Automatic Clustering of Chinese Characters and Words

Chao-Huang Chang (張照煌) and Cheng-Der Chen (陳正德)
Advanced Technology Center (E000)
Computer & Communication Research Laboratories
Industrial Technology Research Institute
Chutung, Hsinchu 31015, Taiwan, R.O.C.
E-mail: changch@e0sun3.ccl.itri.org.tw

Abstract

Research on Chinese part-of-speech tagging has been very active recently. However, there are several problems the research must confront before a successful tagger can be realized. Among them are word definition, segmentation, lexicon, tag set, tagging guideline, and tagged corpora. We propose machine-clustered word classes as an alternative for part-of-speech to be used in class n-gram models. Chinese characters and words are automatically clustered into a predefined number of classes using a simulated annealing approach. The 1991 United Daily text corpus of approximately 10 million characters is used to collect the statistics of character and word collocation. We will show and discuss some preliminary experimental results, which are considered promising and interesting.

1 Introduction

Word n-gram models are useful in many NLP applications. However, they have a lot of parameters, which need huge training data to estimate and take very much memory and disk space to store. Thus, class n-gram models [1] have been proposed to reduce the number of parameters. One of well-known class n-gram models is the Tri-POS model [6], which defines word classes based on syntactic categories, i.e., parts-of-speech. It has been successfully used in many western languages [5]. However, Chinese language models using part-of-speech information have only a very limited success [4,11] due to the following difficulties encountered in Chinese part-of-speech tagging [2]:

1. To define an appropriate Chinese part-of-speech tag set [10,15,16];

2. To find a Chinese lexicon with complete part-of-speech informations;

3. To solve the word segmentation problems [3], e.g., word definition, unregistered words, and compounds;

4. For human to do Chinese part-of-speech tagging; the parts-of-speech of numerous words are either arguable or difficult to decide.

5. To find manually tagged Chinese corpora, counterparts of the Brown and LOB corpora in Chinese.

In the paper, machine-generated *disjoint word classes* are proposed as an alternative for parts-of-speech. The five problems listed above are solved at the same time.

1. Automatic clustering of Chinese words is used to generate the disjoint word classes. Thus, the size of the tag set and the definition of classes are decided automatically.

2. Each word in the lexicon is assigned to one of the machine-generated disjoint word classes.

3. Any raw output of a word segmentation program can be used to train the automatic clustering system. Words are defined as any segmented character strings.

4. There is no need for human to tag the disjoint word classes.

5. Any unsegmented, untagged text corpus can be used for training and/or testing.

The main concept is: Let machines do things in their own way. Chinese words, parts-of-speech are not well-defined concepts even for human, not to mention machines. We consider that this is why machines can not do word segmentation and part-of-speech tagging satisfactorily.

The other advantages include: (1) scalable: The number of classes (i.e., granularity) can be adjusted easily according to applications; (2) adaptable: The word classes can be retrained on corpora of different domains.

2 Automatic Clustering of Words

2.1 The Problem

Let

$T = w_1, w_2, ..., w_L$ be a text corpus with L words;

$V = v_1, v_2, ..., v_{NV}$ be the vocabulary composed of the NV distinct words in T;

$C = C_1, C_2, ..., C_{NC}$ be the set of classes, where NC is a predefined number of classes.

The word clustering problem can be formulated as follows:

Given V and C (with a fixed NC), find a class assignment ϕ from V to C which maximizes the estimated probability of T, $\hat{p}(T)$, according to a specific probabilistic language model.

For a bigram class model, find

$$\phi : V \rightarrow C$$

to maximize

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

Alternatively, *perplexity* [7] or *average mutual information* [1] can be used as the characteristic value to optimize.

Perplexity, PP, is a well-known quality metric for language models in speech recognition. It is also called the average word branching factor of the model.

$$PP = \hat{p}(T)^{-\frac{1}{L}}$$

In a sense, the language model reduces the difficulty of recognition task from distinguishing NV words to distinguishing PP words. The perplexities PP for the word and class bigram models are:

$$PP = exp(-\frac{1}{L}\sum_{i=1}^{L} ln(p(w_i|w_{i-1})))$$

and

$$PP = exp(-\frac{1}{L}\sum_{i=1}^{L} ln(p(w_i|\phi(w_i))p(\phi(w_i)|\phi(w_{i-1}))))$$

respectively, where w_j is the j-th word in the text and $\phi(w_j)$ is the class that w_j is assigned to.

For class N-gram models with fixed NC, lower perplexity indicates better class assignment of the words. The word classification problem is thus defined: Find the class assignment of the words to *minimize* the perplexity of the training text.

2.2 Disjoint Word Classes

Linguistic objects in natural languages can be classified into four categories (Table 1): (I) linguistically defined, ambiguous; (II) linguistically defined, disjoint; (III) artificially defined, ambiguous; and (IV) artificially defined, disjoint.

	ambiguous	disjoint
linguistic	Chinese word, part-of-speech	Chinese character, English word
artificial	-	machine-generated word cluster, word equivalence class

Table 1: Taxonomy of Linguistic Objects

Most of Chinese NLP researchers have dealt with the problems from the linguistic point of view. We have been trying to identify Chinese words from text, to tag the part-of-speech for the words. However, these concepts (Chinese words, part-of-speech) are often poorly defined or highly ambiguous (Type I). The computer is not good at resolving ambiguity according to linguistic criteria. Thus, Type II objects are much easier to process than Type I objects. The concept of word classes has been proposed to reduce the number of parameters in statistical language models. Lee *et al.* [12] approximates Chinese word bigram by the idea of word-lattice-based character bigram, because Chinese characters are disjoint, i.e., unambiguous (Type II)

while Chinese words are boundary ambiguous (Type I). We can not find examples of Type III objects. Kupiec[9] defined unambiguous word equivalence class (Type IV) based on possible tags of a word for the part-of-speech tagging problem.

