Corpus-Based Methods in Chinese Morphology

中文構詞法語料庫方法

Richard Sproat
AT&T Labs — Research
rws@research.att.com
`http://www.research.att.com/~rws`

August 24, 2002
COLING
Taipei, Taiwan, R.O.C.

To the memory of William A. Gale (1939 – 2002)

Overview

1. **Intro to Chinese Morphology**: What is a Word?
2. **Statistical Methods**: Word Frequency Distributions, Measures of Productivity, Measures of Association
3. **Applications**

Most of the material in this tutorial is based on lectures given at the 2001 LSA Summer Institute in Santa Barbara. The course notes for that course are available:

`http://www.research.att.com/∼rws/newindex/notes.pdf`

There are some exercises related to this tutorial:

`http://www.research.att.com/∼rws/exercises`

Please have a look at these in your spare time.

These slides are:

`http://www.research.att.com/∼rws/newindex/coling.pdf`

Overview of Part 1: Introduction to Chinese Morphology

- What is a Word?

- Some Chinese Morphological Phenomena

- Segmentation Standards

- Preview: What Corpus-Based Methods Can Do for Morphology; What Morphology can do for Corpus-Based Methods

What is a Word?

- Having a good definition of what is a word seems like a prerequisite to any study of words in any language.
- In Chinese the problem is exacerbated by the lack of word boundaries in orthography.
- Human judges disagree: in (Sproat et al., 1996) we only measured 0.76 (out of 1.00) agreement among human judges.
- Proposed segmentation standards don't even agree.

What is a Morpheme?

- At least it is agreed that Chinese words are made up of morphemes, but: what is a morpheme?

- Morpheme = single syllable, single character?

- Problems:
  - Polysyllabic morphemes such as 東西 *dōngxī* (east west) 'things'; 馬上 *mǎshàng* (horse on) 'immediately'
  - Borrowed morphemes: 巴基斯坦 *bājīsītǎn* 'Pakistan'

A Clear Class of Disyllabic Morphemes

Orthography	Analysis	Pronunciation	Gloss
蹉跎	<**foot+cuōtuō** >	*cuōtuó*	‘procrastinate’
踉蹌	<**foot+liángcāng** >	*lángqiāng*	‘hobble’
蹂躪	<**foot+róulìn** >	*róulìn*	‘trample’
躊躇	<**foot+shòuzhù** >	*chóuchú*	‘hesitate’
躑躅	<**foot+zhèngshǔ** >	*zhízhú*	‘hesitate’
萵苣	<**grass+guǎjù** >	*wōjù*	‘lettuce’
菡萏	<**grass+hánxiàn** >	*hàndàn*	‘lotus’
蒹葭	<**grass+jiānjiǎ** >	*jiānjiā*	‘type of reed’
苜蓿	<**grass+mùsù** >	*mùsù*	‘clover’
慫慂	<**heart+cóngyǒng** >	*sǒngyǒng*	‘egg on’
忸怩	<**heart+niuní** >	*niǔní*	‘coy’
慇懃	<**heart+yīnqín** >	*yīnqín*	‘attentively’
蝙蝠	<**insect+biǎnfù** >	*biānfú*	‘bat’
蜉蝣	<**insect+fúyóu** >	*fúyóu*	‘mayfly’
蚯蚓	<**insect+qiūyǐn** >	*qiūyǐn*	‘earthworm’

Pairs of characters that only cooccur with each other, and occur at least three times in a 20 million character corpus. Note that in each case both characters share the same semantic radical. See, for instance, the examples with the **foot** radical, underlined in the table above. The second column gives the component analysis following the schema in (Sproat, 2000).

Packard's (2000) Notions of Word: I

- **Orthographic word:** Words as defined by delimiters in written text. This appears to have no relevance in Chinese since there are no such written delimiters (apart from punctuation marks, which mark ends of phrases, not words).

- **Sociological word:** Following (Chao, 1968, pp. 136–138), these are 'that type of unit, intermediate in size between a phoneme and a sentence, which the general, non-linguistic public is conscious of, talks about, has an everyday term for, and is practically concerned with in various ways.' In English this is the lay notion of 'word', whereas in Chinese this is the character (字 *zì*).

- **Lexical word:** This corresponds to Di Sciullo and Williams' (1987) *listeme*.

Packard's (2000) Notions of Word: II

- **Semantic word:** Roughly speaking, this corresponds to a 'unitary concept'.
- **Phonological word:** A word-sized entity that is defined using phonological criteria. (Yes, that's circular.) Packard goes on to note that Chao considers prosodic issues, such as when one can pause in a sentence, to provide useful tests for phonological wordhood. In many languages, phonological words are the domain of particular phonological processes: in Finnish, for example, vowel harmony is restricted to phonological words. In dialects of Mandarin that have (stress) reduction, such phenomena are generally restricted to words, as is consonant lenition.

Packard's (2000) Notions of Word: III

- **Morphological word:** following Di Sciullo and Williams' notion, a morphological word is anything that is the output of a word-formation rule.
- **Syntactic word:** These are all and only those constructions that can occupy X^0 slots in the syntax. (The syntactic word is the definition of word that Packard uses as the basis for the bulk of his discussion.)
- **Psycholinguistic word:** This is the "'word' level of linguistic analysis that is ... salient and highly relevant to the operation of the language processor" (Packard, 2000, page 13).

To these definitions could be added the extrinsically defined notions of word that have formed the basis of the various segmentation standards for Chinese that have been proposed.

Morphological Types

- Reduplication
- Affixation
- Compounding
- Proper names
- Abbreviations (縮寫 *suōxiě*)

Verbal reduplication

(1)

說說	shuō-shuō	(speak speak)	‘speak a little’
看看	kàn-kàn	(look look)	‘have a little look’
走走	zǒu-zǒu	(walk walk)	‘have a little walk’
磨磨	mó-mó	(rub rub)	‘rub a little’
討論討論	tǎolùn-tǎolùn	(discuss discuss)	‘discuss a little’
請教請教	qǐngjiào-qǐngjiào	(ask ask)	‘ask a little’
研究研究	yánjiù-yánjiù	(research research)	‘research into’

Verbal reduplication

(2)

說一說	shuō-yi-shuō	(say one say)	‘go ahead and say it’
看一看	kàn-yi-kàn	(look one look)	‘have a look’
走一走	zǒu-yi-zǒu	(walk one walk)	‘have a little walk’
磨一磨	mó-yi-mó	(rub one rub)	‘rub a little’

(3)

說說看	shuō-shuō-kàn	(speak speak look)	‘talk about it and see’
走走看	zǒu-zǒu-kàn	(walk walk look)	‘walk and see’

Adjectival Reduplication

(4)	紅紅	*hóng-hóng*	(red red)	‘red’
	慢慢	*màn-màn*	(slow slow)	‘slow’

(5)

舒服	*shūfú*	舒舒服服	*shūshū-fúfú*	‘comfortable’
乾淨	*gānjìng*	乾乾淨淨	*gāngān-jìngjìng*	‘clean’
胡塗	*hútú*	胡胡塗塗	*húhú-tútú*	‘muddle-headed’
快樂	*kuàilè*	快快樂樂	*kuàikuài-lèlè*	‘happy’
漂亮	*piàoliàng*	漂漂亮亮	*piàopiào-liàngliàng*	‘pretty’

Adjectival Reduplication

(6)	亮晶	*liàngjīng*	亮晶晶	*liàng-jīngjīng*	‘bright’
	白花	*báihuā*	白花花	*bái-huāhuā*	‘white’

(7)	∗重重要要	*zhòngzhòng-yàoyào*	‘important’
	∗偉偉大大	*wěiwěi-dàdà*	‘majestic’
	∗粉粉紅紅	*fēnfēn-hónghóng*	‘pink’
	∗?美美麗麗	*měiměi-lìlì*	‘beautiful’
	∗透透明明	*tòutòu-míngmíng*	‘transparent’

Measure word reduplication

(8)	磅磅	*bàngbàng*	‘every pound’
	條條魚	*tiáotiáo yú*	‘every fish’
	套套西裝	*tàotào xīzhuāng*	‘every suit’
	件件衣服	*jiànjiàn yīfú*	‘every piece of clothing’
(9)	∗雙雙手	*zhīzhī shǒu*	‘every hand’
	∗冊冊書	*cècè shū*	‘every book’
	∗打打蛋	*dádá dàn*	‘every dozen eggs’
	?座座山	*zuòzuò shān*	‘every mountain’
	?隻隻雞	*zhīzhī jī*	‘every chicken’

Measure word reduplication

(10) 我們的雞,隻隻都病了
wǒmen-de jī, zhīzhī dōu bìng le
(our-DE chicken CL-CL all sick ASP)
'Every one of our chickens is sick.'

(11)	?日日	*rìrì*	'every day' (okay in poetic language)
	天天	*tiāntiān*	'every day'
	年年	*niánnián*	'every year'

(12) ∗公里公里 *gōnglǐ gōnglǐ* 'every kilometer'

Prefixation

(13)

老	*lǎo-*	老王	*lǎo wáng*	‘old Wang’
小	*xiǎo-*	小張	*xiǎo zhāng*	‘little Zhang’
第	*dì-*	第一	*dìyī*	‘first’
初	*chū-*	初三	*chū sān*	‘the third’
好/難	*hǎo-/nán-*	好吃/難吃	*hǎochī/nánchī*	‘tasty/bad-tasting’

(14)	可	*kě-*	可愛	*kě-ài*	‘lovable, cute’
			可靠	*kě-kào*	‘reliable’

Suffixation

(15) “Diminutive” suffixes:

兒	*-er*	鳥兒	*niǎo-er*	‘bird’
子	*-zi*	猴子	*hóu-zi*	‘monkey’
頭	*-tou*	饅頭	*mán-tou*	‘steamed bread’

(16) Other deriational suffixes:

學	*-xué*	心理學	*xīnlǐ-xué*	‘psychology’
家	*-jiā*	物理學家	*wùlǐxué-jiā*	‘physicist’
化	*-huà*	美化	*měi-huà*	‘Americanization’
率	*-l`ü*	生蛋率	*shēng-dàn-l`ü*	‘egg production rate’
主義	*-zhǔyì*	馬克斯主義	*mǎkèsī-zhǔyì*	‘Marxism’

Suffixation

(17)	了	-*le*	吃了	*chī-le*	'have eaten'
	過	-*guò*	吃過了	*chī-guò-le*	'have eaten'
	們	-*men*	孩子們	*háizi-men*	'children'

Compounding

(18) **location**

客廳 沙發	*kètīng shāfā*	'living room sofa'
河 馬	*hémǎ*	'hippopotamus (river horse)'
海 狗	*hǎigǒu*	'seal (sea dog)'

(19) **used for**

指甲油	*zhǐjiǎ yóu*	'nail polish'
乒乓球	*pīngpāng qiú*	'pingpong ball'
太陽眼鏡	*tàiyáng yǎnjìng*	'sunglasses'
飯碗	*fànwǎn*	'rice bowl'

(20) **material**

大理石 地板	*dàlǐshídìbǎn*	'marble (Dali rock) floor'
紙老虎	*zhǐlǎohǔ*	'paper tiger'

Compounding

(21) **powered by**

電 燈	*diàn dēng*	‘electric light’
風 車	*fēng chē*	‘windmill (wind cart)’

(22) **organization**

大學 校長	*dàxué xiàozhǎng*	‘university president’
北京 大學	*běijīng dàxué*	‘Beijing University’

(23) **coordinate**

爸爸 媽媽	*bàbà māmā*	‘father and mother’
風水	*fēng shuǐ*	‘fengshui (wind water)’
紙筆	*zhǐ bǐ*	‘paper and pen’
牛羊	*niú yáng*	‘livestock (cattle sheep)’

