A Multivariate Gaussian Mixture Model for Automatic Compound Word Extraction

***Jing-Shin Chang and +Keh-Yih Su**

***Behavior Design Corporation, Hsinchu, Taiwan, ROC.**

+Department of EE, National Tsing-Hua University
Hsinchu, Taiwan, ROC.

August 23, 1997, Academia Sinica, Taipei, ROCLING-X

Table of Contents

- ☞ **Why Automatic Lexicon Acquisition**
- ☞ **P-R (Precision-Recall) Maximization Problems**
- ☞ **English Compound Word Extraction Problems**

  Two-Stage Strategy -

  - ☆ **Minimizing Classification Error**
  - ☆ **Non-linear Learning for Precision- Recall Maximization (not addressed in the conference paper)**

What is Lexicon Acquisition (English)

For information about installation, see Microsoft Word Getting Started. To choose a command from a menu, point to a menu name and click the left mouse button (滑鼠左鍵). For example, point to the File menu and click to display the File commands. If a command name is followed by an ellipsis, a dialog box (對話框) appears so you can set the options you want. You can also change the shortcut keys (快捷鍵) assigned to commands. (Microsoft Word User Guide)

(1996/10/29 CNN) Microsoft Corp. announced a major restructuring Tuesday that creates two worldwide product groups and shuffles the top ranks of senior management. Under the fourth realignment ..., the company will separate its consumer products from its business applications, creating a Platforms and Applications group and an Interactive Media group. ... Nathan Myhrvold, who also co-managed the Applications and Content group, was named to the newly created position of chief technology officer.

What is Lexicon Acquisition (Chinese)

China Times 1997/7/26:

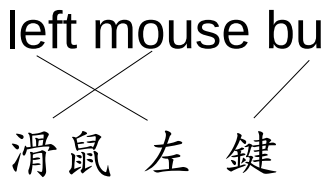
台經院指出，隨著股市活絡與景氣回溫，第一季車輛及零件營業額成長十六・八一％，顯示民間需求回升。再加上爲加入ＷＴＯ，開放進口已是時勢所趨，也將帶動消費成長。台經院預測今年民間消費全年成長率可提昇至六・七四％。

在投資方面，第一季國內投資出現回升走勢，固定資本形成實質增加六・五六％，其中民間投資實質增加八・九五％。在持續有民間大型投資計畫進行、國內房市 回溫、與政府開放投資、加速執行公共工程等多項因素下，預測今年全年民間投資將成長十一・八％。

台經院表示，口蹄疫 連鎖效應在第二季顯現，使第二季出口貿易成長率比預期低，出口年增率二・一％，比去年低。而進口年增率爲七・三八％，因此第二季貿易出超僅十七・一四億美元，比去年第二季減少四十三・六五％。不過，由於第三、四季爲出口旺季，加上國際組織均預測今年世界貿易量擴大，台經院認爲我國商品出口應可轉趨順暢。

Why Automatic Lexicon Acquisition

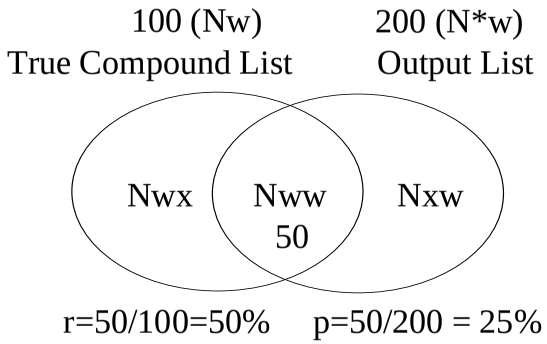
1. A large-scale electronic dictionary is important to many NLP applications
   - machine translation, spoken language processing, spelling check, associated input methods
2. New (unknown) words && compound words increase rapidly
   - vary with *time*
   - vary with *domain*
3. Prefer to lexicalize for easier: disambiguation (analysis), compositionality (generation)
   e.g., book (n, vi, vt) + store (n, vt) <=> book store (n)
   e.g., green house ≠ 'green' + 'house'



dialog: 對話//交談
box: (方)框//盒子
⇔ 對話框

4. Human construction is costly, time consuming and inconsistent
5. Electronic text is becoming widely available

Precision-Recall Optimization Criteria



$$p = \frac{N_{ww}}{N_{ww}+N_{xw}} = \frac{1}{1+N_{xw}/N_{ww}}$$

$$r = \frac{N_{ww}}{N_{ww}+N_{wx}} = \frac{1}{1+N_{wx}/N_{ww}}$$

p = Nww/(Nww+Nxw) = #correct_identification / #output_words

r = Nww/(Nww+Nwx) = #correct_identification / #all_words

(Nij: # of class-i n-grams which are classified as class-j)

(i, j= w - word//compound ; x - non-word//non-compound)

