
目 录

第一章 群众留言分类	3
1.1 数据分析	3
1.2 RoBERTa 模型	3
1.3 RoBERTa 文本分类过程	4
1.3.1 自定义数据集加载	4
1.3.2 PaddleHub 一键加载 RoBERTa	4
1.3.3 选择优化策略和运行配置	5
1.3.4 构建网络并创建分类迁移任务进行 Fine-tune	6
1.3.5 Finetune 微调	6
1.5 实验结果	6
第二章 热点问题挖掘	8
2.1 数据分析	8
2.2 DBSCAN 聚类算法	8
2.3 实验结果	9
第三章 答复意见的评价	11
3.1 评价指标介绍	11
3.2 基于 ROUGE 的答复评价	13
总 结	14
参考文献	14

前 言

目前，大部分电子政务系统还是依靠人工根据经验对留言进行分类，存在工作量大、效率低，且差错率高等问题。某一时段内群众集中反映的某一问题可称为热点问题，及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。

本组基于预训练模型 RoBERTa 和百度 AI Studio 平台支持的 Tesla V100 深度学习环境，经过数据分析和预处理、构建网络、生成预测数据的 Reader、切换到预测的 Program、加载预训练的参数、运行 Program 进行预测，得到群众留言内容的一级标签分类结果。

本组对附件 3 给出的群众留言数据进行数据分析，基于 DBSCAN 聚类算法实现设定给定的邻域距离 ϵ 和邻域最小样本个数 MinPts 的方式热点问题的聚类。

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。本组借鉴“对话系统”评价指标进行分析，介绍 BLEU、ROUGE 和 METEOR 几种评价方法，并实现基于 ROUGE-L 评价结果。

第一章 群众留言分类

基于预训练模型 RoBERTa 和百度 AI Studio 平台支持的 Tesla V100 深度学习环境，经过数据分析、构建网络、生成预测数据的 Reader、切换到预测的 Program、加载预训练的参数、运行 Program 进行预测，得到群众留言内容的一级标签分类结果。

1.1 数据分析

参考附件 1 提供的内容分类三级标签体系，读取附件 2 给出的群众留言数据，显示前三行数据信息并获取数据维度，如图 1-1 所示。

	留言编号	留言用户	留言主题	留言时间	
0	24	A00074011	A市西湖建筑集团占道施工有安全隐患	2020/1/6 12:09:38	\n\\t\\t\\t\\t\\t\\t\\t\\t\\tA3区大道西行便道，未管所路口至加
1	37	U0008473	A市在水一方大厦人为烂尾多年，安全隐患严重	2020/1/4 11:17:46	\n\\t\\t\\t\\t\\t\\t\\t\\t\\t位于书院路主干道的在水一方大厦一
2	83	A00063999	投诉A市A1区苑物业违规收停车费	2019/12/30 17:06:14	\n\\t\\t\\t\\t\\t\\t\\t\\t\\t尊敬的领导：A1区苑小区位于A1区火

图 1-1 群众留言数据基本情况

1.2 RoBERTa 模型

RoBERTa 模型（A Robustly Optimized BERT，即鲁棒性强力优化的 BERT 方法）是 BERT 的改进版，在模型规模，算力和数据上，与 BERT 相比主要有以下几点改进：

- 1) 更大的模型参数量。论文提供的训练时间来看, 模型使用 1024 块 V100 GPU 训练 1 天的时间;
- 2) 更大 batch size。RoBERTa 在训练过程中使用更大的从 256 到 8000 不等的 batch size;
- 3) 更多的训练数据。包括 CC-NEWS 等在内的 160GB 纯文本, 而最初的 BERT 使用 16GB BookCorpus 数据集和英语维基百科进行训练。

