Corpus-Based Methods in Chinese Morphology and Phonology

中文構詞法和音韻學語料庫方法

Richard Sproat

AT&T Labs — Research

rws@research.att.com

www.research.att.com/∼rws

Chilin Shih

Bell Labs

cls@research.bell-labs.com

www.bell-labs.com/project/tts/cls.html

Lecture 3: Applications

2001 LSA Institute – University of California, Santa Barbara

Subinstitute on Chinese Corpus Linguistics

Overview of Lecture 3

- Segmentation methods
- Linguistic Studies
- Other Applications

Segmentation Methods

Approaches to Chinese word segmentation fall into three categories:

- Non-stochastic lexicon-based methods.
- Purely statistical methods.
- Methods that combine statistical corpus-based methods with information derived from lexica.

Segmentation Methods: General issues

- Purely statistical methods have been fairly limited.
- Two primary issues:
  - How to deal with ambiguities:

    馬路上生病了

    (*mǎ-lù-shàng shēng-bìng le*

    '(He) got sick on the road'

    mǎ lù-shàng shēng-bìng le

    'The horse got sick on the road.'

    (cf. Gan 1995)
  - How to deal with unknown words. Two issues here:
    - Identifying (hence segmenting) unknown words *as* words
    - Identifying the category of word.

      馬魁乾 versus 馬肉乾

Segmentation Methods: General issues

- Performance is typically reported in terms of precision and recall

- Size of base dictionary is clearly an important predictor of performance

Purely Statistical Approaches: Sproat & Shih (1990)

1. For a string of characters $c_1 \ldots c_n$, find the pair of adjacent characters with the largest mutual information greater than some threshold (optimally 2.5), and group these.

2. Iterate 1 until there are no more characters to group.

Purely Statistical Approaches: Sproat & Shih (1990)

	我	弟	弟	現	在	要	坐	火	車	回	家
MI	0	10.4	0	4.2	-2.8	0	703		2.1	4.7	
1	我	弟	弟	現	在	要	坐	火	車	回	家
2	我	[弟	弟]	現	在	要	坐	火	車	回	家
3	我	[弟	弟]	現	在	要	坐	[火	車]	回	家
4	我	[弟	弟]	現	在	要	坐	[火	車]	[回	家]
5	我	[弟	弟]	[現	在]	要	坐	[火	車]	[回	家]

Table 1: Stages in the derivation of *wǒ dìdì xiànzài yào zuò huǒchē huíjiā*. The second line shows the mutual information between adjacent characters.

Purely Statistical Approaches: Others

- Sun, Shen and Tsou (1998) propose various enhancements of the Sproat and Shih method, using various association measures besides mutual information. They propose a *t* score for determining whether to group *xyz* as *xy z* or *x yz*:

(1) $$t_{x,z(y)} = \frac{p(z|y) - p(y|x)}{\sqrt{var(p(z|y)) - var(p(y|x))}}$$

- Ge, Pratt and Smyth propose a method based on expectation maximization.

Statistically-Aided Approaches: WFST-Based Approach

- Represent dictionary D as Weighted Finite State Transducer, mapping **from** $ChinChar \cup \epsilon$ **to** $PinyinSyllable \cup POS$. (Note that since the primary application of the Sproat et al system was to text-to-speech, the system was not merely a segmenter, but also a transducer mapping from text into phonemic transcription.)