⇒ Typical Joint Criteria for Precision (p) and Recall (r) Maximization:

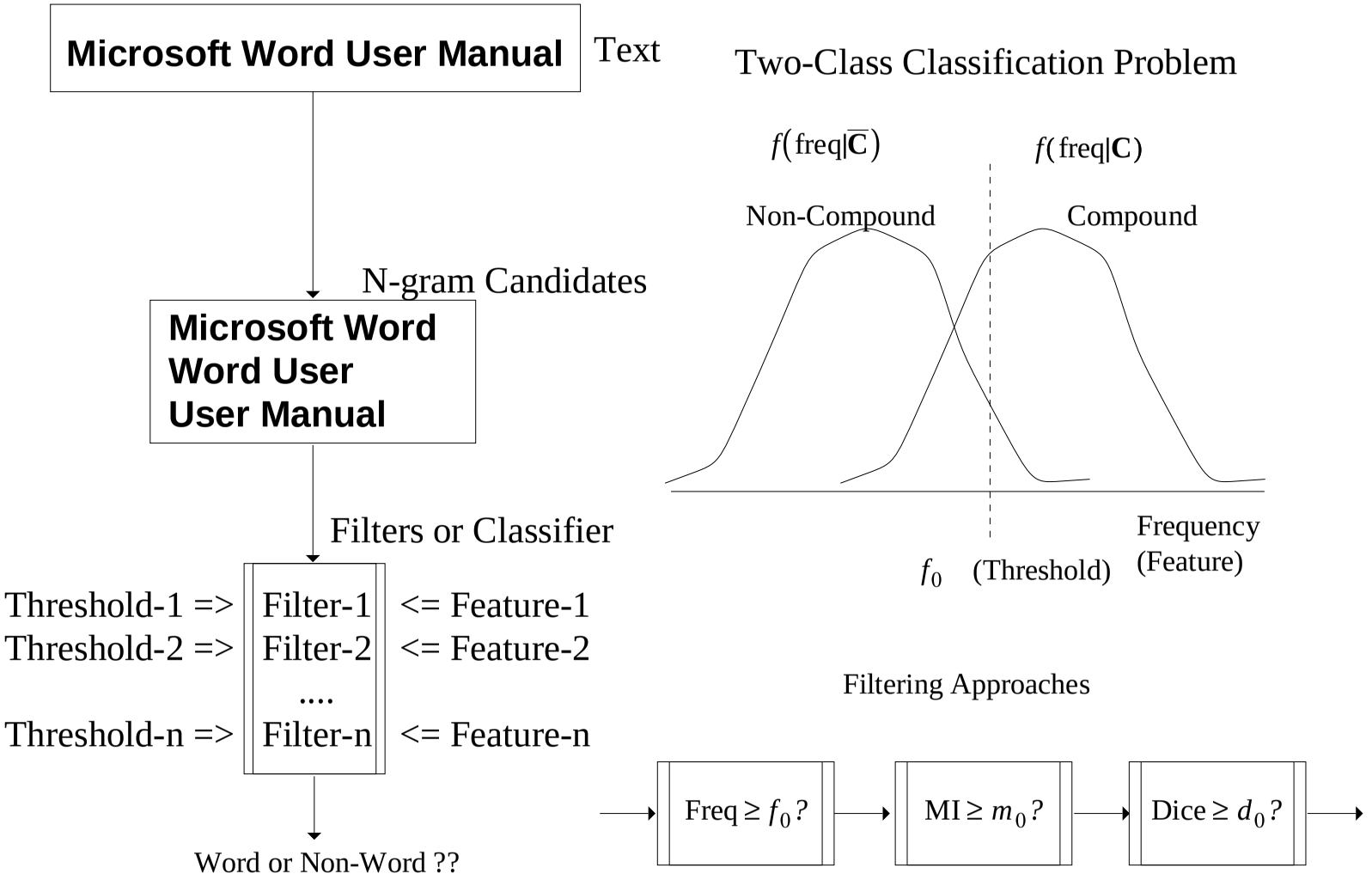
⇛ WPR: Wp*p+Wr*r (weighted Precision/Recall)

[Wp, Wr: weights (Wp+Wr=1)]

– A weighting sum of precision and recall.

