#### 4.3.3 稀疏 Softmax

在概要模型中, 还用到了一个 Softmax 及交叉熵代替品——Sparse Softmax, 我们发现 Sparse Softmax 可以在相当多的分类问题(包括常规分类问题和文本生成等)中替换掉 Softmax, 并且效果能得到一定的提升。

值得注意的是, Sparse Softmax 只适用于有预训练的场景, 因为预训练模型已经训练得很充分了, 因此 finetune 阶段要防止过拟合; 但是如果你是从零训练一个模型, 那么 Sparse Softmax 会造成性能下降, 因为每次只有 k 个类别被学习到, 反而会存在学习不充分的情况(欠拟合)。

### 4.4 效果展示

我们需要将原文作为输入, 通过抽取模型输出抽取摘要, 然后把抽取摘要作为生成模型的输入, 来输出最终摘要。但是, 这有一个问题, 训练的数据我们都是见过的, 但我们真正预测的是未见过的数据, 如果直接训练一个抽取模型, 然后用该模型抽取训练集的摘要, 那么很明显由于都被训练过了, 抽取出来的摘要分数肯定会偏高, 而新样本的效果则会偏低, 造成训练预测的不一致性。

这时候的解决方案就是交叉验证了。具体来说, 我们将标注数据分为 n 份, 其中  $n - 1$  份训练抽取模型, 然后用这个抽取模型预测剩下的那份数据的抽取摘要, 如此重复 n

遍，就得到全部数据的抽取摘要，并且尽可能地减少了训练和预测阶段的不一致性。

The screenshot shows a comparison between the original input text and its generated summary. The input text discusses a vehicle transfer issue involving a police officer's negligence. The summary highlights the key points about the vehicle's history and the officer's handling of the case.

输入正文	摘要
<p>标题： [来电]&lt;徐艳女士&gt;浙江省杭州市拱墅区&lt;机动车和驾驶人管理&gt;</p> <p>正文： 其（徐艳 身份证号：330*****250328）名下的车辆浙A5W2W9于2020年12月在西湖区古墩路699号杭州市公安局交通警察支队车辆管理所办理了车辆过户转移到绍兴，反映当时本人并未到场，也未有委托书，售卖合同也是假的，但是被办理了过户，认为工作人员做事不严谨，导致车辆被不法人士骗走，故来电投诉，希望给予合理的解释。（市公安交警局反馈：车辆在元通交通管理服务站办理车辆转移登记，其代办人员资料齐全，符合机动车转移登记相关规定。根据《二手车流通管理办法》（2005年8月29日商务部、公安部、国家工商行政管理总局、国家税务总局）第十五条：二手车卖方应当拥有车辆的所有权或者处置权。二手车交易市场经营者和二手车经营主体应当确认卖方的身份证明，车辆号牌、《机动车登记证书》、《机动车行驶证》，有效的机动车安全技术检验合格标志、车辆保险单、交纳税费凭证等。第三十五条：商务主管等部门、工商行政管理部门应当在各自的职责范围内采取有效措施，对二手车交易市场经营者和经营主体的监督管理，依法查处违法违规行为，维护市场秩序，保护消费者的合法权益。其在元通服务站办理车辆转移登记，资料齐全，符合机动车转移登记相关规定。车管部门办理车辆登记是上路行驶准可，并非对车辆物（产）权归属进行变更。我支队车管部门通过电话：88945680于2021-08-09 09:06与反映人联系，其称其车辆是当时一朋友通过借车的方式在其不知情的情况下通过杭州元通机动车二手车市场将其车辆过户，反映人认为当时市场也未跟其确认车辆是否过户就在其不知情的情况下对其车辆进行过户）</p>	<p>摘要：</p> <ul style="list-style-type: none"><li>其（徐艳 身份证号：330*****250328）名下的车辆浙A5W2W9于2020年12月在西湖区古墩路699号杭州市公安局交通警察支队车辆管理所办理了车辆过户转移到绍兴，反映当时本人并未到场，也未有委托书，售卖合同也是假的，但是被办理了过户，认为工作人员做事不严谨，导致车辆被不法人士骗走，故来电投诉，希望给予合理的解释。（市公安交警局反馈：车辆在元通交通管理服务站办理车辆转移登记，其代办人员资料齐全，符合机动车转移登记相关规定。</li></ul>

图 4-3 摘要示例

如图为从政府数据中随机抽取的一组正文与标题，可以从生成的摘要示例中看出身份证号码、车牌号等关键信息都能忠实保存；《管理办法》等内容则被省略。由此业务员可以快速理解信访件的内容，提高效率。

## 5. 信访件信息的自动比对

我们参考目前的 SOTA 模型 Match-Ignition，对信访件进行相似度判别，对于怀疑相似的信访件不仅返回相似度，同时还将相似处高亮展示，辅助业务员进行判别。

### 5.1 Match-Ignition 相似度判别算法总体架构

Match-Ignition 是一种分层噪声过滤模型，通过著名的链接分析算法 PageRank 提取匹配（即相似）的文本。PageRank 利用图上的随机游走来确定每个节点（句子或单词）的重要性。通过这种方式，可以消除噪声（即不太重要的文本）并加速算法。

如图 5-1 所示，考虑到长文本匹配问题中的两级结构，Match-Ignition 模型包含两个层次结构，即句子级和单词级。在句子级噪声过滤过程中，节点被定义为来自一对长文本的句子，链接被定义为每对句子之间的相似性。也就是说，每个长文本内部和两个长文本之间的相似性都在图中捕获。然后可以通过 PageRank 的重要性分数识别噪声句子，并直接去除。

单词级噪声过滤过程则与相似度判别共同进行。因为每个单词都依赖于其上下文来表达其具体含义，因此在相似度判别中需要动态估计噪声单词。为此，模型首先将最先进的 Transformer 模型应用于文本以捕捉单词之间的上下文信息，随后将 Transformer 自注意力块中的注意力矩阵视为全连接的词级相似度图，并应用 PageRank 过滤掉每一

