Part V: Suggested Strategies for Better Performance

- Lesson from POS Tagging
- Basic Principle
- Essential Elements
  - Adopting Appropriate Language Model
  - Educating Initial Guess
  - Enhancing Discrimination Power
  - Enhancing Robustness
- Suggested Unsupervised Learning Steps

Lesson from POS Tagging (I)

- A look at why unsupervised training works for POS tagging
  - Tri-gram model works in the supervised-learning mode
    - The performance of supervised-learning is the upper bound for the unsupervised-learning (under the same model)
    - The prediction power of the model (measured by perplexity) is high (remember, the performance upper bound is determined by the adopted feature set)
  - 62% of the words in Brown Corpus have only one tag
    - 62% of the words are also directly observable (human preference is unveiled) in the training set; virtually only 38% ambiguous
    - Not much difference between the complete data space and the incomplete data space.

Lesson from POS Tagging (II)

- Why unsupervised works for POS tagging (cont.)
  - During unsupervised-learning iterations:
    - Errors are randomly distributed (uniformly distributed in the first iteration)
    - Those 62% uni-tag words make the correct sub-patterns dominate (appear more times)
      - Many word bi-grams (or even tri-grams, 4-grams, …) can be tagged unambiguously due to those adjacent anchors; thus, they enhance the dominance of those correct patterns
      - For example, [det n] are unambiguous in some cases (e.g., The (det) dog (n)), although they would be randomly selected in other cases (e.g., The (det) design (n/v) of (prep) ..)
    - Once those correct bi-grams dominate the tagged corpus in terms of their numbers, the dominance will be enhanced iteratively.
    - Those anchor points also impose constrains on their adjacent words; thus, significantly help reducing the task complexity

Lesson from POS Tagging (III)

- What lesson can we learn ?
  - The model at least should work under the supervised-learning mode in other similar cases
    - The effectiveness of the model can thus be verified, as the supervised-learning provides the upper bound
  - The perplexity of the model (task difficulty) should be low
    - Make the problem to be solved easier in the adopted feature space
    - If different density functions of various classes highly overlap over each other in the adopted feature space, which implies that the inherited performance upper bound would be low, then it is no hope to solve this problem.
  - Human preference should be told explicitly or implicitly

Basic Principle for Performance Improvement

- Correlate Likelihood Value with Error Rate as much as possible
  - Coupling the Stochastic Language Model with the Human Preference Model
- Make Model robust in the testing set

Essential Elements

- Adopting Appropriate Language Model
  - Adopting Appropriate Feature Set
  - Adopting Appropriate Form
- Educating Initial Guess
- Enhancing Discrimination Power
- Enhancing Robustness

