

机器翻译基础作业

—英文动词原型解析与获取

刘天伟 21009306

1.问题概述

对字符串中规则动词的词型变化分析与原型还原，同时适当的能对部分不规则动词进行处理，部分以 ed 结尾的形容词、名词区别处理等等。

2.实现情况

Input: 输入文件(*.txt), 包含可能存在单词的字符串

Output: 包含对输入文件的分析: 包含动词数量, 合法词汇数量, 原型-变化对应, 非动词集合。

完成目标:

1. **预处理过程**: 对*.txt 文件进行处理, 过滤非法字符 (标点, 数字, 下划线, 中文), 以空格为间隔符提取字符串, 放入字符串数组中, 此为词形处理基本来源。同时, 将大写字母都转成小写。程序中打印出改过程。同时程序中利用 Linux shell 缓冲区原理, 实现输入字符串动态数组。

2. **处理词筛选过程**: 将过去式, 过去分词筛选, 并对以这些标志结尾的非动词进行区分。本程序对与规则动词变化实现完成完整, 具体程序参见“规则动词”文件夹, 对于不规则动词, 采用查找词典的方式。

改成中对查找词典, 参见“词库检索”, 利用 linux 平台开源词典软件“星际译王”公开的词库及工具, 制作了目前包含 60 个词的词库, 通过程序实现对词库文件的解析, 能够实现不规则动词的直接查找-映射, 该部分对词库根据字母表进行排序, 利用 2 分法查找, 最坏情况查找 $\log(n)$ 次, 速度可以接受, 目前已经实现了输入词, 进行查找的功能。