Root Compounding

(24)

螞蟻 *mǎyǐ* 'ant'	蟻王 *yǐwáng* 'queen ant'	工蟻 *gōngyǐ* 'worker ant'
腦子 *nǎozi* 'brain'	腦水腫 *nǎoshuǐzhǒng* 'hydrocephaly'	後腦 *hòunǎo* 'hindbrain'
蒼蠅 *cāngyíng* 'fly'	蠅屍 *yíngshī* 'fly corpse'	地中海蠅 *dìzhōnghǎiyíng* 'Mediterranean fly'
蘑菇 *mógū* 'mushroom'	菇傘 *gūsǎn* 'pileus'	金菇 *jīngū* 'golden mushroom'

(25) 蟻有蟻國,蜂有蜂國

yǐ yǒu yǐguó, fēng yǒu fēngguó

(ant have ant country, bee have bee country)

'Ants have Antland, bees have Beeland'

Resultative Compounding

(26) 上 *shàng* ‘up’
下 *xià* ‘down’
進 *jìn* ‘enter’
出 *chū* ‘(go) out’
起來 *qǐlái* inchoative
回 *huí* ‘return’
過 *guò* ‘pass, across, beyond’
開 *kāi* ‘open’
完 *wán* ‘finish’
好 *hǎo* ‘good, complete’
死 *sǐ* ‘die’
見 *jiàn* ‘see, perceive’
破 *pò* ‘break’
清楚 *qīngchǔ* ‘clearly’

Resultative Compounding

(27)	**result**	打破	*dǎpò*	'break by hitting'
		拉開	*lākāi*	'open'
	achievement	寫清楚	*xiěqīngchǔ*	'write clearly'
		買到	*mǎidào*	'succeed in buying'
	direction	跳過去	*tiàoguòqù*	'jump across'
		走進來	*zǒujìnlái*	'come walking in'

(28)	跳得過去	*tiàodéguòqù*	'able to jump across'
	跳不過去	*tiàobùguòqù*	'not able to jump across'

Parallel Verbal Compounds

(29)	購買	*gòu-mǎi*	(buy-buy)	‘buy’
	建築	*jiàn-zhù*	(build-build)	‘build’
	檢查	*jiǎn-chá*	(examine-examine)	‘examine’
	治療	*zhì-liáo*	(treat-treat)	‘treat (a sickness)’

Subject-Verb Compounds

(30)	頭疼	*tóu-téng*	(head hurt)	'(have a) headache'
	嘴硬	*zuǐ-yìng*	(mouth hard)	'stubborn'
	眼紅	*yǎn-hóng*	(eye red)	'covet'
	心酸	*xīn-suān*	(heart sour)	'feel sad'
	命苦	*mìng-kǔ*	(life bitter)	'tough straits'

(31) 我頭疼.

wǒ tóu téng

(I head hurt)

'I have a head ache.' (= 'My head hurts.')

(32) 你真嘴硬!

nǐ zhēn zuǐ-yìng

(you really mouth hard)

'You are really stubborn!' (≈ 'Your mouth is really hard!')

Verb-Object Compounds

(33)

出版	*chū-bǎn*	(emit edition)	'publish'
睡覺	*shuì-jiào*	(sleep sleep)	'sleep'
畢業	*bì-yè*	(finish course of study)	'graduate'
開刀	*kāi-dāo*	(operate knife)	'operate'
開玩笑	*kāi-wánxiào*	(operate joke)	'make fun of'
照像	*zhào-xiàng*	(shine image)	'take a photo'

Verb-Object Compounds

(34) 他們出版了那本書

tāmen chūbǎn-le nèiběn shū

(they publish-PERF that-CL book)

'They published that book.'

(35) *他們開刀心臟

tāmen kāidāo xīnzàng

(they operate heart)

'They are operating on the heart'

Verb-Object Compounds

(36) 我睡了覺

wǒ shuì-le jiào

(I sleep-PERF sleep)

'I fell asleep'

(37) 開一個玩笑

kāi yīge wánxiào

(operate one-CL joke)

'make fun of'

照兩張像

zhào liǎngzhāng xiàng

(shine two-CL photo)

'take two photos'

Verb-Object Compounds

(38) 我睡了一個小覺

wǒ shuì-le yīge xiǎo jiào

(I sleep-PERF one-CL little sleep)

'I took a little nap.'

(39) 開他的玩笑

kāi tāde wánxiào

(operate his joke)

'make fun of him'

(40) 一個便我都沒有大

yīge biàn wǒ dōu méiyǒu dà

(one-CL convenience i all NEG have big)

'I haven't defecated at all.'

(cf. 大便 *dàbiàn* (big convenience) 'defecate')

(41) 你開甚麼刀?

nǐ kāi shénmo dāo

(you operate what knife)

'What kind of operation are you having?'

Personal Names.

(42)

1	word	$\Rightarrow$	name
2	name	$\Rightarrow$	1-char-family 2-char-given
3	name	$\Rightarrow$	1-char-family 1-char-given
4	name	$\Rightarrow$	2-char-family 2-char-given
5	name	$\Rightarrow$	2-char-family 1-char-given
6	1-char-family	$\Rightarrow$	char$_i$
7	2-char-family	$\Rightarrow$	char$_i$ char$_j$
8	1-char-given	$\Rightarrow$	char$_i$
9	2-char-given	$\Rightarrow$	char$_i$ char$_j$

(Would need to expand this to cover “double-barreled” family names, which are still commonly used to refer to married women: thus 張王金蘭 *zhāng wáng jīnlán* ‘Mrs. Jinlan (Wang) Zhang’)

Abbreviations.

(43)	亞洲乒乓球聯盟	亞乒聯
	yǎzhōu pīngpāng qiú liánméng	*yǎ pīng lián*
	Asian Ping-Pong Association	
(44)	工業研究院	工研院
	gōngyè yánjiù yuàn	*gōng yán yuàn*
	Industrial Research Center	
(45)	以色列巴勒斯坦和談	以巴和談
	yǐsèliè bālèsītǎn hètán	*yǐ bā hètán*
	Israel-Palestinian discussions	

Abbreviations.

(46) 台灣香港 台港

táiwān xiānggǎng *tái gǎng*

Taiwan-Hong Kong

中國石油 中油

zhōngguó shíyóu *zhōng yóu*

China Oil

(47) 上海杭州鐵路 滬杭鐵路

shànghǎi hángzhōu tiělù *hù háng tiělù*

Shanghai-Hangzhou Railway

Segmentation Standards

- Deciding what is a word in Chinese has practical consequences for a wide variety of applications both “sociological” and technological.

- The sociological applications arose early in the twentieth century in the context of various attempts to romanize Chinese orthography; see (Pan, Yip, and Han, 1993):
  - 國語羅馬字運動 *guǒyǔ luómǎzì yùndòng* ‘Mandarin Romanization Movement’ (Y. R. Chao was involved in this)
  - 拉丁化運動 *lādīnghuà yùndòng* ‘Latinization Movement’

- Technological applications include IR, TTS, and ASR.

The Moral

$\boxed{\text{The "correct" segmentation depends on the intended purpose of the segmentation}}$

Proposed Standards

- Mainland Standard (GB/T 13715–92, 1993).

- ROCLING Standard (Huang et al., 1997).

- University of Pennsylvania/Chinese Treebank (Xia, 1999).

Mainland Standard

Lots of specific rules or principles, not all of them explained.

- Nouns derived in 家 *jiā*, 手 *shǒu*, 性 *xìng*, 員 *yuán*, 子 *zi*, 化 *huà*, 長 *zhǎng*, 頭 *tóu*, 者 *zhe* are segmented as one word: 科學家 *kēxuéjiā* ‘scientist’
- This seems reasonable enough, but for some reason 們 *men* is segmented separately, except in 人們 *rénmen* ‘people’, and words with *erhua*: 哥儿們 *gērmen* ‘brothers’.
- Personal names are also split: 毛 澤東 ‘Mao Zedong’.
- AABB reduplicants are one word: 高高興興 *gāogāoxìngxìng* ‘happy’. On the other hand ABAB reduplicants are two words: 雪白 雪白 *xuěbáixuěbái* ‘snow white’.

ROCLING Standard: I

The ROCLING standard seeks a standard that is “linguistically felicitous, computationally feasible, and must ensure data uniformity.” More of a meta-standard than a standard

- A string whose meaning cannot be derived by the sum of its components should be treated as a segmentation unit.
- A string whose structural composition is not determined by the grammatical requirements of its components,or a string which has a grammatical category other than the one predicted by its structural composition should be treated as a segmentation unit.

More specific guidelines:

- Bound morphemes should be attached to neighboring words to form a segmentation unit when possible.

ROCLING Standard: II

- A string of characters that has a high frequency in the language or high cooccurrence frequency among the components should be treated as a segmentation unit when possible.
- Strings separated by overt segmentation markers should be segmented.
- Strings with complex internal structures should be segmented when possible.

A reference lexicon, based on a reference corpus is assumed.

U Penn Standard

The University of Pennsylvania standard is more akin in spirit to the Mainland standard in that it has less philosophy than the ROCLING standard, and more detailed rules.

- Sensitivity to phonological criteria: 室内 *shìnèi* 'in the room' is considered to be a single word, 中午 以後 *zhōngwǔ yǐhòu* 'after noon' is considered to be two words.

- Internal structure is marked: 打得破 *dǎ-dé-pò* 'able to break' is tagged as [打/V 得/DER 破/V]/V

Comparison of Standards

Form	UPenn	Mainland	ROCLING	Example
ABAB	ABAB	AB-AB	ABAB	研究研究 ‘research (a bit)’
AA-看	[AA/V kàn/V]/V	AA kàn	AA kàn	說說看 ‘talk about it and see’
Pers. Names	One Seg	Two Segs	One Seg	史立璿 Shi Lixuan
Noun + 們	One Seg	Two Segs	Two Segs	朋友們 ‘friends’
Ordinals	One Seg	Two Segs	Two Segs	第一 ‘first’

Table 1: Some differences between the segmentation standards, from (Xia, 1999).

Segmented Corpora

- ROCLING Standard: 5 million words

- Penn Standard (Penn Chinese Treebank): 100K Words

Preview: What Corpus-Based Methods can do for Morphology, and Vice Versa.

- Statistical basis for claims about wordhood.
- Measures of morphological productivity.
- Practical applications in various natural-language processing tasks.
- On the other hand morphological knowledge can help guide or supplement statistical methods:

  cf. Hui et al's (2002) result that morphological models of unknown verbs have high precision (91.7%) but fairly low coverage (23.2%).