层的噪声词，最后利用过滤后剩下的关键信息进行相似性判别。

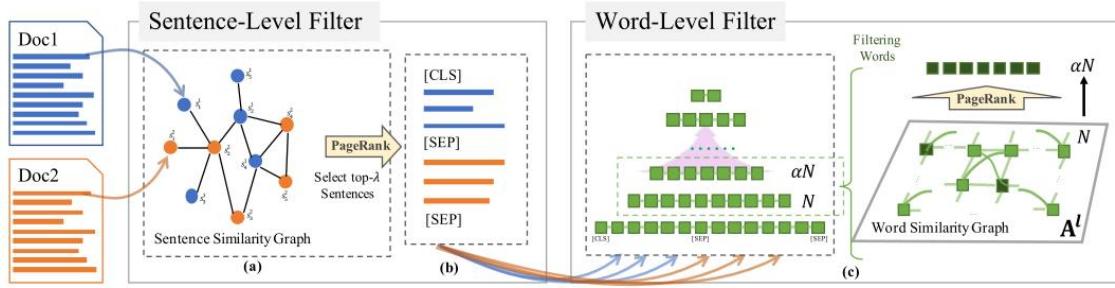


图 5-1 Match-Ignition 的整体架构 (a) 句子级过滤器 (b) 句子级过滤器的输出 (c) 单词级过滤器

## 5.2 句子级噪声过滤

### 5.2.1 PageRank 算法

PageRank 是一种基于图的排名算法，本质上是一种确定图中某个顶点重要性的方法。形式上，给定一个图  $G(V, E)$ ，其中  $V = \{v_1, v_2, \dots, v_N\}$  是一组节点， $E$  是这些节点之间的链接。每个节点  $v_i$  上的 PageRank 值  $u_i$  表示节点  $v_i$  的重要性。定义  $\mathbf{A}$  为邻接矩阵，即  $A_{ij}$  表示  $v_i$  与  $v_j$  具有权重  $A_{ij}$  的链接。 $\mathbf{A}$  也是一个随机矩阵，因为每列总和为 1。在初始步骤中，所有  $u_i$  具有相同的值  $\frac{1}{N}$ ，表示所有节点都同等重要。在接下来的每个步骤中，使用其他节点指向  $v_i$  的链接来更新 PageRank 值  $u_i$

$$u_i = \sum_{v_j \in V} A_{ij} u_j$$

经过多次迭代，PageRank 值  $u_i$  会收敛到一组稳定值，这就是 PageRank 的解。令  $\mathbf{u}^t = [u_1^t, u_2^t, \dots, u_N^t]$  是一个长度为  $N$  的向量，代表所有节点在时间  $t$  的 PageRank 值。然后，PageRank 可以重写为

$$\mathbf{u}^{t+1} = \mathbf{A}\mathbf{u}^t$$

为了解决孤立节点的问题，Match-Ignition 使用了一个更稳定版本的 PageRank：

$$\mathbf{u}^{t+1} = d\mathbf{A}\mathbf{u}^t + \frac{1-d}{N} \cdot \mathbf{1}$$

其中  $d \in [0, 1]$  是一个实数值，用于确定两部分的比例， $\mathbf{1}$  是一个长度为  $N$  的向量，其所有值都为 1。因子  $d$  通常设置为 0.85。在实践中，为了计算效率，将迭代步数  $T$  设置为固定值。因此， $\mathbf{u}^t$  是每个  $v_i \in V$  的最终 PageRank 分数，PageRank 分数越大表示该节点在当前图中的重要性越高，因此我们可以过滤掉 PageRank 值小的节点。

### 5.2.2 句子级噪声过滤实现

在句子级噪声过滤中，首先，将两个待判别相似度的长文本中的句子合并在一起，形成一个联合句子集合。首先将两个长文本拆分成句子，记为  $d_s = [s_1^1, s_2^1, \dots, s_{L1}^1]$  和

$dt = [s_1^2, s_2^2, \dots, s_{L_2}^2]$ , 其中  $L_1$  和  $L_2$  分别是  $ds$  和  $dt$  中的句子数。联合句子集合  $S = \{s_1^1, s_2^1, \dots, s_{L_1}^1, s_1^2, s_2^2, \dots, s_{L_2}^2\}$  有  $L_1 + L_2$  个元素。因此, 可以通过评估联合句子集合  $S$  中的句子对之间的相似度来构建句子相似度图。句子相似度定义为两个句子之间的重叠词比率:

$$Sim(s_i, s_j) = \frac{|\{w_k | w_k \in s_i, w_k \in s_j\}|}{\log(|s_i|) + \log(|s_j|)}, s_i, s_j \in S$$

其中  $w_k$  表示句子中的单词,  $|\cdot|$  表示句子或词集的长度,  $s_i, s_j$  是联合句子集合  $S$  中的两个句子。将 PageRank 算法应用于构建的句子相似度图就可以得到每个句子的重要性。为了在接下来的步骤中平衡来自不同长篇文本的信息, 分别为每个长篇文本提取前  $\lambda$  个句子。因此, 两个文本都包含  $\lambda$  个句子作为它们的摘要。

### 5.3 单词级噪声过滤

为了过滤单词级的噪声, 首先需要在 Transformer 结构中构建一个单词级的相似度图, 随后在 Transformer 模型中嵌入 PageRank, 用于单词级噪声过滤。

#### 5.3.1 图 Transformer 模型

自注意力块是 Transformer 架构中的主要组件, 它计算出句子中所有其他单词相对于它上下文单词的重要性, 这恰好与 PageRank 相符合。自注意力块建立了单词之间的关系, 可以将其视为单词之间的全连接图。自注意力函数可以表达如下:

$$\mathbf{H}^{l+1} = Attn(\mathbf{Q}^l, \mathbf{K}^l, \mathbf{V}^l) = \text{Softmax}\left(\frac{\mathbf{Q}^l(\mathbf{K}^l)^T}{\sqrt{E}}\right)\mathbf{V}^l = \mathbf{A}^l\mathbf{V}^l$$

其中  $\mathbf{Q}^l = \mathbf{W}_Q^l\mathbf{H}^l \in \mathbb{R}^{N \times E}$  表示注意力 query 矩阵,  $\mathbf{K}^l = \mathbf{W}_K^l\mathbf{H}^l \in \mathbb{R}^{N \times E}$  表示 key 矩阵,  $\mathbf{V}^l = \mathbf{W}_V^l\mathbf{H}^l \in \mathbb{R}^{N \times E}$  表示 value 矩阵。  $N$  表示文本中的单词数,  $E$  表示词表示的维度。注意力机制可以解释为: 对于  $\mathbf{Q}$  中的每个注意力 query 向量, 它首先计算 query 与所有 key 的点积, 旨在评估注意力 query 与每个 key 之间的相似性。然后将其除以  $\sqrt{E}$ , 并应用 SoftMax 函数来获得 value 的权重, 表示为  $\mathbf{A}^l$ 。同时还可以通过增加多个注意力头来提高稳定性:

$$\begin{aligned} \mathbf{H}^{l+1} &= \text{Concat}(\text{head}_1, \dots, \text{head}_H)\mathbf{O}^l, \\ \text{head}_k &= Attn(\mathbf{Q}^{kl}, \mathbf{K}^{kl}, \mathbf{V}^{kl}) = \mathbf{A}^{kl}\mathbf{V}^{kl} \end{aligned}$$