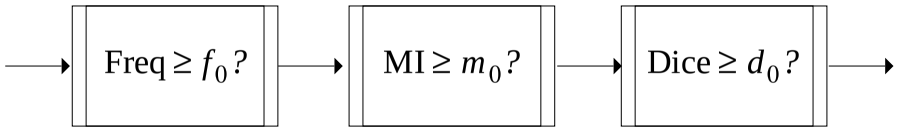
- F-metric (F-measure): $F(\beta) = \frac{(\beta^2+1)pr}{\beta^2 p + r} = \frac{pr}{p\beta^2/(\beta^2+1) + r/(\beta^2+1)}$
  - A metric that appreciate a balance between precision and recall.
    [Maximal at p=r if β=1 and p+r is a constant.]
    (Prefer maximal product of p and r for a given weighted P/R)

General Scheme in English Compound Extraction



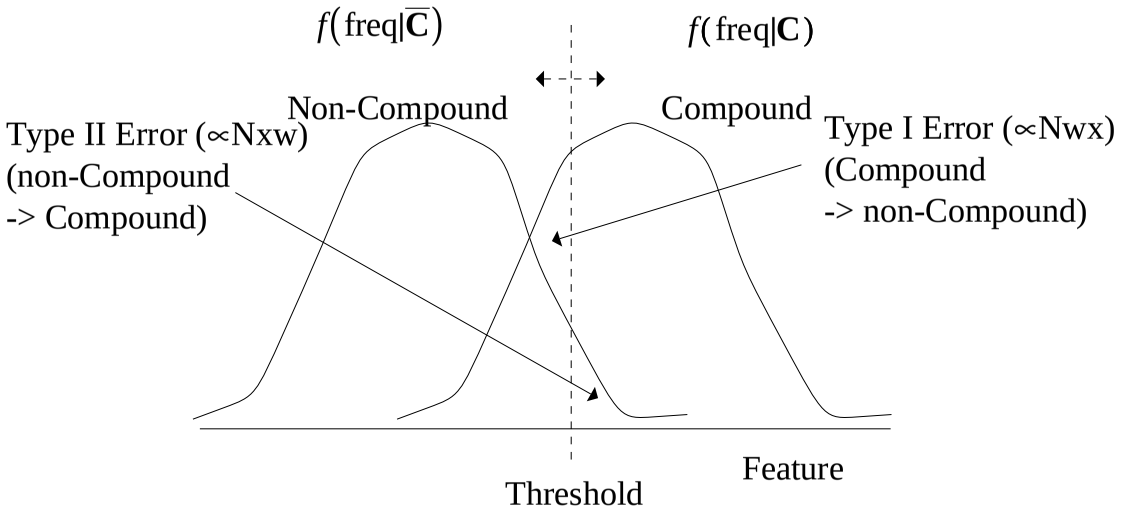
General Problems in Lexicon Acquisition

- Use simple *filtering* approaches and heuristic thresholds in extracting lexicon entries
  - ☆ mostly based on step-by-step filtering approaches which filter out inappropriate candidates with one feature per step



  - ☆ thresholds are determined by trial-and-error
  - ☆ no unified method for integrating various features jointly
    - features are used *independent* of one another
    - no automatic method for identifying the best feature set

Precision-Recall Maximization Problems



- Precision and Recall cannot be tuned in an appropriate manner
  - ☆ precision and recall are nonlinear functions of error counts
  - ☆ *adaptation* to maximize different *joint* P/R preferences (such as F-metric) in different tasks had not been addressed
  - ☆ precision and recall cannot be improved *at the same time*
    - ☆ important thresholds for features are determined arbitrarily

Two-Stage P-R Maximization

- Why two stages?
  - ☆ No simple analytical decision rules that are capable of achieving any user-specified criterion function of precision and recall
    - precision and recall are nonlinear functions of error counts
- Which two stages?
  - ☆ minimize classification error:

    $p = (1 + n_{xw}/n_{ww})^{-1}$; $r = (1 + n_{wx}/n_{ww})^{-1}$

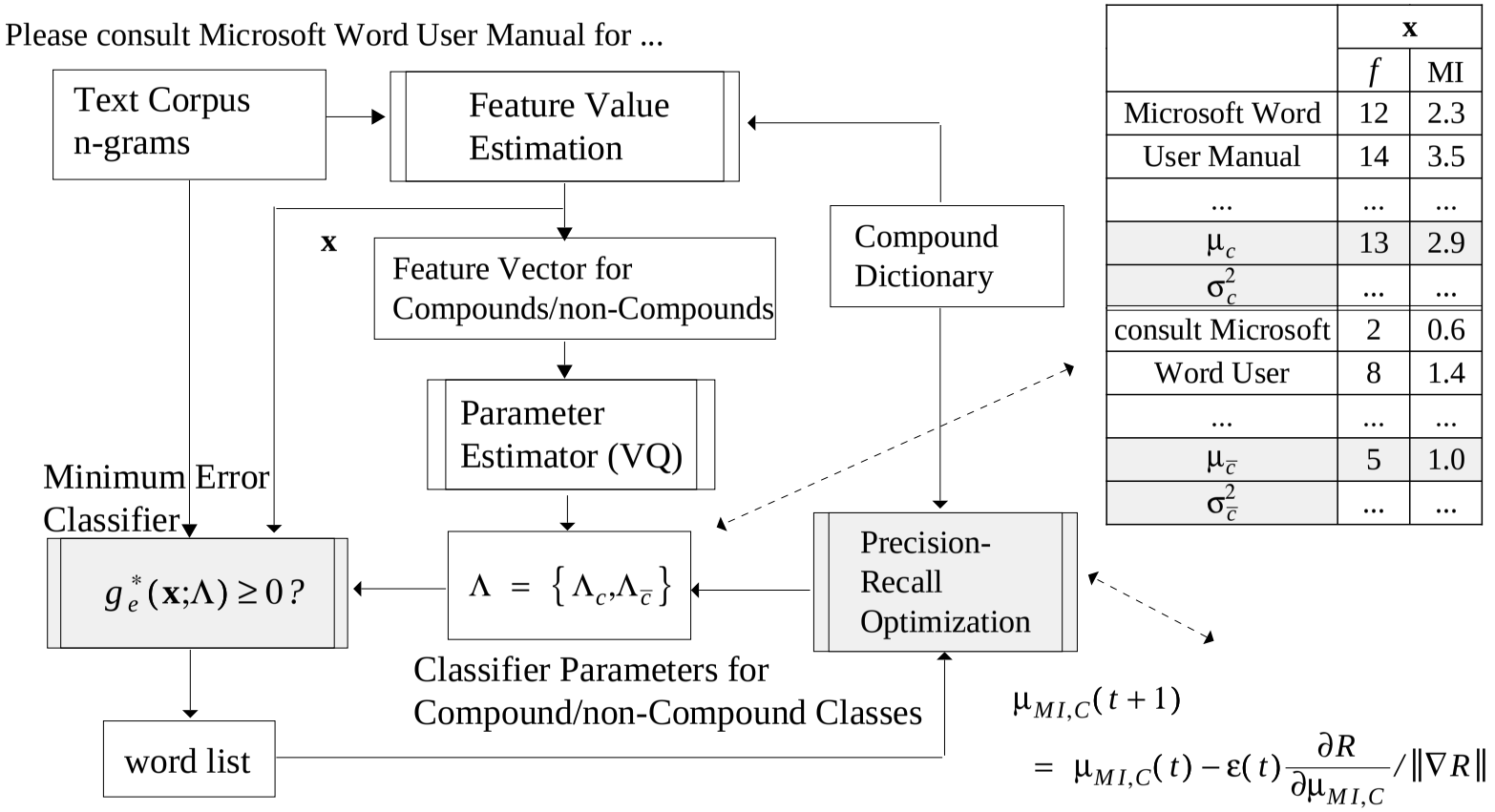
    reduce error rate (Nwx+Nxw) generally improve P, R and other joint functions (Note: Maximize FM == Minimize $(n_{wx} + n_{xw})/n_{ww}$)
  - ☆ maximize precision-recall:

    Min error classification ≠ MaxPR classification

- How to?
  - ☆ minimum error classification: better features, better models for jointly combining all features, better estimation

  - ☆ maximize precision-recall: by parameter learning (nonlinear!!)

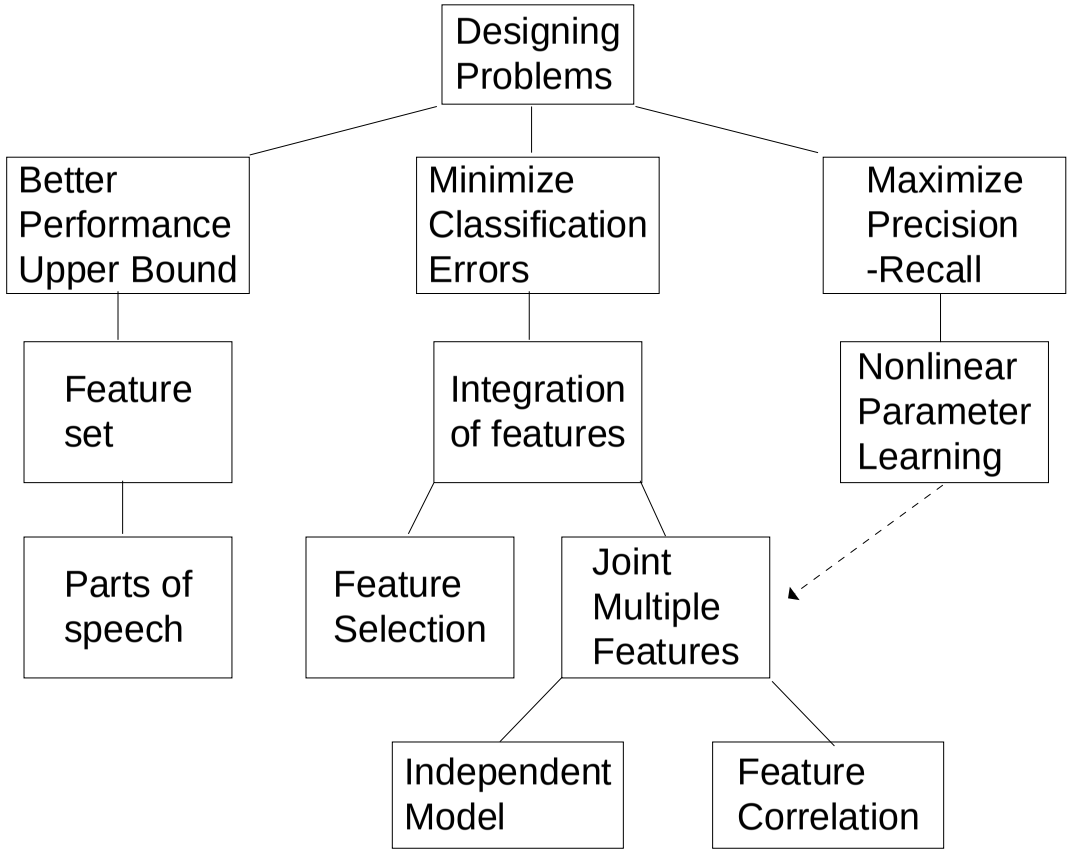
MinErr Classifier+MaxPR Learning Approach



