# 自然语言实验报告

指导老师：杨沐昀

|  |  |  |  |
| --- | --- | --- | --- |
| 学号 | 姓名 | 手机 | 邮箱 |
| 1140310519 | 贺成权 | 18245032253 |  |
| 11403105 | 沈沾健 |  |  |
| 11403105 | 陈安辉 |  |  |
| 11303 | 张环鑫 |  |  |
|  |  |  |  |

## Part one问题描述

1.Task Specification

Given a question and it's corresponding document, your system should select one or more sentences as answers from the document.

2.Data Format

1).An example:

芝加哥国际电影节最佳女主角是谁？ \t 芝加哥国际电影节(Chicago International Film Festival)是北美历史最久的评奖电影节。 \t 0

芝加哥国际电影节最佳女主角是谁？ \t 最佳女主角奖：余男《图雅的婚事》（中国） \t 1

芝加哥国际电影节最佳女主角是谁？ \t 芝加哥电影节每年10月举办，自1965年第一届电影节至今，芝加哥国际电影节已成为世界知名的年度电影盛会。 \t 0

芝加哥国际电影节最佳女主角是谁？ \t 我国著名电影导演张艺谋的《菊豆》曾于1990年首获该电影节最高奖--金雨果奖。 \t 0

芝加哥国际电影节最佳女主角是谁？ \t 芝加哥电影节组织是由电影制作人、形象艺术家麦克尔.库泽于1964年发起成立的。 \t 0

芝加哥国际电影节最佳女主角是谁？ \t 其宗旨为：通过电影和录像艺术手段加强不同文化背景人民之间的理解和沟通。 \t 0

芝加哥国际电影节最佳女主角是谁？ \t 评委会特别大奖：《图雅的婚事》（中国） \t 0

芝加哥国际电影节最佳女主角是谁？ \t 2002年，来自30多个国家和地区的90多部故事片、40多部短片参加了电影节，吸引了世界各地的6万多观众。 \t 0

2).Explanation

A question (the 1st column), question’s corresponding document sentences (the 2nd column), and their answer annotations (the 3rd column) are provided.

If a document sentence is the correct answer of the question, its annotation will be 1, otherwise its annotation will be 0.

The three columns will be separated by the symbol ‘\t’.

All the dataset file are encoded in UTF-8.

3.Data Statistics

dataset # of unique questions

train set 7895

development set 878

test set 5997

4.Evaluation Metrics

MRR, MAP, and ACC@1.