3. **单词还原**: 对每一个单词处理, 根据相应规则进行还原。此处需要对规则进行详细划分, 将各种特殊词区别处理。

4. **结果输出**: 将统计分析结果以文件*.txt 的形式输出, 注意格式化文本。

3.目前实现情况及存在问题

1. 对任意文件分析, 提取合法单词
2. 对规则动词 (规则参见第 6 部分) 变化
3. 实现字典查找, 对不规则动词进行形式变化

存在问题

1. 由于时间仓促, 对“字典查找+文件分析”结合的处理不好, 程序存在 BUG, 接下来时间进行修改, 实现一个完备的分析。
2. 词库规模较小

4.开发平台

系统环境: ubuntu 10.04

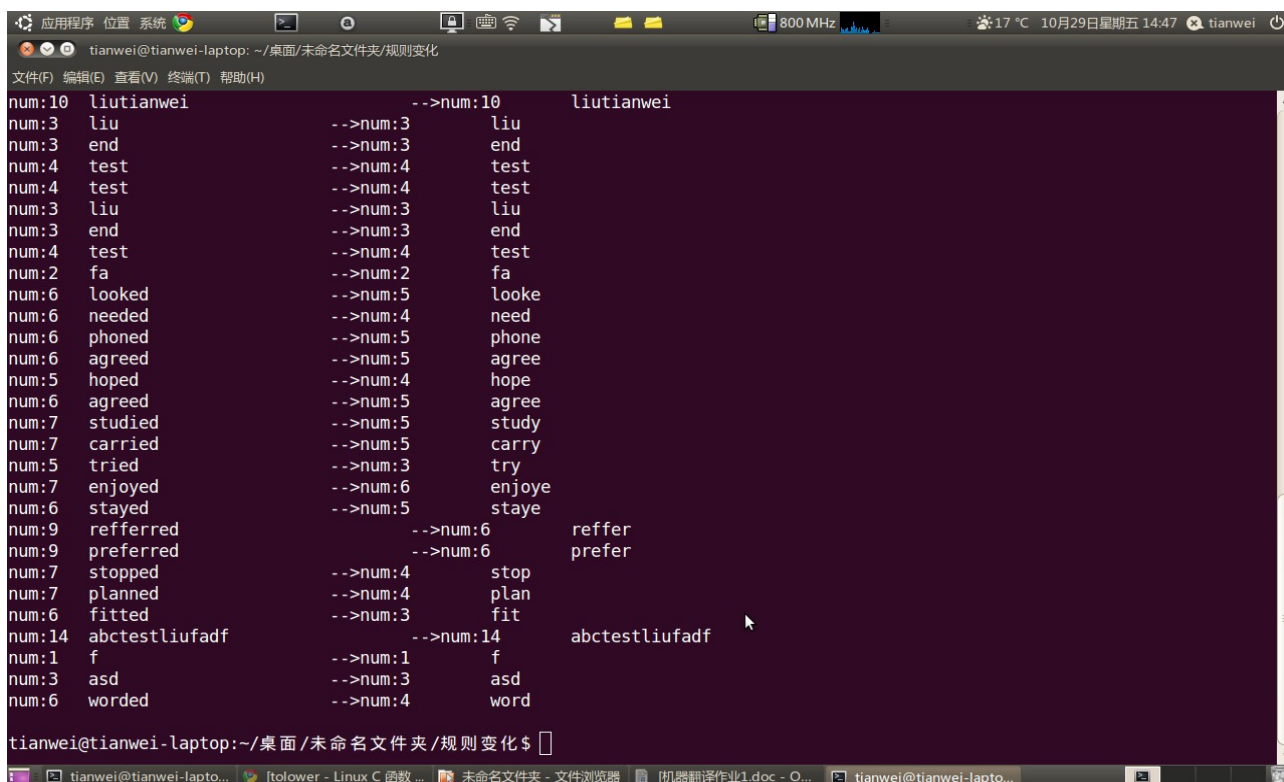
编译器: GCC 4.4.3

调试器: GDB 7.1

编程语言: Linux C

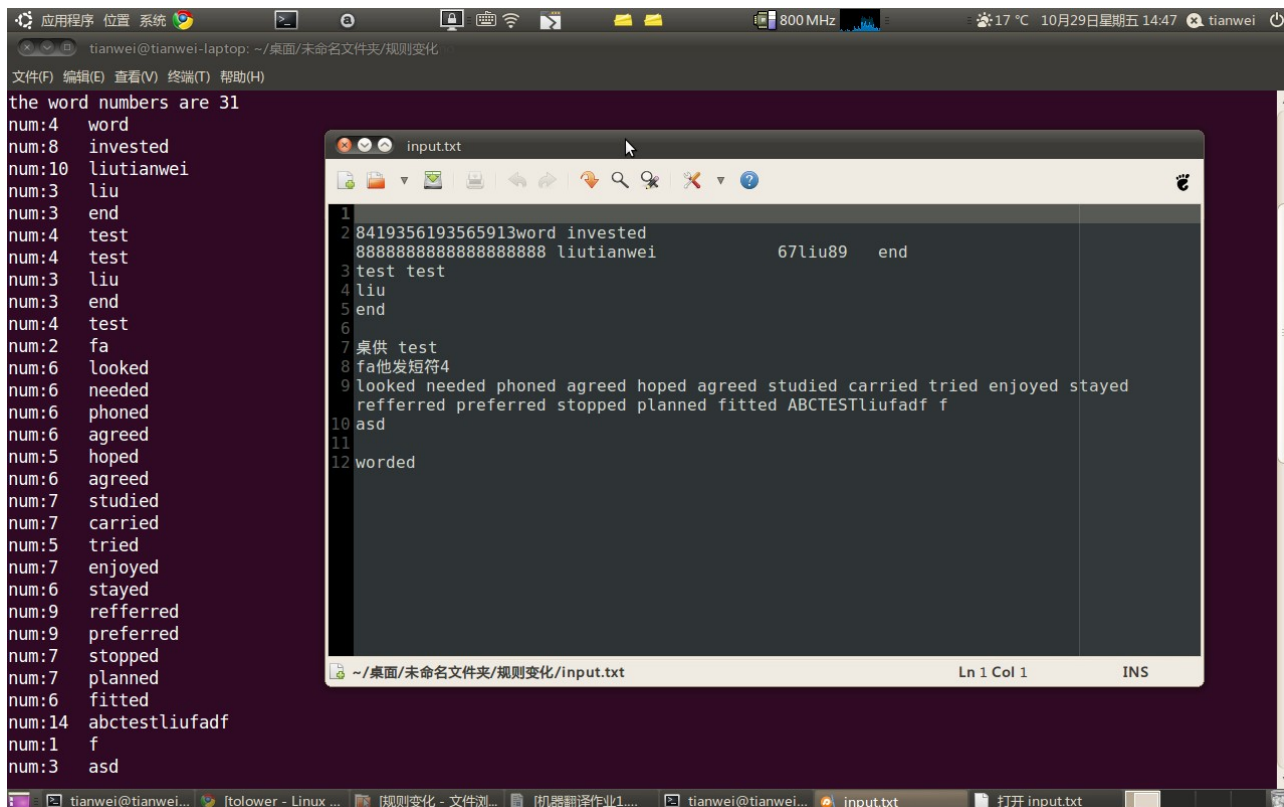
程序运行方法: 通过终端进入文件加, 输入 ./程序执行名

5.运行结果与总结



```
num:10 liutianwei -->num:10 liutianwei
num:3 liu -->num:3 liu
num:3 end -->num:3 end
num:4 test -->num:4 test
num:4 test -->num:4 test
num:3 liu -->num:3 liu
num:3 end -->num:3 end
num:4 test -->num:4 test
num:2 fa -->num:2 fa
num:6 looked -->num:5 looke
num:6 needed -->num:4 need
num:6 phoned -->num:5 phone
num:6 agreed -->num:5 agree
num:5 hoped -->num:4 hope
num:6 agreed -->num:5 agree
num:7 studied -->num:5 study
num:7 carried -->num:5 carry
num:5 tried -->num:3 try
num:7 enjoyed -->num:6 enjoye
num:6 stayed -->num:5 staye
num:9 referred -->num:6 reffer
num:9 preferred -->num:6 prefer
num:7 stopped -->num:4 stop
num:7 planned -->num:4 plan
num:6 fitted -->num:3 fit
num:14 abctestliufadf -->num:14 abctestliufadf
num:1 f -->num:1 f