另外，RoBERTa 在训练方法上有以下改进：

- 1) 去掉下一句预测(NSP)任务;
- 2) 动态掩码。BERT 依赖随机掩码和预测 token。原版的 BERT 实现在数据预处理期间执行一次掩码, 得到一个静态掩码; 而 RoBERTa 使用动态掩码, 即每次向模型输入一个序列时都会生成新的掩码模式。这样, 在大量数据不断输入的过程中, 模型会逐渐适应不同的掩码策略, 学习不同的语言表征;

- 3) 文本编码。Byte-Pair Encoding 是字符级和词级别表征的混合，支持处理自然语言语料库中的众多常见词汇。原版的 BERT 实现使用字符级别的 BPE 词汇，大小为 30K，是

在利用启发式分词规则对输入进行预处理之后学得的。Facebook 研究者没有采用这种方式,而是考虑用更大的 byte 级别 BPE 词汇表来训练 BERT,这一词汇表包含 50K 的 subword 单元,且没有对输入作任何额外的预处理或分词。

RoBERTa 建立在 BERT 的语言掩蔽策略的基础上,修改 BERT 中的关键超参数,包括删除 BERT 的下一个句子训练前目标,以及使用更大的 batch size 和学习率进行训练。RoBERTa 也接受比 BERT 多一个数量级的训练,时间更长。这都使得 RoBERTa 表示能够比 BERT 更好地推广到下游任务。

1.3 RoBERTa 文本分类过程

1.3.1 自定义数据集加载

使用飞桨预训练模型应用工具 PaddleHub 库加载自定义数据集,用户仅需要继承 HubDataset 类,替换数据集存放地址即可,如图 1-2 所示。

```
from paddlehub.dataset.base_nlp_dataset import BaseNLPDataset

class DemoDataset(BaseNLPDataset):
    """DemoDataset"""
    def __init__(self):
        # 数据集存放位置
        self.dataset_dir = "path/to/dataset"
        super(DemoDataset, self).__init__(
            base_path=self.dataset_dir,
            train_file="train.tsv",
            dev_file="dev.tsv",
            test_file="test.tsv",
            # 如果还有预测数据 (不需要文本类别label), 可以放在predict.tsv
            predict_file="predict.tsv",
            train_file_with_header=True,
            dev_file_with_header=True,
            test_file_with_header=True,
            predict_file_with_header=True,
            # 数据集类别集合
            label_list=["0", "1"])
        dataset = DemoDataset()
```

图 1-2 加载自定义数据集

具体使用如下:

1. 数据集文件编码格式建议为 utf8 格式;
2. 如果相应的数据集文件没有上述的列说明,如 train.tsv 文件没有第一行的 text_a label,则 train_file_with_header=False;
3. 针对预测数据(没有文本类别),可以将预测数据存放在 predict.tsv 文件,去掉 label 一列即可;

1.3.2 PaddleHub 一键加载 RoBERTa

使用图 1-3 所示的方式实现 PaddleHub 一键加载 RoBERTa。

```
import paddlehub as hub
module = module = hub.Module(name="chinese-roberta-wwm-ext")

[2020-05-04 16:55:52,214] [ INFO ] - Installing chinese-roberta-wwm-ext module
Downloading chinese-roberta-wwm-ext
[=====] 100.00%
Uncompress /home/aistudio/.paddlehub/tmp/tmpso2gcels/chinese-roberta-wwm-ext
[=====] 100.00%
[2020-05-04 16:56:07,870] [ INFO ] - Successfully installed chinese-roberta-wwm-ext-1.0.0
```