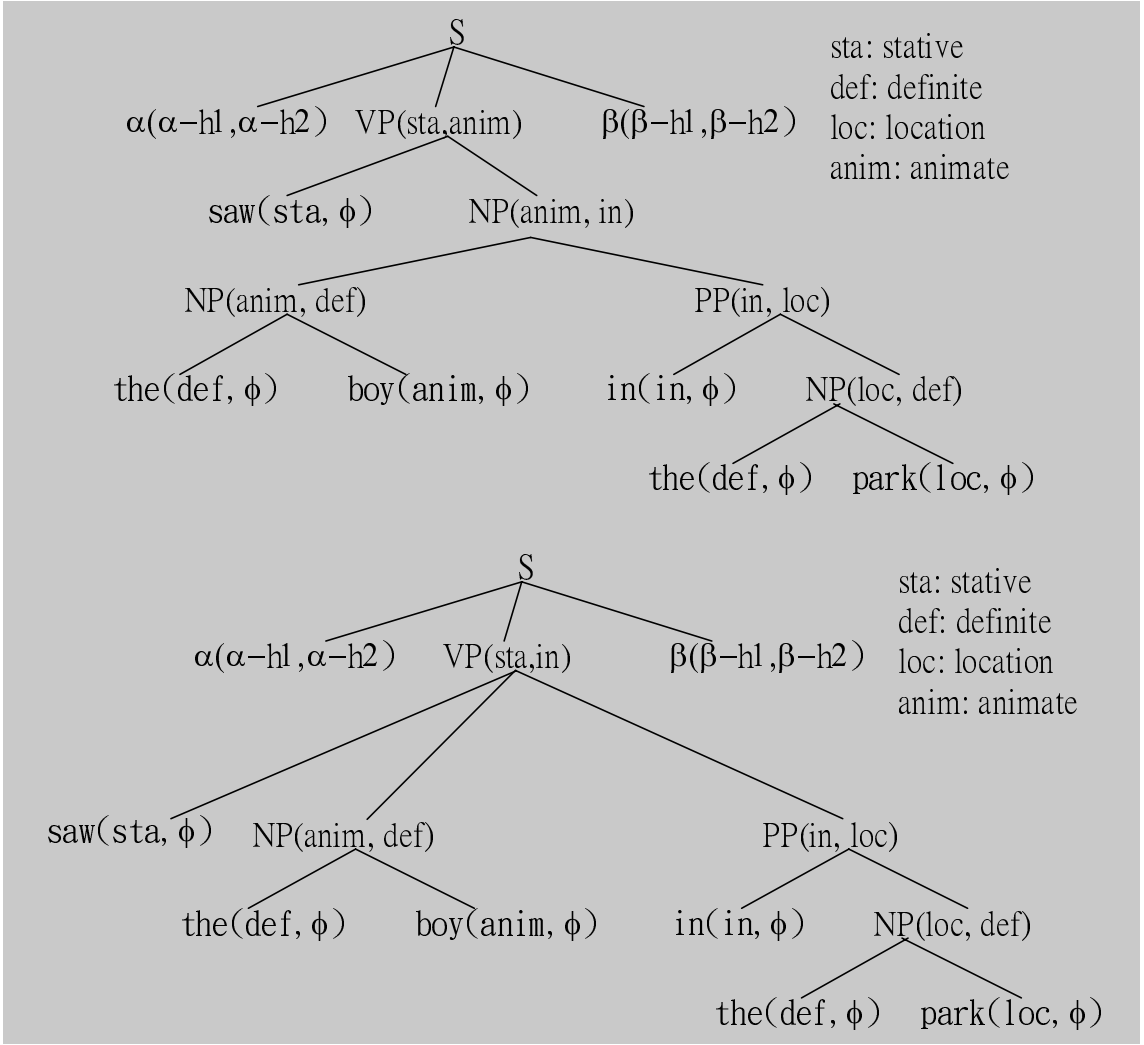
Adopting Appropriate Feature Set (I)

- Heuristically collecting the initial feature set
  - Adopt those existing Explicit Linguistic Features (based on which that human make preference)
  - Add all features that can impose constrains on your task
    - Example: the corresponding target sentences in aligned bilingual sentence pairs
  - Add other helpful Implicit Features:
    - They can be probed by using the statistical measures such as Mutual Information (or the correlation dependency test) to probe

Adopting Appropriate Feature Set (II)

- Heuristically collecting the initial feature set (cont.)
  - Add feature percolation mechanism to dynamically provide the required context-sensitive features
    - Build stochastic language model on top of those non-terminal symbols
    - Non-terminal symbols can be used to provide the percolation mechanism required for projecting those head-features
    - Example: Subject-Verb agreement is a long-distance dependency problem in the surface level; however, it is a local dependency problem in the level of NP and VP
      - As NP and VP are adjacent to each other, bi-gram model is enough to handle the agreement problem in this level
  - You can also add the information provided from other competitive classes as features to enhance the discrimination power
    - Probability measures from each class can be used as features [Su 94]

Example for Head-Feature Percolation



Adopting Appropriate Feature Set (III):