Overview of Part 2: Statistical Methods

- Properties of Word Frequency Distributions
- Measures of Morphological Productivity and a Case Study
  - Mandarin Root Compounds: A Case Study in Productivity
- Measures of Association
  - Probability Estimates
  - (Pointwise) Mutual Information
  - Frequency-Weighted Mutual Information
  - Pearson's χ-Square
  - Likelihood ratios
  - Extracting Non-Binary Collocations

Zipf's Law

(48) $f(w) \;\propto\; \frac{1}{r}$

1995 AP Newswire

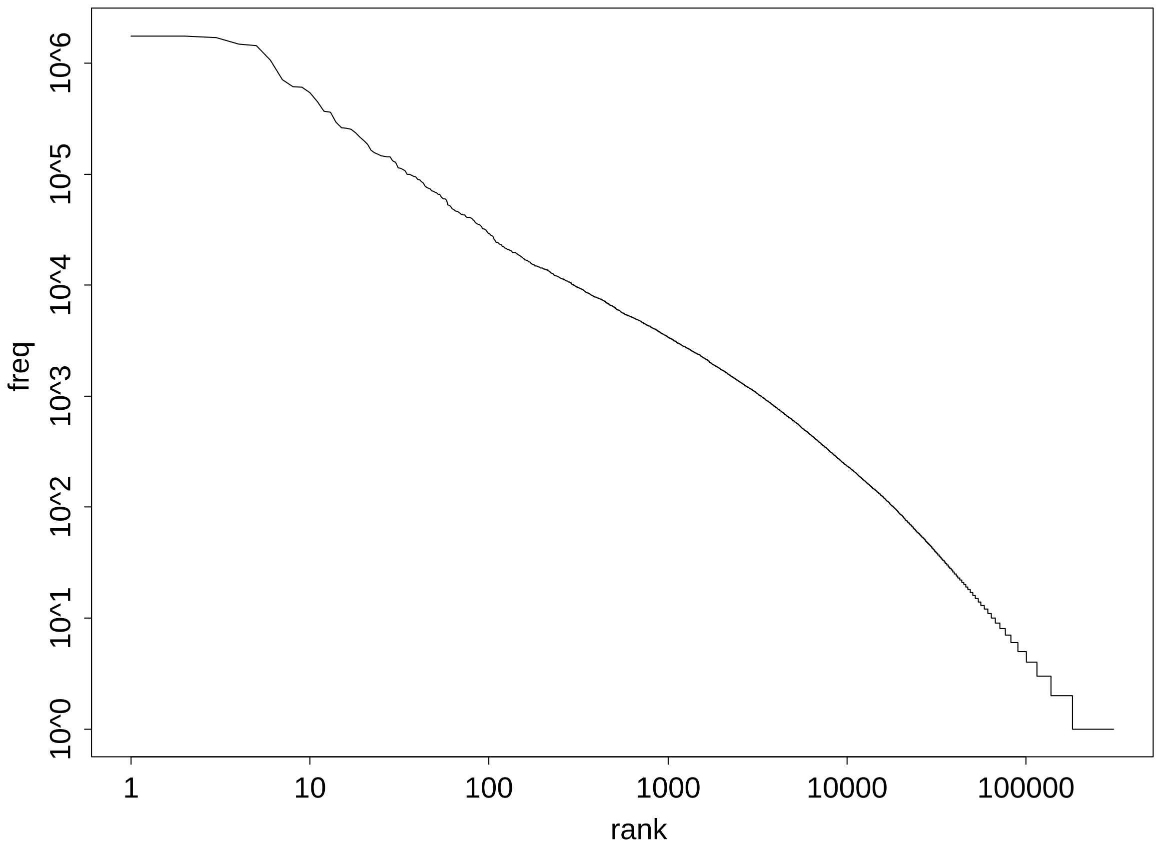


Figure 1: Rank (x) plotted against frequency (y) on a log-log scale for the 1995 Associated Press.

ROCLING Corpus (Characters)

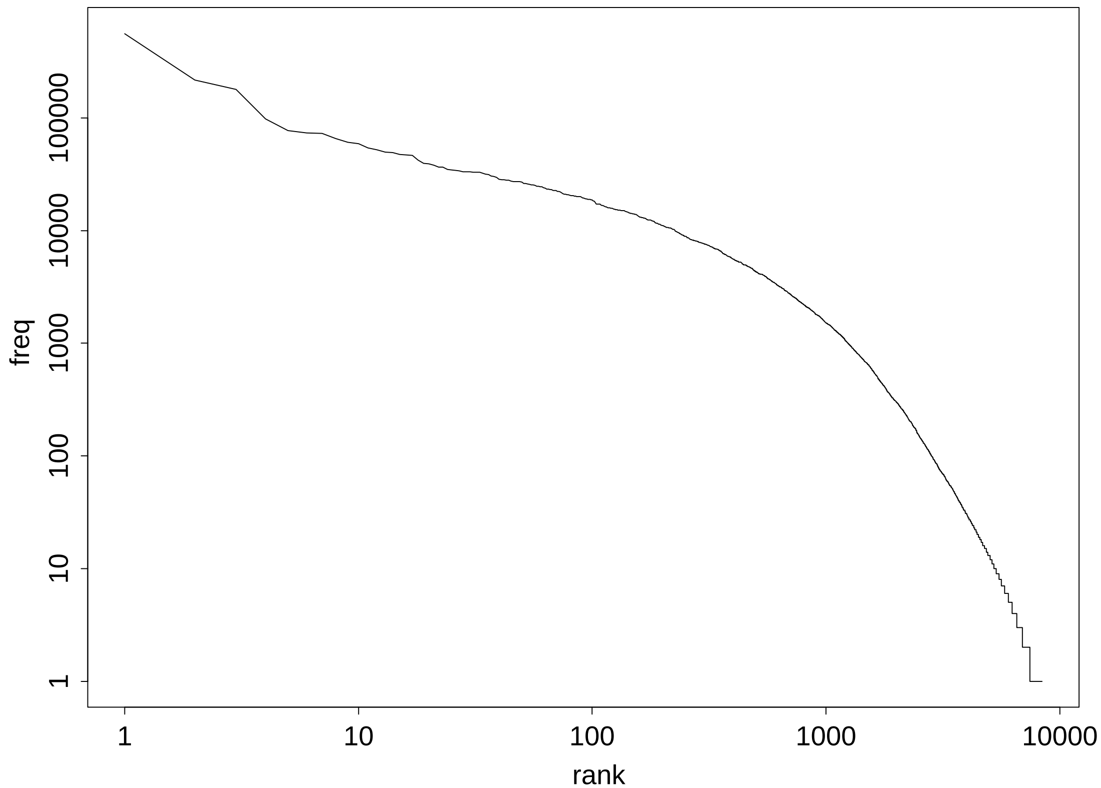


Figure 2: Rank (x) plotted against frequency (y) on a log-log scale over *characters* for the 10 Million character ROCLING corpus.

Important Properties of Word-Frequency Distributions

- Large Number of Rare Events:

  For the 1995 Associated Press corpus 40% of the word types occur just once. (In contrast among the 10 Million character ROCLING corpus, only 11% of the *characters* occur once.) For a smaller corpus, such as the Brown corpus (1 Million words), the amount will be closer to 50%.

  Corollary: words are not normally distributed. Statistical measures that depend on normality (e.g. mutual information) are suspect.

- Statistical parameters change with sample size:

  For increasing corpus N the size of the vocabulary $V(N)$ increases. So does the mean frequency $\frac{N}{V(N)}$.

Change of Parameters of Vocabulary with Corpus Size

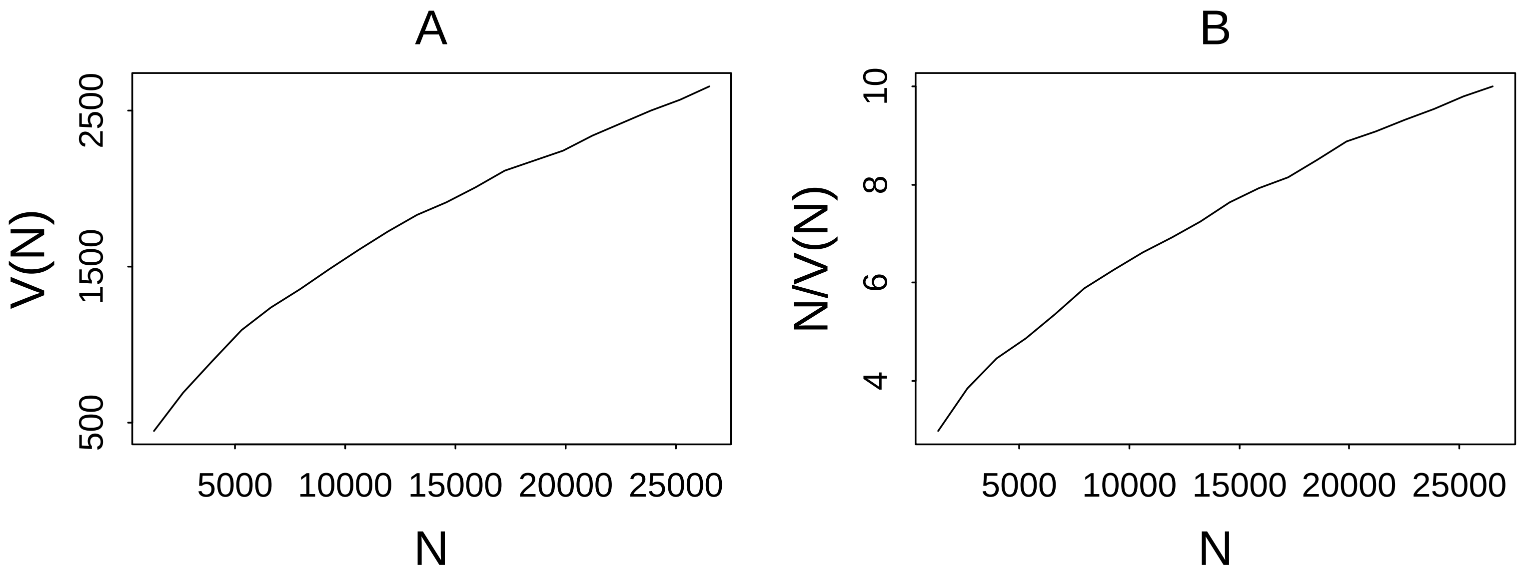


Figure 3: Vocabulary size $V(N)$ (Panel A) and mean word frequency $\frac{N}{V(N)}$ (Panel B) as a function of sample size N in *Alice in Wonderland*, measured at 20 equally spaced intervals. From (Baayen, 2001), figure and caption kindly provided by Harald Baayen.

Measures of Morphological Productivity

(49) Aronoff's (1976) proposal: $I = \frac{V}{S}$

(50) Baayen's (1989) proposal: $\mathcal{P} = \dfrac{n_1}{N}$.

n_1 is the number of words that have been seen just once – the *hapax legomena*.

Intuition behind Good-Turing Measure

Figure 4: A collection of balls where all ball types have been seen.

Intuition behind Good-Turing Measure

Figure 5: A collection of balls where many ball types have not been seen.

Some sample $\mathcal{P}$ scores from Dutch and English

(51)

Morpheme	Gloss	$\mathcal{P}$
-*ing*	'-ion, -ing'	0.038
-*heid*	'-ity'	0.114
noun compounds	—	0.225

(52)

Morpheme	$\mathcal{P}$
-*ness*	0.0044
-*ish*	0.0034
-*ation*	0.0006
-*ity*	0.0007
simplex nouns	0.0001

Some Measures from Mandarin

(53)

Morpheme	Gloss	$\mathcal{P}$
們 *-men*	noun plural	$\frac{247}{5948} = 0.04$
過 *-guò*	experiential	$\frac{330}{1666} = 0.20$

See the first lab exercise for this segment.

Mandarin Root Compounds (Sproat and Shih, 1996)

Some morphological theories, e.g. (Anderson, 1992; Dai, 1992), have a restrictive notion of what can be a compound: only those words formed from two or more other *words*.