图 1-3 PaddleHub 一键加载 RoBERTa

其他语义模型（如 ERNIE, BERT 等），如图 1-4 所示，只需要更换 Module 中的 name 参数即可。

模型名	PaddleHub Module
ERNIE, Chinese	<code>hub.Module(name='ernie')</code>
ERNIE 2.0 Tiny, Chinese	<code>hub.Module(name='ernie_tiny')</code>
ERNIE 2.0 Base, English	<code>hub.Module(name='ernie_v2_eng_base')</code>
ERNIE 2.0 Large, English	<code>hub.Module(name='ernie_v2_eng_large')</code>
RoBERTa-Large, Chinese	<code>hub.Module(name='roberta_wwm_ext_chinese_L-24_H-1024_A-16')</code>
RoBERTa-Base, Chinese	<code>hub.Module(name='roberta_wwm_ext_chinese_L-12_H-768_A-12')</code>
BERT-Base, Uncased	<code>hub.Module(name='bert_uncased_L-12_H-768_A-12')</code>
BERT-Large, Uncased	<code>hub.Module(name='bert_uncased_L-24_H-1024_A-16')</code>
BERT-Base, Cased	<code>hub.Module(name='bert_cased_L-12_H-768_A-12')</code>
BERT-Large, Cased	<code>hub.Module(name='bert_cased_L-24_H-1024_A-16')</code>
BERT-Base, Multilingual Cased	<code>hub.Module(name='bert_multi_cased_L-12_H-768_A-12')</code>
BERT-Base, Chinese	<code>hub.Module(name='bert_chinese_L-12_H-768_A-12')</code>

图 1-4 PaddleHub 支持的语义模型

1.3.3 选择优化策略和运行配置

1) 优化策略

针对 BERT 类任务，PaddleHub 封装迁移学习优化策略 AdamWeightDecayStrategy:

- `learning_rate`: Fine-tune 过程中的最大学习率;
- `weight_decay`: 模型的正则项参数，默认 0.01，如果模型有过拟合倾向，可适当调高这一参数;
- `warmup_proportion`: 如果 `warmup_proportion>0`，例如 0.1，则学习率会在前 10%的 steps 中线性增长至最高值 `learning_rate`;
- `lr_scheduler`: 有两种策略可选(1) `linear_decay` 策略学习率会在最高点后以线性方式衰减; `noam_decay` 策略学习率会在最高点以多项式形式衰减;

2) 运行配置

RunConfig 主要控制 Fine-tune 的训练，包含以下可控制的参数:

- use_cuda: 是否使用 GPU 训练，默认为 False;
- checkpoint_dir: 模型 checkpoint 保存路径，若用户没有指定，程序会自动生成;
- num_epoch: Fine-tune 的轮数;
- batch_size: 训练的批大小，如果使用 GPU，请根据实际情况调整 batch_size;
- strategy: Fine-tune 优化策略;

1.3.4 构建网络并创建分类迁移任务进行 Fine-tune

基于 PaddleHub 提供的预训练模型 RoBERTa 和准备要迁移的数据集，组建 Task，如图 1-5 所示。

```
inputs, outputs, program = module.context(
    trainable=True, max_seq_len=128)

# Use "pooled_output" for classification tasks on an entire sentence.
pooled_output = outputs["pooled_output"]

feed_list = [
    inputs["input_ids"].name,
    inputs["position_ids"].name,
    inputs["segment_ids"].name,
    inputs["input_mask"].name,
]

cls_task = hub.TextClassifierTask(
    data_reader=reader,
    feature=pooled_output,
    feed_list=feed_list,
    num_classes=dataset.num_labels,
    config=config,
    metrics_choices=["f1"])

[2020-05-04 16:56:10,207] [ INFO ] - Load pretraining parameters from /home/aistudio/.paddlehub/modules/chinese_roberta
```