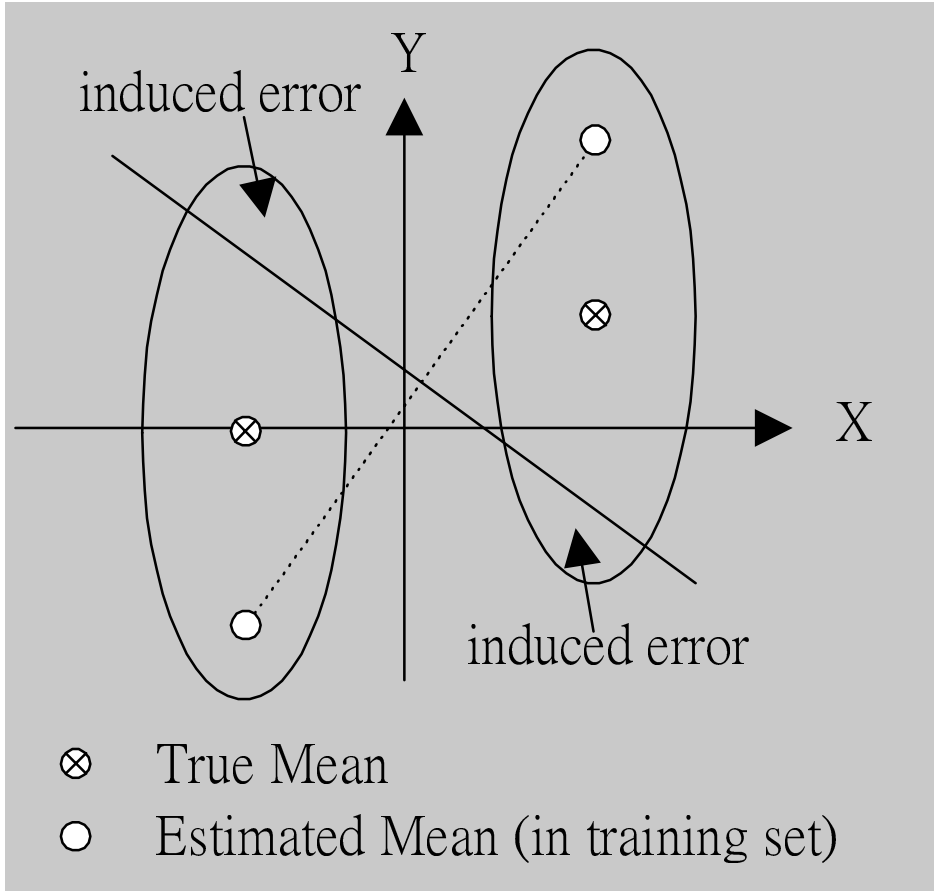
- Select Robust Features from the Initial Feature Set
  - Discrimination Information always increases through adding nontrivial features
    - Discrimination Information: $$L(\mathbf{q_0};\mathbf{q_1}) = \sum_{k=1}^{K} q_{0k} \log \frac{q_{0k}}{q_{1k}} = \sum_{k=1}^{K} q_{0k} l(k)$$
      - Expected value of log-likelihood ratio between two sets of probability vectors of observing features from two hypotheses H_0 and H_1
      - k: an observed feature, *l(k):* log-likelihood ratio for feature k
      - q_{ok}: the probability that k is from H_0, q_{1k}: the probability that k is from H_1
    - Symmetrical form is known as the Divergence: $L(\mathbf{q_o}; \mathbf{q_1}) + L(\mathbf{q_1}; \mathbf{q_o})$
    - The discrimination Information of each feature can be evaluated with the aid of the seed corpus (or the validation set)

Adopting Appropriate Feature Set (IV):

- Select Robust Feature Set (cont.)
  - However, some features are vulnerable (easy to be contaminated) and are considered harmful in the testing set ; therefore, they should be discarded
    - Features possessing Large Discrimination Power are more robust
      - Which are features that have large inter-classes distance & small intra-class variance
    - Discarding Non-discriminative Features to enhance robustness
      - by Sub-space Projection approach [Su and Lee 1994]
    - Use suitable method for Feature Selection (see appendix)
      - Cross-Validation set is used
      - Selection is performed based on the adopted objective measuring function (e.g., error rate, divergence, etc.)
      - Feature selection is actually executed through the adopted model; therefore, it is an iterative design process (feature decide the model; however, the model is also used to select features)

Select Robust Feature Set *via.* Subspace Projection

- Projecting observations into subspace to reduce error rate



Adopting Appropriate Form (I):

- Closely Echo Human Preference Network
  - Follow Human Preference Network and introduce required intermediate form
    - Example:

$$P\left(T_i|S_j\right) = \sum_{SubTree} P(T_i, SubTree|S_j)$$

  - Adopting Non-terminal Symbols to Handle Long-distance Dependency
    - Class-based Modeling: e.g., (NP, VP) versus N-gram Markov Chain
  - Drop Terms according to their Relevant Ranking
    - e.g., using Pearson's Chi-square Test for testing (and ranking) degree of independence, drop features that are most independent w.r.t. the outcome

Adopting Appropriate Form (II):