We propose the following directions for Chinese class n-gram models:

1. Using disjoint classes instead of ambiguous (overlapping) classes;
2. Using artificial (machine-generated) classes rather than linguistically defined classes.

In other words, Type IV objects are preferred over Type-I and Type-II objects.

2.3 A Simulated Annealing Approach

The word classification problem can be considered as a combinatorial optimization problem to be solved with a simulated annealing algorithm. Jardino and Adda [7] used a simulated annealing approach to automatically classify words in a French corpus of 40,000 words and a German corpus of 100,000 words. A simulated annealing algorithm needs four components [8]:

(1) a specification of **configuration**, (2) a random **move generator** for rearrangements of the elements in a configuration, (3) a **cost function** or objective function to evaluate a configuration, (4) an **annealing schedule** that specifies time and duration to decrease the control parameter (or temperature).

For the word classification problem, the configuration is clearly the class assignment ϕ. The move generator is also straightforward – randomly choosing a word to be reassigned to a randomly chosen class. Perplexity can serve as the cost function to evaluate the quality of word classification[7]. The annealing schedule follows that of Metropolis algorithm. Thus, the clustering procedure is:

1. *Initialize*: Assign the words randomly to the predefined number of classes to have an initial configuration;
2. *Move*: Reassign a randomly selected word to a randomly selected class (Monte Carlo principle);

3. *Accept or Backtrack*: If the perplexity is changed within a controlled limit (decreases or increases within limit), the new configuration is accepted; otherwise, undo the reassignment (Metropolis algorithm, see below);

4. *Loop*: Iterate the above two steps until the perplexity converges.

Metropolis algorithm [7]: The original Monte Carlo optimization accepts a new configuration only if the perplexity decreases, suffers from the local minimum problem. Metropolis *et al.* (1953) proposed that a worse configuration can be accepted according to the control parameter *cp*. The new configuration is accepted if $exp(\Delta PP/cp)$ is greater than a random number between 0 and 1, where ΔPP is the difference of perplexities for two consecutive steps. *cp* is decreased logarithmically after a fixed number of iterations.

In the following two sections, we use similar simulated annealing techniques to automatically cluster Chinese characters and words in the 1991 United Daily corpus.

3 Clustering Chinese Characters

3.1 The Corpus and Character Bigrams

The statistics of Chinese character bigram is based on the 1991 UD corpus (`1991ud`) of approximately 10,000,000 characters. There are totally 5,403 character types: the 5401 commonly used characters in the Big-5 character set, a type for all 7,650 (13,051-5,401) other Chinese characters in Big-5, and another type for special symbols, such as punctuation marks and foreign characters (Arabics, English). There are 723,681 nonzero entries in the full 5403x5403 bigram and 9,529,107 ($= L$) occurrences of character types.

To keep the clustering experiments running within reasonable time, using the whole UD corpus for full character or word bigrams is not satisfactory. A smaller subcorpus, `day7`, containing one day of news, was extracted from the 1991 UD corpus. There are 147,976 nonzero entries in the full 5403x5403 bigram and 540,561 ($= L$) occurrences of character types.

3.2 Experimental Results: Clustering 100 Simple Characters

To illustrate the clustering process, the first 100 Big-5 characters are chosen as objects to classify (Table 2).

一乙丁七乃九了二人儿入八几刀刁力匕十卜又
三下丈上丫丸凡久么也乞于亡兀刃勺千叉口土
士夕大女子孑孓寸小尢尸山川工己已巳巾干廾
弋弓才丑丐不中丰丹之尹予云井互五亢仁什仃
仆仇仍今介仄元允内六兮公冗凶分切刈勻勾勿

Table 2: The 100 Characters to be Classified

The statistics for these 100 characters are extracted from the full 5403-character bigram for `1991ud`. That is, a 100x100 submatrix is extracted from the 5403x5403 matrix. This is an approximation of a text composed of these 100 characters. There are 1,968 nonzero entries in the 100x100 bigram and 144,261 ($= L$) occurrences of them.

The control parameter cp is initialized to 0.1 and divided by two after every 1,000 iterations.

First, we tried to cluster the 100 characters into 10 classes (see Figure 1 for the converging process). One of the classes, Class-0, is used to represent unknown characters and characters with zero frequency (i.e., never occurs in the training text). We observe that:

- Initial configuration: Class-0 contains the zero-frequency characters; all other characters are assigned to Class-1. This practice follows the suggestion made by Jardino and Adda [7]. Initial perplexity is 32.950.
- Perplexity decreases quickly at first several runs (each run corresponds to a fixed cp and 1,000 iterations), from 32.950, 22.052, 19.396, to 18.608 after the third run ($cp = 0.025$).
- If the classical Monte Carlo method is used – a new configuration is accepted only if the perplexity decreases, it will get stuck at a much worse local minimum.
- After only 12 runs ($cp = 4.88 \cdot 10^{-5}$), the perplexity converges to its final value 17.719.
- The clustering result is very encouraging. The final configuration is shown in Table 3.
  - Class-3 consists of six digits.