(54) 哈密瓜 *hāmìguā* (Hami melon) 'cantaloupe'

(55) 香腸 *xiāngcháng* (fragrant intestine) 'sausage'

Mandarin Root Compounds

But bound roots seem to be active in Mandarin morphology:

(56) 螞蟻 *mǎyǐ* ‘ant’

工蟻 *gōngyǐ* ‘worker ant’

(57) 腦子 *nǎozi* ‘brain’

腦水腫 *nǎoshuǐzhǒng* (brain water swelling) ‘hydrocephaly’

(58) 蒼蠅 *cāngyíng* ‘fly’

地中海蠅 *dìzhōnghǎiyíng* ‘Mediterranean fly’

(59) 蘑菇 *mógū* ‘mushroom’

菇傘 *gūsǎn* (mushroom umbrella) ‘pileus’

Mandarin Root Compounds

These formations cannot be an instance of affixation

(60)

蟻 *yǐ*	蟻王	工蟻
	yǐwáng ‘queen ant’	*gōngyǐ* ‘worker ant’
腦 *nǎo*	腦水腫	後腦
	nǎoshuǐzhǒng ‘hydrocephaly’	*hòunǎo* ‘hindbrain’
蠅 *yíng*	蠅屍	地中海蠅
	yíngshī ‘fly corpse’	*dìzhōnghǎiyíng* ‘Mediterranean fly’
菇 *gū*	菇傘	金菇
	gūsǎn ‘pileus’	*jīngū* ‘golden mushroom’

Absent another category, these presumably must be compounds. (NB: Packard (2000) calls them *bound root words*)

Mandarin Root Compounds

224	香菇	104	蘑菇	102	洋菇
23	菇類	16	菇農	9	磨菇
9	草菇	8	菇寮	8	冬菇
8	塊菇	7	金針菇	6	出菇
5	菇價	4	菇傘	4	菇柄
4	鮑魚菇	4	慈菇	3	夏塊菇
3	裸蓋菇	3	茨菇	3	冬季菇
2	菇舍	2	菇木	2	鮮菇
2	發菇	2	採菇	2	夏季菇
2	食用菇	2	食菇	2	春季菇
2	出菇期	2	三菇菜膽	1	菇體
1	菇醇	1	菇腳	1	菇菌
1	菇湯	1	菇場	1	鮑菇
1	褐菇	1	種菇	1	黑菇
1	菌菇	1	產菇量	1	乾菇
1	金菇	1	早發菇	1	生菇
1	可食菇	1	包菇	1	山菇

Table 2: Distribution for the Mandarin nominal root $g\bar{u}$ 'mushroom' collected from a 40 million character corpus.

Mandarin Root Compounds

Root	Whole Word	Meaning	n_1	N	V	P_{max}	$\mathcal{P}$
石 *shí*	石頭 *shítou*	‘rock’	75	583	147	57	0.129
盒 *hé*	盒子 *hézi*	‘box’	82	1205	167	257	0.068
蟻 *yǐ*	螞蟻 *mǎyǐ*	‘ant’	21	322	35	173	0.065
蛙 *wā*	青蛙 *qīngwā*	‘frog’	22	407	49	199	0.054
龜 *guī*	烏龜 *wūguī*	‘turtle’	21	414	46	137	0.051
餃 *jiǎo*	餃子 *jiǎozi*	‘dumpling’	8	167	19	66	0.048
蠅 *yíng*	蒼蠅 *cāngyíng*	‘fly’	14	325	35	142	0.043
棉 *mián*	棉花 *miánhuā*	‘cotton’	52	1283	138	147	0.041
菇 *gū*	蘑菇 *mōgū*	‘mushroom’	19	598	49	224	0.032
木 *mù*	木頭 *mùtou*	‘wood’	265	8904	617	701	0.030
腦 *nǎo*	腦子 *nǎozi*	‘brain’	34	1077	75	791	0.032
駝 *tuó*	駱駝 *luòtuó*	‘camel’	6	210	16	104	0.029
腸 *cháng*	腸子 *chángzi*	‘intestine’	62	2268	148	373	0.027
蜂 *fēng*	蜜蜂 *mìfēng*	‘bee’	23	858	63	104	0.027
肚 *dù*	肚子 *dùzi*	‘belly’	36	1434	83	734	0.025

Table 3: Productivity measures of some Mandarin nominal roots, measured over a 40 million character corpus.

More on Good-Turing

See http://www.research.att.com/∼rws/exercises/gt-utf8.html for some $\mathcal{P}$ measures for other Chinese morphological processes.

Measures of Association

- Probability estimates
- Pointwise Mutual Information
- Frequency-Weighted Mutual Information
- Pearson's χ^2
- Dunning's likelihood ratios
- Non-binary collocations

Probability Estimates

Basic measure of probability is the *maximum likelihood estimate*: $f(t)/N$. This is quite reasonable for frequent words.

	AP92		AP93		AP94	
	f	p	f	p	f	p
the	1,659,949	0.039	1,451,984	0.039	1,311,237	0.039
United	30,883	0.00072	28,336	0.00076	25,456	0.00075
country	21,225	0.00050	18,304	0.00050	15,919	0.00047
night	12,422	0.00029	10,328	0.00028	10,566	0.00031
dog	1,048	2.44e-05	1,045	2.80e-05	992	2.91e-05

Table 4: Maximum likelihood estimates for five common words from three years of the Associated Press (1992–1994). N is 43,012,596, 37,386,960, and 34,041,151, respectively, for these three years.

Probability Estimates

Maximum likelihood estimate becomes less reliable for infrequent words, so typically some *smooth* of the frequency distribution is necessary.

(61) Good-Turing estimate: $r^* = (r+1)\frac{E(n_{r+1})}{E(n_r)}$

The estimates $E(n_r)$ and $E(n_{r+1})$ can be made in various ways, including using the empirical values:

(62) $r^* = (r+1)\frac{n_{r+1}}{n_r}$

For frequencies other than zero this will result in a reestimation downwards. The probabilities will then be reestimated as $\frac{r^*}{N}$, with the probability mass $\frac{n_1}{N}$ reserved for unseen items.

(Pointwise) Mutual Information

Mutual Information was originally proposed as an information-theoretic measure of channel capacity (Fano, 1961).

(63) $I(x;y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$

1995 AP Newswire Collocations

M.I.	f(1,2)	f(1)	f(2)	pair
18.3908	101	104	106	Picket Fences
18.3280	101	101	114	Highly Effective
18.2061	119	121	122	Ku Klux
18.1722	124	124	127	alma mater
18.1574	115	124	119	SPACE CENTER
18.1480	118	127	120	JUDGE LANCE
18.1275	127	131	127	Phnom Penh
18.1266	124	126	129	Velika Kladusa
18.0901	95	103	124	Ginny Terzano
18.0627	134	136	135	Notorious B.I.G
18.0580	116	119	134	Spiritual Laws
18.0486	123	134	127	Deepak Chopra
18.0271	80	105	107	Myriam Sochacki
17.9417	147	149	147	TEL AVIV
17.8421	96	117	131	Reba McEntire
17.7610	108	152	120	Dollar Spin
17.7551	117	124	160	SALT LAKE

Table 5: Sample of highly associated adjacent word pairs from the 37 million words of the 1995 Associated Press. Shown are, from left to right: the mutual information (M.I.); the frequency of the pair f(1,2); the frequency of the first word f(1) and the frequency of the second word f(2). Note that f(1) and f(2) are both greater than 100 for this sample.

1995 AP Newswire Non-Collocations

M.I.	f(1,2)	f(1)	f(2)	pair
-0.000104371	2	44293	1694	But 32
-0.000104371	12	44293	10164	But public
-0.000108717	1	5938	6318	big reports
-0.000122066	7	540363	486	in heroin
-0.000122066	7	486	540363	determination in
-0.000123791	1	20716	1811	says Hall
-0.000135193	2	567	132335	Building from
-0.000135827	1	6639	5651	water given
-0.000135827	1	5651	6639	given water
-0.000137212	1	8365	4485	programs enough
-0.000137212	1	2275	16491	village children
-0.000144883	2	5283	14203	study most
-0.000151325	1	14452	2596	her includes
-0.000163745	1	645	58167	Delaware this

Table 6: Sample of poorly associated adjacent word pairs from the 1995 Associated Press.

Problems with Mutual Information

- It is unreliable for small counts. (But this is really a problem with the MLE)

- The second, and more serious problem is that mutual information relates to estimated probability in a counterintuitive way:

$$I(w_1; w_2) = \log_2\left(\frac{\frac{100}{10{,}000{,}000}}{\frac{100}{10{,}000{,}000}\frac{100}{10{,}000{,}000}}\right) = \log_2(100,000) = 16.6$$

$$I(w_1; w_2) = \log_2\left(\frac{\frac{1{,}000}{10{,}000{,}000}}{\frac{1{,}000}{10{,}000{,}000}\frac{1{,}000}{10{,}000{,}000}}\right) = \log_2(10,000) = 13.3$$

Frequency-Weighted Mutual Information

(64) $$I_{fw}(x;y) = f(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)}$$

1995 AP Newswire Collocations

F.W.M.I.	f(1,2)	f(1)	f(2)	pair
469631	174742	708948	1435262	of the
341288	129132	540363	1435262	in the
184196	16797	16816	18733	Associated Press
182881	17287	24135	17564	United States
181021	66059	258389	1435262	to the
144801	31575	447529	110206	to be
140778	51336	200524	1435262	on the
136748	12956	24177	13364	New York
135747	19968	112171	60003	have been
135610	12452	17858	13778	All Rights
134177	28748	144059	294607	he said
133324	11684	13778	11684	Rights Reserved
120476	17592	95448	60003	has been
118810	50201	254405	1435262	for the
114737	17820	69930	110206	will be
109856	15576	261695	16816	The Associated
107843	36922	127433	1435262	at the
97455	9561	10928	28038	White House
96982	15799	76338	110206	would be

Table 7: Sample of highly associated adjacent word pairs from the 1995 Associated Press. Shown are, from left to right: the frequency weighted mutual information (F.W.M.I.); the frequency of the pair f(1,2); the frequency of the first word f(1) and the frequency of the second word f(2). Again, f(1) and f(2) are both greater than 100 for this sample.

Problems with Frequency-Weighted Mutual Information

Main problem is that it tends to over-reward frequency.

Pearson's χ-Square

- χ-square provides a confidence measure for rejecting an assumption of independence between events.
- χ-square applies to tables:

	$w_1 = new$	$w_1 \neq new$
$w_1 = companies$	8 (*new companies*)	4,667 (e.g. *old companies*)
$w_1 \neq companies$	15,820 (*new machines*)	14,287,181 (e.g. *old machines*)

Table 8: A 2×2 table showing the distribution of bigrams in a corpus (from (Manning and Schütze, 1999, Table 5.8, page 169). There were 8 instances of *new companies*, 4,667 instances of *X companies*, where *X* is different from *new*, 15,820 instances of *new Y*, where *Y* is different from *companies*, and 14,287,181 instances of *XY*, where *X* and *Y* are different, respectively, from *new* and *companies*.

Pearson's χ-Square

General statement:

(65) $$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Special case for 2×2 table:

(66) $$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

1995 AP Newswire Collocations

χ^2	f(1,2)	f(1)	f(2)	pair
999319	600	6262	2156	attorney general
996401	47	280	297	Connie Mack
993054	192	552	2522	20th century
991404	36	299	164	OR MUSIC
991243	306	818	4330	atomic bomb
990132	145	1709	466	loan guarantees
986765	18	113	109	Geographic Explorer
985647	375	2599	2058	Catholic Church
985038	67	1540	111	racial slur
984217	84	1379	195	highly publicized
983361	147	284	2902	plead guilty
982709	23	159	127	Justices Antonin
975478	289	7415	433	Sen Arlen
972147	22	116	161	Celtic Journey
970830	322	2808	1426	illegal immigrants
968693	126	2984	206	flight attendant
968378	261	1571	1679	pleaded innocent
968056	20	124	125	Joey Buttafuoco
967695	37	147	361	silicone implants

Table 9: Sample of highly associated adjacent word pairs from the 1995 Associated Press, using chi-square. Shown are, from left to right: the chi square value; the frequency of the pair f(1,2); the frequency of the first word f(1) and the frequency of the second word f(2). Again, f(1) and f(2) are both greater than 100 for this sample.

