DICT：在 Stata 中进行中英文互译

程振兴

版本：*3.0.0*

更新：*2019* 年 *5* 月 *28* 日

摘　　要

Stata 是由 StataCorp LLC 开发的统计编程软件，有着非常强大的数据管理、统计、绘图、编程和报告排版的功能。Stata 公司还为用户编写了长达 14000 页的用户文档。不过都是英语的。所以中文用户在使用的过程中需要一定的英语阅读能力。尽管如此，中文用户在阅读的过程中还是会经常遇到一些生疏的单词。使用本文介绍的 dict 命令可以方便中文用户直接在 Stata 中查阅单词。

1　导论

在我刚刚开始学 Stata 的时候我就深深为 Stata 的英语帮助文档所困扰，那个时候我就在想，如果有一个命令，可以直接把 Stata 的英文帮助文档翻译成中文的就好了，就像我们可以直接把英文网页翻译成中文网页一样。虽然这个命令到现在还是没有人写出来，自己也没有写这个命令的好思路。但是自己找到了如何写一个可以在 Stata 中进行中英文互译的命令的方法。

有一次我使用必应词典查单词的时候，发现必应词典查单词的网页请求非常容易构造，例如查询单词 hello 的中文释义只需要访问 https://cn.bing.com/dict/search?q=hello 即可。打开网页你就会看到下面图1中的解释：

图 1: 必应词典中 hello 的释义

所以只须在 Stata 中发起对 https://cn.bing.com/dict/search?q=hello 的请求，然后将返回的源码整理成需要的样式即可。

将中文翻译成英文也可以使用必应词典完成，不过在构造 URL 前需要先把中文编码成 URL 码，这个可以使用 percentencode 命令实现，例如我想查询 **你好**的英文释义，我需要访问 https://cn.bing.com/dict/search?q= 你好，如果你把这个链接复制到浏览器的地址栏里，浏览器会自动把其中的中文转码成 URL 码。但是 Stata 的 copy 命令不能自动完成这一过程。首先我们需要使用 percentencode 命令对该链接进行 URL 转码，percentencode 命令来源于 Matsuoka，该博客的作者编写了一个 mata 函数进行 URL 转码（链接：http://www.wmatsuoka.com/uploads/2/1/4/6/21469478/twitter-programs-lib.do）：

```
*****************************************************
* Percent Encode
*****************************************************
mata:
  string scalar percentencode(string scalar s)
  {
    lc = "abcdefghijklmnopqrstuvwxyz"
    uc = "ABCDEFGHIJKLMNOPQRSTUVWXYZ"
    no = "1234567890"
    re = "-._~"
    str = lc + uc + no + re
    asc = ascii(s)'
    sel = rowmax(asc :== J(rows(asc), 1, ascii(str)))
    chr = sel :* strofreal(asc)
    enc = !sel :* ("%" :+ inbase(16, asc))
    final = strupper(chr :+ enc)
    for(i=1; i<=rows(final); i++) {
      if (substr(final[i], 1, 1)!="%") {
      final[i] = char(strtoreal(final[i]))
      }
    }
    return (invtokens(final', ""))
    }
end
```

只需要简单修改就能打包成 ado 命令使用了：

```
*! URL 转码
cap program drop percentencode
program define percentencode, rclass
mata: st_local("percentencode", percentencode(`""`1'"'))
return local percentencode "`percentencode'"
di "`percentencode'"
end
```

```
mata:
string scalar percentencode(string scalar s){
  lc = "abcdefghijklmnopqrstuvwxyz"
  uc = "ABCDEFGHIJKLMNOPQRSTUVWXYZ"
  no = "1234567890"
  re = "-._~"
  str = lc + uc + no + re
  asc = ascii(s)'
  sel = rowmax(asc :== J(rows(asc), 1, ascii(str)))
  chr = sel:* strofreal(asc)
  enc = !sel :* ("%" :+ inbase(16, asc))
  final = strupper(chr :+ enc)
  for(i=1; i<=rows(final); i++) {
    if (substr(final[i], 1, 1)!="%") {
      final[i] = char(strtoreal(final[i]))
    }
  }
  return (invtokens(final', ""))
}
end
```

使用示例:

```
. percentencode 你好
%E4%BD%A0%E5%A5%BD

. ret list
macros:
      r(percentencode) : "%E4%BD%A0%E5%A5%BD"
```

说明 你好 的 *URL* 代码为 %E4%BD%A0%E5%A5%BD。转码之后就可以使用 *copy* 命令获取 https://cn.bing.com/dict/search?q=%E4%BD%A0%E5%A5%BD 的源代码再进行整理得到 你好 的英文示例了。

最开始的 *dict* 命令只能用于中英文词语的互译，判断使用者输入的是中文词语还是英文单词的方法是使用正则表达式进行匹配，如果匹配到了中文就运行中文翻译为英文的程序，反之则运行英文翻译为中文的程序。

