

Introduction to SRILM Toolkit



Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University



Available Web Resources

- SRILM: “ <http://www.speech.sri.com/projects/srilm/> ”
 - A toolkit for building and applying various statistical language models (LMs)
 - Current version: 1.5.6(stable)
 - Can be executed in Linux environment
- Cygwin: “<http://www.cygwin.com/>”
 - Cygwin is a Linux-like environment for Windows
 - Current version: 1.5.25-11

Steps for Installing Cygwin

1. Download the cygwin installation file “**setup.exe**” from the website
2. Run *setup.exe*
3. Choose “Install from Internet” (or others)
4. With a default setting, it will be installed in “**c:\cygwin**”
5. “Local Package Directory” means the temporary directory for packages
6. Choose a downloadable (mirror) website

Steps for Installing Cygwin (cont.)

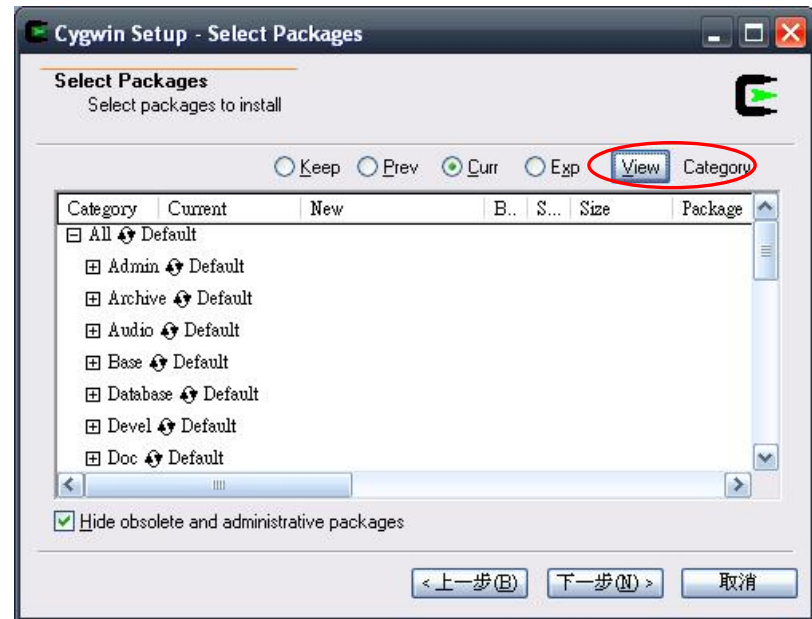
7. Note that:

If you want to compile original source code

Change Category “View” to Full

Check if the packages “**binutils**”, “**gawk**”, “**gcc**”, “**gzip**”, “**make**”, “**tcltk**”, “**tcsh**” are selected