Problems with χ-Square

- χ-square still assumes normality, which can be violated for small counts.
- χ-square is a symmetric measure: it does not distinguish between events that are much more likely than chance from events that are much *less* likely than chance.

  tape-recorded conversations — a plausible collocation — you find that *tape-recorded* occurs 100 times, *conversations* 639 times, and the collocation 7 times, with a high χ-square value of 28,752.8.

  Similarly for *sickening reminder* the breakdown is, 17 for *sickening*, 307 for *reminder* and 2 for the pair, with a χ-square value of 28,747.7.

  However a very similar χ-square value is obtained for *the of*, where *the* occurs 1,435,262 times, *of* 708,948 times, the pair exactly once (due to a typo in the data), yielding a χ-square value of 28,744.5.

Dunning's (1993) Likelihood Ratios

- Hypothesis 1: $p(w_2|w_1) = p = p(w_2|\neg w_1)$
- Hypothesis 2: $p(w_2|w_1) = p_1 \neq p_2 = p(w_2|\neg w_1)$
- $p = \frac{c_2}{N}$, $p_1 = \frac{c_{12}}{c_1}$, $p_2 = \frac{c_2 - c_{12}}{N - c_1}$
- Assuming a binomial distribution: $b(k; n, p_x) = \binom{n}{k} p_x^k (1 - p_x)^{(n-k)}$

$L(H_1) = b(c_{12}; c_1, p) b(c_2 - c_{12}; N - c_1, p)$

$L(H_2) = b(c_{12}; c_1, p_1) b(c_2 - c_{12}; N - c_1, p_2)$

$$\begin{aligned}
\log \lambda &= \log \frac{L(H_1)}{L(H_2)} \\
&= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) \\
&\quad - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2)
\end{aligned}$$

(where $L(k, n, x) = x^k (1 - x)^{n-k}$)

1995 AP Newswire Collocations

$-2\log\lambda$	f(1,2)	f(1)	f(2)	pair
587215.02	174742	708948	1435262	of the
417624.29	129132	540363	1435262	in the
403795.56	16797	16816	18733	Associated Press
387423.99	17287	24135	17564	United States
281913.17	12956	24177	13364	New York
279548.21	12452	17858	13778	All Rights
230483.10	19968	112171	60003	have been
223206.08	66059	258389	1435262	to the
218798.60	31575	447529	110206	to be
211761.70	15576	261695	16816	The Associated
203728.79	17592	95448	60003	has been
200819.84	28748	144059	294607	he said
192063.84	9561	10928	28038	White House
190007.67	17820	69930	110206	will be
173072.45	51336	200524	1435262	on the
161027.74	194	1435262	1435262	the the
157326.20	15799	76338	110206	would be
143998.60	9530	24496	40710	more than
143592.45	10387	19586	89982	did not
142904.19	18933	1435262	24135	the United

Table 10: Sample of highly associated adjacent word pairs from the 1995 Associated Press, using log likelihood ratios. Shown are, from left to right: the value of $-2\log\lambda$ (which is asymptotically χ-square distributed; the frequency of the pair f(1,2); the frequency of the first word f(1) and the frequency of the second word f(2). Again, f(1) and f(2) are both greater than 100 for this sample.

Problems with Likelihood Ratios

Shares with χ-square the problem that it is symmetric

$-2\log\lambda$	f(1,2)	f(1)	f(2)	pair
4988.08	340	14203	2247	most powerful
1959.01	420	607952	2247	a powerful
1336.07	131	24496	2247	more powerful
1093.76	89	7967	2247	most powerful
532.11	57	14183	2247	very powerful
527.49	36	2247	1410	powerful earthquake
438.62	285	1435262	2247	the powerful
363.11	31	2247	3345	powerful lower
309.31	22	1053	2247	politically powerful
293.39	20	2247	776	powerful storms
249.88	30	10575	2247	so powerful
237.09	1	2247	1435262	powerful the
225.00	31	15989	2247	as powerful

Table 11: Possible collocations of *powerful*, along with their log likelihood ratios, from the 1995 Associated Press.

Association Measures for Finding 2-Character Words in Chinese

See http://www.research.att.com/∼rws/exercises/assoc-utf8.html for examples from the ROCLING corpus (10 million characters): 500 most highly associated collocations according to each measure, where each collocation occurs at least five times.

Mutual Information

20.8914	5	5	5	枇 杷
20.6283	6	6	6	踉 蹌
20.6283	6	6	6	酩 酊
20.6283	6	6	6	蚯 蚓
20.6283	6	6	6	氤 氳
20.6283	5	5	6	窈 窕
20.6283	5	5	6	砥 礪
20.406	7	7	7	萵 苣
20.406	6	7	6	蜥 蜴
20.2133	7	7	8	襤 褸
20.2133	7	7	8	袈 裟
20.2133	6	6	8	桎 梏
20.2133	6	6	8	蹉 跎
20.2133	5	8	5	檸 檬
20.0434	9	9	9	邂 逅

Weighted Mutual Information

125677	15536	27178	20411	表 示
78771.1	10234	19371	24761	昨 天
75138.6	10817	25679	33218	警 方
74098.2	9464	33208	12185	公 司
67031	8402	31653	10239	台 灣
60718	6396	10708	8061	問 題
60269.5	8089	13962	32206	目 前
59121.7	8527	16109	42116	指 出
54095.2	7280	27179	15099	政 府
53548.1	7673	26206	22576	因 此
53388.6	8299	24614	37947	由 於
48128.2	5605	9173	15457	單 位
47026.8	5862	13035	16822	萬 元
45690.5	4500	5473	7021	希 望
42142.7	6228	10669	52139	認 為

Chi Square

998246	15	137	16	狼 狽
997035	40	289	54	宇 宙
995487	171	420	680	撲 滅
990862	774	2470	2378	紀 錄
990639	29	32	258	翡 翠
986131	572	5411	596	流 氓
984730	120	1185	120	罰 鍰
984573	7	22	22	嚥 嚥
984430	33	326	33	疲 憊
982982	13	88	19	茄 苳
982425	2129	4250	10528	服 務
981128	5387	34912	8186	民 眾
972268	520	664	4071	丈 夫
971444	32	51	201	懊 惱
968743	45	66	308	嘔 吐

Dunning's Likelihood Ratios

247542.11	15536	27178	20411	表 示
144931.99	10234	19371	24761	昨 天
140756.48	9464	33208	12185	公 司
132653.49	10817	25679	33218	警 方
128505.03	8402	31653	10239	台 灣
120992.83	6396	10708	8061	問 題
109109.52	8089	13962	32206	目 前
104500.02	8527	16109	42116	指 出
96486.77	7280	27179	15099	政 府
92924.23	7673	26206	22576	因 此
92063.38	4500	5473	7021	希 望
90385.71	8299	24614	37947	由 於
89910.47	5605	9173	15457	單 位
85074.76	5862	13035	16822	萬 元
74228.04	6228	10669	52139	認 為

And the Winner Is . . .

Measure	Error Rate
Mutual Information	92/500
Chi Square	69/500
Weighted Mutual Information	35/500
Likelihood Ratios	34/500

Extracting Non-Binary Collocations

- Can extend approaches for the binary case. For example, mutual information can be defined for triples (Chang and Su, 1997):

  $$I(x, y, z) = \log \frac{P(x, y, z)}{P_I(x, y, z)}$$

  where

  $$P_I(x, y, z) = P(x)P(y)P(z) + P(x)P(y, z) + P(x, y)P(z)$$

- Or, more commonly, different approaches have been used.

Smadja's (1993) *Xtract*

- Compute a concordance of all instances of a given word w in a corpus.

- For each w_i found in the context of w, a profile is computed of how often each w_i occurs in each position in a window ranging from five to the left to five to the right of w.

Smadja's *Xtract*

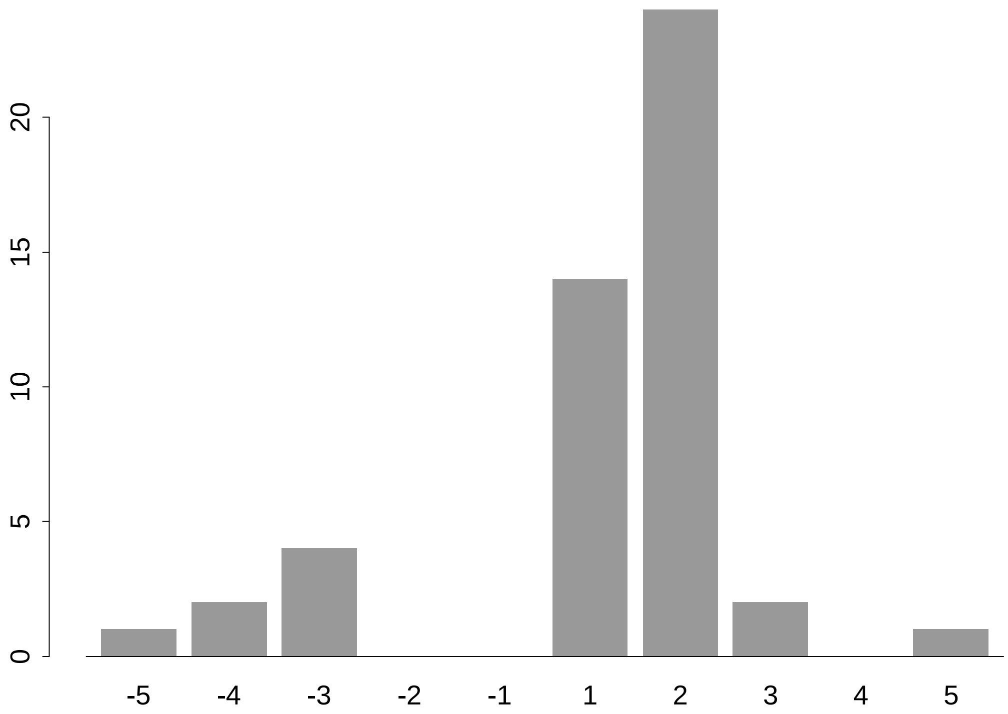


Figure 6: Occurrences of *forecast* in a window around *weather*, computed on the the 1995 Associated Press corpus.

Smadja's *Xtract*

- Remove "uninteresting" words based on property of profile. e.g.: the profile must show at least one peak. A flat distribution suggests a pair of words that is only semantically associated, such as *doctor* and *nurse*.

- Once interesting two-word collocates are found, compute further concordances for each pair of words found in the first phase, for each relative position. Recurring sequences of words are kept as potential collocates. For example, a collocation *blue. . . stocks* discovered during the first phase would be replaced in the second phase by *blue* **chip** *stocks*, since *chip* would occur frequently in this context.

- Fung and Wu (1994) have applied the Xtract system to Chinese.

Overview of Part 3: Applications

- 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.

Segmentation Methods: General issues

馬魁乾

Segmentation Methods: General issues

馬魁乾 *Mǎ Kuíqián* (personal name)

Segmentation Methods: General issues

馬肉乾

Segmentation Methods: General issues

馬肉乾 *mǎròu gān* ‘horsemeat jerky’

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 and 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 and Shih, 1990)

	我	弟	弟	現	在	要	坐	火	車	回	家
MI	0	10.4	0	4.2	-2.8	0	70.3		2.1	4.7	
1	我	$\boxed{弟}$	$\boxed{弟}$	現	在	要	坐	火	車	回	家
2	我	[弟	弟]	現	在	要	坐	$\boxed{火}$	$\boxed{車}$	回	家
3	我	[弟	弟]	現	在	要	坐	[火	車]	$\boxed{回}$	$\boxed{家}$
4	我	[弟	弟]	$\boxed{現}$	$\boxed{在}$	要	坐	[火	車]	[回	家]
5	我	[弟	弟]	[現	在]	要	坐	[火	車]	[回	家]

Table 12: 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, 1990) approach, using various association measures besides mutual information. They propose a t score for determining whether to group *xyz* as *xy z* or *x yz*:

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

- Ge, Pratt and Smyth (1999) propose a method based on expectation maximization.