图 1-5 构建网络并创建分类迁移任务进行 Fine-tune

1.3.5 Finetune 微调

选择 finetune_and_eval 接口来进行模型训练，周期性的进行模型效果的评估并了解整个训练过程的性能变化，如图 1-6 所示。

```
In[11]: run_states = cls_task.finetune_and_eval()

[2020-05-04 16:57:08,072] [ INFO ] - Strategy with warmup, linear decay, slanted triangle learning rate, weight decay
/opt/conda/envs/python35-paddle120-env/lib/python3.7/site-packages/paddle/fluid/executor.py:804: UserWarning: There are r
warnings.warn(error_info)

[2020-05-04 16:57:10,516] [ INFO ] - Try loading checkpoint from hub_roberta_text_cls_demo/ckpt.meta
[2020-05-04 16:57:10,517] [ INFO ] - PaddleHub model checkpoint not found, start from scratch...
[2020-05-04 16:57:10,599] [ INFO ] - PaddleHub finetune start
[2020-05-04 16:57:16,312] [ TRAIN ] - step 10 / 230: loss=1.92618 f1=0.43333 [step/sec: 1.78]
[2020-05-04 16:57:21,635] [ TRAIN ] - step 20 / 230: loss=1.62800 f1=0.00000 [step/sec: 1.91]
[2020-05-04 16:57:26,868] [ TRAIN ] - step 30 / 230: loss=1.04570 f1=0.51613 [step/sec: 1.94]
[2020-05-04 16:57:32,074] [ TRAIN ] - step 40 / 230: loss=0.67254 f1=0.86842 [step/sec: 1.95]
```

图 1-6 Finetune 训练微调过程显示

1.5 实验结果

当 Finetune 完成后，我们使用模型来进行预测，整个预测流程大致可以分为以下几步：

构建网络、生成预测数据的 Reader、切换到预测的 Program、加载预训练的参数、运行 Program 进行预测,得到预测结果并显示如图 1-7 所示,另见附件“1-群众留言分类 RoBERTa 结果.xlsx”。

```
> 1 import numpy as np
2
3 inv_label_map = {val: key for key, val in reader.label_map.items()}
4
5 # Data to be predicted
6 data = [[d.text_a, d.text_b] for d in dataset.get_test_examples()]
7
8 index = 0
9 run_states = cls_task.predict(data=data)
10
11 predict = pd.DataFrame(columns=['留言主题', '留言详情', '一级标签'])
12
13 results = [run_state.run_results for run_state in run_states]
14 for batch_result in results:
15     # get predict index
16     batch_result = np.argmax(batch_result, axis=2)[0]
17     for result in batch_result:
18         predict = predict.append(pd.DataFrame({'留言主题': [data[index][0]],
19                                               '留言详情': [data[index][1]],
20                                               '一级标签': [inv_label_map[result]]},
21                                           ignore_index=True))
22         print("%s\tpredict=%s" % (data[index][0], inv_label_map[result]))
23     index = index + 1
24 writer = pd.ExcelWriter('result-ERNIE.xlsx')
25 predict.to_excel(writer, index=False)
26 writer.save()
```

```
[2020-05-04 16:21:25,569] [ INFO] - PaddleHub predict start
[2020-05-04 16:21:25,571] [ INFO] - The best model has been loaded
[2020-05-04 16:21:30,431] [ INFO] - PaddleHub predict finished.
M5市二中乱收费请管一管! predict=教育文体
```

图 1-7a 群众留言分类预测结果显示

留言主题	留言详情	一级标签
关于尽快修复A市沿江风光带亮化惠民工程的请求	市民承诺：“关于修复广场灯、	城乡建设
J市金龙小区乱象谁来管？	量要开到最大才能听到。这些乱	城乡建设
K市华南大酒店的烟尘污染让十多万人深受其害！	工河、垃圾不上街面、噪声不扰	环境保护
请B市教育局对小哈佛幼儿园进行调查	发现，如果家长有问题只能通过	教育文体
K5县各乡镇农机员强烈要求补缴事业单位劳动保障	、无地、无保险、无工作、无住	劳动和社会保障
河西五金机电大市场二期房产证何时能办下来？	一名业主，房子交房已经有3年	城乡建设
我超生要罚31万，我一家七口还吃饭吗？	签字，并要我们签下放弃复议的	卫生计生
望G市人社局为本人办理失业证，简化手续	生，准备好了所有资料办失业证	劳动和社会保障
请有关部门整治一下H3县的民办教育市场	违规办学校的办学压力。就目	教育文体
再婚双方各自有一孩，咨询婚后能否再生一孩？	婚双方各自有一孩子，婚后能不	卫生计生