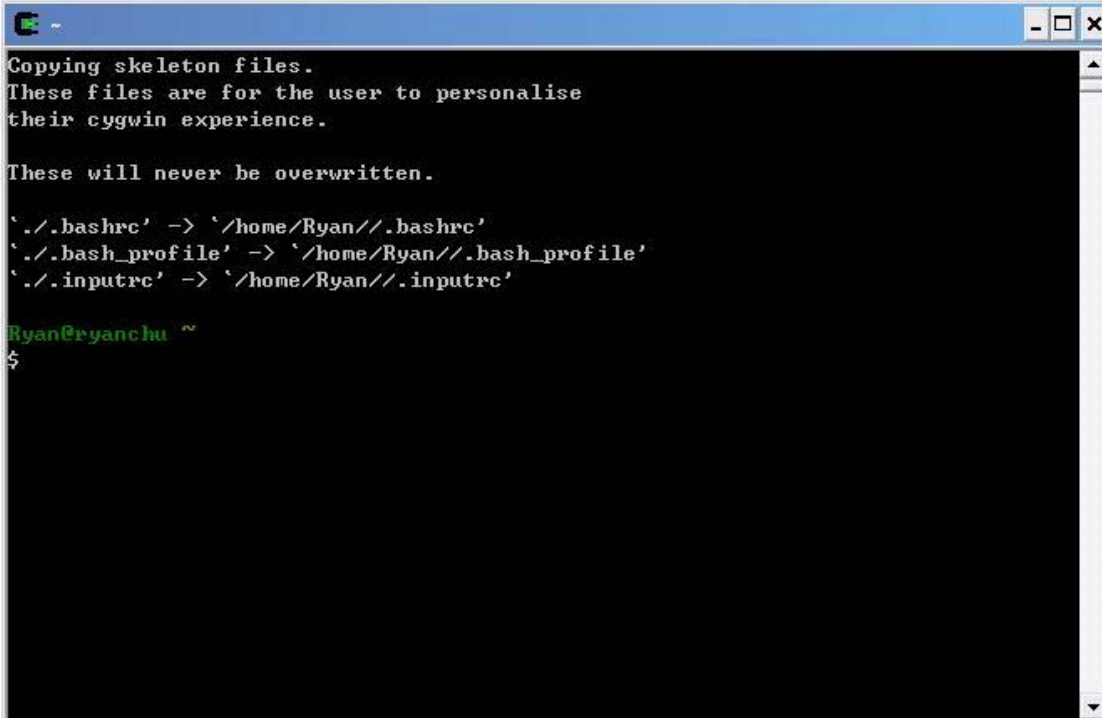
If not, use the default setting



Steps for Installing Cygwin (cont.)

8. After installation, run cygwin

It will generate **“.bash_profile”**, **“.bashrc”**, **“.inputrc”** in **“c:\cygwin\home\yourname\”**



```
Cygwin -  
Copying skeleton files.  
These files are for the user to personalise  
their cygwin experience.  
  
These will never be overwritten.  
  
'./.bashrc' -> '/home/Ryan/./.bashrc'  
'./.bash_profile' -> '/home/Ryan/./.bash_profile'  
'./.inputrc' -> '/home/Ryan/./.inputrc'  
  
Ryan@ryanchu ~  
$
```

Steps for Installing SRILM Toolkit

Now we then install “**SRILM**” into the “**Cygwin**” environment

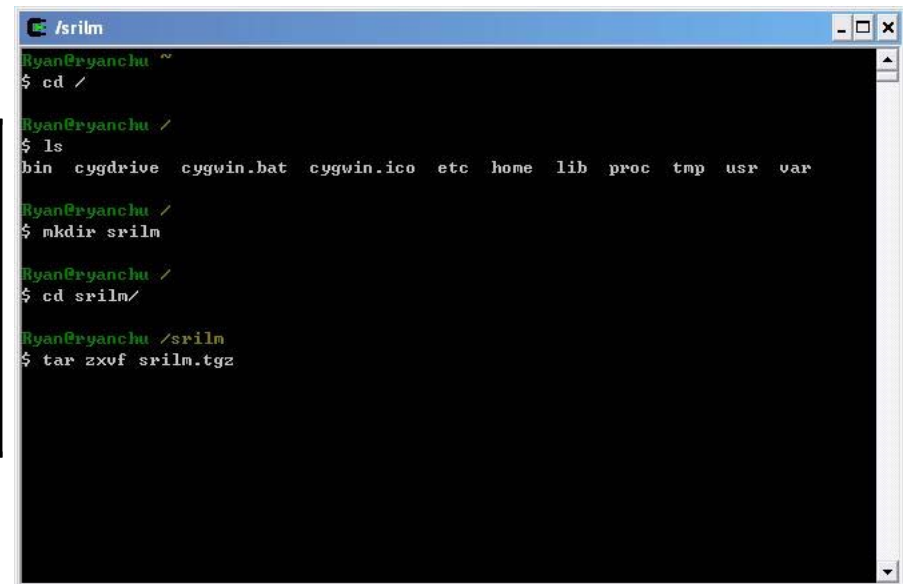
1. Copy “**srilm.tgz**” to “**c:\cygwin\srilm**”

- Create the “**srilm**” directory if it doesn’t exist
- Or, merely copy “**srilm.zip**” to c:\cygwin