  (EM has also been used in other work: e.g. (Peng and Schuurmans, 2001) who estimate probabilities for words in a lexicon using an EM-based algorithm, and use mutual information to weed out proposed words whose components are not strongly associated.)

Statistically-Aided Approaches: WFST-Based Approach (Sproat et al., 1996)

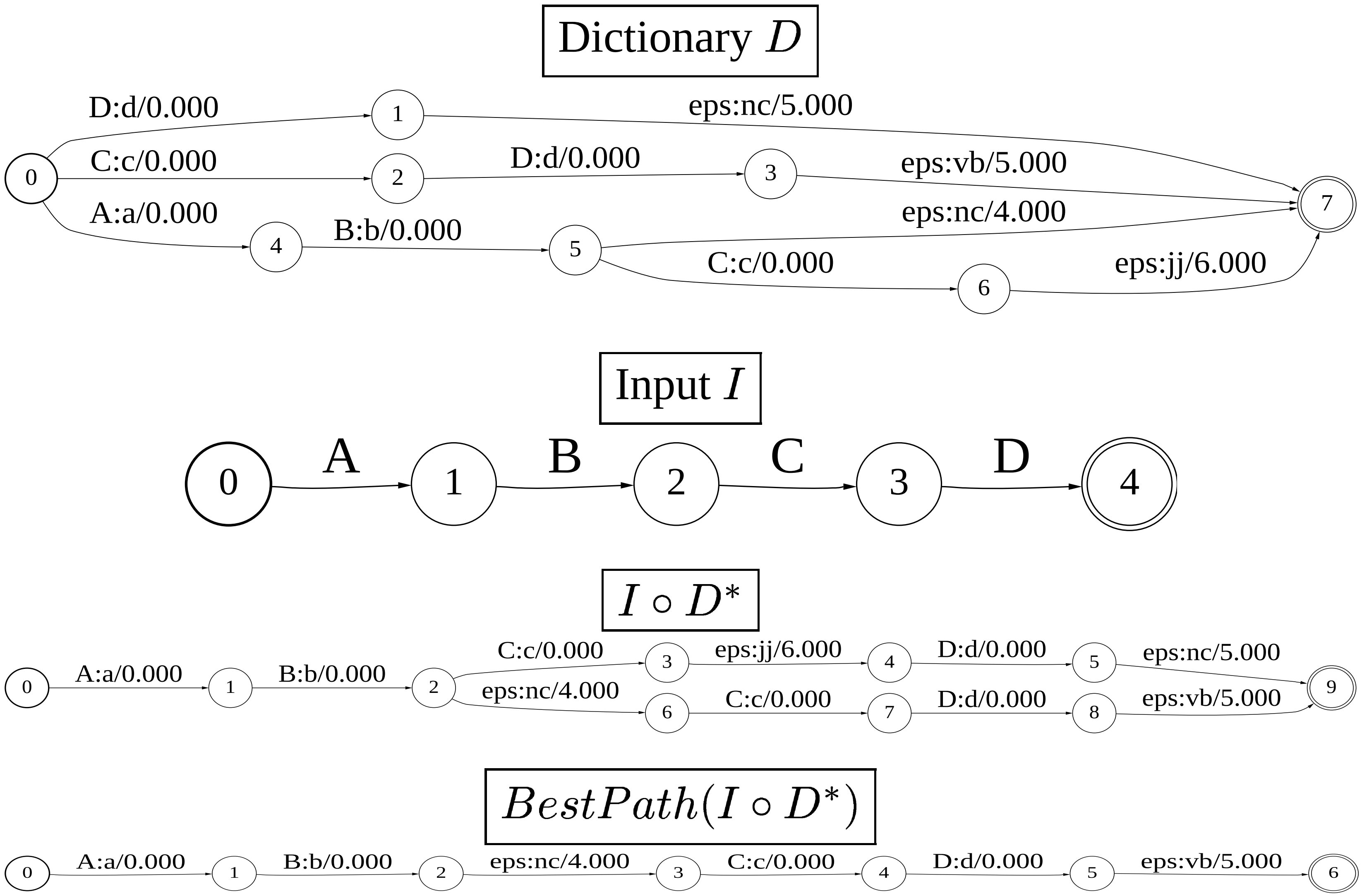
- Represent dictionary D as Weighted Finite State Transducer, mapping **from** $ChinChar \cup \epsilon$ **to** $PinyinSyllable \cup POS$. (Note that since the primary application of the Sproat et al system was to text-to-speech, the system was not merely a segmenter, but also a transducer mapping from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



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:

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

Statistically-Aided Approaches: WFST-Based Approach

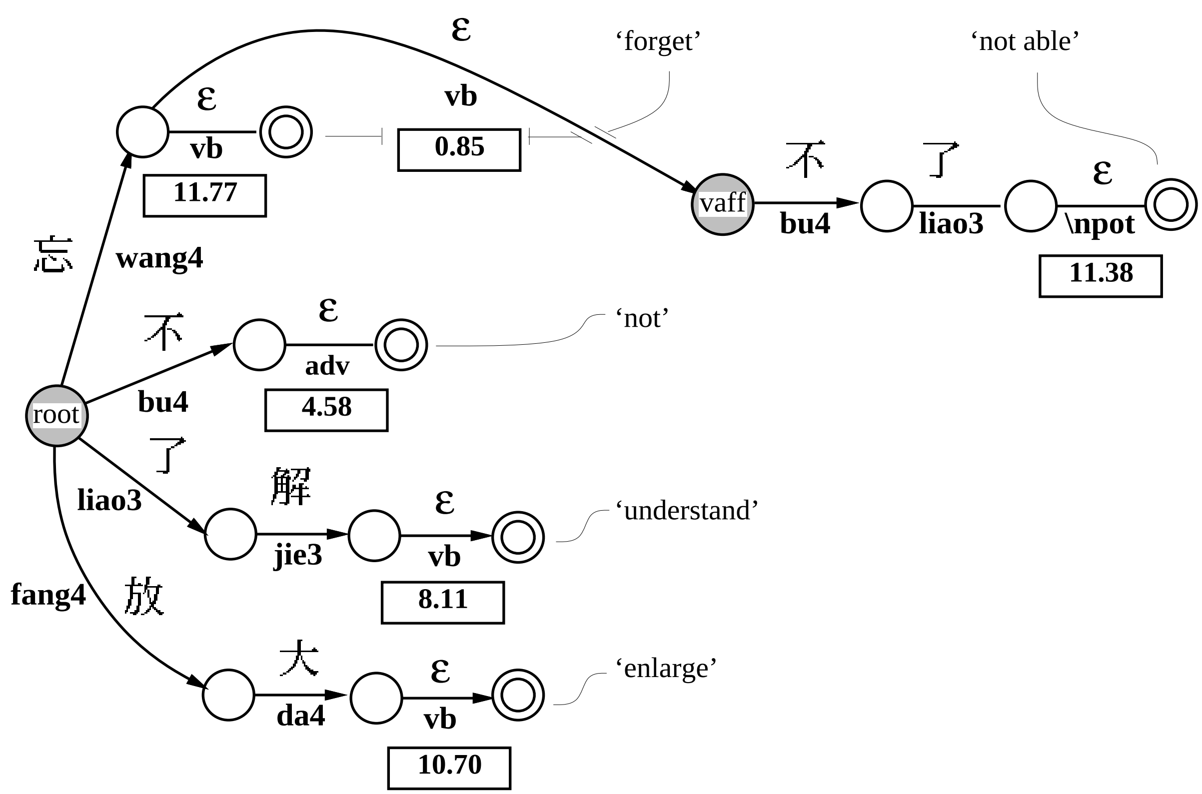
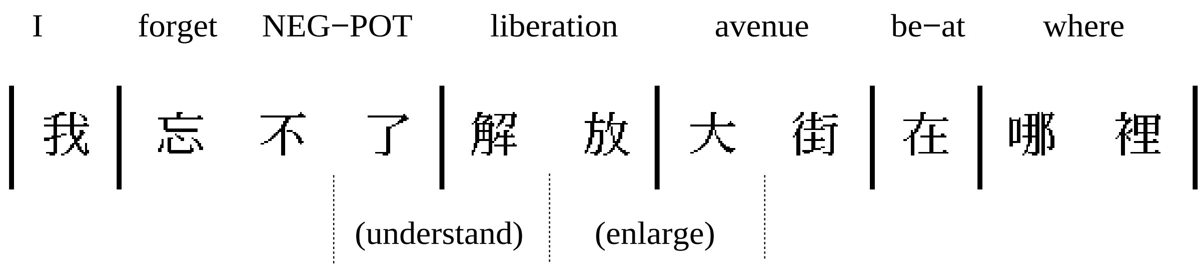
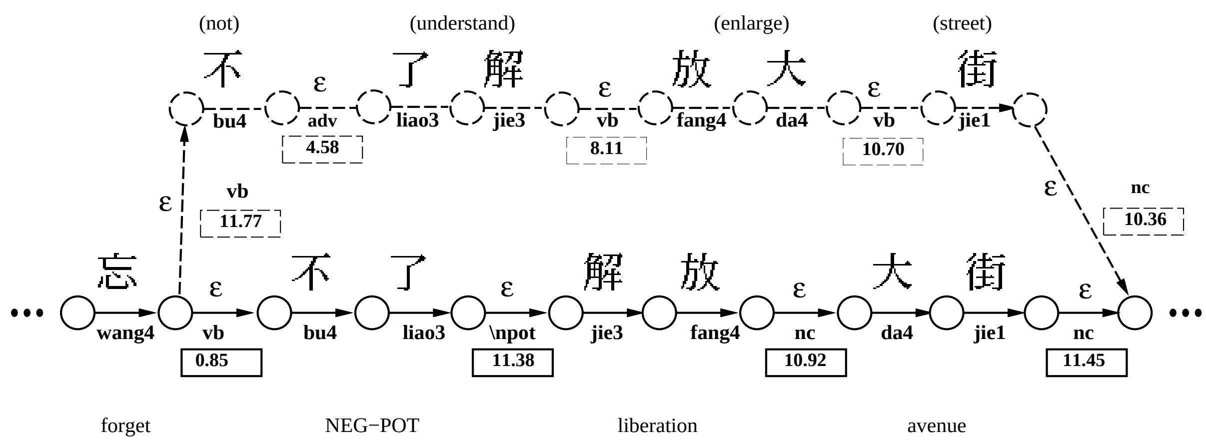


Figure 7: A fragment of the dictionary represented as a WFST.