其中  $\mathbf{Q}^{kl}, \mathbf{K}^{kl}$  和  $\mathbf{V}^{kl}$  是第  $l$  层的第  $k$  个注意力头, 具有不同的可学习权重,  $\mathbf{O}^l$  通过下投影以匹配跨层的维度,  $H$  是每层中头的数量,  $L$  是层数。

如果我们将每个单词视为图中的一个节点, 它们会通过聚合所有其他上下文单词的

注意力权重来更新其自身权重，就像从图神经网络中的其他相邻节点传递的消息一样。因此，自注意力块可以被视为一个全连接的词图，其中它的邻接矩阵是单词间相似度矩阵 $\mathbf{A}^{kl}$ 的转置。

### 5.3.2 PageRank 嵌入 Transformer

Match-Ignition 将 PageRank 嵌入到 Transformer 模型中，以过滤单词级别的噪声。在每个自注意力块中，利用嵌入的 PageRank 算法动态过滤噪声词，可以逐层减少序列长度。

标准的 Transformer 结构有 $L$ 层多头自注意力块，一层一层堆叠，每一层都保持相同的序列长度 $N$ 。单词级噪声过滤过程是每层一次，因此需要对第  $l$  层中不同注意力头的所有邻接矩阵进行平均，得到

$$\mathbf{A}^l = \frac{1}{H} \sum_{k=1}^H \mathbf{A}^{kl}$$

因为 $\mathbf{A}^l$ 是逐行 SoftMax 函数的输出，所以 $\mathbf{A}^l$ 的每一行之和为 1。因此， $(\mathbf{A}^l)^T$  是一个随机矩阵，可以将其视为图中的邻接矩阵。因此，将  $(\mathbf{A}^l)^T$  代入 PageRank 得到：

$$\mathbf{u}^{t+1} = d(\mathbf{A}^l)^T \mathbf{u}^t + \frac{1-d}{N} \cdot \mathbf{1}$$

迭代求解上面的等式，然后得到第 $(l-1)$ 层中所有单词/节点的 PageRank 值，表示为 $\mathbf{u}$ 。因此， $\mathbf{u}$ 表示第 $(l-1)$ 层中各个单词的重要性。在将注意力机制应用于第 $(l-1)$ 层中的单词之后，得到一个新的单词重要性列表作为第  $l$  层的输入。为了过滤噪声词，必须估计第  $l$  层中单词/节点的重要性，这可以通过在分布 $\mathbf{A}^l$ 下重新分配第 $(l-1)$ 层中词的重要性来评估，因此单词重要性是  $\mathbf{r} = \mathbf{A}^l \mathbf{u}$ 。最后，可以通过删除在 $\mathbf{r}$ 中具有较小值的单词节点来减少第  $l$  层的序列长度。在单词级噪声过滤后就可以利用提取出的关键信息进行相似性判别。

## 5.4 效果展示

如图 5-2 所示，为了展示 Match-Ignition 模型的效果，我们给出了一个来自 CNSE 数据集的示例。图中以不同颜色显示了单词重要性，其中颜色越深表示重要性得分越高。具体来说，重要性分数是根据保留单词的层数计算的，它显示了每个单词在整个匹配过程中的重要性。结果非常直观。例如，地点“佛山”和政策名称“征求意见稿”对于新闻的匹配程度很重要，它们确实在模型中获得了更高的重要性得分，如红框标注的那样。

**Doc1:**

[CLS] 佛 山 网 约 车 拟 规 定 : 车 辆 缴 为 本 地 牌 照 司 机  
 需 有 佛 山 户 籍 或 居 住 证 10 月 31 日 , 佛 山 市 交 通  
 运 输 局 在 其 官 网 及 市 政 府 网 正 式 发 布 《 佛 山 市  
 网 络 预 约 出 租 汽 车 经 营 服 务 管 理 暂 行 办 法 ( 征  
 求 意 见 案 ) 》 ( 以 下 简 称 《 暂 行 办 法 》 ) 以 及

**Doc2:**

[SEP] 佛 山 网 约 车 新 政 零 过 渡 期 全 国 首 创 暂 行 办 法  
 今 日 上 午 , 佛 山 市 交 通 运 输 局 在 官 网 上 发 布 了  
 《 佛 山 市 网 络 预 约 出 租 汽 车 经 营 服 务 管 理 暂 行  
 办 法 ( 征 求 意 见 案 ) 》 , 即 佛 山 版 网 约 车 新 政

图 5-2 相似度判别结果, 颜色越深表示单词越重要, 红框表示两段文本的相似处

## 6. 信访件所属的智能分类

### 6.1 信访四级分类类别信息

信访 4 级分类标准中的一级分类, 对应的类别及类别数如下表 6-1 所示, 可视化统计结果如下图 6-1 所示。

表 6-1 信访一级分类

序号	一级	计数
1	城乡建设	572
2	党务政务	82
3	纪检监察	24
4	交通运输	97
5	教育	41
6	经济管理	59
7	军队事务	34
8	科技与信息产业	25
9	劳动和社会保障	45
10	民政	54
11	农村农业	93
12	其他	11
13	生态环境	107
14	市场监管	97
15	卫生健康	49
16	文体旅游	48
17	应急	54
18	政法	128
19	自然资源	114
20	组织人事	53