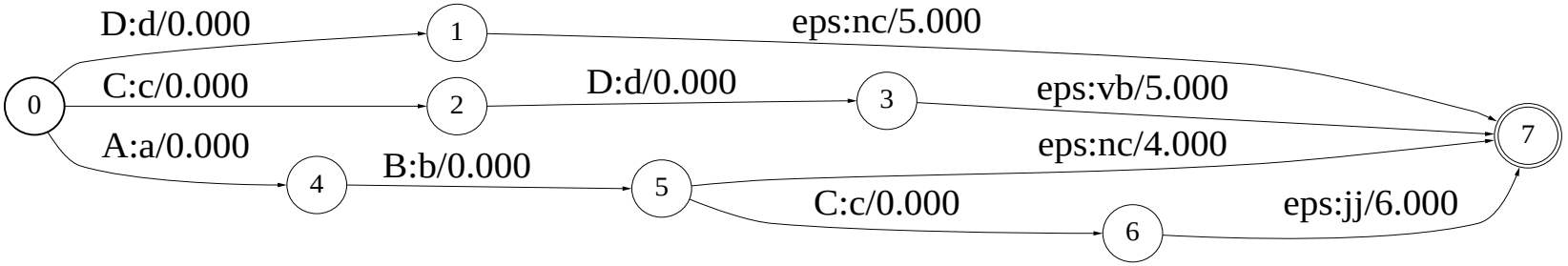
- Weights on word-strings are derived from frequencies of the strings in a 20M character corpus.

- Represent input I as unweighted acceptor over the set $ChinChar$

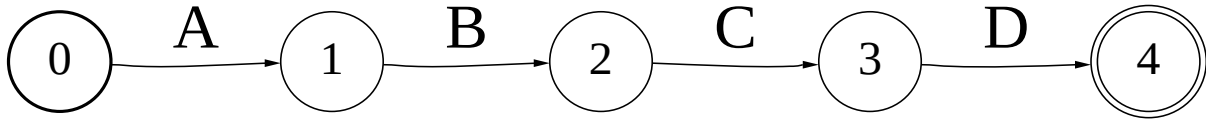
- Choose segmentation as $BestPath(I \circ D^*)$

Statistically-Aided Approaches: WFST-Based Approach

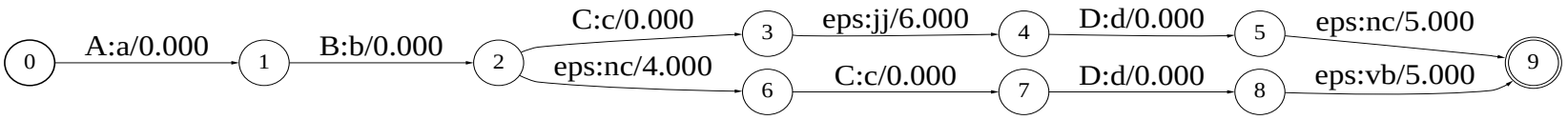
Dictionary D



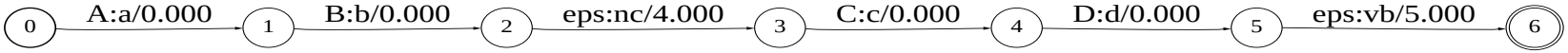
Input I



$I \circ D^*$



$BestPath(I \circ D^*)$



Statistically-Aided Approaches: WFST-Based Approach

- Estimate probabilities of dictionary words from unlabeled corpus.
- For morphologically derived words:
  - Estimate probabilities for examples in the corpus (青蛙們 *qīngwā*+*men* (frog+PL) '(anthropomorphized) frogs') as for underived words.
  - For unattested formations, use the Good-Turing estimate:

(2) $P(\text{南瓜們}) = P(unseen(\text{們}))P(\text{南瓜})$

Statistically-Aided Approaches: WFST-Based Approach

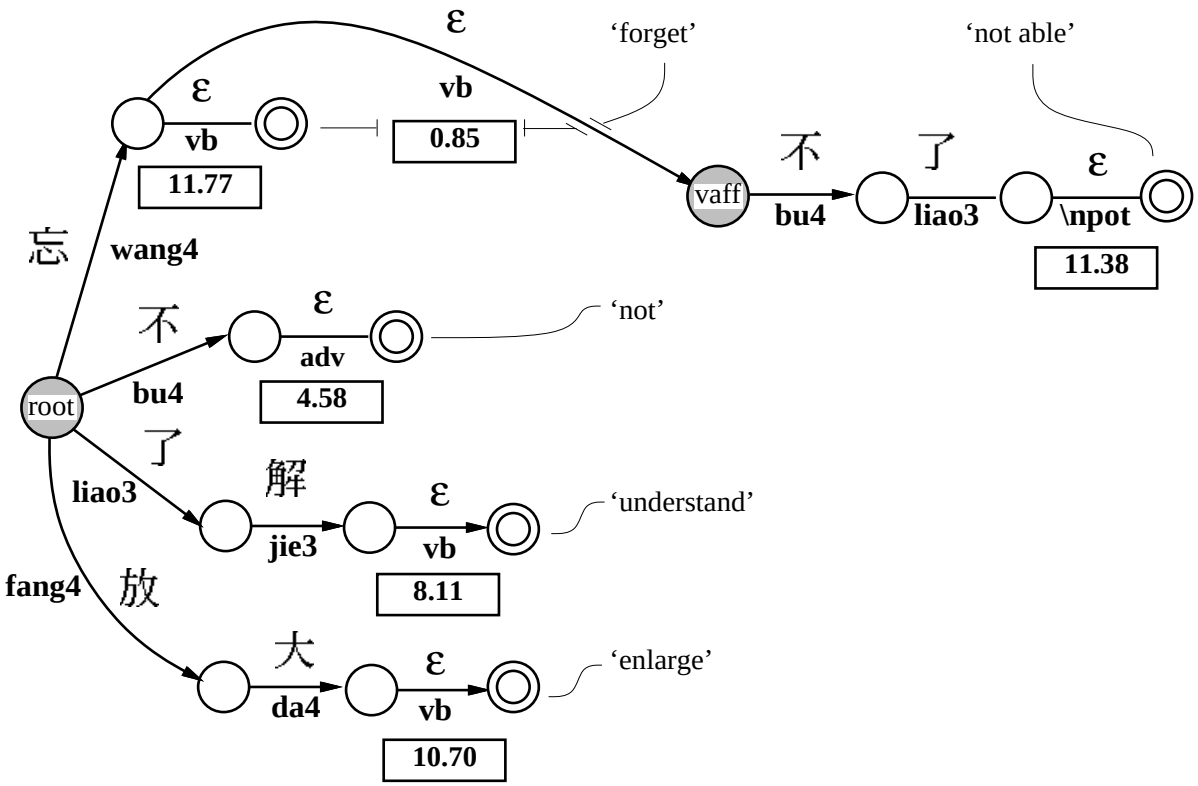
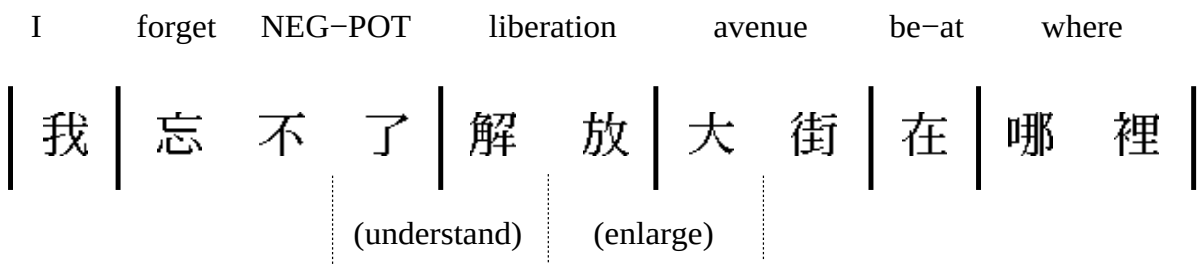
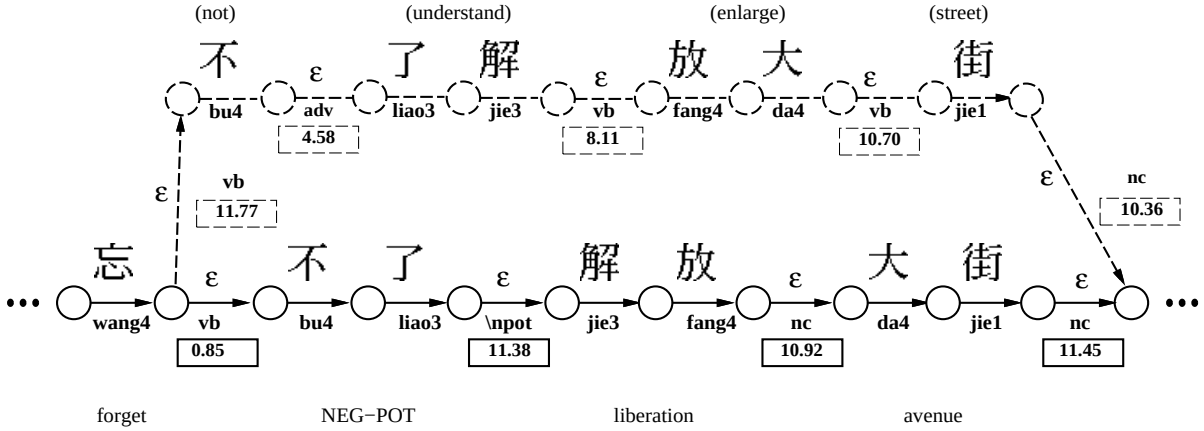