图 1-7b 群众留言分类预测结果显示

本组首先对附件 3 给出的群众留言数据进行数据分析，基于 DBSCAN 聚类算法实现设定给定的邻域距离 ϵ 和邻域最小样本个数 MinPts 的方式对热点问题的聚类。

读取附件 3 给出的群众留言数据，显示前三行数据信息并获取数据维度，如图 2-1 所示。

图 2-1 群众留言数据情况

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)是典型的基于密度的聚类算法。与基于划分和层次的聚类方法不同，该算法将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。以下是 DBSCAN 算法中的几个重要概念：

核心对象：如果给定对象 E 邻域内的样本点数大于等于 $MinPts$ ，则称该对象为核心对象；

密度可达：对于样本集合 D ，给定一串样本点 $p_1, p_2 \cdots p_n$ ， $p=p_1, q=p_n$ ，假如对象 p_i 从 p_{i-1} 直接密度可达，那么对象 q 从对象 p 密度可达。

密度相连：存在样本集合 D 中的一点 o ，如果对象 o 到对象 p 和对象 q 都是密度可达的，那么 p 和 q 密度相联。

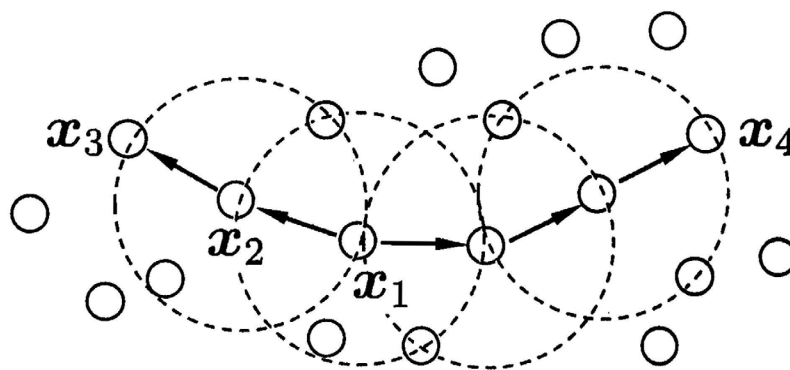


图 2-2 DBSCAN 概念解释图

上图直观显示 DBSCAN 中这几个概念：当 $\text{MinPts}=3$ 的时候，虚线圆圈为 ϵ 邻域， x_1 是核心对象， x_2 由 x_1 密度直达， x_3 由 x_1 密度可达， x_3 与 x_4 密度相连。

DBSCAN 算法步骤大致描述如下：

对于给定的邻域距离 ϵ 和邻域最小样本个数 MinPts ：

1. 遍历所有样本，找出所有满足邻域距离 ϵ 的核心对象的集合；
2. 任意选择一个核心对象，找出其所有密度可达的样本并生成聚类簇；
3. 从剩余的核心对象中移除 2 中找到的密度可达的样本；
4. 从更新后的核心对象集合重复执行 2-3 步直到核心对象都被遍历或移除。

DBSCAN 的优点在于：

- 1) 与 K-means 方法相比，DBSCAN 不需要事先知道要形成簇类的数量；
- 2) 与 K-means 方法相比，DBSCAN 可以发现任意形状的簇类；
- 3) DBSCAN 能够识别出噪声点；
- 4) DBSCAN 对于数据库中样本的顺序不敏感。

2.3 实验结果

本组首先对附件 3 给出的群众留言数据进行数据分析，设定邻域距离 $\epsilon=0.5$ 和邻域最小样本个数 $\text{MinPts}=3$ 时，得到基于 DBSCAN 聚类算法热点问题的文本聚类，结果如图 2-3 所示。