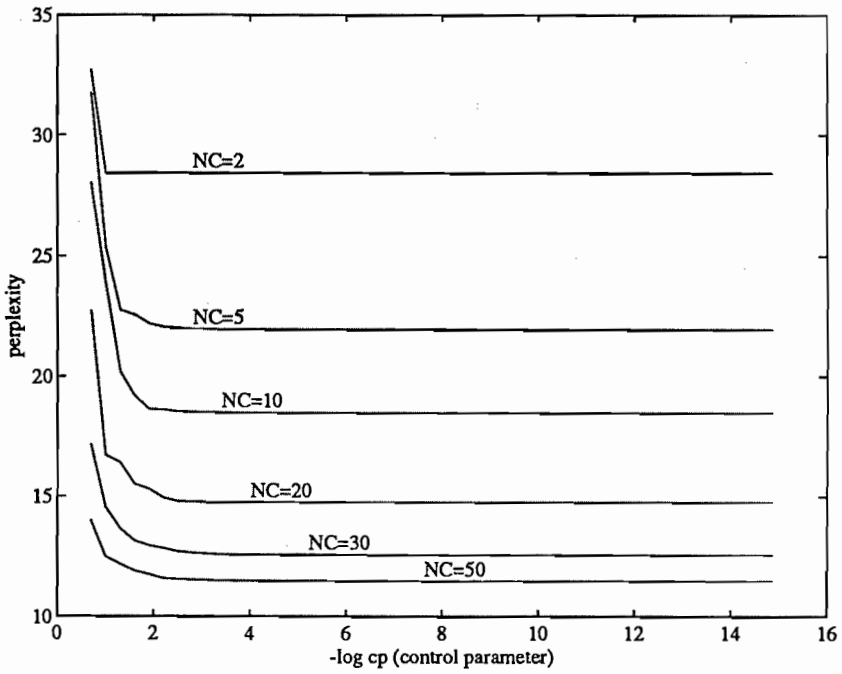

Figure 1: Clustering 100 Chinese characters: different NCs

- Class-9 consists of two characters 十 and 千, which are quite similar syntactically and semantically.
- Class-1 contains several first characters of measure words, 丈, 寸, 元, 公, 分, etc. Note that the classification is based on character bigram, so 公can be considered as a part of measures such as 公分.
- Class-0 is artificially made for unseen characters.
- Class-6 has a large family; somehow, members in the meaningful groups (1) 上中下 (2) 大中小 (3) 女子 (4) 乙丁 (5) 山川 are together.
- The other classes are less meaningful.
- Two digits 一, 九 are not assigned to Class-3.

Figure 1 also compares the converging processes for different values of NC and Table 4 shows the converged perplexities. As expected, the perplexity decreases when the number of classes increases. The perplexity of the character bigram model, i.e., $NC = 100$, is 11.250. If we

Class	Members
Class-0 :	儿刁匕兀尤尸巳廾弋彳仆冗刈
Class-1 :	丈夕寸弓丰元公分勻
Class-2 :	一孑什仄
Class-3 :	七二八三五六
Class-4 :	人又也乞刃叉己井互仍
Class-5 :	了巾之凶
Class-6 :	乙丁乃入几刀力卜下上丫丸凡久亡勺土大女子 小山川工已干中丹予云亢仁仇今介内兮
Class-7 :	九孓切勿
Class-8 :	么于口士才丑丐不尹允勾
Class-9 :	十千

Table 3: 10 Output Character Clusters for the 100 Characters

classify the 100 characters into 30 (NC) classes, the perplexity is just 12.556. There is not much difference. This shows the feasibility of class n-gram models: reducing the number of parameters significantly, having competitive performance, requiring much less resources, and robust.

NC	Perplexity
1	32.959
2	28.442
5	21.947
10	17.719
20	14.760
30	12.556
50	11.472
100	11.250

Table 4: Perplexities for different NCs

3.3 Experimental Results: Clustering 5401 Chinese Characters

Running the full 5401-character bigram for a large corpus ($L > 1,000,000$) takes huge amount of time. Note that Jardino and Adda [7] used much smaller corpora (40,000 and 100,000 words, respectively). However, the algorithm has a time complexity proportional to L. They reported that it took 7 hours on a 486-33 PC to classify 14,000 words into 120 classes using a 75,000-word training set. We have designed an incremental version of the system which is much faster than the original version which recomputes all probabilities in each trial. In our experience, it took 20.06 CPU hours on a DEC 3000/500 AXP workstation to classify 5403 characters into 200

classes (with 50,000 trials in each of 64 iterations) using 540,000-character day7 corpus.

We have conducted three experiments on the day7 corpus.