Figure 1: A fragment of the dictionary represented as a WFST.

Statistically-Aided Approaches: WFST-Based Approach



“I couldn’t forget where Liberation Avenue is.”



Statistically-Aided Approaches: WFST-Based Approach

- Personal names handled following Chang et al (1992):

(3)

$$p(name|FG_1G_2) = p(|fam| = 1, |giv| = 2) \times p(Fam = F) \times p(Giv_1 = G_1) \times p(Giv_2 = G_2) \times p(name)$$

- Foreign names in transliteration handled with a weighted finite-state recognizer over characters commonly used to transliterate foreign words.

Statistically-Aided Approaches: WFST-Based Approach

亞	0.0383	斯	0.0365
拉	0.0334	爾	0.0267
克	0.0218	巴	0.0205
尼	0.0187	利	0.0178
馬	0.0169	阿	0.0151
西	0.0151	蘭	0.0138
羅	0.0134	加	0.0134
里	0.0129	達	0.0125
特	0.0125	德	0.0111
維	0.0102	波	0.0102
哥	0.0089	多	0.0089
布	0.0089	格	0.0076
比	0.0076	倫	0.0071

Statistically-Aided Approaches: WFST-Based Approach

- Six human judges were asked to segment by hand a test corpus of 100 sentences (4,372 characters total).
- The human segmentations were compared with the algorithm just sketched (judge "ST"), as well as:
  1. A **greedy** algorithm (or 'maximum matching' algorithm) (judge "GR"): proceed through the sentence, taking the longest match with a dictionary entry at each point.
  2. An **anti-greedy** algorithm, (judge "AG"): instead of the longest match, take the shortest match at each point.
- Pairwise comparison between each of the human and automatic judges, computing the precision and recall in each case, and then computing the arithmetic mean between the two:

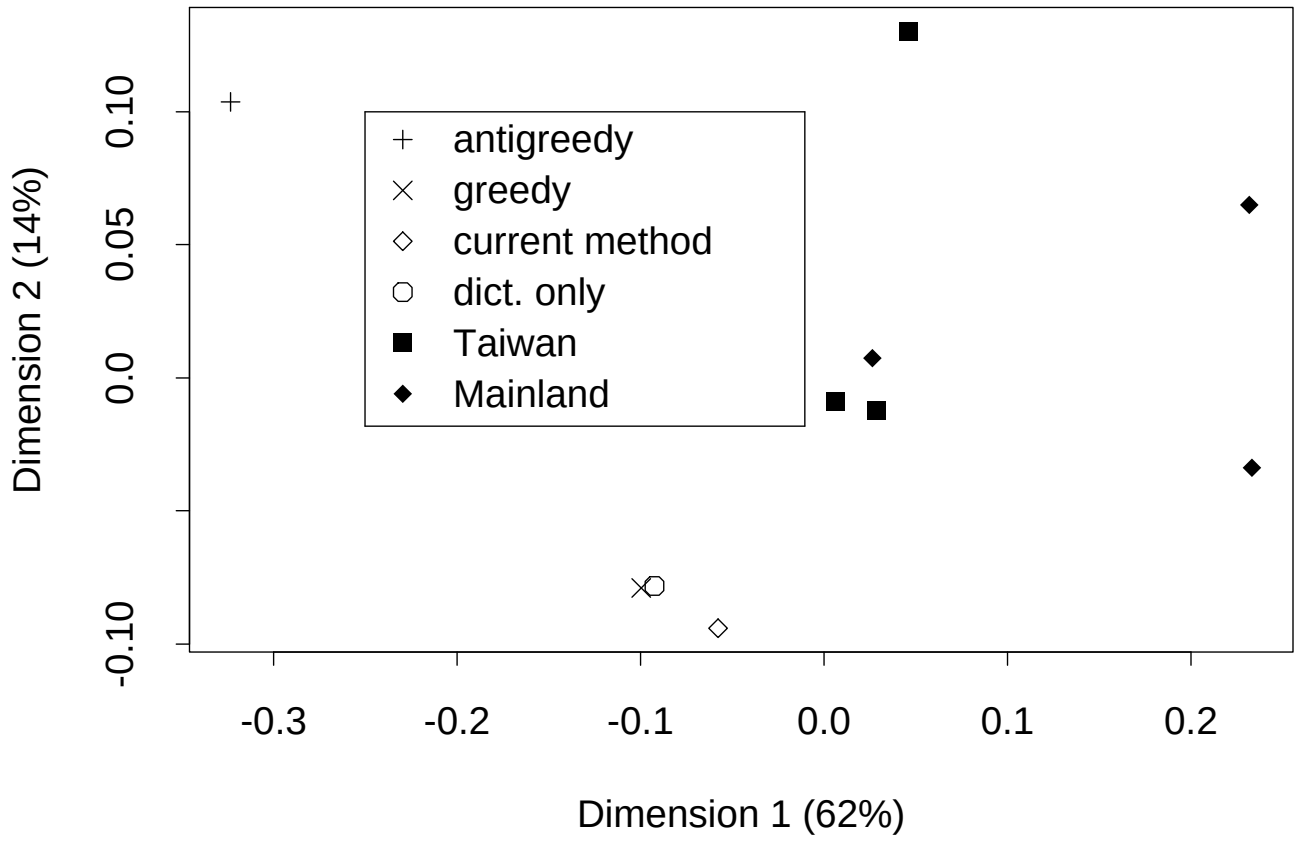
(4) $$Similarity = \frac{Precision + Recall}{2}$$

Statistically-Aided Approaches: WFST-Based Approach

Judges	AG	GR	ST	M1	M2	M3	T1	T2	T3
AG		0.70	0.70	0.43	0.42	0.60	0.60	0.62	0.59
GR			0.99	0.62	0.64	0.79	0.82	0.81	0.72
ST				0.64	0.67	0.80	0.84	0.82	0.74
M1					0.77	0.69	0.71	0.69	0.70
M2						0.72	0.73	0.71	0.70
M3							0.89	0.87	0.80
T1								0.88	0.82
T2									0.78

Table 2: Similarity matrix for segmentation judgments

Statistically-Aided Approaches: WFST-Based Approach



Transformation-Based Learning

Transformation-based Error-driven Learning (TBL), due to Brill (1993) has been applied to Chinese word segmentation by Palmer (1997) and Hockenmaier and Brew (1998).

- Start with a reference corpus (e.g. a segmented corpus of Chinese), and an initial tagger.

  The initial tagger can be simple: e.g. identify every character as a separate word.