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):

(69)

$$
\begin{aligned}
p(name|FG_1G_2) &= p(|fam| = 1, |giv| = 2) \times p(Fam = F) \times \\
&\quad p(Giv_1 = G_1) \times p(Giv_2 = G_2) \times p(name)
\end{aligned}
$$

- 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:

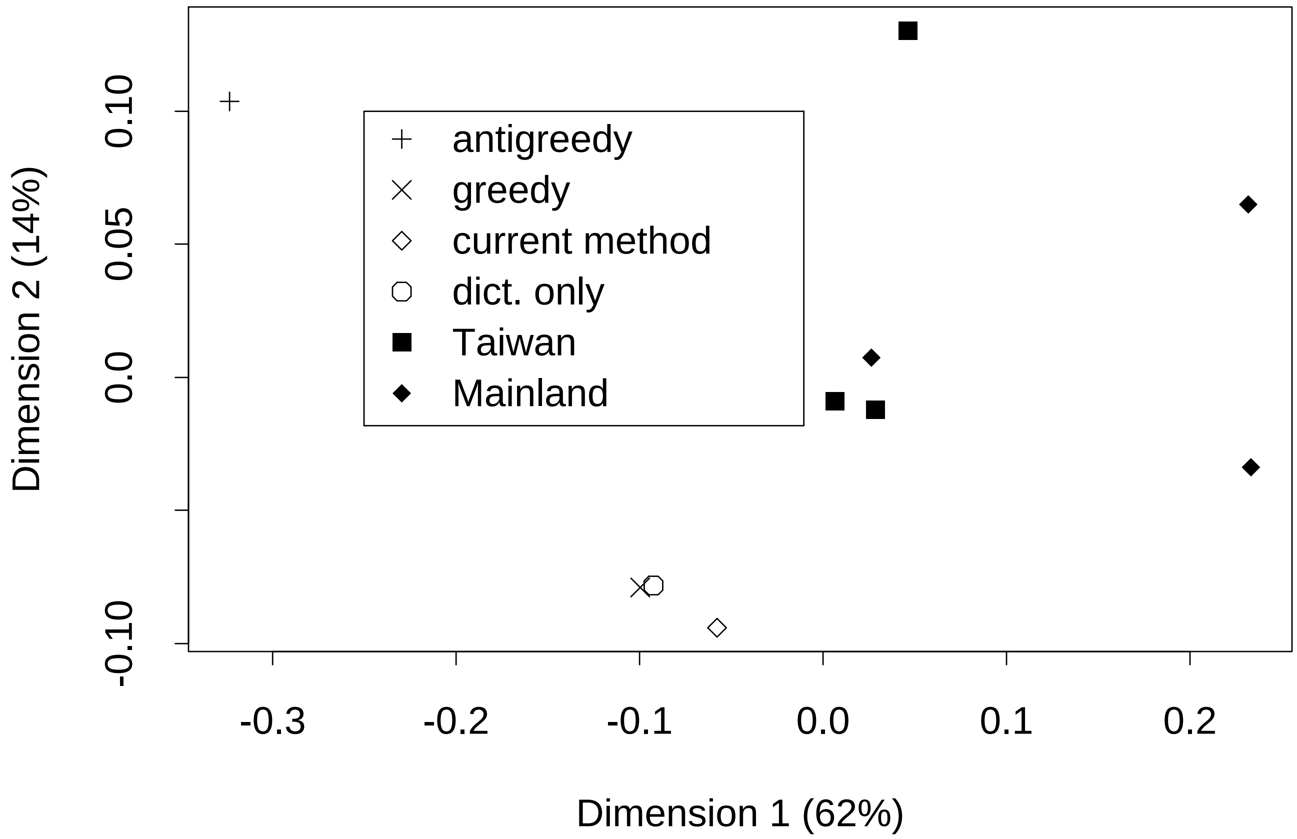
(70) $$Similarity = \frac{Precision + Recall}{2}$$

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 13: 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: $F = \dfrac{(1+\beta)PR}{\beta P + R}$

Gan et al's (1996) "Statistically Emergent Approach"

A simulated annealing approach

- Large pot of associations between words/characters linked into *conceptual network*. Relations include: character 'affinity', affix relation, classifier relation, demonstrative relation, agent relation, ... These have associated *quanties*, which is the number of codelets posted to the *coderack* (dependent on the length of the sentence).

Gan et al's (1996) "Statistically Emergent Approach"

- Each relation is associated with a *codelet* which builds a relation of the specified type. Probability of selecting a codelet C_j at *temperature* t is given as:

  $$P_t(C_j) = U_{j,t}Q_j / \sum_{i=1}^{n}(U_{i,t}Q_i)$$

  where $U_{j,t}$ is the urgency of the jth codelet at temperature t, and Q_j is the quantity of the jth codelet, and n is the number of codelet types.

- Strengths of competing structures are updated by a formula that minimizes differences in strengths at high temperatures and maximizes them at low temperatures:

  $$S_t = S^(120 - t)/40$$

Yao and Lua (1998) 'Splitting-Merging' model

1. Split C into $C_{1,k}$ and $C_{k+1,n}$ using the minimum mutual information between adjacent characters

2. Stop on $C_{1,k}$ or $C_{k+1,n}$ if can be found in the dictionary.

3. Else do merging operation on $C_{1,k}$ to form $C_{1,j-1}C_{j,j+1}C_{j+2,k}$ where $C_{j,j+1}$ has the highest mutual information; similarly for $C_{k+1,n}$

4. If $C_{j,j+1}$ is in the dictionary repeat process on $C_{1,j-1}$ and $C_{j,j+1}$; else repeat process on $C_{1,k}$

Error rate on "outside" dataset with 8% OOV is 10.6% by sentence.

Yao and Lua: Unknown Word Detection

- Use a probabilistic morphological model to detect new words, training probability of a character being independent, word-final, word-initial or word medial. E.g.:

  P(independent|是) = 0.7893
  P(head|是) = 0.0275
  P(tail|是) = 0.1789
  P(middle|是) = 0.0043

- Method includes heuristics to group reduplicated characters

Error rate on “outside” dataset drops to 1.2% by sentence.

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.

More on Unknown Words: (Chang and Su, 1997)

- Initial segmentation.
  - Start with dictionary augmented with all n-grams in a corpus up to a given length that occur some specified minimum number of times.
  - Segment corpus using augmented dictionary with n-gram probabilities from the corpus.
- Feed found words into "Likelihood Ratio Ranking Module" to remove unlikely words:
  - Compare estimate of the likelihood of the n-gram being in the word class, versus the likelihood of its being in the non-word class:

$$\lambda = \frac{f(x|W)p(W)}{f(x|\bar{W})p(\bar{W})}$$

(x = feature vector associated with a given n-gram; $f(x|W)$ and

$f(x|\bar{W})$ are the density functions for x being in the word class, versus the non-word class; $p(W)$ and $p(\bar{W})$ are of words and non-words)

  - Features used are n-way mutual information between characters, and the average entropy ($H = -\Sigma_{i=1}^{n} p(x_i) \log(p(x_i))$)
  - The density functions are modeled as mixtures of multivariate Gaussian distributions.

- Repeat process with remaining words.

Yields approximately 80% recall and 70% precision for new two-character words.

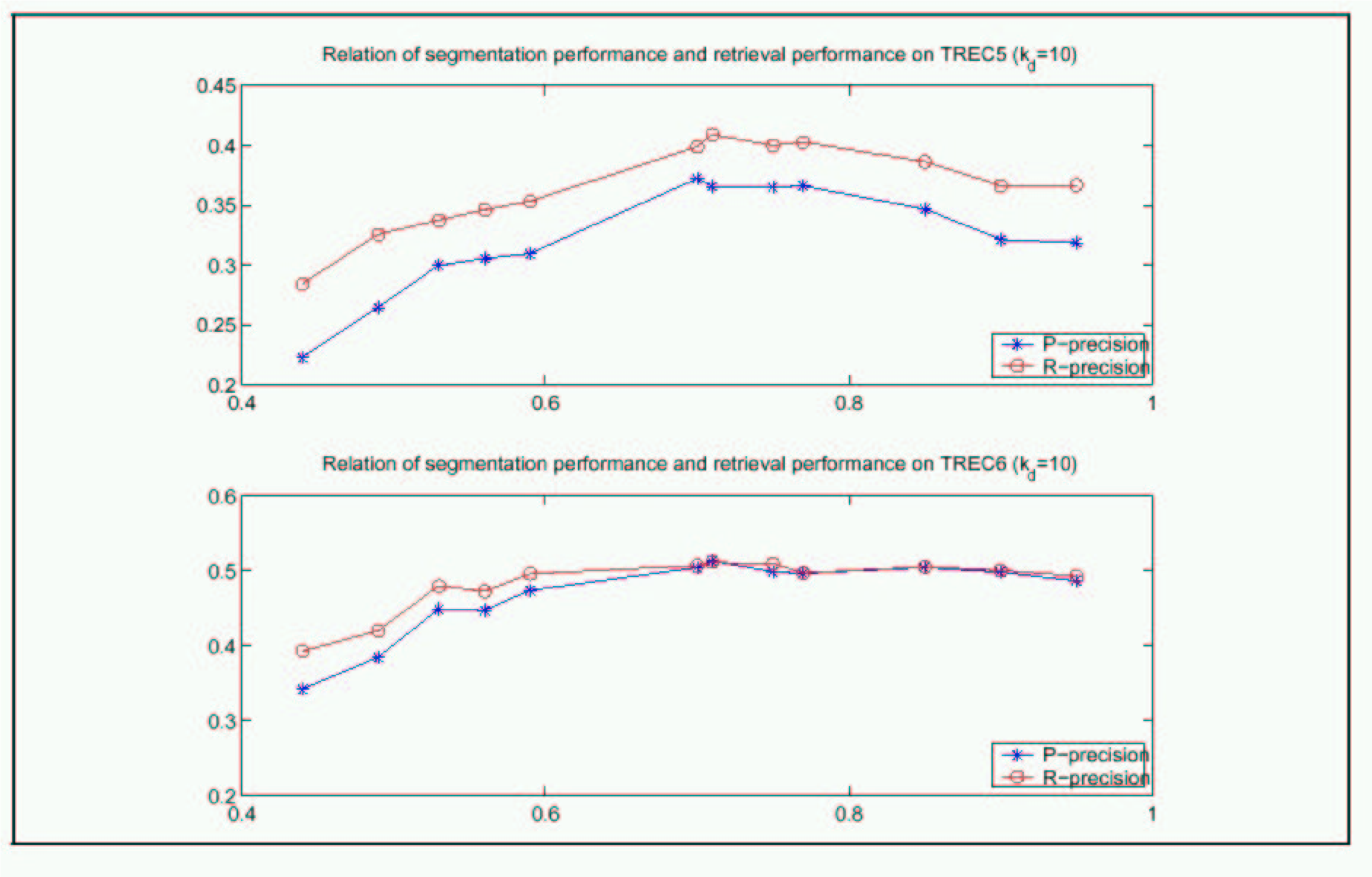
Segmentation and Information Retrieval

What are the effects on IR of no segmentation, bad segmentation, better segmentation . . . ?

- Peng et al (2002) show that, broadly speaking, better segmentation correlates with better precision and recall on TREC tasks.

- Chen (2001) argues more broadly that identifying noun phrases is useful in IR.

Tradeoff between Segmentation Performance and IR Performance (from (Peng et al., 2002))



(Peng et al., 2002) speculate that a weaker segmenter may tend to break long words, which can be better for IR

Linguistic Studies

- Properties of morphological constructions

- Suoxie

- Historical morphological change

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 14: Monomorphemic words and their mutual information.

Properties of Morphological Constructions

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

Table 15: 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 16: Versatile polymorphemic words and their mutual information.

Properties of Morphological Constructions

看魚	kàn yú	'look at fish'	-1.27
一第	yī dì	'one PREFIX'	-4.40
性人	xìng rén	'-ity human'	-1.44

Table 17: Phrases and non-words.

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

Table 18: Some borrowed words, and their mutual information.

See http://www.research.att.com/∼rws/exercises/huangmi-utf8.html

Analysis of Suoxie: (Huang, Ahrens, and Chen, 1994; Huang, Hong, and Chen, 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, Ahrens, and Chen, 1994; Huang, Hong, and Chen, 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.