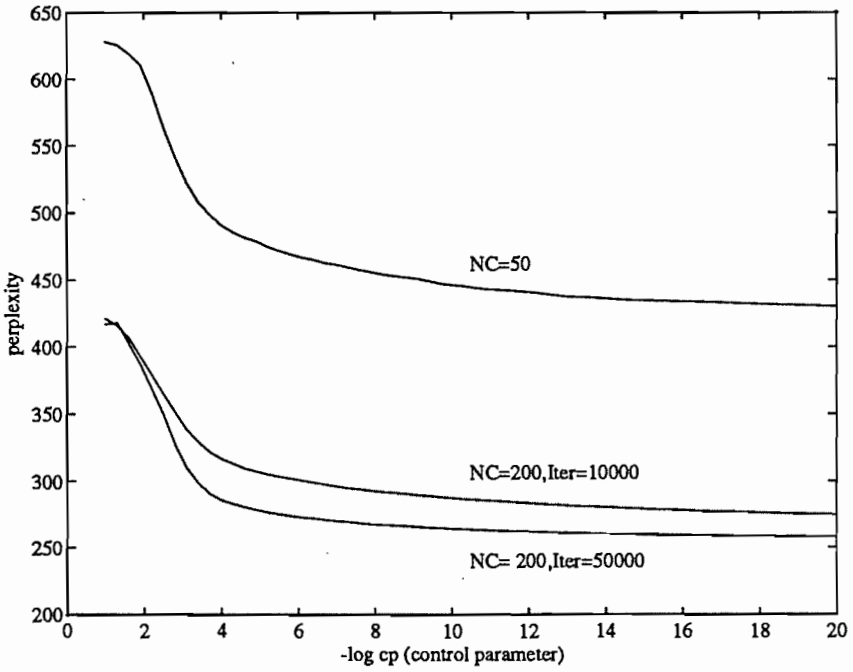

Figure 2: Clustering 5,401 Chinese characters: the converging process

Experiment 1:

- Number of Classes: 50
- Initial cp: 0.1
- Initial perplexity: 675
- Characters with frequency less than 3 are assigned to Class-0
- Number of iterations in each run: 10000
- The perplexity converging process (Figure 2): 675, 628, 626, 619, 610, 589, 562, 541, 522, ..., 430 (*cp* too small, less than 10^{-20})
- The final configuration:

- Class-0 contains 2,293 characters.
- Class-1 contains human-related characters, 人于大女予仁及夫巴引比火父王仔冊卉古史尼母田全, ...
- Class-14: 一今每赤兩昔昨晝翌規逐傍幾瑕睦逾鐮
- Class-18 contains: (numbers) 七九二八三五六四彷零綽銖嘰墩餘.
- Class-20 : 凡允尤乍另因如汎而但否坊迄呵咋俟染倘唯捨毫嗯搽寥漂誰縫譬. Many of them are conjunctions.
- Class-21 consists of 中什他它全各夷她舟你我杜牠那妳拒信俄帝怎某洪美耐范苗英哥哪栽泰爹翁您這敞粵豪緯薰蟹驅. Almost all pronouns are included in this class.
- Class-22 and Class-24 are composed of measure words among others:
  (Class-22) 元分斗冊呎坪姐秒哩畝措楞歲號蝶噸點;
  (Class-24) 天日月乎旦兆吋年次佰岸枚肢頁個株般隻晚棵番週塊漸磅篇趟艘顆曜釐藩蹴
- Class-26 contains several function words, 又也才仍勿只亦刪即均希更並尚彼則卻既皆盼祇若俾豈焉都竟就廁猶滲頗躺還雖
- Class-28 contains several surnames: 于戈王古吉安江余吳呂宋李貝阮林邱侯姜柯柳胡韋唐孫徐桂張梁郭陳麥傅游雲馮黃楊鄒廖趙劉潘蔡鄭黎墨燕盧蕭薛鍾簡魏羅, among others.
- Class-37 consists of orders of ten, 十千卅廿仟百彿巷第萬億蝴
- Somehow, we consider that 50 is not enough for character class discrimination.

Experiments 2 and 3:

- Number of Classes: 200
- Initial cp: 0.1
- Initial perplexity: 675
- Characters with frequency less than 3 are assigned to Class-0
- Number of iterations in each run: 10000 and 50000
- The perplexity converged to 274 and 258, respectively (see Figure 2).

1	山川井卉古丞后州臣村谷里坡河城娃拷泉洛皇厝哥埔島...
3	乙丁丹匹丙仙尼玉瓜甲仲仰吉宇汝亨佛佐君廷...
11	次批屈套座陣貫番項種樣樁篇趟輛輩
19	元呎辰坪畝歲銖噸鎊
23	巴冬寺春玻秋胡迪夏徐桂秦細荷陰紫詠嘉蜜撙澎踴蟬羅藤蘇
25	于王丘朱江汎艾余吳呂吟巫李沈狄貝邢阮卑庚林邵邱侯姜柯柳洪范郁韋倪唐孫袁娼寇張曹梁莫莊貪郭陳麥傅馮黃楊鄒廖蒼蜿赫趙劉潘蔣蔡褒鄭黎儒墾盧穆蕭賴鮑薛謝魏譚龔蠶
26	天日晚溢
51	尺斤司室頃寓債
52	士員
57	屯市禾京省郊莞橙縣韓
63	予於握鄰
69	下上濤
70	倆們搔
72	午月周巷秒週
73	各專殖農漁闔
78	千仟百佰炫
81	北竹西東股南園墓魁糖轄曜壢藩
87	七九八
94	台左屏桃焊菩臺閩緬墨
106	十卅廿第
114	兆萬億
124	每那某哪這
128	才只皆祇都就僅還
135	他它她你我牠妳茨您渥嗯誰懊蟹
138	年屆
139	今去昔拜昨翌傍
140	一兩幾
141	向在迄披拘赴從猜被喘趁跟憑
144	令把使垂堵逢替給雇僱對蓋讓
160	名位段隻梯瓶舶啼磅
173	乃凡尤而但倘俾惟譬
189	二三五六四零餘
190	又也仍亦均並尚卻既若豈竟猶滲
196	至初迎近牲剩逾厲囂

Table 5: Part of Output Character Clusters

- See Table 5 for part of the final configuration with $Iter = 50000$.