Transformation-Based Learning

- Give the system an inventory of possible transformations. E.g.:
  - **Insert** – place a new boundary between two characters
  - **Delete** – remove an existing boundary between two characters
  - **Slide** – move an existing boundary from its current location between two characters to a location 1, 2, or 3 characters to the left or right

- Iterate, at each stage choosing a transformation that gives the greatest local improvement according to some defined error function, until no further improvement is obtained.

Transformation-Based Learning

For example, if the current segmenter gives:

我 愛 吃 西 瓜

and the reference segmentation is

我 愛 吃 西瓜
wǒ ài chī xīguā
'I like to eat watermelon'

the system might learn that it should delete a boundary between 西 and 瓜

Transformation-Based Learning

- "Character as Word": initial F-measure[a] of 40.3, increased to 78.1 after learning 5,903 transformations.

- Maximum matching: the initial F score was 64.4, increased to 84.9 after acquiring 2,897 transformations.

[a] F is defined in terms of precision P and recall R as follows: $F = \frac{(1+\beta)PR}{\beta P + R}$

More on Unknown Words

- Wang, Li and Chang (1992) propose a method for dealing with names based on titles (e.g. 先生 *xiānshēng* 'Mr.') and other key words. They also use "adaptive dynamic word-formation"
- Chen and Chen (2000) use a lexicon of known organizations along with Chinese texts spidered from the WWW.

  Process the lexicon to find common suffixes: e.g. 女青年會 *nǚ qīngnián huì* 'Women's Youth Assocation'

  Texts from the WWW in particular topic domains are collected; high frequency key words extracted following (Smadja 1993).

  Check key words for suffixes that match the organization suffixes collected in the first phase. If the prefix of the name is not an ordinary word and if the suffix is found with at least n distinct prefixes in the corpus, then the prefix-suffix combination is assumed to be a name.

  Precision between 90% and 100%, for n = 2 or 3, respectively.

Linguistic Studies

- Properties of morphological constructions

- Suoxie

Properties of Morphological Constructions: Huang (1999)

- Character-based mutual information is a good indicator of wordhood.
- Mutual information also correlates with morphological type:

鞠躬	júgōng	‘to bow’	17.15
躊躇	chóuchú	‘to hesitate’	20.30
蜘蛛	zhīzhū	‘spider’	18.25
霓虹	níhóng	‘neon’	15.66

Table 3: Monomorphemic words and their mutual information.

Properties of Morphological Constructions: Huang (1999)

母親	mǔqīn	‘mother’	9.78
教會	jiàohuì	‘to teach’	9.63
游泳	yóuyǒng	‘to swim’	11.77

Table 4: Non-versatile polymorphemic words and their mutual information.

第一	dìyī	‘first’	5.11
免職	miǎnzhí	‘to dismiss’	3.29
唱歌	chànggē	‘to sing’	8.42

Table 5: Versatile polymorphemic words and their mutual information.

Properties of Morphological Constructions: Huang (1999)

看魚	kàn yú	'look at fish'	-1.27
打人	dǎ rén	'to hit people'	-0.91
一第	yī dì	'one PREFIX'	-4.40
性人	xìng rén	'-ity human'	-1.44

Table 6: Phrases and non-words.

Non-Versatile			
咖啡	kāfēi	'coffee'	14.86
拷貝	kǎobèi	'copy'	12.77
Versatile			
伏特	fútè	'volt'	4.99
攝氏	shèshì	'Celsius'	9.43

Table 7: Some borrowed words, and their mutual information.

Analysis of Suoxie: Huang et al (1993, 1994)

- Why is 北京大學 *běijīng dàxué* abbreviated to 北大 *běidà* and not 京大 *jīngdà*?

- 高雄 *gāoxióng* 'Kaohsiung'
  → 高 *gāo* (高縣 *gāoxiàn* 'Kaohsiung County')
  → 雄 *xióng* (雄中 *xióngzhōng* 'Kaohsiung High School')

- "Informativeness criterion" can't explain this

- "Morphological blocking": 高中 *gāozhōng* 'high school' exists as a word and so blocks its use in for 'Kaohsiung High School'

Analysis of Suoxie: Huang et al (1993, 1994)

Huang et al propose a mutual-information based model. for a two-character word AB, to form a *suoxie* compound with X from one element of AB:

1. If A and B are both *associated* with X (e.g., as measured on a corpus with a 5-character window to the left of X) then pick A (the lefthand member) to form AX. As in (Huang 1999) a pair of characters are assumed to be associated if the mutual information exceeds 2.