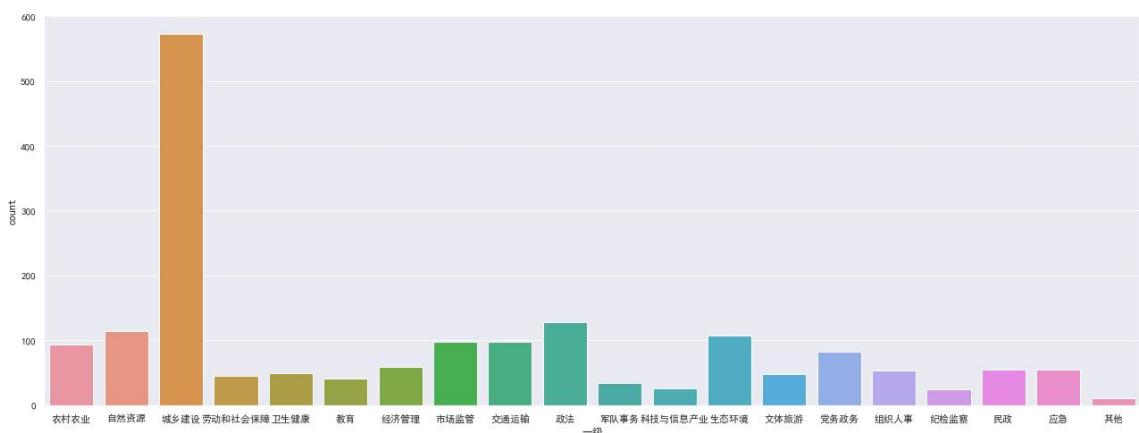


图 6-1 信访一级分类

根据信访件的四级内容分类标准，信访 4 级分类标准中的一级分类类别总数共计 20 类。而当一级分类为城乡建设时，对应的二级分类类别及其数目如下。此时，主要的二级分类类别有：城市建设与管理、城乡规划、村镇建设、工程管理、国有土地上房屋征收与补偿、集体土地上房屋拆迁与补偿、建筑市场、其他、住房保障与房地产，共计 9 类，且对应数目依次为：430、10、5、6、15、11、16、4、75。统计结果，如下图所示。

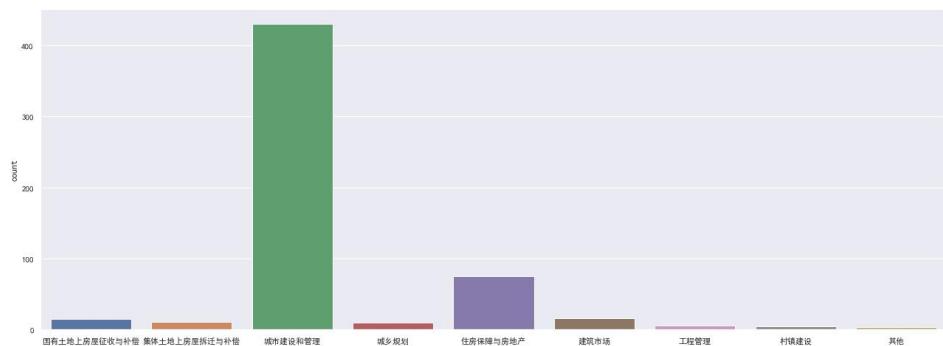


图 6-2 信访二级分类

在一级分类为城乡建设时，且对应的二级分类为城市建设与管理时，对应的三级分类类别及其数目如下。此时，主要的三级分类类别有：城管执法、城市公共设施、居民服务设施、园林绿化环卫，共计 4 类，且对应数目依次为：97、186、13、134。统计结果，如下图所示。

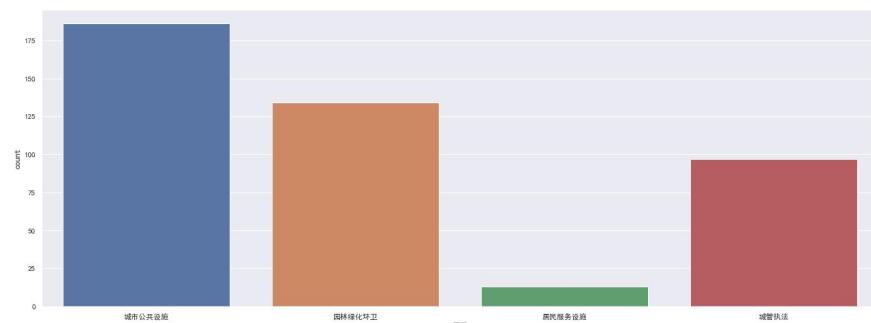


图 6-3 信访三级分类

在一级分类为城乡建设时，且对应的二级分类为城市建设与管理，对应的三级分类类别为城市公共设施时，四级分类对应的信息如下。

表 6-2 信访四级分类

1	城市自助缴费机破损、缺失、倾斜或功能缺失
2	城市牌匾标识许可设施破损、断字缺亮、倾斜
3	部门管养范围外坐椅破损、缺失、倾斜
4	燃气经营者的职责
5	部门管养范围外坐椅破损、缺失、倾斜（无产权、管理维护单位或产权、管理维护单位不清晰）
...	...
182	部门管养范围外绿地护栏破损、脱落、缺失
183	部门管养范围外其它交接箱破损、倾斜、箱体严重锈蚀
184	部门管养范围外雨水井盖出现破损、移位或丢失的（代维）（无产权、管理维护单位或产权、管理维护单...
185	部门管养范围外绿化平侧石破损、缺失、移位（无产权、管理维护单位或产权、管理维护单位不清晰）
186	城市施工告示牌破损、倾斜

在一级分类为城乡建设，且对应的二级分类为城市建设与管理，对应的三级分类类别为城市公共设施时，对应的四级分类对应的信息中的关键词及其权重，通过可视化大小的方式展现。如下图所示。



图 6-4 信访四级分类

## 6.2 基于 Bert-BiGRU 的信访分类模型

设计了基于 Bert-BiGRU 的文本分类模型，利用 Bert 预训练模型和 BiGRU 神经网络搭建模型。将处理好的文本向量作为输入，得到输出结果，通过对结果分析，获取历史信访数据中有价值信息。