  The listed classes have obvious meanings, the others are less clear. However, we observe that function-word and content-word characters are usually grouped separately.

Test Set Perplexity

To evaluate the performance of classification, we use another subcorpus `day5` (544,606 characters) to compute a test set perplexity. For character clustering, smoothing is not necessary since the character set is mapped to a fixed set of 5403 character types. Table 6 shows the test set perplexities for different NC and number of trials per iteration (Iter). In general, classifications with higher NC have lower test set perplexity because the character set is closed. Clustering with higher Iter also reduces test set perplexity but with limit.

NC	Iter	Train PP	Test PP
50	10000	430.266	461.355
200	10000	274.916	290.208
200	50000	258.124	274.470

Table 6: Test Set Perplexities

4 Clustering Chinese Words

4.1 The Corpus and Word Bigrams

The statistics of Chinese word bigram is based on the above-mentioned (`day7`) corpus. The corpus is segmented automatically into clauses, then into words by our Viterbi-based word identification program VSG. There are totally 42,537 clauses, 355,347 ($= L$) words (189,838 1-character, 150,267 2-character, 10,783 3-character, 4,460 4-character), belonging to 23,977 word types (3,377 1-character, 16,004 2-character, 2,461 3-character, 2,135 4-character). There are 203,304 nonzero entries in the full 23,977x23,977 bigram, which is stored in compressed form.

4.2 Experimental Results: Clustering Chinese words

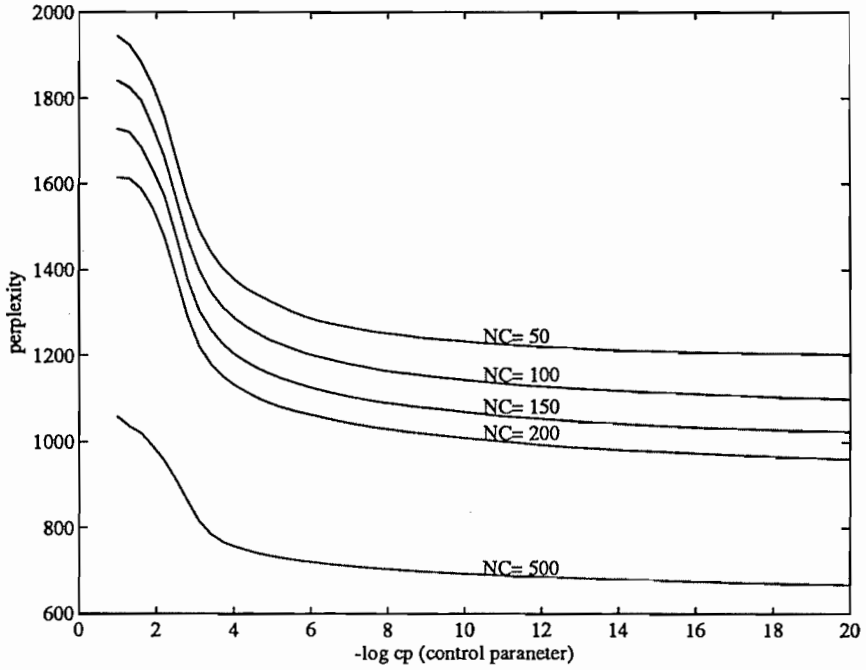

Figure 3: Clustering Chinese words: the converging process

Initial Configuration

Words with frequency less than m (default value: 5) are assigned to Class-0, the unseen word class [7]. Punctuation marks are assigned to a special class Class-1. 1–4 character number words are assigned to Classes 2–5, respectively. (Note that the numbers are composed by lexical rules in our word segmentation program VSG.) All other words are assigned to Class-6. Initial perplexity is 2,048.

Word Clustering: Changing NC and Iter

Numbers of experiments have conducted on the `day7` corpus with different parameters. The following table shows the converged training set perplexity.

As expected, classifications with higher NC or higher Iter have lower training set perplexities. However, it is not the case for test set perplexity. Figure 3 shows the perplexity converging

NC	*10,000 iter.*	*20,000 iter.*	*50,000 iter.*
50	1305	1234	1203
100	1212	1140	1099
200	1068	1019	960
500	723	699	667

Table 7: Training Set Perplexities

processes for $iter = 50000$.

Testing Set Perplexities

For the problem of unseen words and bigrams, we adopt a similar linear smoothing scheme to that of Jardino and Adda [7]. (For details, see the original paper.) The interpolation parameters α and β are set to $1 - 10^{-5}$ and 0.1, respectively.

Four subcorpora are used: `day7` for training, `day5`, `day8`, `day9` for testing. Simple statistics are summarized in Table 8.

corpus	#clauses	#vocabulary	#words
day7	42,539	23,977	355,347
day5	44,334	24,706	360,464
day8	27,946	18,948	232,818
day9	26,579	19,111	221,105

Table 8: Four Subcorpora: `day5`, `day7`, `day8`, `day9`

The test set perplexities are summarized in Table 9 (for $Iter = 50,000$).

It appears that the most appropriate number of classes is about 150 to 200 for the size of training corpus. The clustering with $NC = 500$ is apparently overtrained (Figure 4).

Word Classification Results

As we all know, the smallest meaningful unit in Chinese is words rather than characters. The classification results for words are also more meaningful, if the clustering is well trained. When

NC	day5	day8	day9
50	1543	1548	1489
100	1491	1495	1437
150	1482	1487	1427
200	1478	1489	1436
500	1655	1637	1594

Table 9: Test Set Perplexities: day5, day8, day9

we set Iter to 10,000, the result did not have clear meaning. However. we now set Iter to 50,000, the classification results are encouraging.