2. Extract “**srilm.tgz**” (src files) or “**srilm.zip**” (executable files)

commands in cygwin:

```
$ cd /  
$ mkdir srilm //create the “srilm” directory  
$ cd srilm  
$ tar zxvf srilm.tgz //extract srilm.tgz
```



```
srilm  
Ryan@ryanchu ~  
$ cd /  
Ryan@ryanchu /  
$ ls  
bin  cygdrive  cygwin.bat  cygwin.ico  etc  home  lib  proc  tmp  usr  var  
Ryan@ryanchu /  
$ mkdir srilm  
Ryan@ryanchu /  
$ cd srilm/  
Ryan@ryanchu /srilm  
$ tar zxvf srilm.tgz
```

Steps for Installing SRILM Toolkit (cont.)

3. Edit “c:\cygwin\home\yourname\.bashrc”

- Add the following several lines into this file

```
export SRILM=/srilm
export MACHINE_TYPE=cygwin
export PATH=$PATH:$pwd:$SRILM/bin/cygwin
export MANPATH=$MANPATH:$SRILM/man
```

4. Restart “Cygwin”

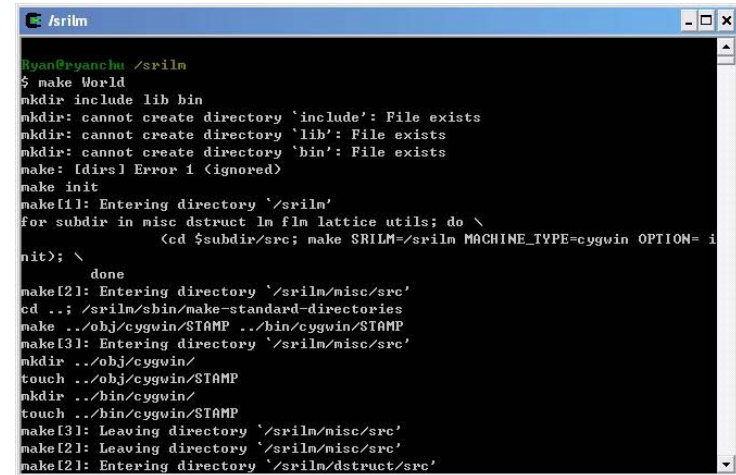
- We can start to use the SRILM if the precompiled files (e.g., those extracted from “**srilm.zip**”) are installed/copied into the desired directory
- Or, we have to compile the associated source code files (e.g., those extracted from “**srilm.tgz**”) manually (See **Steps “5”**)

Steps for Installing SRILM Toolkit (cont.)

5. Compile the SRILM source code files

- Run cygwin
- Switch current directory to “/srilm”
- Modify “/srilm/Makefile”
 - Add a line: “**SRILM = /srilm**” into this file
- Execute the following commands

```
$ make World  
$ make all  
$ make cleanest
```



```
Ryan@ryanclu /srilm  
$ make World  
mkdir include lib bin  
mkdir: cannot create directory 'include': File exists  
mkdir: cannot create directory 'lib': File exists  
mkdir: cannot create directory 'bin': File exists  
make: [dirs] Error 1 (ignored)  
make init  
make[1]: Entering directory '/srilm'  
for subdir in misc dstruct lm flm lattice utils; do \  
  (cd $subdir/src; make SRILM=/srilm MACHINE_TYPE=cygwin OPTION= i  
nit); \  
done  
make[2]: Entering directory '/srilm/misc/src'  
cd ../srilm/shin/make-standard-directories  
make ../obj/cygwin/STAMP ../bin/cygwin/STAMP  
make[3]: Entering directory '/srilm/misc/src'  
mkdir ../obj/cygwin/  
touch ../obj/cygwin/STAMP  
mkdir ../bin/cygwin/  
touch ../bin/cygwin/STAMP  
make[3]: Leaving directory '/srilm/misc/src'  
make[2]: Leaving directory '/srilm/misc/src'  
make[2]: Entering directory '/srilm/dstruct/src'
```