## Part two 解决方案

### 2.1 开发环境

编程语言： python 3.6.1

IDE: Anaconda Jupyter notebook

用的第三方库: thulac，用于分词，获得词性等

### 2.2 解决方案步骤

#### 2.2.1 分析数据特征

根据Part one的2.1）的数据格式描述，可知，数据文件的每一行是问题Q，答案A，和答案的标签L

Q，A，都是utf-8的字符串，L是数字0或1。0表示答案错误，1表示答案正确



Figure 2.2.1

#### 2.2.2 分析问题和正确答案的特征

提取每个问题和对应的正确答案：



Figure 2.2..2.a

基本思路：

Step 1问题分类，从上图可以看出，问题存在类别。

类别：

WHO

WHERE

WHEN

WHY

AMOUNT

…

Step 2 问题和正确答案之间，问题类别，问题关键词，问题关键词词性，正确答案的关键词有很大的关联

提取问题关键词和关键词词性



Figure 2.2.2.b

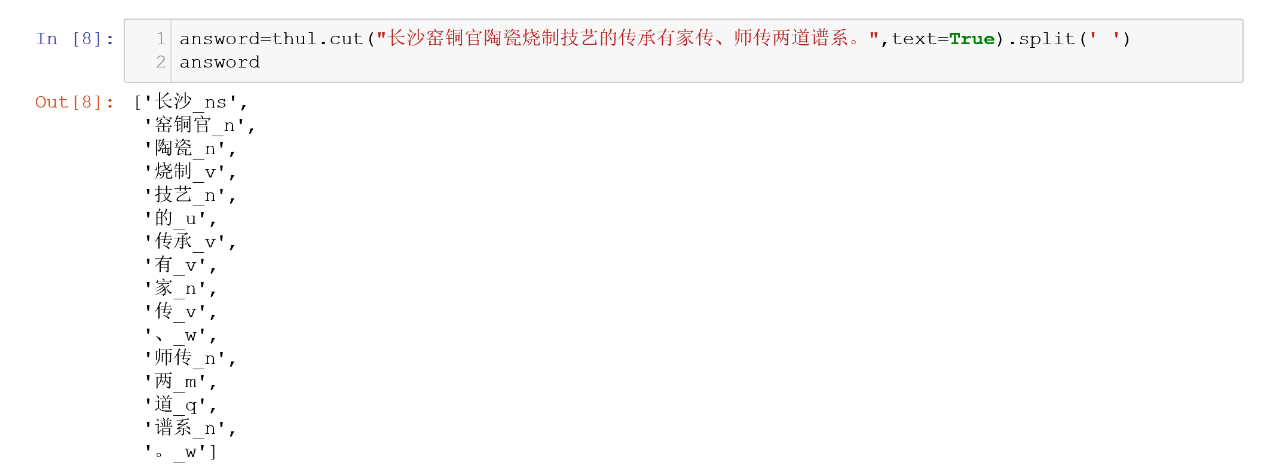


Figure 2.2.2.c

Step 3 代码实现部分