For $NC = 200$, 15,900 words are assigned to the six special classes (15070, 5, 17, 131, 266, and 411 words, respectively). The other words are assigned to the other 194 classes, approximately according to part-of-speech. Part of the output clusters are shown in Table 10. The following are some observations:

- Title nouns: Class-7
- Place nouns: Class-10 (counties, cities, towns), Class-25 (nations, capitals, etc.), Class-79 (organizations)
- Time nouns: Class-12 (seasons), Class-44 (weekdays), Class-196
- Personal Names: Class-14, Class-42 (including some foreign names)
- Common Nouns: Class-6, Class-9, Class-11, Class-13, Class-27, Class-33
- Verbs: Class-8, Class-22 (是-related), Class-37 (有-related), Class-38 (count-related), Class-45 (1-character verbs)
- Adverbs: Class-15
- Prepositions: Class-16, Class-26,
- Adjectives: Class-17
- Conjunctions: Class-23, Class-29, Class-34
- Modals: Class-67
- Number nouns: Class-2, Class-3, Class-4, Class-5, Class-91
- Measure nouns: Class-36, Class-70, Class-74

6	中華文化/內心/公務/巴勒斯坦/心靈/文教/水箱/人力/大自然/...
7	分析師/民航局/主任/主委/主持人/主席/召集人/次長/局長/助理/巡官/抱/委任/拜訪/指認/ 前任/秋山/紀/首相/負責人/祕書/祕書長/秘書/副處長/強暴/組長/理事長/ 部長/處長/發言人/ 隊長/會長/董事長/署長/監委/總書記/總裁/總經理/鎮民/邊界
8	分贓/大戰/干預/代表會/住戶/似/局勢/攻擊/供應/花費/指示/相對/修訂/...
9	心意/人/人選/老婦/見面/角落/例/門窗/巷口/冒險/故鄉/時效性/...
10	中衛/斗六/三重/大甲/白河/台中/台北/台南/宜蘭/花蓮/東勢/板橋/恆春/ 南投/南京/苗栗/桃園/埔里/ 草屯/馬公/高雄/貢寮/通霄/雲林/新竹/嘉義/彰化/龜山/豐原/蘭陽/觀音
11	份子/缸/時/娘娘/創刊/學期
12	乙/末/白/光華/冬/出任/辰/協/帕/奈/昇/娃/冠/秋/美/春/胖/英/ ...
13	公職/勾結/手法/手續/支票/方針/比重/人數/上車/生動/成長率/收入/...
14	布/弗/主/古/宋/希/余/克拉/卓/周/坦/姜/姚/恆/柳/倪/范/夏/ 徐/特/莊/郭/傅/游/渥/湯/程/馮/黃/翡翠/潘/謝/韓/蘇/灌
15	引人/十分/大為/大幅/必經/有何/伺機/呈/具/受/招/欣賞/派人/...
16	不惜/大年/已於/去年度/自/名列/於/原定/將於/對付/糖果/...
17	大/可怕/多/多人/年長/形/看好/高點/淡/煙花/僵持/適合/遲/舉/懷念/轟轟烈烈
22	不至於/才是/正是/何況/係/是/是有/將是/就是/顯得
23	不但/上將/老是/似乎/並/並且/究竟/固然/始終/便/則/皆/既/根本/殊/送完/從來/猶/確/還將/雖
25	中美/中華民國/太平洋/巴國/.../日本/上海/本市/本地/本國/白宮/全美/全球/ 加拿大/加國/外地/米蘭/多倫多/我國/東方/東亞/東京/波士頓/帝國/科威特/ 美國/政大/海地/國內/國民黨/麻州/華盛頓/菲律賓/越南/新加坡/新華社/ 詹氏/歐美/澎湖/廣大/廣西/德州/德國/墾丁/澳洲/雙子星/羅馬尼亞/...
26	引用/大肆/充當/加入/向/在於/存入/呈報/享譽/往/委託/返回/按/前往/洽/查出/致函/留在/...
27	中餐/午餐/文章/日子/水果/牙刷/人潮/口袋/口號/生命/交往/仲裁/...
29	不料/尤其/凡是/也就是/未料/由於/以防/另一方面/外傳/而且/而後/至於/但/但是/因此/否則/並稱/例如/...
33	夫妻/一則/女子/母雞/目前為止/半/外匯存底/有心/年底/攻勢/車子/車主/典型/男友/男人/男子/...
34	尤其是/一旦/乃是/且/以及/以免/以使/以便/以致/有如/而/而是/自從/合計/如/直至/直到/卻因/俾/致使/...
36	寸/句/外國語/次/個/套/埠/條/塊/層樓/輛/艘
37	不無/不敷/充滿/有/自成/均有/具有/沒/祇有/倍感/起出/將有/創下/處於/連續/曾有/無/...
38	.../共有/共計/多達/法例/珍禽/約/降至/高出/高溫/高達/售/款項/超過/達/逾/...
42	于/伊/朱/吳/呂/孟/林/林森/威廉/約翰/洪/柴/馬/郝/梅傑/寇/現任/棻/鼎/維/赫/蔣/蔡/黎/穆/鮑/羅/魏/鶯
44	今天/日前/凡/早年/次日/即日/明天/周四/初一/前天/昨/昨天/昨日/除夕/ 週五/週一/當日/...
45	切/引/丟/住/坐/忘記/扮/改/判/走/刺/忽/花錢/阻礙/指控/限/乘/哭/...
67	不必/不可能/不能/不得不/勿/一再/一度/已經/必須/本著/未/未曾/可/可以/可望/有沒有/並未/沒有/...
70	公斤/分鐘/斤/月分/人次/千年/小時/世紀/年/年中/年代/年後/頁/餘年/擊敗
74	元/公分/公升/公尺/公克/公里/公頃/公噸/二元/平方公尺/立方公尺/冊/...
79	公司/公所/公會/出口區/加油站/局/周轉/值/計程車/商/酒樓/教育局/會館/誠/署/戲院/餐廳/黨部
91	千億/平方/百億/尾/巷/億/數百萬/餘/點/厘
196	中午/午/午夜/升/一時/下午/四時/至/年初/初/凌晨/耗資/起至/晚上/晚間/傍晚/墩