num:3 asd -->num:3 asd
num:6 worded -->num:4 word
```

图 1：规则动词分析



```
the word numbers are 31
num:4 word
num:8 invested
num:10 liutianwei
num:3 liu
num:3 end
num:4 test
num:4 test
num:3 liu
num:3 end
num:4 test
num:2 fa
num:6 looked
num:6 needed
num:6 phoned
num:6 agreed
num:5 hoped
num:6 agreed
num:7 studied
num:7 carried
num:5 tried
num:7 enjoyed
num:6 stayed
num:9 referred
num:9 preferred
num:7 stopped
num:7 planned
num:6 fitted
num:14 abctestliufadf
num:1 f
num:3 asd
```

图 2 单词提取（小图为文本，背景图为处理后的）

```
tianwei@tianwei-laptop: ~/桌面/未命名文件夹/词库检索
文件(F) 编辑(E) 查看(V) 终端(T) 帮助(H)

tianwei@tianwei-laptop:~/桌面/未命名文件夹/词库检索$ ./dict
input:looked
looked,288,4
look
tianwei@tianwei-laptop:~/桌面/未命名文件夹/词库检索$ ./dict
input:a
a,0,9
test
test
tianwei@tianwei-laptop:~/桌面/未命名文件夹/词库检索$ ./dict
input:haven
对不起, 此词未收录
tianwei@tianwei-laptop:~/桌面/未命名文件夹/词库检索$ ./dict
input:given
given,217,4
give
tianwei@tianwei-laptop:~/桌面/未命名文件夹/词库检索$ ./dict
input:came
came,131,8
come 来
tianwei@tianwei-laptop:~/桌面/未命名文件夹/词库检索$ ./dict
input:blown
blown,91,8
blow 吹
tianwei@tianwei-laptop:~/桌面/未命名文件夹/词库检索$ fly
未找到 'fly' 命令, 您要输入的是否是:
命令 'fld' 来自于包 'kon2' (universe)
命令 'fls' 来自于包 'sleuthkit' (universe)
fly: 找不到命令
tianwei@tianwei-laptop:~/桌面/未命名文件夹/词库检索$
```