预处理后的信访件诉求数据，输入到如下图所示的模型中。

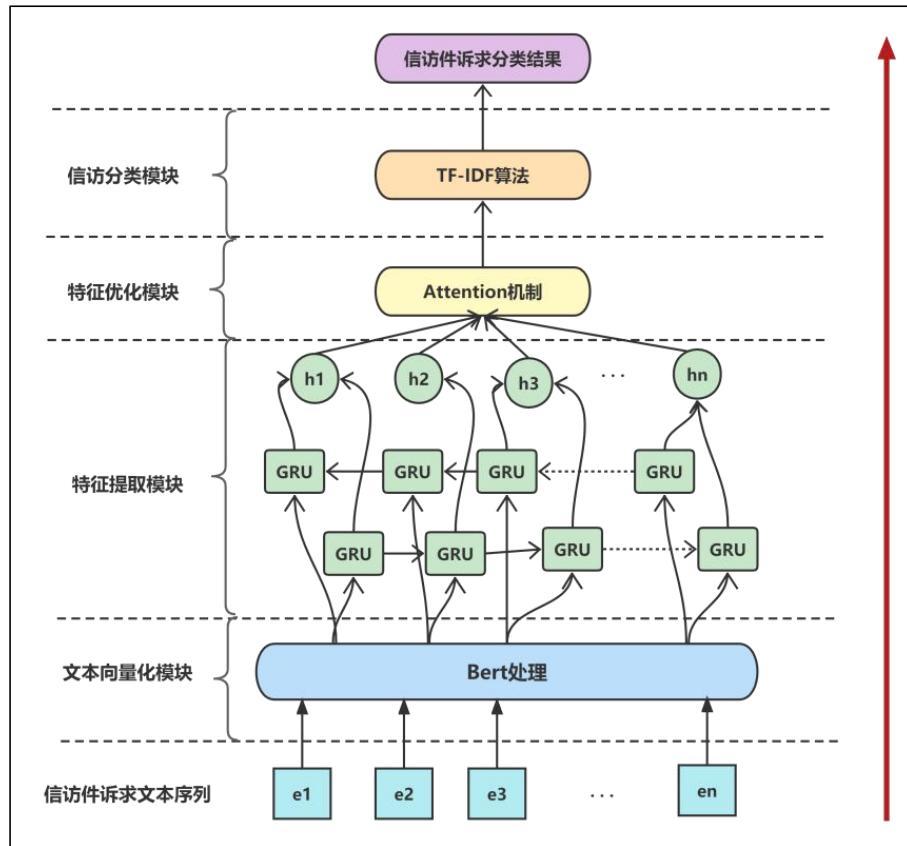


图 6-5 信访分类模型

传统循环神经网络模型存在长期依赖，所造成以下问题：梯度消失、模型复杂参数、训练时间长等。我们设计了一种基于注意力机制的 Bert-BiGRU 文本处理方法。

- (1) 使用 Bert 预训练模型生成词向量，通过 BiGRU 提取文本特征；
- (2) 采用 Attention 机制对 BiGRU 输出的每个时序向量计算加权，作为特征向量；
- (3) 通过 TF-IDF(term frequency-inverse document frequency) 算法对数据进行分类。

值得注意的是，Attention 机制解决了长期依赖问题，并且 BiGRU 减少了模型的复杂度，提高了分类效率。

### 6.2.1 信访文本向量化

使用 Bert 模型进行文本向量化表示。Bert 除了可以像 Word2vec 一样对文本中的每个字符都建立对应的向量输出，还可以处理句子级别的向量。

实验过程中，我们设计使用交叉熵损失函数，模型训练优化使用 Adam 方法，实验结果验证阶段使用 K 折交叉验证方法，K 值设置为 10。

同时，将 Bert 训练模型的相关重要参数设计如下。

表 6-3 Bert 模型参数

BERT 参数	参数值
最大序列长度	1000

每批训练集数据量	64
初始学习率	1e-5
模型迭代轮次	5
Dropout 随机失活率	0.5

训练好的模型可以将任何一个词语转换成 60 维的向量，从而将非结构化的文本数据转化成向量，将其作为自然语言处理算法的输入，进而对数据进行分类。

### 6.2.2 信访文本特征提取

我们设计使用 BiGRU 完成文本特征提取的工作，BiGRU 的输入是经 Bert 转化而来的向量，它可以把任意长度的句子转化为特定维度的浮点数向量，同时保留向量中比较重要的单词，让记忆保存比较长的时间。

由于 RNN 是一个有偏模型，导致越后置的信息占比越大，所以使用双向 BiGRU，即一层从左向右，一层从右向左，保证了处理数据时上下文都能兼顾。

GRU 是 LSTM 的一个变体，相对于 RNN 和 LSTM，GRU 改进了记忆处理结构，选择什么作为输出、具体遗忘哪些信息显得并不重要了，真正重要的是这类模型中具有记忆结构，同时模型能够根据当前输入对记忆进行修改，并最终结合记忆和当前输入进行计算、处理并输出，这是 GRU 能够取得成功的根本所在，GRU 比 LSTM 和 RNN 的记忆处理机制更加合理、更加高效。GRU 模型如下图所示。

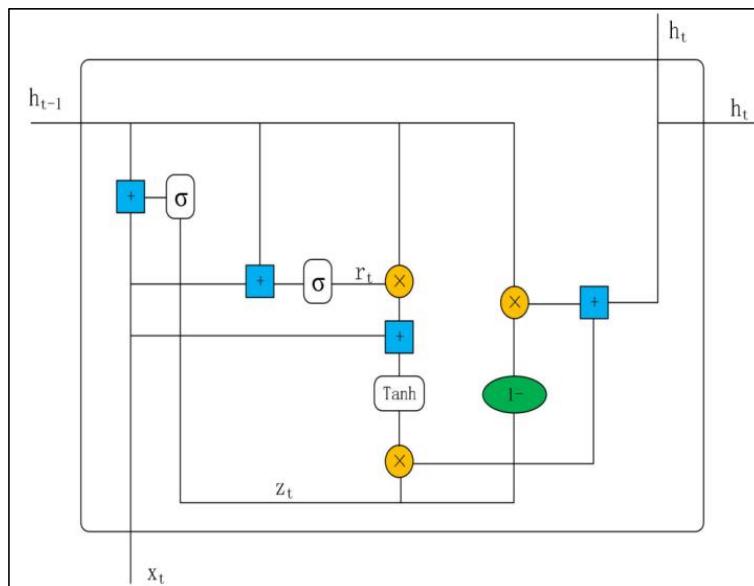


图 6-6 GRU 模型图

### 6.2.3 信访文本特征优化

对于 BiGRU 提取的特征，经过 Attention 的筛选会获得最优特征。

Attention 是一种权重参数的分配机制，协助模型捕捉重要信息，弥补了 BiGRU 在

捕捉特征信息时受到序列长度限制的缺陷。模型使用的是多头自注意力机制，self-Attention 指的是 Q (Query)、K (Key)、V (Value) 三个矩阵均来自同一输入，首先计算 Q 与 K 之间的相似度，再将其结果归一化为概率分布，然后再乘以矩阵 V 就得到权重求和表示。Multi-head Attention 本质是多个 self-Attention 的结合，这样给了模型更大的容量。

Attention 机制能够处理 GRU 模型的输出结果，然后通过训练计算出一组权值，将所有结果按照权值进行分配占比，并行计算的同时兼顾了任意时刻的信息保留程度，如下图所示。

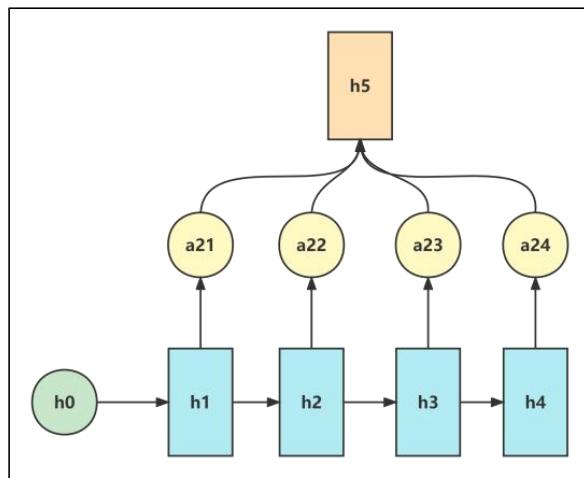


图 6-7 Attention 模型

在深度学习模型中，输入的每个文字是由一系列成对的<地址 Key，元素 Value>构成，而目标中的每个文字是 Query，那么，可以用 Key、Value、Query 重新解释如何计算 context vector。