Results 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

See http://www.research.att.com/~rws/exercises/suoxie-utf8.html

Historical Morphological Change

- Feng (1997) claims that there was a *dramatic* increase in the number of disyllabic compounds between Chinese of the pre-Han period and Han Chinese. He compares the text of 孟子 Mengzi (Mencius, c. 372–289 BC) and the text of his main Han commentator 趙岐 Zhao Qi (c. 107–201 AD), noting many cases where Zhao Qi uses two character words to gloss single character words in Mengzi: e.g. 過 *guò* 'mistake' glossed as 謬誤 *miùwù* 'mistake'. Feng argues that this commonly observed increase in the disyllabicity of Chinese words was for prosodic reasons, related to syllable-structure simplification and the minimal prosodic word.
- Is this a claim about types or tokens?
  - Feng talks about *tokens* in the corpus of Mencius and Zhao Qi
  - But the theoretical claim seems relate to the structure of the vocabulary — hence *types*

  We'll assume it's types.

Was there an Increase in Disyllabicity?

See http://www.research.att.com/∼rws/exercises/historical-utf8.html

- Compare disyllabic terms in texts from the pre-Han, Han and later (Jin/Song/Ming) periods to see if there is a general increase of disyllabic forms over time. (See the website for the list of the texts from the Academia Sinica historical corpora.)

- How to estimate the number of disyllabic types:
  - Use Good-Turing estimate: but you'd need a large sample to have a robust estimate of $\mathcal{P}$.
  - Use an association measure to generate lists of highly associated forms and compare the "yield" across different periods.

  NB: either of these approaches need to be done on equivalent-sized samples.

Most Associated Terms Using Likelihood Ratios: pre-Han Period

32287.75	2853	6668	6940	天 下
17801.29	870	904	1422	岐 伯
16298.28	931	1641	1444	諸 侯
15275.01	2238	27105	5299	不 可
†11096.48	983	1175	16421	對 曰
10155.46	647	720	4406	桓 公
9840.58	613	1156	1728	黃 帝
†9817.85	1849	10888	16421	子 曰
9551.15	1485	27105	3885	不 能
9535.83	724	784	10888	晏 子
†9202.19	982	1728	16421	帝 曰
†9130.62	2900	22889	27105	而 不
†8903.81	2101	46720	6762	之 所
8691.55	509	2433	604	百 姓

Most Associated Terms Using Likelihood Ratios: Han Period

79890.48	5090	6666	5949	師 古
†58603.83	5078	5949	18719	古 曰
21301.79	1537	4148	2804	將 軍
18157.66	1592	4456	5275	天 下
16579.27	738	768	917	匈 奴
13692.44	1243	6822	2987	大 夫
13388.57	1097	2587	4753	諸 侯
10802.83	661	975	2307	殿 本
10611.92	816	3953	1773	太 后
9936.40	582	740	2294	單 于
8441.76	549	821	2968	丞 相
†8296.61	862	1821	12363	頁 一
7808.62	503	541	5275	陛 下
7352.19	362	929	432	景 祐

Most Associated Terms Using Likelihood Ratios: Jin-Song-Ming Period

19859.07	1577	4567	3651	將 軍
12242.32	839	1645	2778	尚 書
11465.54	827	3720	1524	太 守
10329.42	542	624	1520	刺 史
8073.00	1063	12537	3977	一 個
†7685.49	456	2109	624	州 刺
7576.58	422	835	943	散 騎
6881.40	721	3516	4831	如 此
6782.76	714	3191	5365	天 下
5969.37	334	427	1667	員 外
5723.30	359	366	5365	陛 下
5717.88	1036	10373	9137	以 為
5497.58	395	1724	1305	司 馬

Yield Across Periods

Period	Number of Words	% Yield	% of Corpus
Pre-Han	318/500	64%	5%
Han	401/500	80%	6%
Jin-Song-Ming	430/500	86%	5%

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 ...

What we've Covered

1. Introduction to Chinese Morphology:
   (a) What are words in Chinese?
   (b) What are some of the morphological operations of Chinese?
   (c) Proposed segmentation standards.
2. Some statistical methods:
   (a) Word frequency distributions.
   (b) Measures of productivity.
   (c) Measures of association.
3. Applications of statistical methods:
   (a) Segmentation.
   (b) Properties of some morphological constructions.
   (c) Historical morphological change.

Future Work: Some Examples

- Segmentation standards:
  How to incorporate the reality that different tasks may require different segmentations (cf. (Peng et al., 2002)).

- Unknown words in segmentation:
  Several good methods for dealing with known words. There are some promising methods for *detecting* unknown words. Some work on *classification* (e.g. in named entity extraction), but more is needed.

- Historical morphology:
  Linguistic discussions of the "monosyllabicity" of Ancient Chinese, and the transition to polysyllabic words are often imprecise. Historical corpora are now available (e.g. Academia Sinica historical corpus), so it should be possible to improve these discussions.

References

Anderson, Stephen. 1992. *A-Morphous Morphology*. Cambridge University Press, Cambridge.

Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. MIT Press, Cambridge, MA.

Baayen, Harald. 1989. *A Corpus-Based Approach to Morphological Productivity: Statistical Analysis and Psycholinguistic Interpretation*. Ph.D. thesis, Free University, Amsterdam.

Baayen, Harald. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Brill, Eric. 1993. *A Corpus-Based Approach to Language Learning*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Chang, Chao-Huang and Cheng-Der Chen. 1995. A study on corpus-based classification on chinese words. *Communications of COLIPS*, 5(1-2).

Chang, Jing-Shin and Keh-Yih Su. 1997. An unsupervised iterative method for Chinese new lexicon extraction. *International Journal of Computational Linguistics and Chinese Language Processing*.

Chang, Jyun-Shen, Shun-De Chen, Ying Zheng, Xian-Zhong Liu, and Shu-Jin Ke. 1992. Large-corpus-based methods for Chinese personal name recognition. *Journal of Chinese Information Processing*, 6(3):7–15.

Chao, Yuen-Ren. 1968. *A Grammar of Spoken Chinese*. University of California Press, Berkeley, CA.

Chen, Hongbiao. 2001. *Looking for Better Chinese Indexes: A Corpus-based Approach to Base NP Detection and Indexing*. Ph.D. thesis, Guangdong University of Foreign Studies, Guangzhou.

Chen, Keh-Jiann and Chao-jan Chen. 2000. Knowledge extraction for identification of chinese organization names. In *Second Chinese Language Processing Workshop*, Hong Kong.

Dai, John X.-L. 1992. *Chinese Morphology and its Interface with the Syntax*. Ph.D. thesis, The Ohio State University, Columbus, OH.

Di Sciullo, Anna Maria and Edwin Williams. 1987. *On the Definition of Word*. MIT Press, Cambridge, MA.

Dunning, Ted E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Fano, R. 1961. *Transmission of Information*. MIT Press, Cambridge, MA.

Feng, Shengli. 1997. Prosodic structure and compound words in Classical Chinese. In Jerome Packard, editor, *New Approaches to Chinese Word Formation: Morphology Phonology and the Lexicon in Modern and Ancient Chinese*, number 105 in Trends in Linguistics: Studies and Monographs. Mouton de Gruyter, Berlin, pages 197–260.

Fung, Pascale and Dekai Wu. 1994. Statistical augmentation of a chinese machine-readable dictionary. In *WVLC-94: Second Annual Workshop on Very Large Corpora*.

Gan, Kok Wee. 1995. *Integrating Word Boundary Identification with Sentence Understanding [?]*. Ph.D. thesis, National University of Singapore.

Gan, Kok-Wee, Martha Palmer, and Kim-Teng Lua. 1996. A statistically emergent approach for language processing: Application to modeling context effects in ambiguous Chinese word boundary perception. *Computational Linguistics*, 22(4):531–553.

GB/T 13715–92. 1993. Contemporary Chinese language word-segmentation specification for information processing. Technical report, , Beijing.

Ge, Xianping, Wanda Pratt, and Padhraic Smyth. 1999. Discovering Chinese words from unsegmented text. In *SIGIR '99*, pages 271–272.

Hockenmaier, Julia and Chris Brew. 1998. Error driven segmentation of Chinese. *Communications of COLIPS*, 8(1):69–84.

Huang, Chu-Ren. 1999. From quantitative to qualitative studies: Developments in computational and corpus linguistics of Chinese. Institute of Linguistics, Academia Sinica.

Huang, Chu-Ren, Kathleen Ahrens, and Keh-Jiann Chen. 1994. A data-driven approach to psychological reality of the mental lexicon: Two studies on chinese corpus linguistics. In *Language and its Psychobiological Bases*, Taipei.

Huang, Chu-Ren, Keh-Jiann Chen, Chang Lili, and Feng-yi Chen. 1997. Segmentation standard for Chinese natural language processing. *International Journal of Computational Linguistics and Chinese Language Processing*, 2(2):47–62.

Huang, Chu-Ren, Wei-Mei Hong, and Keh-Jiann Chen. 1994. Suoxie: An information based lexical rule of abbreviation. In *Proceedings of the Second Pacific Asia Conference on Formal and Computational Linguistics II*, pages 49–52, Japan.

Hui, Huihsin (會慧馨), Chaolin Liu (劉昭麟), Zhaomin Gao (高照明), and Keh-Jiann Chen (陳克健). 2002. 以構詞律與相似法為本的中文動詞自動分類研究 (A Hybrid Approach for Automatic Classification of Chinese Unknown Verbs). *International Journal of Computational Linguistics and Chinese Language Processing*, 7(1):1–28.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Packard, Jerome. 2000. *The Morphology of Chinese: A Linguistics and Cognitive Approach*. Cambridge University Press, Cambridge.

Palmer, David. 1997. A trainable rule-based algorithm for word segmentation. In *Proceedings of the Association for Computational Linguistics*.

Pan, Wenguo (潘文國), Po-Ching Yip (葉步青), and Yang Saxena Han (韓洋). 1993. 漢語的構詞法研究 *(Studies in Chinese Word-Formation)*. Student Book Company, Taipei.

Peng, Fuchun, Xiangji Huang, Dale Schuurmans, and Nick Cercone. 2002. Investigating the relationship of word segmentation performance and retrieval performance in Chinese IR. In *Proceedings of the 19th International Conference on Computional Linguistics (COLING 2002)*, Taipei, ROC.

Peng, Fuchun and Dale Schuurmans. 2001. Self-supervised Chinese word segmentation. In F. Hoffman et al., editor, *Advances in Intelligent Data Analysis, Proceedings of the Fourth International Conference (IDA-01)*, number 2189 in LNCS, pages 238–247, Heidelberg. Springer-Verlag.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Sproat, Richard. 2000. *A Computational Theory of Writing Systems*. Cambridge University Press, Stanford, CA.

Sproat, Richard and Chilin Shih. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4:336–351.

Sproat, Richard and Chilin Shih. 1996. A corpus-based analysis of Mandarin nominal root compounds. *Journal of East Asian Linguistics*, 4(1):1–23.

Sproat, Richard, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22(3).

Sun, Maosong, Dayang Shen, and Benjamin K. Tsou. 1998. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of COLING-ACL '98*, pages 1265–1271.

Wang, Liang-Jyh, Wei-Chuan Li, and Chao-Huang Chang. 1992. Recognizing unregistered names for mandarin word identification. In *Proceedings of COLING-92*, pages 1239–1243. COLING.

Xia, Fei. 1999. Segmentation guideline, Chinese Treebank Project. Technical report, University of Pennsylvania. `http://morph.ldc.upenn.edu/ctb/`.

Yuan, Yao and Kim-Teng Lua. 1998. Splitting-merging model of Chinese word tokenization and segmentation. *Natural Language Engineering*, 4(4):309–324.