图 3 不规则检索

```
tianwei@tianwei-laptop: ~/source_code/mt_2010/demo
文件(F) 编辑(E) 查看(V) 终端(T) 标签页(b) 帮助(H)

tianwei@tianwei-laptop: ~/source_code/mt_2010/demo
input.c (~/.source_code/mt_2010/demo) - VIM

num:9 preferred
num:7 stopped
num:7 planned
num:6 fitted
num:14 abctestliufadf
num:1 f
num:3 asd
num:7 quwhere
num:5 qwliu

i!!!!
*** glibc detected *** ./dict: malloc(): memory corruption: 0x09ae3040 ***
===== Backtrace: =====
/lib/tls/i686/cmov/libc.so.6(+0x6b591)[0xa63591]
/lib/tls/i686/cmov/libc.so.6(+0x6e395)[0xa66395]
/lib/tls/i686/cmov/libc.so.6(+0x5c37f)[0xa5437f]
/lib/tls/i686/cmov/libc.so.6(+0x5c37f)[0xa5437f]
/lib/tls/i686/cmov/libc.so.6(+0x5c37f)[0xa5437f]
./dict[0x80489a1]
./dict[0x804932f]
/lib/tls/i686/cmov/libc.so.6(__libc_start_main+0xe6)[0xa0ebd6]
./dict[0x8048911]
===== Memory map: =====
009a9000-009c4000 r-xp 00000000 08:03 131053 /lib/ld-2.11.1.so
009c4000-009c5000 r--p 0001a000 08:03 131053 /lib/ld-2.11.1.so
009c5000-009c6000 rw-p 0001b000 08:03 131053 /lib/ld-2.11.1.so
009f8000-00b4b000 r-xp 00000000 08:03 144780 /lib/tls/i686/cmov/libc-2.11.1.so
00b4b000-00b4c000 ---p 00153000 08:03 144780 /lib/tls/i686/cmov/libc-2.11.1.so
00b4c000-00b4e000 r--p 00153000 08:03 144780 /lib/tls/i686/cmov/libc-2.11.1.so
00b4e000-00b4f000 rw-p 00153000 08:03 144780 /lib/tls/i686/cmov/libc-2.11.1.so
00b4f000-00b52000 rw-p 00000000 00:00 0
00d69000-00d6a000 r-xp 00000000 00:00 0 [vdso]
00f8c000-00fa9000 r-xp 00000000 08:03 130380 /lib/libgcc_s.so.1
00fa9000-00faa000 r--p 0001c000 08:03 130380 /lib/libgcc_s.so.1
00faa000-00fab000 rw-p 0001d000 08:03 130380 /lib/libgcc_s.so.1
08048000-0804a000 r-xp 00000000 08:07 7902 /home/tianwei/source_code/mt_2010/demo/dict
0804a000-0804b000 r--p 00001000 08:07 7902 /home/tianwei/source_code/mt_2010/demo/dict
0804b000-0804c000 rw-p 00002000 08:07 7902 /home/tianwei/source_code/mt_2010/demo/dict
09ae2000-09ae4000 rw-p 00000000 00:00 0 [heap]
b7600000-b7621000 rw-p 00000000 00:00 0
b7621000-b7700000 ---p 00000000 00:00 0
b77d3000-b77d4000 rw-p 00000000 00:00 0
b77f0000-b77f4000 rw-p 00000000 00:00 0
bfa0f000-bfa24000 rw-p 00000000 00:00 0 [stack]
已放弃
tianwei@tianwei-laptop:~/source_code/mt_2010/demo$
```

图 4 整合后程序存在存在的问题

6.相关知识

0.元音字母: A、E、I、O、U, 最短动词长度: 3 act

1.规则动词：过去式，过去分词

1.1 一般在动词原型+ed: look-looked-looked

1.2 以 e 结尾的+d: phoned,agreed,hoped

1.3 以辅音字母+y 结尾的词，变 y 为 i+ed : carry-carried,try-tried,study-studied

1.4 以原音字母+y 结尾的词，直接加+ed: enjoy-enjoyed,stay-stayed

1.5 以-r 结尾的词，双写 r 字母+ed: refer-referred,preferred-preferred

1.6 末尾只有一个辅音字母结尾的重度闭音节，双写该辅音字母：
planned,fitted,stopped