	x	
	f	MI
Microsoft Word	12	2.3
User Manual	14	3.5
...	...	...
μ_c	13	2.9
σ_c^2	...	...
consult Microsoft	2	0.6
Word User	8	1.4
...	...	...
$\mu_{\bar{c}}$	5	1.0
$\sigma_{\bar{c}}^2$	...	...

Figure 1 Supervised Training of Classifier Parameters for English Compound Extraction

General Problems in Classifier Design



General Problems in Classifier Design

- Feature Extraction: (will not be addressed)
  - ✓ extract most discriminative features for the task
- Better Feature Set:
  - ✓ including high level features such as parts of speech
- Automatic Feature Selection:
  - ✓ adopt a unified feature selection mechanism for all available features so that (1) complementary features are used jointly, instead of being applied independently, and (2) the most appropriate features are selected automatically and less discriminative or redundant features are rejected

General Problems in Classifier Design (cont.)

- Classifier Design:
  - ✓ design appropriate decision rules for qualifying word candidates using known features jointly
- Parameter Estimation:
  - ✓ estimate statistical parameters for the classifier to fit particular *estimation criteria* (e.g., maximum likelihood estimation)
- Performance Maximization:
  - ✓ adjust statistical parameters to maximize desired *performance criteria* (e.g., a joint precision-recall performance such as F-metric)

MinErr Classifier: Two-Class Classifier for Identifying New Words or Compound Words

Input: n-grams (n-word compounds, n-character words) in the text corpus

Output: assign a class label ("word" or "non-word") to each n-gram

Classifier: a log-likelihood ratio (LLR) tester (minimum error classifier)

$$g(\mathbf{x}) = LLR(\mathbf{x}) = \log \frac{f(\mathbf{x}|\mathbf{W})P(\mathbf{W})}{f(\mathbf{x}|\overline{\mathbf{W}})P(\overline{\mathbf{W}})}$$

Decision Rules:

$$class(w) = \begin{cases} +w \quad (word) & if \quad LLR(\bullet) \geq \lambda_0 \\ -w \quad (non-word) & if \quad LLR(\bullet) < \lambda_0 \end{cases}$$

Advantage: ensure minimum classification error (with λ_0 =0) if the distributions are known.

Features for the Classifier

— Normalized Frequency $f(x) = \text{freq/avrg\_freq}$: a character n-gram, x, is likely to be a word if it appears more frequently than the average.

— Mutual Information: characters x and y with high mutual information tend to have high association [Church 90]

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

— Entropy: random distribution of the left/right neighbors (C_i) of an n-gram x implies a natural break at the n-gram boundary [Tung 94]:

$$H(x) = -\sum_{c_i} P(c_i;x) log P(c_i;x)$$

— Dice: similar to mutual information with non-occurring events (x=0,y=0) ignored [Smadja 96]:

$$D(x,y) = \frac{P(x=1,y=1)}{\frac{1}{2}[P(x=1)+P(y=1)]}$$

Features for the Classifier (cont.)

Part-of Speech Discrimination:

$$D_{pos}(x_i;\{P_{ij}\},\{P_j\}) = \sum_j P_{ij} log \frac{P_{ij}}{P_j}$$

$$P_{ij} \equiv P(j|w_i), \qquad P_j \equiv P(j)$$

An n-gram, Xi, is likely to be a word if its parts-of-speech (詞類) distribution is "close to" the parts-of-speech distribution of the n-grams in the word-class, where closeness is measured in terms of the discrimination between two probability distributions.

Pij: probability for Xi to be tagged with part-of-speech pattern j (e.g., j = [n n] for a noun-noun compound word).