Table 10: Part of Output Word Clusters

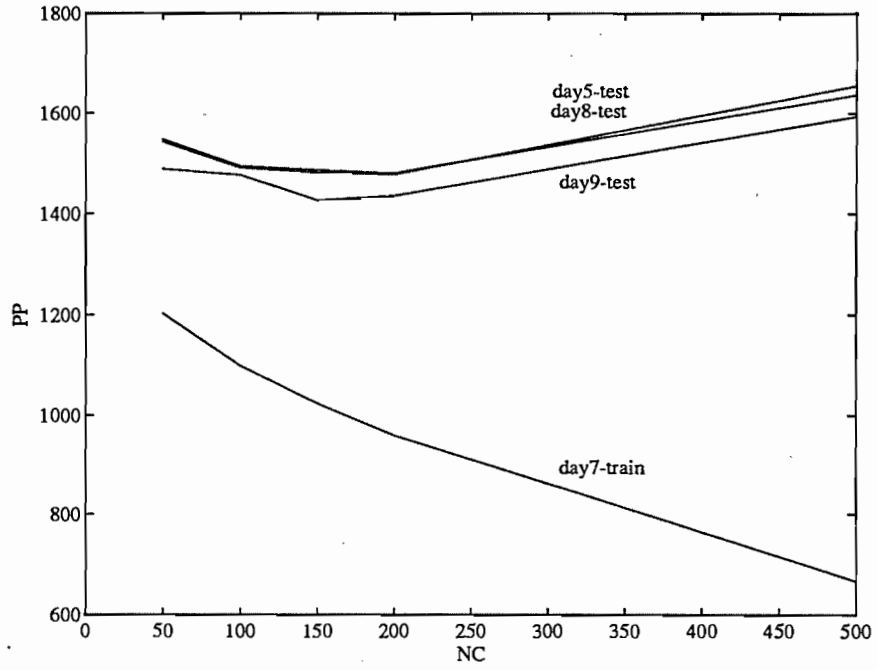

Figure 4: Training vs. Test Set Perplexity

- Equi-length words are usually grouped together. For example, one-character verbs, two-character verbs, one-character nouns, two-character nouns are grouped separately.

4.3 How to Use the Classification Results

The classification results of words (or characters) can be used in language models for speech recognition or OCR postprocessing. The class-ids for the words can be stored in the system dictionary. Words in a new input sentence are just automatically mapped to the classes through dictionary look-up. Thus, a class n-gram language model can be easily built up based on the machine-generated classes. As we mentioned, the number of classes can be adjusted according to the size of training corpus and application needs. As a common rule of thumb, the size of training data should be at least 10 times the number of parameters. Thus, if we have a corpus of size L for an N-gram model, the appropriate number of classes NC can be computed using the equation:

$$L = 10 \cdot NC^N$$

For example, if bigram ($N = 2$) models are used, NC is 100 for $L = 100,000$ and 1000 for $L = 10,000,000$.

For evaluating the performance of such language models, the test set perplexity can be used. In a sense, the language model using the day7-trained character clustering reduces the difficulty of recognition task from 5403 (character types) to 461 ($NC = 50$) and 274 ($NC = 200$). Similarly, the day7-trained word clustering reduces task difficulty from 24,706 to 1,543 ($NC = 50$) and 1,478 ($NC = 200$).

5 Other Approaches for Automatic Word Clustering

5.1 Brown *et al.* (1992)

Brown *et al.* [1] presented several statistical algorithms for automatic word classification. They use the average mutual information $I(C_i, C_j)$ as the characteristic value to maximize for a class bigram model:

$$I(C_i, C_j) = \sum_{C_i C_j} P(C_i C_j) \log \frac{P(C_j|C_i)}{P(C_j)}$$

They proposed two word clustering algorithms [1]:

Greedy-style Merging (1) Assign each word to a unique class and compute $I(C_i, C_j)$; (2) Merge two classes if the loss in $I(C_i, C_j)$ is least; (3) C classes remains after $V - C$ times of merging; (4) For each word in vocabulary, move it to a class to maximize the average mutual information. The merging steps produce a binary tree according to statistical similarity. However, they reported that this algorithm is not practical for a vocabulary with more than 5,000 words.

Add-One Merging For a larger vocabulary, (1) Assign each of the NC most frequently used words to a unique class; (2) Assign the next unclassified word with largest frequency to

a new class $C_{(NC+1)}$, and merge two classes if the loss in $I(C_i, C_j)$ is least; (3) After $V - C$ steps, the NV words in the vocabulary are assigned to NC classes. A 260,741-word vocabulary had been classified into 1,000 classes this way.