通过计算 Query 和各个 Key 的相似性，得到每个 Key 对应 Value 的权重系数，权重系数代表信息的重要性，亦即 attention score，Value 则是对应的信息。

再对 Value 进行加权求和，得到最终的 context vector。Attention 的计算公式为：

$$\text{Attention}(\text{Query}, \text{Source}) = \sum_{i=1}^{L_x} \text{Similarity}(\text{Query}, \text{Key } i) \times \text{Value}_i$$

利用 Attention 筛选的最优特征，可以完成后续的信访相关下游任务。

#### 6.2.4 信访文本分类算法

在进行文本分类时，使用 TF-IDF 算法将信访数据按照信访四级类别标准，从上至下依次分类。

TF-IDF 是一种常用于信息处理和数据挖掘的加权技术。根据字词在文本中出现的次数和在整个语料中出现的文档频率计算一个字词在整个语料中的重要程度，使用该方法可以提取关键字并计算关键词出现次数。

词频表示词条或者关键字在文本中出现的频率。该值通常会被归一化，以防止它偏向长的文件，如下式所示。

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

$$TF_w = \frac{\text{在某一类中词条 } w \text{ 出现的次数}}{\text{该类中所有的词条数目}}$$

Softmax 用于多分类过程中，它可以将多个神经元的输出，映射到 (0, 1) 区间内，可以将它理解成概率。最后，选取输出结点的时候，可以选取概率最大的结点，作为目标结果。假设有一个数组  $V$ ,  $i$  表示  $V$  中的第  $i$  个元素，那这个元素的 Softmax 值为：

$$S_i = \frac{e^i}{\sum_j e^j}$$

经过 TF-IDF 算法层筛选的特征，使用 Softmax 层会得到文本所属类别的概率值，选取概率最大项对应的分类即该文本所属类别。

### 6.2.5 信访文本分类评价指标

#### (一) 准确率 Precision

计算预测正确样本占被预测所有样本的比例，Precision 体现了模型对负样本的区分能力，Precision 越高，模型对负样本的区分能力越强。计算公式为：

$$Precision = \frac{n_{correct}}{n_{total}}$$

#### (二) 召回率 Recall

计算预测结果的某一分类中预测正确的样本占该分类中所有样本的比例，Recall 体现了模型对正样本的识别能力，Recall 越高，模型对正样本的识别能力越强。计算公式为：

$$Recall = \frac{\text{out\_correct}}{\text{out\_all}}$$

#### (三) F 值

F 值是准确率和召回率的调和平均值，是两者的综合， $F_{score}$  越高，说明模型越稳健。计算公式为：

$$F_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

## 6.3 信访文本四级分类

根据上述对照目前信访的 4 级分类标准的数据挖掘分析，在智能判断信访件所属分

类时，先信访件所属的一级分类，再依次判断信访件的二级分类、三级分类和四级分类。整体的四级分类模型如下图所示。

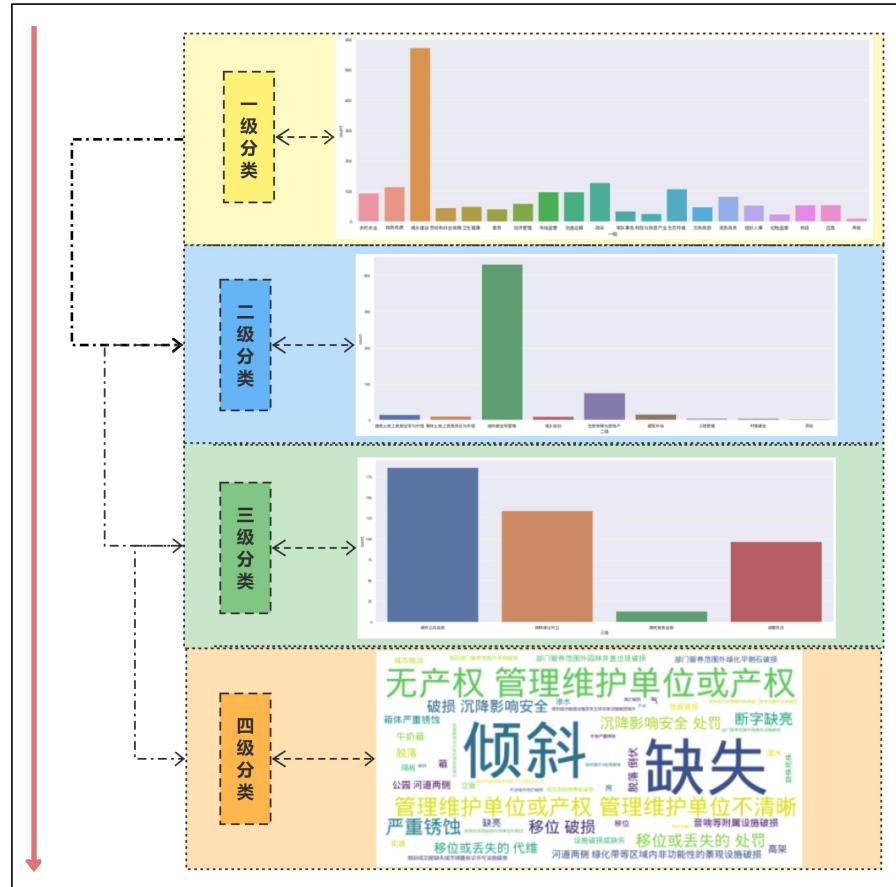


图 6-8 信访文本四级分类模型

同时，观察到信访的 4 级分类标准中的公开数据集中，部分级别的样本数据类别不平衡，例如：一级分类中的“城乡建设”类别明显多于其他类别，而二级分类中的部分类别所拥有的数据样本数量过少。故，针对信访数据样本的不平衡分类问题，我们设计通过以下 2 个方面进行改进。

(1) 改变数据分布，降低不平衡度，包括采样的方法（过采样算法、欠采样算法）和数据合成的方法。其中过采样是把小众类复制多份，而欠采样则是从大众类中剔除一些样本，或者说只从大众类中选取部分样本。