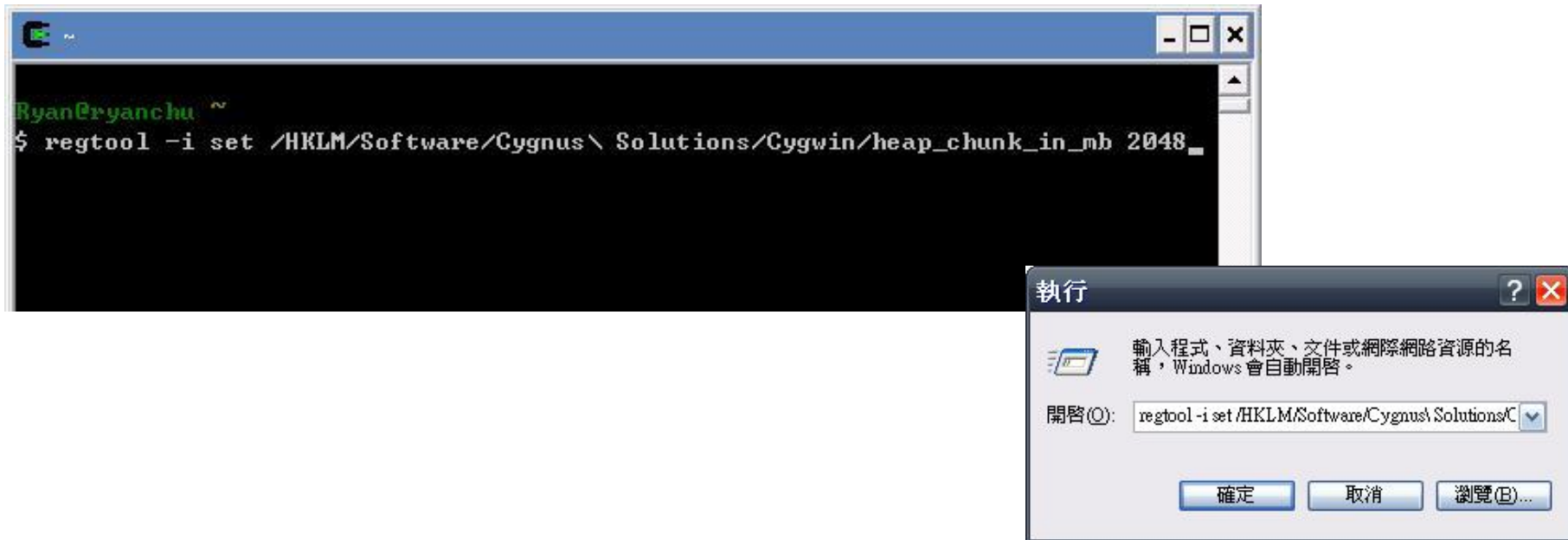
- Check “INSTALL” or “srilm/doc/README.windows” for more detailed information

Environmental Setups - Memory

- Change cygwin's maximum memory(by cygwin or windows cmd mode)

“regtool -i set /HKLM/Software/Cygnus\ Solutions/Cygwin/heap_chunk_in_mb 2048”

– Referred to: “ <http://cygwin.com/cygwin-ug-net/setup-maxmem.html> ”



Environmental Setups – Chinese input

- Use Chinese Input In Cygwin
 - Manually edit the “c:\cygwin\home\yourname**.bashrc**” and “c:\cygwin\home\yourname**.inputrc**” files

.inputrc

```
set meta-flag on
set convert-meta off
set input-meta on
set output-meta on
```