Step 3.1提取所有问题，问题关键词及其词性，问题的答案，答案的标签



Figure 2.2.2.d

Step 3.2 分析问题类型

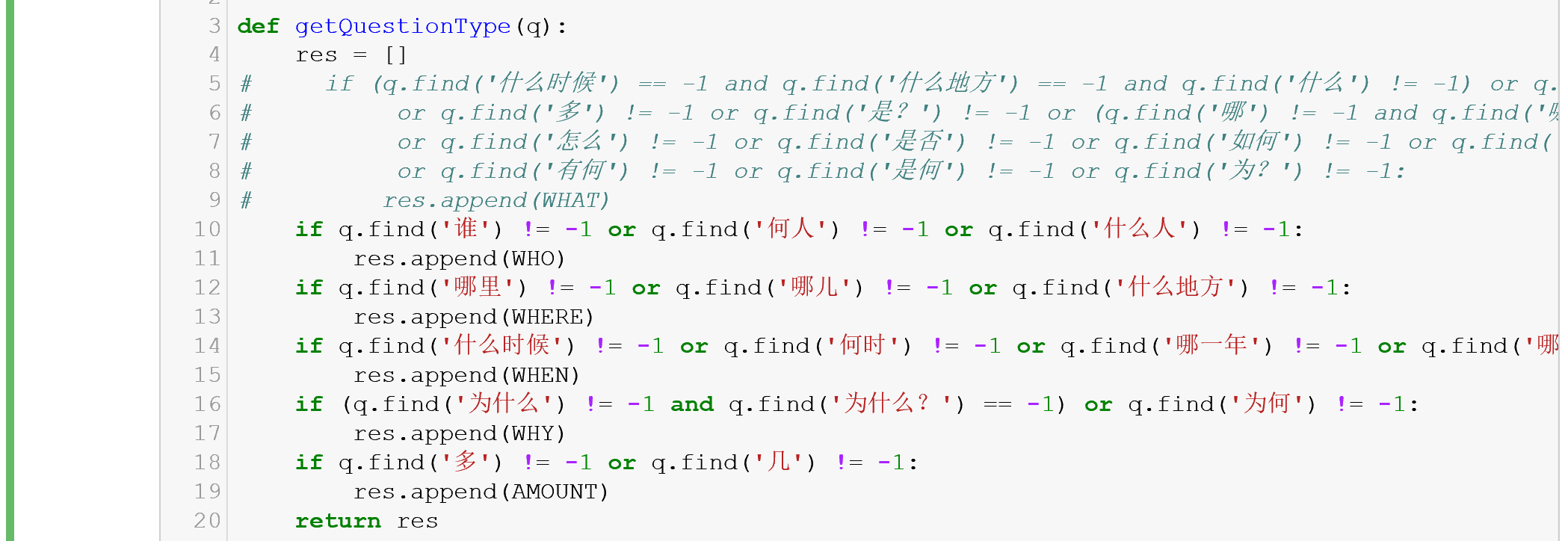


Figure 2.2.2.e

Step 3.2提取答案关键字及其词性

屏幕剪辑

Figure 2.2.2.f

Step 3.3 依照问题类别，关键词，关键词词性寻找最佳答案



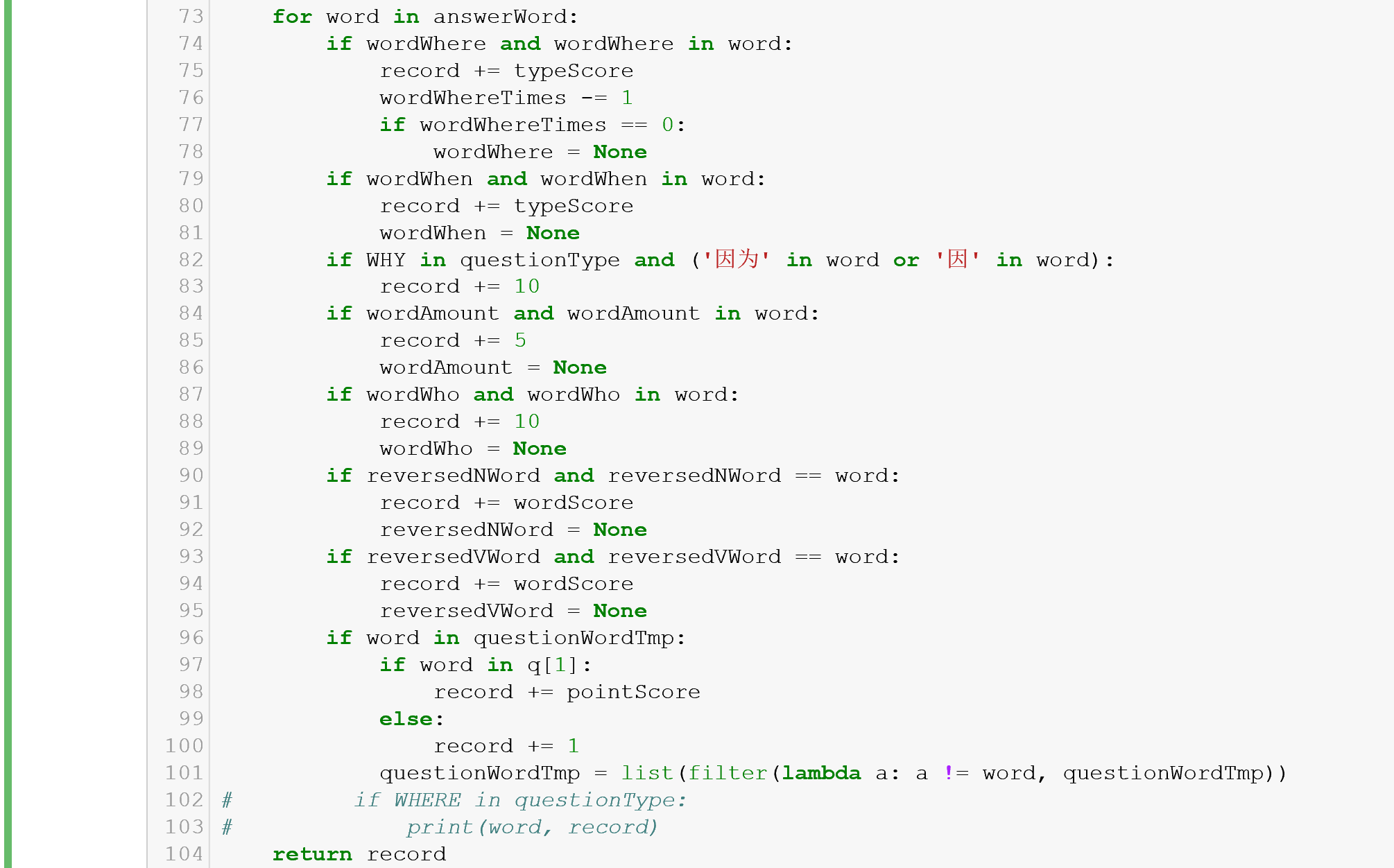


Figure 2.2.2.g