Pj: probability for any n-grams to be tagged with part-of-speech pattern j.

Baseline: Error Rate by Using One Feature

		Training Set					Testing Set				
Feature		Dpos	MI	H	NF	D	Dpos	MI	H	NF	D
2-gram Baseline	Recall	11.09	0.0	4.87	6.01	12.33	8.07	0.0	1.35	2.69	36.77
	Precision	100.0	*	30.92	30.69	37.07	100.0	*	23.08	33.33	57.75
	Error Rate	11.03	12.41	13.15	13.34	13.47	21.20	23.06	23.78	23.68	20.79
	WPR(1:1)	55.54	*	17.90	18.35	24.70	54.03	*	12.22	18.01	47.26
	F-measure	19.97	*	8.41	10.05	18.50	14.93	*	2.55	4.98	44.93
Feature		Dpos	MI	H	NF	D	Dpos	MI	H	NF	D
3-gram Baseline	Recall	0.0	0.0	13.99	10.20	7.58	0.0	0.0	12.07	3.45	39.66
	Precision	*	*	42.11	22.58	25.49	*	*	58.33	66.67	41.07
	Error Rate	4.95	4.95	5.21	6.18	5.67	11.51	11.51	11.11	11.31	13.49
	WPR(1:1)	*	*	28.05	16.39	16.54	*	*	35.20	35.06	40.37
	F-measure	*	*	21.00	14.05	11.69	*	*	20.00	6.56	40.35

Table 1 Error Rate Performance Using only One Feature
(*: undefined, i.e., all candidates are classified as non-compound.).

Use Features Jointly and Select Discriminative Features Automatically for the Classifier

0. Initialize current feature set as empty.

1. Classify training data by jointly (*) using current feature set and one of the remaining features not in the current feature set. Try all the remaining features one-by-one, and include the feature that performs best to the current feature set.

2. Stop including new features whenever the performance of the classifier begins to flatten or degrade due to the inclusion of redundant or contradictory features.

3. Use the selected features for lexicon acquisition.

(*) ⇒ Models for Jointly Integrating Features:

IN: Independent Normal Model

Mx: Mixtures of Gaussian Density Functions

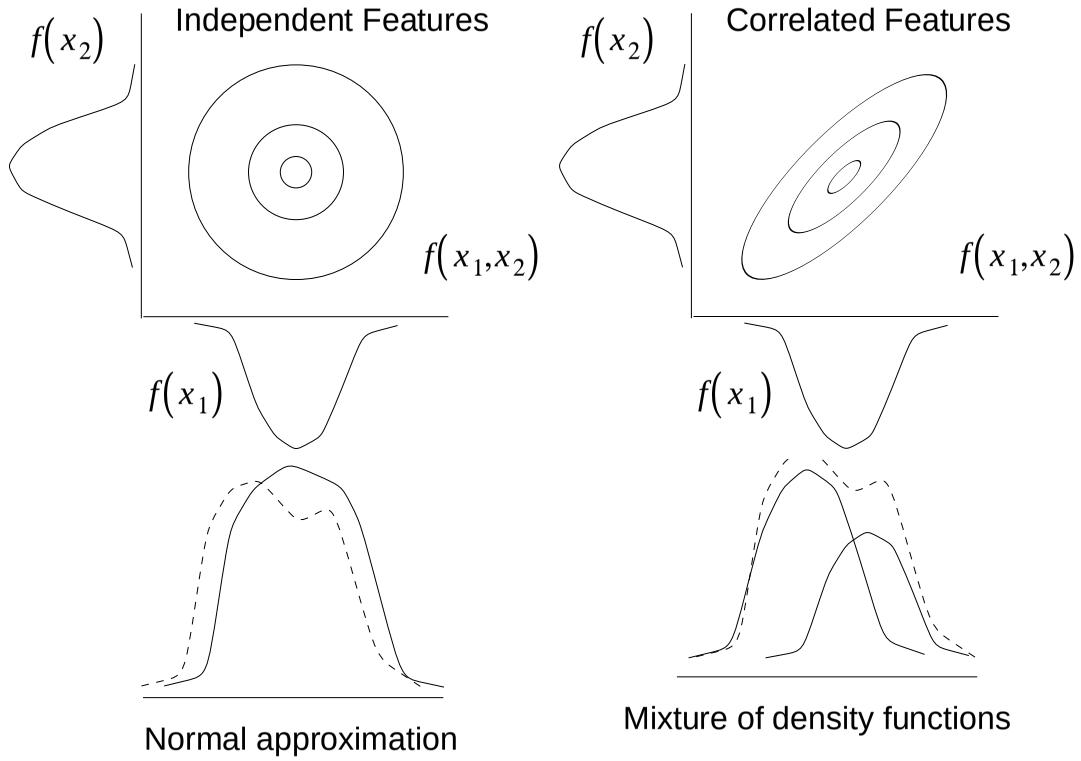
Error Rate by Using Independent Normal Model with Feature Selection for Joint Consideration

