SRILM 說明

- # SRILM

  SRILM is a toolkit for building and applying statistical language models (LMs), and it runs under Linux/UNIX system.

  http://www.speech.sri.com/projects/srilm/

  Current Version 1.4.3 3 December 2004

- # Cygwin

  Cygwin is a Linux-like environment for Windows.

  http://www.cygwin.com/

  Current Version 1.5.12-1 November 2004

SRILM安裝說明

- Cygwin Setup

  step1. 下載 setup file from http://www.cygwin.com/

  step2. 執行setup.exe開始安裝。選擇install from internet。接著一直下一步都不用改，直到要選一個mirror site，選ftp://ftp.nctu.edu.tw/會比較快。接著不用改任何設定，直接default安裝就可以了。

  step3. 安裝完，執行桌面上的捷徑，會在c:\cygwin\home\username\產生三個檔案， .bash_profile, .bashrc, .inputrc.

  step4 .編輯 .bashrc，加入底下幾行:

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

- SRILM Setup

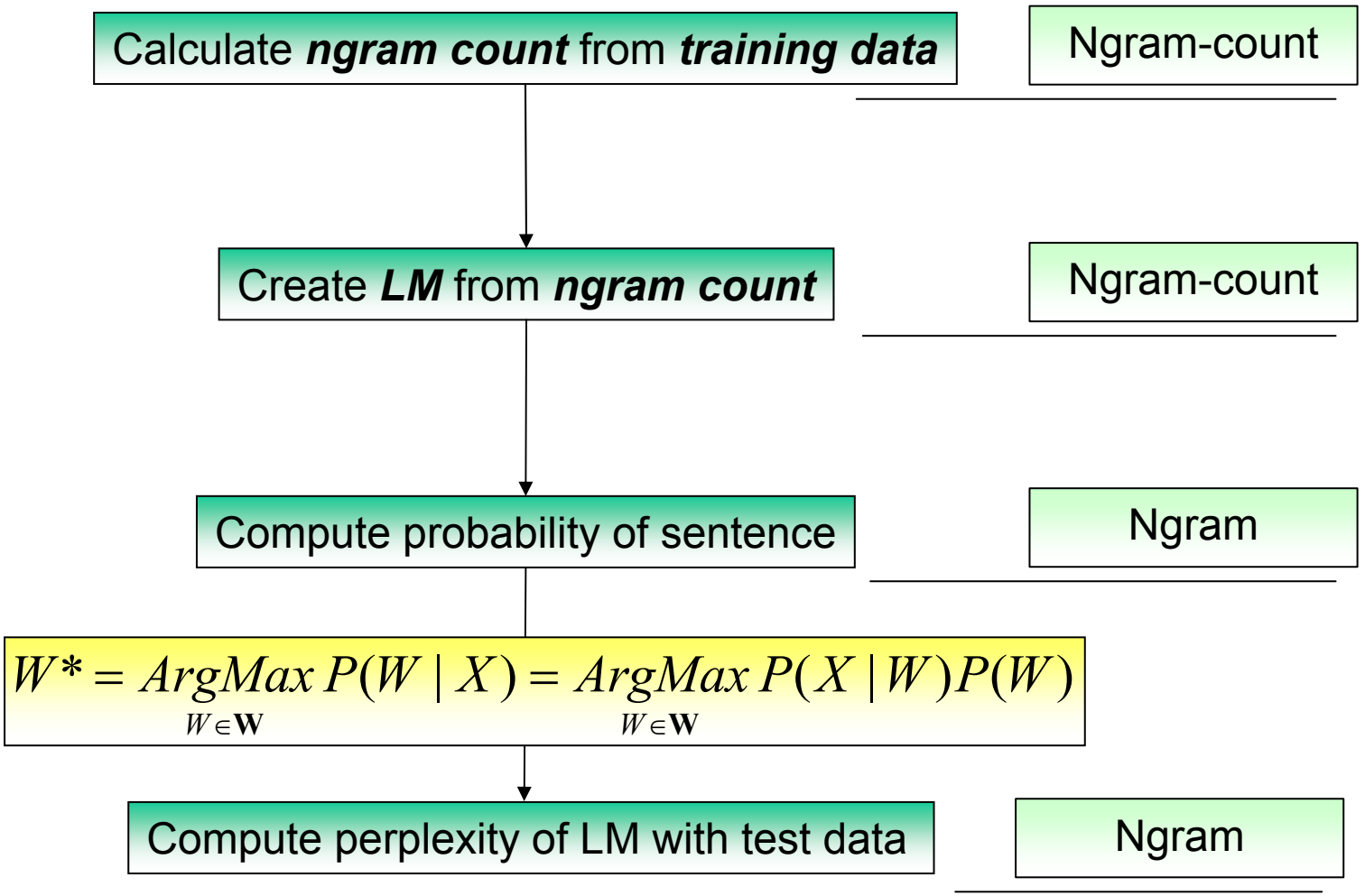
  把已經compile過的srilm.rar解壓縮到c:\cygwin\

- 如果要compile source code，cygwin要安裝gcc, g++, make等工具。
- 修改srilm/Makefile 增加一行SRILM=/srilm
- 在srilm目錄下依序執行

make World、make all、make cleanest

詳細資料參考INSTALL、srilm/doc/README.windows

Main Tasks


Calculate ***ngram count*** from ***training data***
Ngram-count
Create ***LM*** from ***ngram count***
Ngram-count
Compute probability of sentence
Ngram
$W^* = \underset{W \in \mathbf{W}}{ArgMax}\, P(W \mid X) = \underset{W \in \mathbf{W}}{ArgMax}\, P(X \mid W) P(W)$
Compute perplexity of LM with test data
Ngram