热度排名	热度指数	时间范围	地点/人群	问题描述
1	27	2019/07/07至2019/09/01	A市景园滨河苑车位	投诉A市伊景园滨河苑捆绑销售车位
2	20	2019/02/02至2019/12/30	A市限令	对A市禁摩限令处罚不解
3	13	2019/03/06至2019/12/16	A市808公交车线路	建议A市808公交车按原线路行车
4	12	2019/01/02至2019/04/22	市圈力度	请A市加快一圈二场三道建设力度
5	10	2019/01/16至2019/12/02	A市人才补贴通知问题	咨询A市人才购房补贴通知问题
6	10	2019/11/15至2020/01/09	丽新城小区附近站噪音	丽发新城小区附近的搅拌站噪音严重扰民
7	9	2019/01/03至2019/10/23	7号线南A市生态动物园	强烈建议将地铁7号线南延至A市生态动物园
8	9	2019/01/06至2019/09/12	A3区西湖街道茶场村五组规划	咨询A3区西湖街道茶场村五组的拆迁规划
9	8	2019/01/08至2019/09/09	市国家中心城市	请加快A市国家中心城市建刻不容缓
10	8	2019/04/17至2019/09/04	A4区中山路48号阳光城檀工地	A市A4区中山路48号阳光城檀悦工地夜间施工扰民

图 2-3 基于 DBSCAN 聚类算法热点问题聚类结果

第三章 答复意见的评价

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。本组借鉴“对话系统”评价指标进行分析，介绍 BLEU、ROUGE 和 METEOR 几种评价方法，并实现 ROUGE-L 的评价结果。

3.1 评价指标介绍

从下图 3-1 可以看出，黄色标识人类的主观角度，而蓝色表示系统的一些客观评价指标，可以用来衡量主观需求。对于对话而言，可以使用响应的适当性、流畅度、相关性；同时可以涵盖更多的话题、回复真实可信等。借鉴“对话系统”常见的 BLEU、ROUGE 和 METEOR 等评价指标。

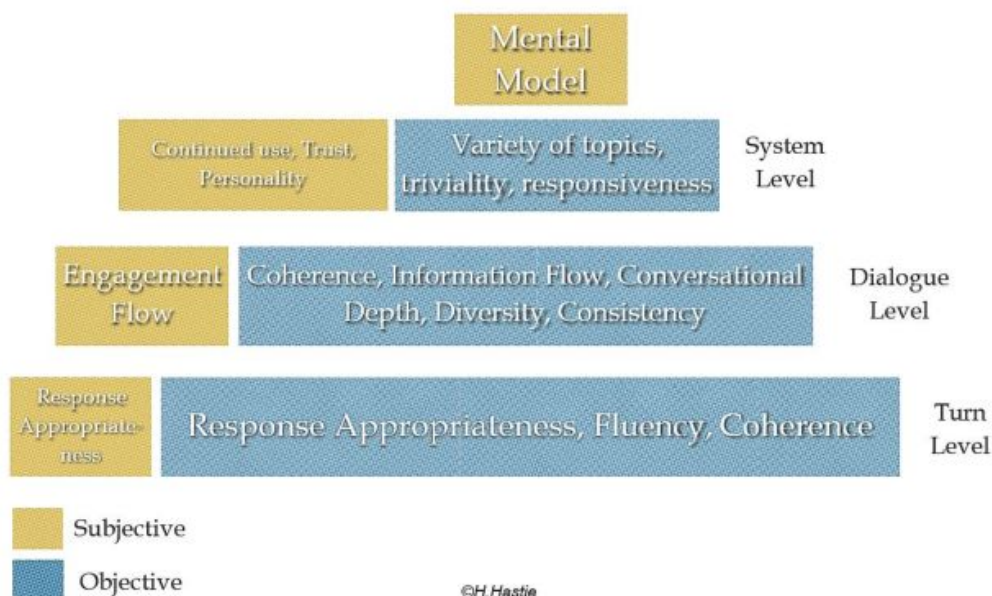


图 3-1 对话系统常用评价指标

BLEU

该评价指标有 IBM 在 2002 年提出，参考论文“BLEU: a Method for Automatic Evaluation of Machine Translation”，常作为机器翻译系统评价指标。其实就是统计生成响应和真实响应中的 n-gram 词组在整个训练语料中出现次数。公式如下所示：