		Training Set					Testing Set				
Feature Sequence		Dpos	H	MI	NF	D	Dpos	H	MI	NF	D
2-gram	Recall	11.09	40.41	54.61	35.34	31.30	8.07	35.43	60.54	33.63	50.67
	Precision	100.0	88.04	77.39	71.04	49.67	100.0	89.77	92.47	82.42	66.47
	Error Rate	11.03	8.07	7.61	9.81	12.46	21.20	15.82	10.24	16.96	17.27
	WPR(1:1)	55.54	64.23	66.00	53.19	40.49	54.04	62.60	76.51	58.03	58.57
	F-measure	19.97	55.39	64.03	47.20	38.40	14.93	50.81	73.17	47.77	57.50
Feature Sequence		Dpos	MI	H	D	NF	Dpos	MI	H	D	NF
3-gram	Recall	0.0	14.29	33.53	29.45	26.24	0.0	17.24	44.83	56.90	48.28
	Precision	*	100.0	70.99	46.98	33.83	*	100.0	86.67	49.25	47.46
	Error Rate	4.95	4.24	3.97	5.14	6.19	11.51	9.52	7.14	11.71	12.10
	WPR(1:1)	*	57.15	52.26	38.22	30.04	*	58.62	65.75	53.08	47.87
	F-measure	*	25.01	45.55	36.20	29.56	*	29.41	59.09	52.80	47.86

Table 2 Error rate performances of the independent normal model.

Joint Consideration of the Features by Considering Feature Correlation

0. Why ?
   - Features are not really independent (have correlation)
   - Features are not really normally distributed (use mixtures)



1. Model the distributions of the features with a k-mixture Gaussian Density Functions to take correlations among features into consideration. [k is to be determined automatically in the feature selection mechanism.]

$$f(x|\Lambda) \equiv \sum_{i=1}^{K} r_i \cdot N(x;\mu_i,\Sigma_i), \quad \sum_{i=1}^{K} r_i = 1$$

$$N(x;\mu,\Sigma) = (2\pi)^{-D/2}|\Sigma|^{-1/2} exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

2. Estimate the parameters of the feature distributions using a clustering algorithm to maximize the likelihood of the input feature vectors.

Fixing K throughout Feature Selection Process

		Training Set					Testing Set				
Feature Sequence		Dpos	H	MI	NF	D	Dpos	H	MI	NF	D
2-gram	Recall	69.84	71.50	71.61	50.67	51.71	69.06	71.30	69.96	67.26	47.09
	Precision	100.0	97.87	88.93	62.93	45.53	100.0	95.78	93.41	80.65	52.24
	Error Rate	3.74	3.73	4.63	9.82	13.67	7.14	7.34	8.07	11.27	22.13
	WPR(1:1)	84.92	84.69	80.27	56.80	48.62	84.53	83.54	81.68	73.95	49.66
	F-measure	82.24	82.63	79.34	56.14	48.42	81.70	81.75	80.00	73.34	49.53

Table 3 The Best Bigram Performance of the Minimum Error Rate Classifier Using a 2-Mixture Multivariate Normal Density Function (K=2).

Feature Sequence		Dpos	H	MI	D	NF	Dpos	H	MI	D	NF
3-gram	Recall	63.27	68.22	67.06	51.90	54.23	75.86	74.14	74.14	36.21	37.93
	Precision	100.0	95.12	90.91	80.91	39.08	100.0	97.73	95.56	95.45	41.51
	Error Rate	1.82	1.75	1.96	2.99	6.45	2.78	3.17	3.37	7.54	13.29
	WPR(1:1)	81.63	81.67	78.98	66.40	46.65	87.93	85.93	84.85	65.83	39.72
	F-measure	77.50	79.45	77.18	63.24	45.43	86.27	84.32	83.50	52.50	39.64

Table 4 The Best Trigram Performance of the Minimum Error Rate Classifier Using a 3-Mixture Multivariate Normal Density Function (K=3).