Training Data Format (CNA0001-2M.Train)

- CNA0001-2M.Train 49.8MB

中華民國 八十 九 年 一月 一 日
中央 通訊社 新聞
行政院 新聞局
局版臺訊字 第
一 四 號 中華 郵政 認爲 第一 類新聞紙類中
華民國 十三 年 四月 一 日 創刊
發行人 行人
汪
萬

Vocabulary Format (Lexicon2003-72k.txt)

巴
八
扒
叭

墨竹
默祝
末梢
沒收
墨守
陌生
莫說

共有 71695 個words

Count

- 計算training data中每個出現的ngram的出現次數

輸出 ngram count檔

要計算count的 Training data

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

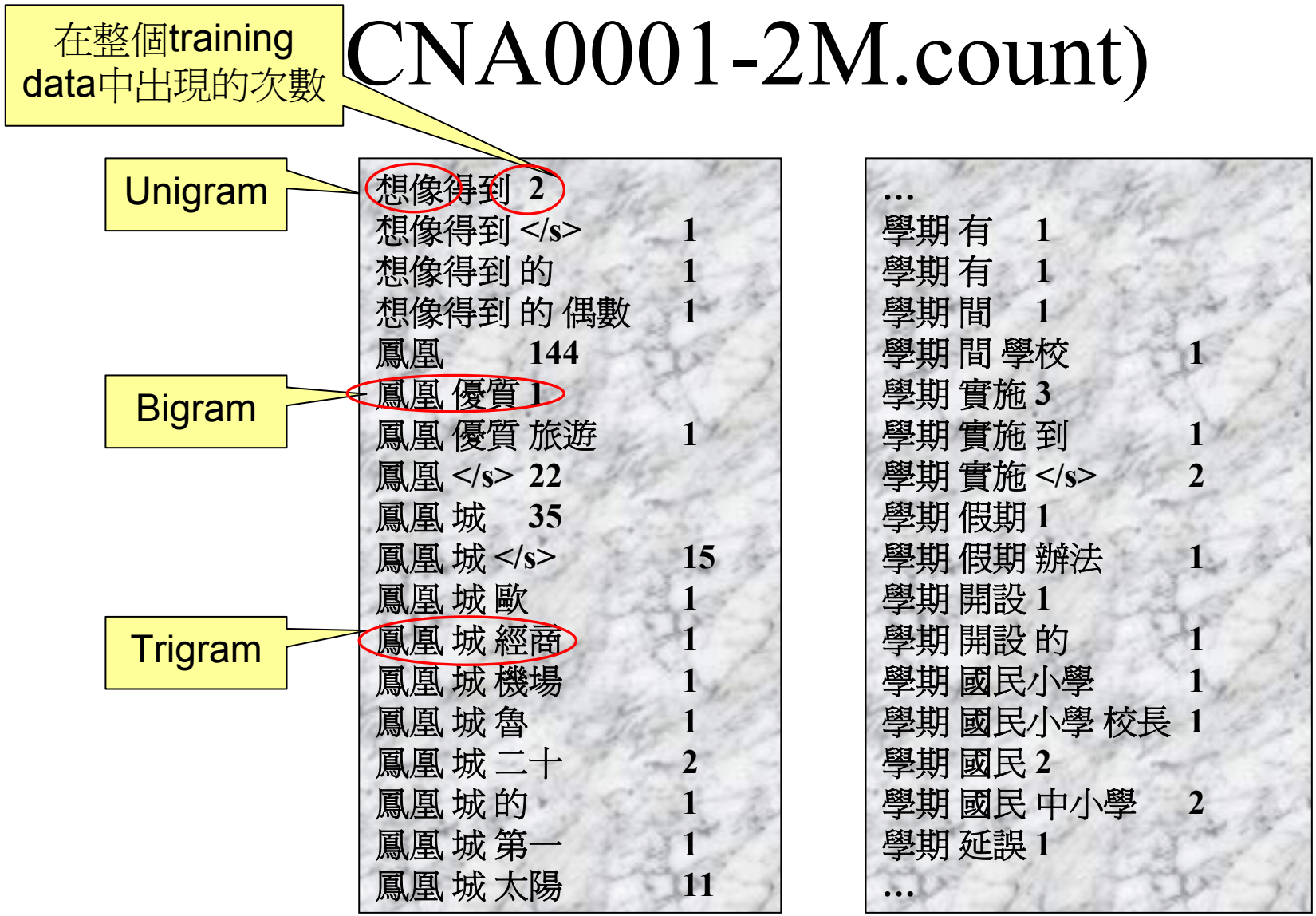
trigram

有-unk：把詞典中沒有的詞(oov)當作一個<unk>符號處理

無-unk :遇到oov就移除

詞典檔

Ngram Count Format (CNA0001-2M.count)