Using a 1001-word window and the concept of semantic stickiness, they had classified English words semantically and had interesting results [1].

5.2 Ney and Essen (1991)

Ney and Essen [13] proposed a decision-directed, iterative unsupervised learning procedure: (1) Choose an initial mapping $\phi : V_i \rightarrow C_j = \phi(V_i)$ (2) Update the bigram and word counts $N(C_jW)$ and $N(C_j)$; (3) Compute the probability estimates $P(W|C_j) = N(C_jW)/N(C_j)$; (4) Find the optimal class $\phi(V_i)$ for each predecessor word V_i

$$\phi(V_i) = argmax_{C_j} \sum_W N(V_iW) \log P(W|C_j)$$

A German corpus of 95,671 words ($NV = 14080$) and a English corpus of 1,157,260 words ($NV = 49615$) have been classified into 128 classes this way.

5.3 Schutze (1993)

Schutze [14] proposed the idea of *category space.* A category of each word is represented by a vector in a multidimensional real-valued space. The category space is built by collecting various word distributional information. Four bigram matrices are built with distances -2, -1, 1, 2. A sparse matrix algorithm SVDPACK is then used to compute fifteen singular values for each word. A word is represented as a 15-dimensional vector in the category space. Close neighbors in the space are grouped into a word class (part-of-speech). A linear-time clustering algorithm called Buckshot was used to cluster the category space. The experiment was conducted on a 5000-word vocabulary. 278 high-frequency closed-class words such as prepositions are assigned distinct classes. The other 4,722 words were clustered into 222 classes. A second singular value decomposition was then performed on 22,771 words from New York Times based on the resulting 500 classes. Finally, a second Buckshot was used to classify the words into 200 output clusters.

6 Concluding Remarks

We have proposed using machine-generated disjoint word classes as an alternative for the popular word class – part-of-speech in Chinese class n-gram models. A simulated annealing approach is used to cluster Chinese characters and words into a predefined number of classes. Encouraging and interesting experiment results on the 1991 UD corpus have been shown and discussed. Future works include (1) more experiments on word clustering with different parameters; (2) studying more efficient algorithm for simulated annealing; (3) using windows to study semantic clustering; (4) applying to language models for speech recognition or OCR; and (5) studying and applying different clustering approaches.

Acknowledgements

This paper is a partial result of the project no. 37H2100 conducted by the ITRI under sponsorship of the Minister of Economic Affairs, R.O.C.

References

[1] P.F. Brown, V.J. Della Pietra, P.V. de Souza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[2] C.-H. Chang and C.-D. Chen. HMM-based part-of-speech tagging for Chinese corpora. In *ACL-93: Workshop on Very Large Corpora*, Ohio, USA, June 1993.

[3] K.J. Chen and S.-H. Liu. Word identification for Mandarin Chinese sentences. In *Proc. COLING-92*, Nantes, France, 1992.

[4] T.-H. Chiang, J.-S. Chang, M.-Y. Lin, and K.-Y. Su. Statistical models for word segmentation and unknown word resolution. In *Proc. of ROCLING V*, pages 121–146, Taipei, Taiwan, September 1992.

[5] K. Church. A stochastic parts program and noun phrase parser for unresticted text. In *Proc. of ICASSP-89*, pages 695–698, Glasgow, Scotland, 1989.

[6] A. Derouault and B. Merialdo. Natural language modeling for phoneme-to-text transcription. *IEEE Trans. PAMI*, 8(5):742–749, 1986.

[7] M. Jardino and G. Adda. Automatic word classification using simulated annealing. In *Proc. of ICASSP-93*, pages II:41–44, Minneapolis, Minnesota, USA, 1993.

[8] S. Kirkpatrick, C.D. Gelatt, Jr., and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[9] J. Kupiec. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, 6:225–242, 1992.

[10] H.-J. Lee and C.-H. Chang Chien. A Markov language model in handwritten Chinese text recognition. In *Proc. of Workshop on Corpus-based Researches and Techniques for Natural Language Processing*, Taipei, Taiwan, September 1992.

[11] H.-J. Lee, C.-H. Dung, F.-M. Lai, and C.-H. Chang Chien. Applications of Markov language models. In *Proc. of Workshop on Advanced Information Systems*, Hsinchu, Taiwan, May 1993.

[12] L.-S. Lee et al. Golden Mandarin (II) - an improved single-chip real-time Mandarin dictation machine for Chinese language with very large vocabulary. In *Proc. of ICASSP-93*, pages II:503–506, April 1993.

[13] H. Ney and U. Essen. On smoothing techniques for bigram-based natural language modelling. In *Proc. of ICASSP-91*, pages 825–828, Toronto, September 1991.

[14] Hinrich Schutze. Part-of-speech induction from scratch. In *Proc. of ACL-93*, pages 251–258, Columbus, Ohio, USA, June 1993.

[15] K.-Y. Su, Y.-L. Hsu, and C. Saillard. Constructing a phrase structure grammar by incorporating linguistic knowledge and statistical log-likelihood ratio. In *Proc. of ROCLING IV*, pages 257–275, Pingtung, Taiwan, 1991.

[16] M.S. Sun, T.B.Y. Lai, S.C. Lun, and C.F. Sun. The design of a tagset for Chinese word segmentation. In *First International Conference on Chinese Linguistics*, Singapore, June 1992.