在版本 *2.0.0* 中我添加了中英文句子互译的功能，这个是通过使用有道词典实现的，实现的原理与使用必应词典类似。

2 安装

Stata 提供了一种安装外部命令的基础命令：net install，你可以在 *Stata* 的命令输出窗口输入下面的命令安装 *dict* 命令：

```
net install dict, from("https://www.czxa.top/dict")
ssc install moss
```

但是由于 *dict* 命令依赖于一些其它的外部命令，例如 *moss*，这是一个进行正则表达式匹配的命令。而上面的 net install 命令运行的时候不会安装这个外部命令，因此推荐使用 *E. F. Haghish* 开发的 *github* 命令安装：

首先你需要安装 *github* 命令：

```
net install github, from("https://haghish.github.io/github/")
```

然后就可以安装这个命令了：

```
github install czxa/dict, replace
```

3 用法

dict contents, [*nosplit sentence*]

contents: 是一列需要查询的英语单词、中文词语或中英文句子。查询句子时需要使用 *sentence* 选项。

1. *nosplit*：可以简写为 *no*。为了便于区分多个查询结果，系统会自动在每个查询结果后面画一条黄线，加上选择项 *nosplit* 可以取消这条线。
2. *sentence*：可以简写 *s*。为表明需要翻译的内容为句子，注意每次只能翻译一个句子，句子需要使用双引号括起来。

4 用法示例

4.1 查询单个英语单词

```
  dict apple
  *> 【单词】:apple
  *> 【读音】:美[æp(ə)l]英['æpl]
  *> 【释义】:
  *> n.:苹果公司; 【植】苹果; 【植】苹果树;
  *> net.:苹果电脑; 美国苹果; 美国苹果公司;
```

4.2 查询单个中文单词

```
  dict 再见
  *> 【词语】:再见
  *> 【拼音】:zài jiàn
```

```
*> 【英语】:
*> na.: 〈客套〉good-bye; see you again;
*> net.: Goodbye; See you; Bye;
```

4.3 翻译英文句子

```
*> dict "It is necessary to learn information and data collection quickly.", s
*> 【原文】: It is necessary to learn information and data collection quickly.
*> 【译文】: 快速学习信息和数据收集是必要的。
```

4.4 翻译中文句子

```
dict "学会信息和数据快速采集都是非常必要的", s
*>【原文】: 学会信息和数据快速采集都是非常必要的
*>【译文】: It is necessary to learn how to collect information and data quickly
```

4.5 翻译一组单词或词语

需要翻译一组单词或词语的时候，需要用空格将各个单词或词语分开:

```
dict apple 苹果 hello 你好
*> 【单词】:apple
*> 【读音】:美[æp(ə)l]英['æpl]
*> 【释义】:
*> n.:苹果公司; 【植】苹果; 【植】苹果树;
*> net.:苹果电脑; 美国苹果; 美国苹果公司;
*> ----------------------------------------------------------
*> 【词语】:苹果
*> 【拼音】:píng gu
*> 【英语】:
*> n.: 【食】apple;
*> net.: Apple; iphone; Apple I
*> nc.:;
*> ----------------------------------------------------------
*> 【单词】:hello
*> 【读音】:美[he lə ]英[hə'lə ]
*> 【释义】:
*> int.:你好; 喂; 您好; 哈喽;
*> net.:哈罗; 哈啰; 大家好;
*> ----------------------------------------------------------
*> 【词语】:你好
*> 【拼音】:n  h o
*> 【英语】:
```

```
*> na.: hello; 〈正式,口〉how do you do?;
*> net.: Hello; Hi; How do you do;
*> ----------------------------------------------------------
```

需要注意，不能同时翻译多组句子。如果你想翻译多组句子，可以在循环中使用该命令。

5 基准

为了记录 *dict* 命令查询单词的速度，运行 *10* 次下面的代码计算平均耗时作为基准。便于其它命令与之比较：

```
dict 你好
```

表 *1* 展示了 *10* 次运行的时间和平均耗时：

表 1: dict 命令运行的耗时

	用时（秒）
第 1 次	0.97
第 2 次	0.84
第 3 次	0.99
第 4 次	1.19
第 5 次	0.95
第 6 次	0.83
第 7 次	0.91
第 8 次	0.89
第 9 次	0.81
第 10 次	0.95
平均	0.93

参考文献

HAGHISH E F. github: a module for building, searching, installing, managing, and mining stata packages from github[EB/OL]. https://github.com/haghish/github.

MATSUOKA W. Stata and the twitter api (part ii)[EB/OL]. http://www.wmatsuoka.com/stata/stata-and-the-twitter-api-part-ii.

ROBERT PICARD N J C, 2016. moss: Find multiple occurrences of substrings[M]. [S.l.]: Durham University.