In [12]:

```
import numpy as np
import pandas as pd
import re
import os
import pymysql
```

In [13]:

```
def getmsg(fname, lab, reg):
    labelset = ['True', 'False', 'Knowledge']
    tmp = list(pd.read_table(fname, header = None)[0])
    rule_match = []
    for i in tmp:
        tmp2 = reg.findall(i)
        tmp3 = list(tmp2[0])
        tmp3.append(labelset[lab])
        rule_match.append(tmp3)
    return rule_match
```

In [14]:

```
pwd = os.getcwd()
ls = os.listdir()
```

In [15]:

```
reg = r'^{"in_node": "(\d.\d)", "out_node": "(\d.\d)", "msg": "(.*?)", "pattern": ".*?"}$'
reg = re.compile(reg)
```

In [16]:

```
rule_match = []
for i in ls:
    if '.txt' in i:
        if 'true' in i:
            lab = 0
        elif 'false' in i:
            lab = 1
        elif 'kld' in i:
            lab = 2
        rule_match += getmsg(i, lab, reg)
```

In [17]:

```
db = pymysql.connect("192.168.162.192", "robot_collection", "robot_collection_20180523", db="robot",
```

In [18]:

```
for i in rule_match:
    cursor = db.cursor()
    sql="select rgl from regular where process=0 and in_node=" + i[0] + " and out_node=" + i[1]
    cursor.execute(sql)
    result = cursor.fetchall()[0][0]
    reg = re.compile('(' + result + ')')
    a = reg.findall(i[2])
    i.append(a)
```

In [19]:

```
rule_match_df = pd.DataFrame(rule_match)
rule_match_df.columns = ['in_node', 'out_node', 'message', 'label', 'keyword_matched']
```

匹配正误情况统计

In [20]:

```
rule_match_df['label'].value_counts()
```

Out[20]:

```
True         233
False         24
Knowledge     14
Name: label, dtype: int64
```

In [21]:

```
rule_match_mistake = rule_match_df[rule_match_df['label'] != 'True'].sort_values(by = ['in_node', 'c
```

错误匹配情况汇总

In [22]:

```
rule_match_mistake
```

Out[22]:

	in_node	out_node	message	label	keyword_matched
27	2.1	3.2	不清楚啊	False	[(清楚, , 清楚)]
28	2.1	3.2	六十致电买来如果卡还没传于两个月原来	False	[(还没, ,), (两个, ,)]
29	2.1	3.2	在这两天	False	[(两天, ,)]
30	2.1	3.2	我知道这个已经付掉了	False	[(知道, ,)]
31	2.1	3.2	啊我知道我已经存上去了可是	False	[(知道, ,)]
33	2.1	3.2	知道多少多少	Knowledge	[(知道, ,)]
32	2.1	3.4	哦我知道知道我现在没拿的钱拿来什么的我的房子去了	False	[什么]
34	2.1	3.4	嗯什么江	Knowledge	[什么]
35	2.1	3.4	嗯什么贷款	Knowledge	[什么]
36	2.1	3.4	是什么公司啊	Knowledge	[什么]
37	2.1	3.4	说什么	Knowledge	[什么]
38	2.1	3.4	说的什么意思我没听懂	Knowledge	[什么]
39	2.1	3.4	怎么什么金那会吉呢	Knowledge	[什么]
151	2.2	3.5	对不愿意	False	[对]
153	2.2	3.5	你说你是谁	Knowledge	[你是]
154	2.2	3.5	你说是哪个	Knowledge	[说是]
155	2.2	3.5	哦我现在是您是	Knowledge	[现在, 是, 您是]
152	2.2	3.6	不是啊买公司啦等一下就行啦	False	[不是]
166	3.2	4.1	呃能跟我们说的那个能够都不知道那个悟明老师他那你得很多	False	[呃能, 个能, 都]
167	3.2	4.1	哎呀明天	False	[明天]
168	3.2	4.1	哦你明天吗	False	[明天]
169	3.2	4.1	明天的话不不	False	[明天]
170	3.2	4.1	超出了三十本三百一十五一分钱都不回还	False	[都]
210	3.4	4.4	我知道有有这款但是它怎么不扣啊	Knowledge	[(我知道, ,)]
207	3.4	4.5	不是我说是不是会英文起来	False	[不是, 不是]
208	3.4	4.5	那个我那就是说就是就是那个手机啦我不是已经还清了吗十二期已经黄金啊	False	[不是]
209	3.4	4.5	没有你说	False	[没有]
211	3.4	4.5	我说是不是慧君	Knowledge	[不是]
218	3.5	4.7	啊没问题没问题	False	[没, 没]
220	4.1	5.1	不能	False	[能]

	in_node	out_node	message	label	keyword_matched
221	4.1	5.1	二十号时间可不可以	False	[间可, 可以]
248	4.4	5.5	还没好	False	[好]
268	5.6	6.5	我的银行卡里面有钱他没有空	Knowledge	[银行]
264	5.6	6.6	哦那行行行我我这几天是我今天没有钱你们那个宽限一个星期时间	False	[没有钱]
265	5.6	6.6	啊可以	False	[]
266	5.6	6.6	我这边啊	False	[]
267	5.6	6.6	不知道我就是快递还是包裹啊	False	[]
270	6.5	7.1	我说可以换多少钱	Knowledge	[可以]

总结

上表中汇总了规则匹配错误的情况，其中包括进入了错误节点的情况（表中label数据项中False的情况）与本不应进入任何一节点的情况（表中label数据项中Knowledge的情况）。一共在9个in_node上有错误匹配的情况，合计38条错误记录，有些问题是有些节点专有的，还有些问题是全局性的，下面详细说明：

1. 只提取了明确肯定词或否定词，而忽略的句中其他内容，可能其他内容中有转折含义或疑问含义，该问题包括27、30、31、151、152、168、169、170、210、207、208、209、211、218、220、221、248、270。尤其是27“不清楚”中只提取出“清楚”，218“没问题”中只提取出“没”，220“不能”只提取出“能”，完全得到了相反的结论。
2. 对方有疑问但没有解答就进入了下级节点，2.1->3.4、2.2->3.5的几乎全部节点都是该原因。
3. 完全不能理解语义的，可能是语音识别问题导致。包括28、166等。

问题1有18条记录，问题2有9条记录（问题2和问题1有交集，只统计不在问题1中的，以避免重复统计），共占到了总错误数的71%，我认为可以优先从这两个问题入手来提高匹配准确率。