Good-Turing Discounting and Katz Backoff

預設的 discounting跟 backoff方法

Ngram count檔

輸出LM 檔

➢ngram-count -read CNA0001-2M.count -lm CNA0001-2M_N3.LM.3-7.gt -unk

> ngram -ppl. 506.PureText -lm CNA0001-2M_N3.LM.3-7.gt -unk – order 3 > NTNU2004hw.ppl

輸出ppl結果

要算ppl的 Test data

LM檔

file TestText.txt: 1096781 sentences, 4564098 words, 0 OOVs
0 zeroprobs, logprob= -1.31635e+07 ppl= 211.521 ppl1= 765.86

$10^{-\frac{\text{logprob}}{\text{句數+字數}}}$

$10^{-\frac{\text{logprob}}{\text{字數}}}$

LM Format (CNA0001-2M_N3.LM.3-7.gt)

```
\data\
ngram 1=71697
ngram 2=571216
ngram 3=578595

\1-grams:
-0.8422701              </s>
-99          <s>        -1.167359
-2.07538     一         -1.257989
-4.031029    一一       -0.6068805
-7.448923    一一恐怖
-7.448923    一一恐怖攻擊
-7.448923    一丁點
-3.339295    一九九     -2.965494
-7.448923    一了百了
-7.448923    一刀兩斷
-4.720053    一下       -0.3730158
-5.184257    一下子     -0.1966212
-6.6661      一口咬定
```

以\data\開頭

Log probability (Base 10)

Log of backoff weight (Base 10)

最高的ngram沒有backoff weight

```
-1.227553  德 碁 </s>
-1.154267  德 碁 半導體
-1.6313    德 碁 合併
-1.154267  德 碁 和
-1.154267  德 碁 的
-1.227553  德 碁 後
-1.227553  德 碁 換
-1.6313    德 碁 普通
-1.6313    德 碁 對
-0.7096938          德 碁 與
-0.1099438          不 銹 鋼
-0.4164561          <s> 裏 </s>
-0.1435432          戈 裏 峰
-0.6717286          那 裏 </s>
-0.1544242          薩 裏 德
-0.5066067          化 粧 </s>

\end\
```

以\end\結尾

Good-Turing Discounting Default Parameters

自己指定mincount跟maxcount :

> ngram-count -read CNA0001-2M.count -lm CNA0001-2M_N3.LM.3-7.gt -unk --gt1min 1 -gt1max 7 -gt2min 1 -gt2max 7 -gt3min 1 -gt3max 7

1-gram GT parameters
mincount 1
maxcount 1

2-gram GT parameters
mincount 1
maxcount 7

3-gram GT parameters
mincount 2
maxcount 7

預設值

maxcount

mincount

$$P_{Katz}(w_i|w_{i-1}) = \begin{cases} C(w_{i-1}w_i)/C(w_{i-1}) & r > k \\ d_r C(w_{i-1}w_i)/C(w_{i-1}) & k \geq r > 0 \\ \alpha(w_{i-1})P(w_i) & r = 0 \end{cases}$$

where $d_r = \dfrac{\dfrac{r^*}{r} - \dfrac{(k+1)n_{k+1}}{n_1}}{1 - \dfrac{(k+1)n_{k+1}}{n_1}}$ and $\alpha(w_{i-1}) = \dfrac{1 - \sum_{w_i: r>0} P_{Katz}(w_i|w_{i-1})}{1 - \sum_{w_i: r>0} P_{Katz}(w_i)}$

Absolute Discounting

- ngram-count -read CNA1999.count -lm CNA1999.cd5.lm -unk -cdiscount1 0.5 -cdiscount2 0.5 -cdiscount3 0.5

> ngram -ppl CNA1999.test -lm CNA1999.cd5.lm -unk -order 3 > CNA1999.cd5.o3.ppl

Modified Kneser-Ney Discounting

> ngram-count -read CNA1999.count -lm CNA1999.knd.lm -unk -kndiscount1 -kndiscount2 -kndiscount3

> ngram -ppl CNA1999.test -lm CNA1999.knd.lm -unk -order 3 > CNA1999.knd.o3.ppl

Witten-Bell Discounting

> ngram-count -read CNA1999.count -lm CNA1999.wbd.lm -unk -wbdiscount1 -wbdiscount2 -wbdiscount3

> ngram -ppl CNA1999.test -lm CNA1999.wbd.lm -unk -order 3 > CNA1999.wbd.o3.ppl

Ristad's Natural Discounting

> ngram-count -read CNA1999.count -lm CNA1999.nd.lm -unk -ndiscount1 -ndiscount2 -ndiscount3

> ngram -ppl CNA1999.test -lm CNA1999.nd.lm -unk -order 3 > CNA1999.nd.o3.ppl