源程序：

# 预处理数据集

def get\_word():

    f = open("BaiDu.txt", encoding="UTF-8")

    lines = f.readlines()

    word = list()

    max\_len = 0 # 这里我们不事先定义词的最大长度，直接从词典中获取最大词长

    for line in lines:  # 数据处理, 根据数据集格式，筛选词语

        if line[2:-1]:

            word.append(line[2:-1])

        max\_len = max(len(line[2:-1]), max\_len)  # 删掉前两个字符和最后的换行符

    return word, max\_len

word, max\_len = get\_word()

# 前向算法

def forward(string):

    left = 0 # 定义需要分词段的左指针

    res = [] # 保存分词结果

    err = [] # 保存词典外的字

    while left <= len(string) - 1: # 当left指针指到句子末尾时，停止分词

        right = left + min(max\_len - 1, len(string) - 1) # 考虑到检测的句子长度可能小于最大词长，所以这里的右指针等于最大词长和句长中的较小值

        for i in range(right, left - 1 , -1):

            # 当该词在词典中，则直接分词

            if string[left: i + 1] in word:

                res.append(string[left: i + 1])

                left += len(string[left: i + 1])

                break

            # 若该词不在词典中，且i移到左指针，则说明没有收录该字

            if (string[left: i + 1] not in word) & (left == i):

                err.append(string[left])

                # 手动将改字加入词典中

                word.append(string[left])

                left += 1

                break

    return res, err

# 后向算法

def backward(string):

    right = len(string) - 1 # 定义需要分词段的右指针

    res = [] # 保存分词结果

    err = [] # 保存词典外的字

    while right >= 0: # 当right指针指到句子开头时，停止分词

        left = right - min(max\_len - 1, len(string) - 1)

        for i in range(left, right + 1):

            if string[i: right + 1] in word:

                res.append(string[i: right + 1])

                right -= len(string[i: right + 1])

                break

            if (string[i: right + 1] not in word) & (right == i):

                err.append(string[right + 1])

                word.append(string[right + 1])

                right -= 1

                break

    return res

def run():

    string = input("请输入句子： ")

    while string != 'q':

        res1, err = forward(string)

        res2 = backward(string)

        print("前向最大匹配： ", res1)

        print("后向最大匹配： ", res2[::-1])

        print("其中未出现在词典中的字有： ", err)

        string = input("请输入句子： ")

    print("------------分词完毕------------")

run()

程序运行结果：