2. Otherwise pick the character that is *least* associated with X.

Analysis of Suoxie: Results of Test on 20 Million Character AS Corpus

Full	Suoxie	MI(c_1,縣)	MI(c_2,縣)
桃園 Taoyuan	桃縣	4.39	3.32
宜蘭 Yilan	宜縣	3.11	3.86
苗栗 Miaoli	苗縣	4.40	5.37
彰化 Changhua	彰縣	4.48	2.13
雲林 Yunlin	雲縣	3.78	2.13
嘉義 Chiayi	嘉縣	3.49	2.43
澎湖 Penghu	澎縣	3.33	1.92
花蓮 Hualian	花縣	0.95	4.08
高雄 Kaohsiung	高縣	1.33	3.21
屏東 Pingtong	屏縣	1.00	1.29
台北 Taipei	北縣	2.64	0.81
新竹 Hsinchu	竹縣	0.67	0.66
台中 Taichung	中縣	2.64	0.19
南投 Nantou	投縣	0.58	0.29
台南 Tainan	南縣	2.64	0.58
台東 Taitong	東縣	2.64	1.29

Inferring Word Classes: Chang and Chen (1995)

- Find a class assignment ϕ for words that maximizes the probability of a corpus. A bigram approximation of this would state that the estimated probability of the text T of length L — $\hat{p}(T)$ — is modeled as the product over each word w_i of the probability of w_i given the inferred class $\phi(w_i)$, along with the probability of $\phi(w_i)$ given the previous $\phi(w_{i-1})$:

$$\hat{p}(T) = \prod_{i=1}^{L} p(w_i|\phi(w_i))p(\phi(w_i)|\phi(w_{i-1}))$$

- A sensible group: 戶, 大類, 件, 次, 位, 名, 多名, 枋, 段, 段式, 隻, 條, 瓶, 塊, 層樓 ...

- An odd group: 中山, 中正, 中興, 仁愛, 文學, 水, 牛, 山, 平等, 打開, 民生, 白蘭地 ...

Readings for Lecture 3

- Sproat and Shih (1990).
- Sproat et al. (1996).
- Huang et al. (1994).
- Chang and Chen (1995).
- Huang (1999) (online: huangpaper.{txt,ps,pdf})
- Chang and Su (1997) (*on reserve*).

Lab Assignment for Lecture 3

1. Implement the segmenter from Sproat and Shih 1990.
2. Huang (1999) proposes that Mutual Information correlates with morphological productivity: a pair of characters (or more generally a pair of morphological components) that has high mutual information is argued to correlate with low productivity, whereas if the components have lower (but not too low!) mutual information, then the productivity of the components is probably higher. Is there a problem with this view? Consider some productive morphological processes, and compute the mutual information between the components in exemplars of those processes.
3. Does the theory of *suoxie* propounded in Huang, Ahrens and Chen (1993) and Huang, Hong and Chen (1994) for Taiwan county names work on other corpora? For extra credit, also consider whether it works for other *suoxie*.

4. Feng (1997) (*on reserve*) claims that there was an increase in the number of disyllabic compounds in the text of 孟子 Mengzi (Mencius, c. 372–289 BC) and the text of his main Han dynasty commentator 趙岐 Zhao Qi (c. 107–201 AD). He argues that this commonly observed increase in the disyllabicity of Chinese words was for prosodic reasons. But is there in fact such an increase? If you compare the text of the pre-Han period with a similarly sized sample of later texts — say the Han period, the Jin-Song-Ming period, and Modern Chinese — do you observe fewer compounds? Use one of the association measures introduced in Lecture 2 to answer this question, comparing pre-Han texts with later texts.