$$P_n(r, \hat{r}) = \frac{\sum_k \min(h(k, r), h(k, \hat{r}_i))}{\sum_k h(k, r_i)} \quad (3-1)$$

$$\text{BLEU-N} := b(r, \hat{r}) \exp\left(\sum_{n=1}^N \beta_n \log P_n(r, \hat{r})\right) \quad (3-2)$$

第一个公式 P_n 用于计算 n -gram 短语词组在整个数据集中的准确度。 $h(k,r)$ 表示每个 n -gram 词组在真实响应中出现的次数，故上式表示每个 n -gram 词组在真实相应和生成响应中出现次数的较小值求和除以其在生成响应中出现次数求和。当然，我们需要考虑 n 的取值（一般取 1-4），得到第二个公式， β 表示各个 n -gram 的权重，而 $b(r,r^*)$ 表示长度惩罚因子，即我们不想让生成的答案长度太短，加一个惩罚因子来改善效果。

ROUGE

该指标常用于文本摘要领域，包含 ROUGE-N, ROUGE-L(最长公共子句, F-measure), ROUGE-W(带权重的最长公共子句,F-measure), ROUGE-S(不连续二元组, F-measure)四种。

ROUGE-L 是计算最长公共子序列的长度，其实就是寻找真是响应和生成响应之间的最长公共子序列长度，然后计算其 F-measure 分数。其与 BLEU 相似，都可以反应词语顺序，但是 ROUGE 的词可以不是连续的，而 BLEU 的 n -gram 要求词语必须连续出现。

$$R = \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|}$$

$$P = \max_j \frac{l(c_i, s_{ij})}{|c_{ij}|}$$

$$ROUGE_L(c_i, s_i) = \frac{(1 + \beta^2)RP}{R + \beta^2 P} \quad (3-4)$$

METEOR

METEOR 是基于 BLEU 进行一些改进，加入生成响应和真实响应之间的对齐关系。使用 WordNet 计算特定的序列匹配，同义词，词根和词缀，释义之间的匹配关系，改善 BLEU 的效果，使其跟人工判别共更强的相关性。同样也是使用 F-measure 的计算方法，如公式 3-5 所示。

$$F_{mean} = \frac{P_m R_m}{\alpha P_m + (1 - \alpha) R_m}$$

$$Pen = \gamma(frag)^\theta$$

$$METEOR = (1 - Pen)F_{mean} \quad (3-5)$$

3.2 基于 ROUGE 的答复评价

基于常用于文本摘要领域的 ROUGE-L(最长公共子句, F-measure)方法计算'rouge-1', 'rouge-2', 'rouge-3', 'rouge-4'的 F-measure 如下图 3-2 所示, 具体见附件“3-答复意见评价”。