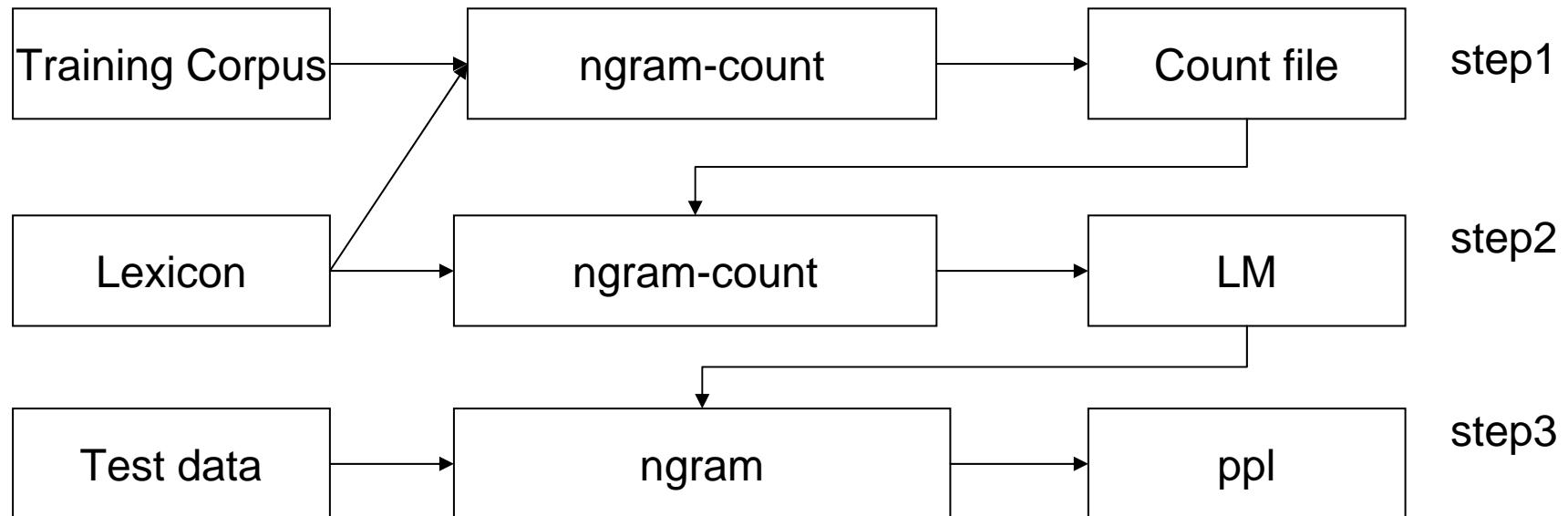
.bashrc

```
export LESSCHARSET=latin1
alias ls="ls --show-control-chars"
```

- Referred to: “ http://cygwin.com/faq/faq_3.html#SEC48 ”

Functionalities of SRILM

- Three Main Functionalities
 - Generate the n-gram count file from the corpus
 - Train the language model from the n-gram count file
 - Calculate the test data perplexity using the trained language model



Format of the Training Corpus

- Corpus: e.g., “CNA0001-2M.Train” (56.7MB)
 - Newswire Texts with Tokenized Chinese Words

中華民國八十九年一月一日
萬
黃兆平
面對這個歷史性的時刻
由中國電視公司
昨晚在中正紀念堂吸引了超過十萬人潮
共同迎接千禧年
勤奮努力
欣欣向榮外
.....

Format of the Lexicon

- Lexicon: “Lexicon2003-72k.txt”

巴
八
扒
叭

墨竹
默祝
末梢
沒收
墨守
陌生
.....

- Vocabulary size: 71695
- Maximum character-length of a word: 10

Generating the N-gram Count File

- Command

```
ngram-count -vocab Lexicon2003-72k.txt  
            -text CNA0001-2M.Train  
            -order 3  
            -write CNA0001-2M.count  
            -unk
```

- Parameter Settings

- vocab: lexicon file name
- text: training corpus name
- order: n-gram count
- write: output countfile name
- unk: mark OOV as <unk>

Format of the N-gram Count File

•E.g., “CNA0001-2M.count”

Counts in training corpus

Unigram

Bigram

Trigram

想像得到	1
想像得到的	1
想像得到的 重大	1
鳳凰	162
鳳凰花	5
鳳凰花 </s>	1
鳳凰花開	4
鳳凰 </s>	23
鳳凰 獎章	2
鳳凰 獎章 </s>	2
鳳凰城	41
鳳凰城 </s>	6
鳳凰城 及	1
鳳凰城 駕駛	1
鳳凰城 以北	1
鳳凰城 舉辦	1
鳳凰城 十八	1
鳳凰城 太陽	28

...	
業界 傷心 </s>	1
業界 統計	1
業界 統計 分析	1
業界 一再	1
業界 一再 提出	1
業界 希望	2
業界 希望 迫切	1
業界 希望 立法院	1
業界 出現	1
業界 出現 一	1
業界 上	1
業界 上 </s>	1
業界 關係	1
業界 關係 良好	1
業界 就	1
業界 就 聚集	1
...	