Step 3.4 比较正确答案和最佳答案的得分，判断最佳答案是否正确

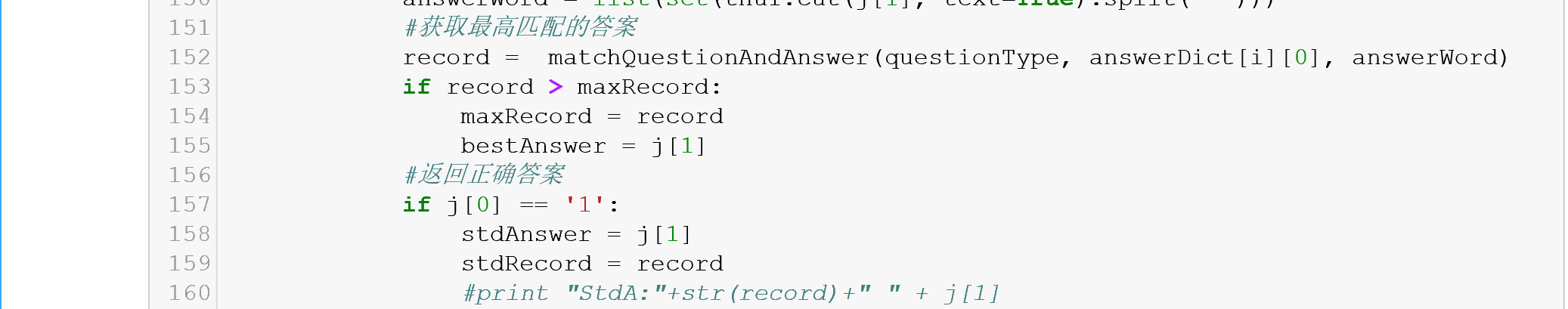


Figure 2.2.2.h

## Part three

运行结果，正确率分析

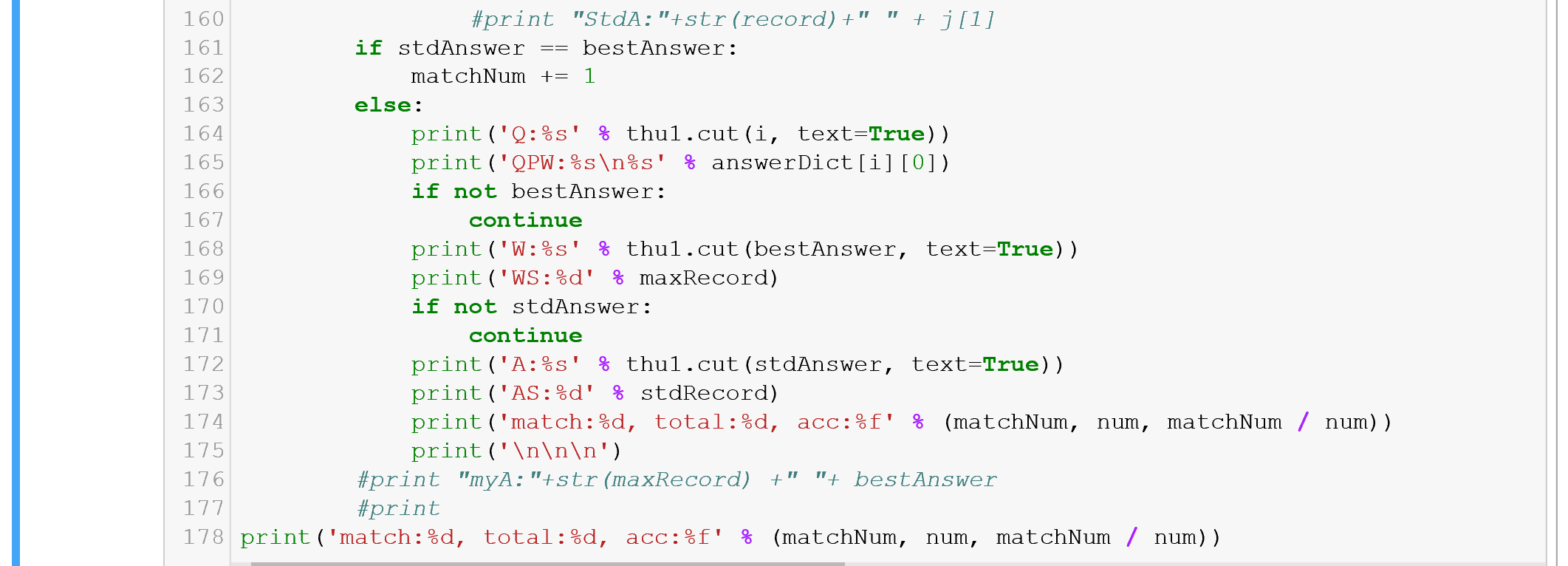


Figure 3.1

在develop.data数据集的运行结果

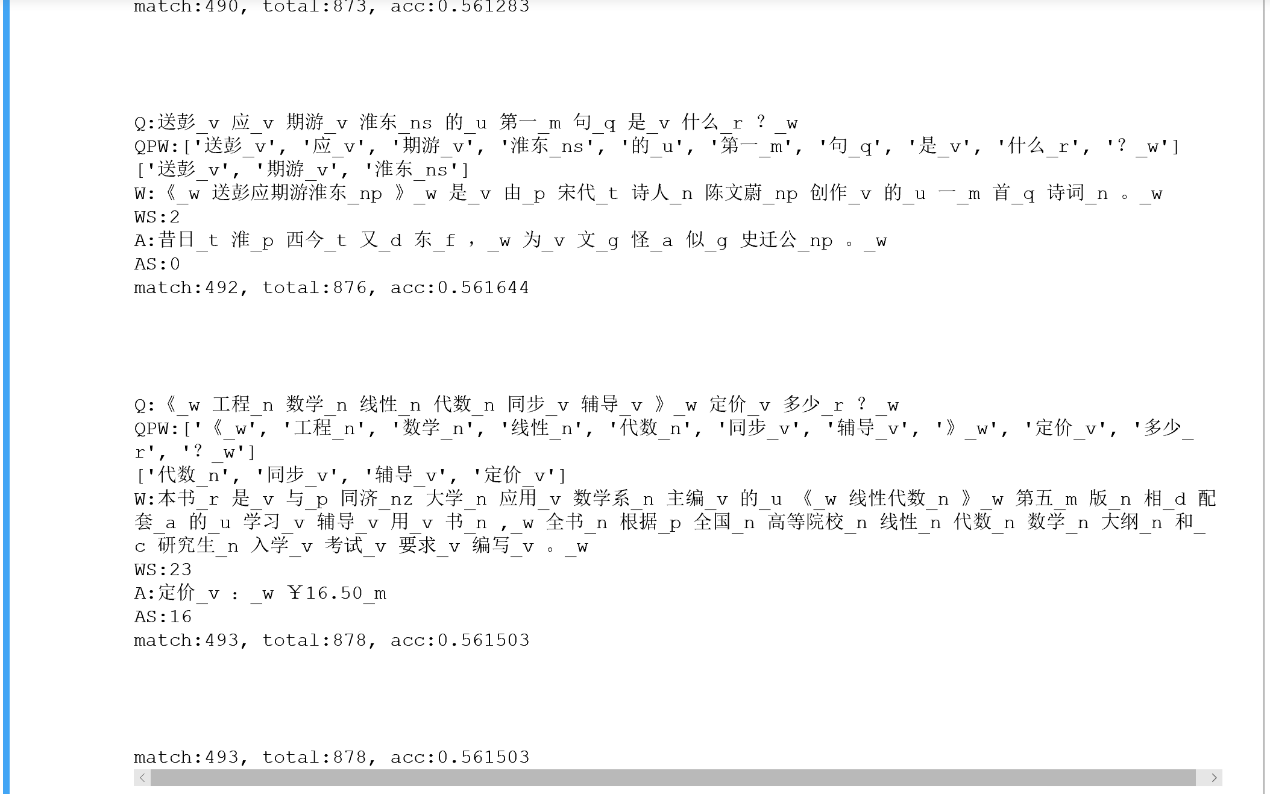


Figure develop\_result

正确率只有56%左右