(2) 使用优化算法，分析已有算法在面对不平衡数据的缺陷，改进算法或者提出新算法来提升少数类的分类准确率，主要通过代价敏感和集成学习方式。代价敏感通过修改损失函数使得模型更加重视少数类，集成学习通过将多个分类器的结果集成提高整体分类准确度。

值得注意的是，在一级分类为农村农业，且对应的二级分类为村务管理，对应的三级分类类别为集体资产管理时，对应的四级分类的信息中的关键词及其权重，如下图所示。



图 6-9 信访文本四级分类

从上图可以明显看到，由于此时对应的四级分类的信息量过少，只有四条对应的四级分类样本数据，我们设计在基于 TF-IDF 分类的基础上，通过建立进一步的精细分类，将只有少量的分类样本数据的四级分类组合成文本向量信息，使用前述计算文本内容相似度的计算，进一步提高此类情况的分类精确度。

## 7. 信访件单位的智能分派

### 7.1 事权单位实体识别

和分词和词性标注一样，命名实体识别也是自然语言处理里面一项非常基础但又非常重要的技术。命名实体识别指的是指从文本中识别出命名性质的称项，为关系抽取等任务做铺垫。

狭义上来说，命名实体识别指的是识别出人名、地名和组织机构名这三类命名实体。因为像时间、货币名称等会构成规律明显的实体类型，因此可以直接用正则表达式等方式来进行识别。一个命名实体识别的例子如下：



图 7-1 命名实体识别

当然，在特定的领域中，会相应地定义领域内的各种实体类型。

我们设计，针对信访的事权单位，建立相应的事权单位名称实体。

假设，任务是识别出信访文本中的人名、地名、机构名。则可以定义以下规则来对句子进行标记。

- (1) B-PER: 人名的第一个字。
- (2) I-PER: 人名的后几个字。
- (3) B-ORG: 机构名的第一个字。
- (4) I-ORG: 机构名的后几个字。
- (5) B-LOC: 地名的第一个字。
- (6) I-LOC: 地名的后几个字。
- (7) O: 非实体名称的其他字。

通过定义上面所述的规则,就可以把一个句子转化成为标注的形式。例如下图所示:



图 7-2 句子转化成标注

而现在的命名实体识别的任务就是输入一个句子,然后输出该句子每个字相应的标注,然后再通过规则转换得到相应的事权单位实体名称。

## 7.2 双向长短时记忆网络

长短时记忆网络,也简称为 LSTM( Long Short Term Memory)。其是简单循环神经网络的一种重要变体。LSTM 的整体架构与标准的循环神经网络一样。与之不同的是,LSTM 每个单元的内部结构更复杂,其主要是通过定义三个门来实现对长期历史信息的记忆。

双向长短时记忆网络,也简称为 BiLSTM,用来进行命名实体识别,得到事权单位名称。

因为神经网络要求输入的是数值向量,所以将每个字都转成对应的向量形式。在这里,我们设计将每个字用一个长度为  $1 \times 20$  的向量进行表示。

转换成为向量之后,就直接进入到 BiLSTM 网络中,设置正向网络和逆向网络的输出为 24。合并两者得到  $1 \times 48$  的向量。然后再乘以一个矩阵形状为  $48 \times 7$  的矩阵得到  $1 \times 7$  的向量。该向量的每个位置负责预测一个标签的值。

定义模型的一些超参数,这里定义为全局变量。值得注意的是,为了便于模型训练,我们将词向量、循环单元等超参数都设置得比较小。可根据实际情况,适当的增大这些值。具体超参数,我们设计如下。

表 7-1 模型超参数

模型超参数	参数值
每次训练使用的样本数	16
每个句子 padding 后的长度	100
词向量的长度	60

标签总数	4
学习率	0.001
LSTM 单元数	2

对于损失函数计算函数。我们设计使用交叉熵损失函数来训练模型，其计算公式如下：

$$L = - \sum_i (y_i \log \hat{y}_i)$$

在上式中， $y_i$ 为真实标签值， $\hat{y}_i$ 为预测值。在训练时需要将标签转换成为独热编码的形式。

### 7.3 BiLSTM 与 CRF 算法

CRF 表示条件随机场，而条件随机场的预测是一个寻找最大路径的问题。也许是给定一个句子，得出的是许多满足要求的标注序列，然后寻找出概率最大的那一条序列。简单来说，条件随机场在预测时，会考虑整个句子之间的上下文关系。

在 BiLSTM 中，最终的输出是每个时刻都对应一个向量。而向量的每个位置负责预测一个类别标签。每个时刻之间的预测都是独立的。因此，BiLSTM 在预测时没有依赖上下文关系。

因此，为了使模型在预测时也能考虑上下文之间的关系，我们设计在 BiLSTM 的输出加上一层 CRF。

条件随机场与隐马尔可夫模型类似，是一种比较重要的统计模型。而且一般都用于序列标注任务当中。一般而言，最简单的是线性链条件随机场。条件随机场可用下图进行表示：

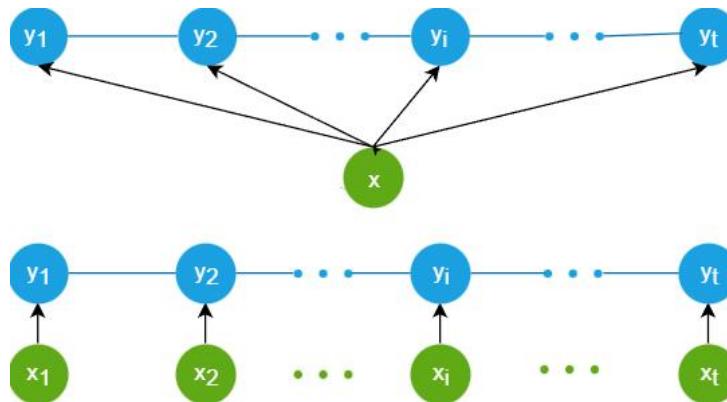


图 7-3 CRF 算法

### 7.4 信访件转送到事权单位

根据智能分类结果及问题属地等信息，我们设计先根据“信访事项问题属地代码 ID”信息，同时结合《民呼我为机构代码表》中的“行政区划代码”信息，进行初步的直接

匹配，若存在匹配且是唯一匹配，则将该信访件选择转送到相应的事权单位。

若在初步的匹配中，存在的匹配不是唯一的，则需要进行进一步的二次匹配。结合前期的四级智能分类结果，以及通过事权单位实体识别的结果，与《民呼我为机构代码表》中的“机构名称”信息、“机构全称”信息和“机构简称”信息，进行匹配。