Generating the N-gram Language model

- Command

```
ngram-count -vocab Lexicon2003-72k.txt  
            -read CNA0001-2M.count  
            -order 3  
            -lm CNA0001-2M_N3_GT3-7.lm  
            -gt1min 3 -gt1max 7  
            -gt2min 3 -gt2max 7  
            -gt3min 3 -gt3max 7
```

- Parameter Settings

- read: read count file

- lm: output LM file name

- gt n min: Good-Turing discounting for n -gram

Format of the N-gram Language Model File

- E.g., “CNA0001-2M_N3_GT3-7.lm”

<pre>\data\ ngram 1=71697 ngram 2=2933381 ngram 3=1205445 \1-grams: -0.8424806 </s> -99 <s> -1.291354 -2.041174 一 -1.287858 -3.804316 一一 -0.8553778 -5.374712 一一恐怖 -1.269383 -4.772653 一一恐怖攻擊 - 0.8950238 -9.690391 一丁點 -3.51804 一九九 -2.89049 -7.180892 一了百了 -0.1229095 -6.481923 一刀兩斷 -0.6672484 -4.802495 一下 -0.4828814</pre>	<p>Log of backoff weight (Base 10)</p>
<pre>-1.38444 <s> 裏 表現 -1.38444 <s> 裏 面 -1.076253 <s> 裏 海 -0.624772 戈 裏 峰 -0.624772 年 裏 </s> -1.198803 那 裏 </s> -0.3165856 哪 裏 去 -0.7112821 家 裏 的 -1.323742 家 裏 開 -0.4998333 時 間 裏 </s> -0.3147101 眼 裏 </s> -0.323742 過 程 裏 </s> -0.721682 <s> 恒 生 -0.323742 億 恒 科 技 -0.1760913 化 粧 品 \end\ Log probability (Base 10)</pre>	

Calculating the Test Data Perplexity

- Command:

```
ngram -ppl 506.pureText  
      -order 3  
      -lm CNA0001-2M_N3_GT3-7.lm
```

- Parameter Settings

- ppl: calculate perplexity for test data

file 506.PureText: 506 sentences, 38307 words, 0 OOVs
0 zeroprobs, logprob= -117172 ppl= 1044.42 ppl1= 1144.86

$$10^{-\frac{\text{logprob}}{\#\text{Sen} + \#\text{Word}}}$$

$$10^{-\frac{\text{logprob}}{\#\text{Word}}}$$

Other Discounting Techniques

- Absolute Discounting

```
nggram-count -vocab Lexicon2003-72k.txt  
-read CNA0001-2M.count  
-order 3  
-lm CNA0001-2M_N3_AD.lm  
-cdiscount1 0.5  
-cdiscount2 0.5  
-cdiscount3 0.5
```

- Witten-Bell Discounting

```
nggram-count -vocab Lexicon2003-72k.txt  
-read CNA0001-2M.count  
-order 3  
-lm CNA0001-2M_N3_WB.lm  
-wbdiscout1  
-wbdiscout2  
-wbdiscout3
```

Other Discounting Techniques (cont.)

- Modified Kneser-Ney Discounting

```
ngram-count -vocab Lexicon2003-72k.txt  
            -read CNA0001-2M.count  
            -order 3  
            -lm CNA0001-2M_N3_KN.lm  
            -kndiscount1  
            -kndiscount2  
            -kndiscount3
```

- Online documentation available at:

“ <http://www.speech.sri.com/projects/srilm/manpages/> ”