Comparison: Joint Consideration of the Features

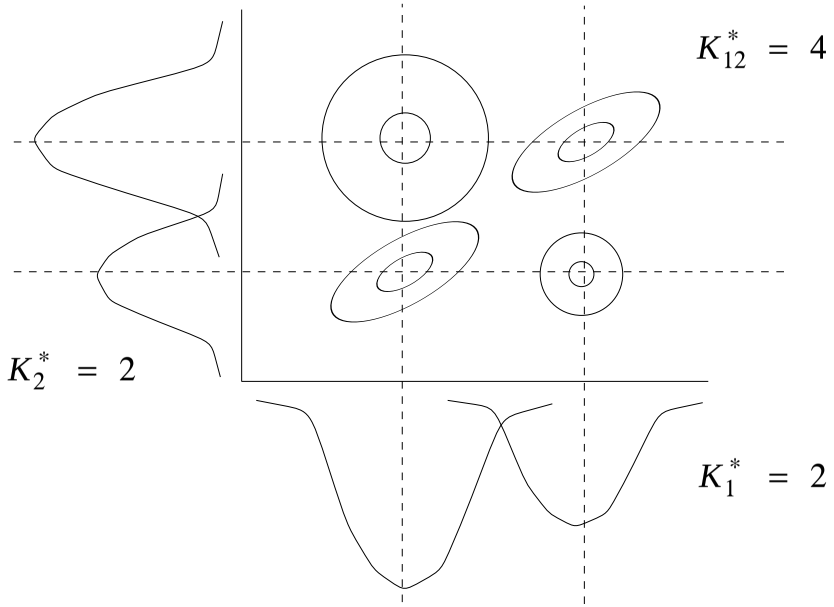
N		Training Set					Testing Set				
	Model & Features	P	R	E	WPR	FM	P	R	E	WPR	FM
2	IN: Dpos+H	88.04	40.41	8.07	64.23	55.39	89.77	35.43	15.82	62.60	50.81
	IN: Dpos+H+MI	77.39	54.61	7.61	66.00	64.03	92.47	60.54	10.24	76.51	73.17
	Mx: Dpos+H (K=2)	97.87	71.50	3.73	84.69	82.63	95.78	71.30	7.34	83.54	81.75
3	IN: Dpos+MI	100.0	14.29	4.24	57.15	25.01	100.0	17.24	9.52	58.62	29.41
	IN: Dpos+MI+H	70.99	33.53	3.97	52.26	45.55	86.67	44.83	7.14	65.75	59.09
	Mx: Dpos+H (K=3)	95.12	68.22	1.75	81.67	79.45	97.73	74.14	3.17	85.93	84.32

Table 5 Comparison between Independent Normal (IN) Model and K-mixture Multivariate Normal (Mx) Model. (2: 2-gram, 3: 3-gram, P: Precision, R: Recall, E: Error Rate, WPR: Weighted Precision/Recall with equal weights, FM: F-measure.)

Searching for the Best Number of Mixtures (K*)

Why ?

(1) As the number of features increases, K*, in general, will increase rapidly



Number of Mixtures increases rapidly with feature dimension

Searching for the Best Number of Mixtures (K*)

(2) The estimation algorithm can only achieve local maximum likelihood

- using a larger K does not guarantee to reach better local maximum likelihood estimate than using a smaller K

& even (global) maximum likelihood ≠> minimum error rate

=> using larger K ≠> smaller error rate

Searching for the Best Number of Mixtures (K*)

		Training Set					Testing Set				
Feature Sequence		Dpos(2)	H(2)	MI(3)	NF(3)	D(1)	Dpos	H	MI	NF	D
2-gram	Recall	69.84	71.50	72.12	67.05	32.12	69.06	71.30	70.40	65.92	44.39
	Precision	100.0	97.87	90.74	83.70	56.78	100.0	95.78	94.01	93.63	68.28
	Error Rate	3.74	3.73	4.37	5.71	11.45	7.14	7.34	7.86	8.89	17.58
	WPR(1:1)	84.92	84.69	81.43	75.37	44.45	84.53	83.54	82.21	79.77	56.34
	F-measure	82.24	82.63	80.37	74.46	41.03	81.70	81.75	80.51	77.37	53.80
Feature Sequence		Dpos(3)	H(3)	MI(3)	D(3)	NF(1)	Dpos	H	MI	D	NF
3-gram	Recall	63.27	68.22	67.06	51.90	24.49	75.86	74.14	74.14	36.21	44.83
	Precision	100.0	95.12	90.91	80.91	33.60	100.0	97.73	95.56	95.45	48.15
	Error Rate	1.82	1.75	1.96	2.99	6.13	2.78	3.17	3.37	7.54	11.90
	WPR(1:1)	81.63	81.67	78.98	66.40	29.04	87.93	85.93	84.85	65.83	46.49
	F-measure	77.51	79.45	77.19	63.24	28.34	86.27	84.32	83.50	52.50	46.43

Table 6 The Performance of the Minimum Error Rate Classifier Using Multivariate Normal Density Function up to 3 Mixtures (Kmax=3).

Concluding Remarks

1. Various association metrics can be used jointly to rank word candidates by using a two-class classification model, which could minimize the classification error.
2. Error rate can be reduced (and precision-recall indirectly improved) by
   - including multiple features jointly
   - examining independence assumptions
   - considering feature correlation in modeling the density function
   - underlying density functions can be better modeled with mixtures of Gaussian density functions (by searching for the best number of mixtures)