留言编号	留言用户	留言主题	rouge-1	rouge-2	rouge-3	rouge-4
2549	A00045581	A00045581	{'f': 0.1839080459770115, 'p': 0.1777777777777778, 'r': 0.19047619047619047}	{'f': 0.03488372093023255, 'p': 0.033707865168539325, 'r': 0.03614457831325301}	{'f': 0.023529411764705882, 'p': 0.022727272727272728, 'r': 0.024390243902439025}	{'f': 0.011904761904761906, 'p': 0.011494252873563218, 'r': 0.012345679012345678}
2554	A00023583	A00023583	{'f': 0.05847953216374269, 'p': 0.056818181818181816, 'r': 0.060240963855421686}	{'f': 0.0, 'p': 0.0, 'r': 0.0}	{'f': 0.0, 'p': 0.0, 'r': 0.0}	{'f': 0.0, 'p': 0.0, 'r': 0.0}
2555	A00031618	A00031618	{'f': 0.2774566473988439, 'p': 0.2727272727272727, 'r': 0.2823529411764706}	{'f': 0.09356725146198829, 'p': 0.09195402298850575, 'r': 0.09523809523809523}	{'f': 0.03550295857988166, 'p': 0.03488372093023256, 'r': 0.03614457831325301}	{'f': 0.011976047904191616, 'p': 0.011764705882352941, 'r': 0.012195121951219513}
2557	A000110735	A000110735	{'f': 0.22429906542056077, 'p': 0.4, 'r': 0.15584415584415584}	{'f': 0.07619047619047618, 'p': 0.13793103448275862, 'r': 0.05263157894736842}	{'f': 0.019417475728155338, 'p': 0.03571428571428571, 'r': 0.013333333333333334}	{'f': 0.0, 'p': 0.0, 'r': 0.0}
2574	A0009233	A0009233	{'f': 0.30952380952380953, 'p': 0.9285714285714286, 'r': 0.18571428571428572}	{'f': 0.2682926829268293, 'p': 0.8461538461538461, 'r': 0.15942028985507245}	{'f': 0.25, 'p': 0.8333333333333334, 'r': 0.14705882352941177}	{'f': 0.23076923076923075, 'p': 0.8181818181818182, 'r': 0.13432835820895522}
2759	A00077538	A00077538	{'f': 0.1168831168831169, 'p': 0.125, 'r': 0.10975609756097561}	{'f': 0.0, 'p': 0.0, 'r': 0.0}	{'f': 0.0, 'p': 0.0, 'r': 0.0}	{'f': 0.0, 'p': 0.0, 'r': 0.0}
2849	A000100804	A000100804	{'f': 0.16666666666666669, 'p': 0.1686746987951807, 'r': 0.16470588235294117}	{'f': 0.012048192771084338, 'p': 0.012195121951219513, 'r': 0.011904761904761904}	{'f': 0.0, 'p': 0.0, 'r': 0.0}	{'f': 0.0, 'p': 0.0, 'r': 0.0}

图 3-2 基于 ROUGE-L 的答复评价

总 结

本组经过数据分析和预处理、构建网络、生成预测数据的 Reader、切换到预测的 Program、加载预训练 RoBERTa 的参数、运行 Program 进行预测，得到群众留言内容的一级标签分类结果。

本组对附件 3 给出的群众留言数据进行数据分析，基于 DBSCAN 聚类算法实现设定给定的邻域距离 ϵ 和邻域最小样本个数 MinPts 的方式对热点问题的聚类。

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。本组借鉴“对话系统”评价指标进行分析，介绍 BLEU、ROUGE 和 METEOR 几种评价方法，并实现基于 ROUGE-L 评价结果。

参考文献

- [1] Yinhan Liu, Myle Ott, Naman Goyal, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692
- [2] Devlin, Jacob and Chang, Ming-Wei, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805
- [3] Ester, M., Kriegel, H. P., Sander, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, Vol. 96, No. 34, pp. 226-231.
- [4] Sander, J., Ester, M., Kriegel, et al. Density-based clustering in spatial databases: The algorithm gbscan and its applications. Data mining and knowledge discovery, 2(2), 169-194.
- [5] PaddlePaddle <https://www.paddlepaddle.org.cn/>
- [6] AI studio <https://aistudio.baidu.com/aistudio/index>
- [7] Paddlehub <https://github.com/PaddlePaddle/PaddleHub>
- [8] 周志华. 机器学习 Machine learning. 清华大学出版社.
- [9] 深度学习对话系统理论篇--数据集和评价指标介绍
<https://zhuanlan.zhihu.com/p/33088748>