此时，匹配方式主要采用两种，首先采用概要提取的结果、单位实体识别的结果和四级智能分类结果，分别依次与机构代码表中的“机构名称”信息、“机构全称”信息和“机构简称”信息进行关键词匹配，并根据词频等统计信息计算相似度值。然后，基于前面所描述的文本相似度计算方法，计算出信访件和事权单位之间的相似度值。综合上述两次计算出的相似度值，确定最终相似度最高的事权单位，并将该信访件选择转送到此事权单位。

整体的智能分派，流程图如下。

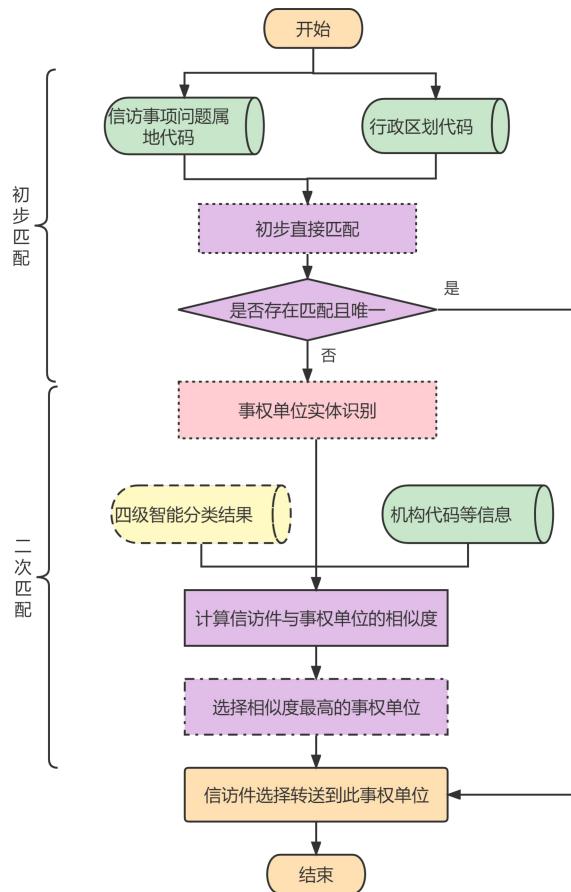


图 7-4 智能分派流程图

## 8. 基于信访大数据的缠访和群体事件预警模型

在赛题前 4 问结果的基础上，我们拓展了作品的内容。在已有的缠访和群体事件预警模型的基础上，利用信访件的相似度判别和意图分析能够有效解决现有模型重复信访

判定不清的问题。

目前针对缠访和群体事件预警模型的实证研究比较少，其中较为流行的是董英杰提出的分解信访过程导出指标体系，以信访发生的整个过程为分析框架将信访指标分为3级。其中包括4个一级指标，7个二级指标，以及16个三级指标。指标具体构成见表8-1。这三个指标层次涵盖了信访的整个过程，从信访的发生、过程到结果都能全面概括。

表8-1 信访引发社会风险程度指标构成

目标	一级指标	二级指标	三级指标
信访引发社会风险的程度	信访环境	社会环境	两会国庆期间来访件次
			两会国庆期间来信件次
	信访意图	直接利益诉求	求决类件次
			申诉类件次
		非直接利益诉求	检举揭发类件次
			批评建议类件次
	信访行为	信访规模	来信、来电、来访总件次
		信访强度	集体访件次
			联名信件次
	信访结果	重复信访程度	个人重信次数
			集体重信次数
			个人重访件次
			集体重访件次
		质变程度	越级上访比重

由表8-1可见，信访意图和重复信访程度是缠访和群体事件预警模型的重要因素，目前已有很多研究基于以上指标进行聚类分析来构建预警模型，但是这些模型都存在重复信访判定不清的问题。为此，我们提出首先对信访件进行相似度判别，随后再进行聚类分析，并建立了缠访和群体事件预警决策树，如图8-1所示。

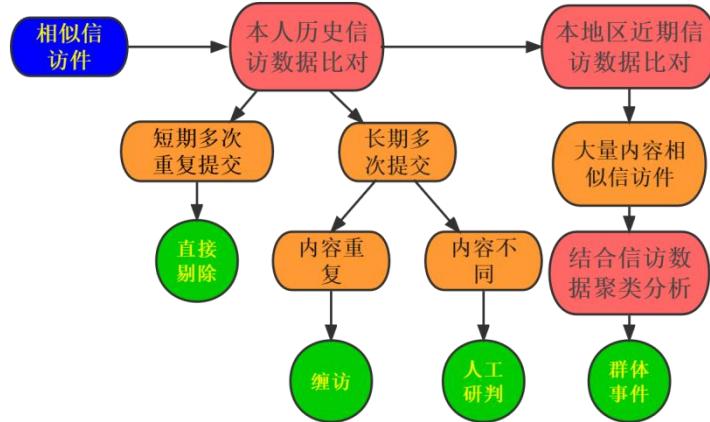


图 8-1 缠访和群体事件预警决策树

## 9. 作品的意义与价值

### 9.1 已有成效

利用 AI 辅助剔除重复和无效的信访数据，并对群众信访诉求进行概要提取，减轻业务员的负担。

有效识别和预警屡次上访、跨区域、越级重复上访以及有过恶意上访记录的缠访者，为接访提供联动方案支持。

提取对群体事件具有预测力的信访事件，利用信访数据建立科学的群体事件预警模型，为政府有效预防信访引起的群体事件提供参考。

全方位提炼发现重要信访信息。从信访大数据中发现隐藏的信访信息和规律，为领导提供涉及城市经济、能源、环境、卫生、城市建设和社会管理等各方面的一手信息，服务领导决策。

准确感受信访群众的诉求需求。对信访文本进行智能分析，让各级领导直接了解到通过信访渠道反映的社情民意，方便领导听到基层最真实的“声音”，为领导提供做好信访人情绪稳定工作的有效策略。

实时、动态提供信访趋势预测。帮助领导实时掌握信访发展趋势预测，从而合理配置有限的行政资源，未雨绸缪，下好社会治理先手棋。

### 9.2 预期成效

积极推进电子政务和智慧城市的建设，将多渠道的数据集中和共享化处理，建设全国一体化的多部门大数据共享中心；

发现新形势下信访治理的不够完善之处，通过主动掌握数据来更大幅度实现信访制度本身价值；

加快各个环节和部门的沟通和协同发展，形成大数据的合力聚集状态，在信访治理

全领域通过大数据技术实现预警、接访、决策和考核的全过程精准化治理；

给国家在信访领域的治理挑战以创新的方式提供更科学的解决方案，实现国家的长治久安与政治稳定，实现人民美好生活的目标。