- Closely Echo Human Preference Network (cont.)
  - Example of adopting different feature dependencies:

$$P\left(x_1^n\right) = \prod_i P\left(x_i | x_1^{i-1}\right)$$

$$= \prod P\left(x_i | x_{i-(n-1)}^{i-1}\right) \qquad (n-gram)$$

$$= \prod P\left(x_i | y_{i-m+1}^{i-1}\right) \quad y_{i-m+1} \xrightarrow{predict} y_{i-m} \cdots \longrightarrow x_i \quad (causal\ chain)$$

$$y_j = H_j\left(x_{i-(m-1)}^{i-1}\right); \quad (non-terminal, or\ head\ \ features)$$

Adopting Appropriate Form (III):

- Closely Echo Human Preference Network (cont.)
  - Example: (use Conditional Independence assumptions carefully)

    If f_A and f_B are highly correlated, then formulate the following probability factor as:

$$P(f_A, f_B | WrdSense_i)$$
$$\approx P(f_B | f_A) \times P(f_A | WrdSense_i)$$

instead of simply assuming them to be conditional independent:

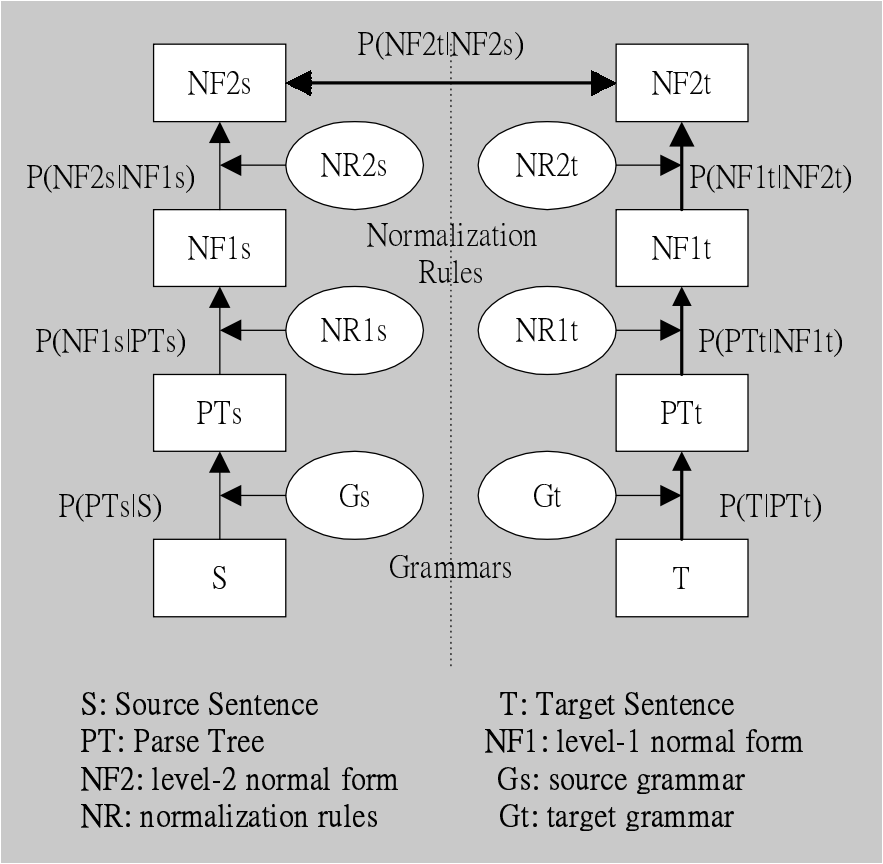
$$P(f_A, f_B | WrdSense_i)$$
$$\approx P(f_A | WrdSense_i) \times P(f_B | WrdSense_i)$$

Adopting Appropriate Form (IV)

- Embedding All Possible Constraints in the Model (to reduce task complexity)
  - Finding Correlated Resources as Training Corpus
  - Modeling with Implicit Constraints Embedded
  - Example: Adopting Bilingual Corpus
    - Build aligned bilingual sentence pairs
    - Sentence in each language side would help to impose the constrains on its corresponding sentence in other side
    - For example, in the task of Sense Disambiguation (in Source Side), the possible lexicon senses of each language is restricted by the possible sense set (listed in the dictionary) of the corresponding lexicon in another language

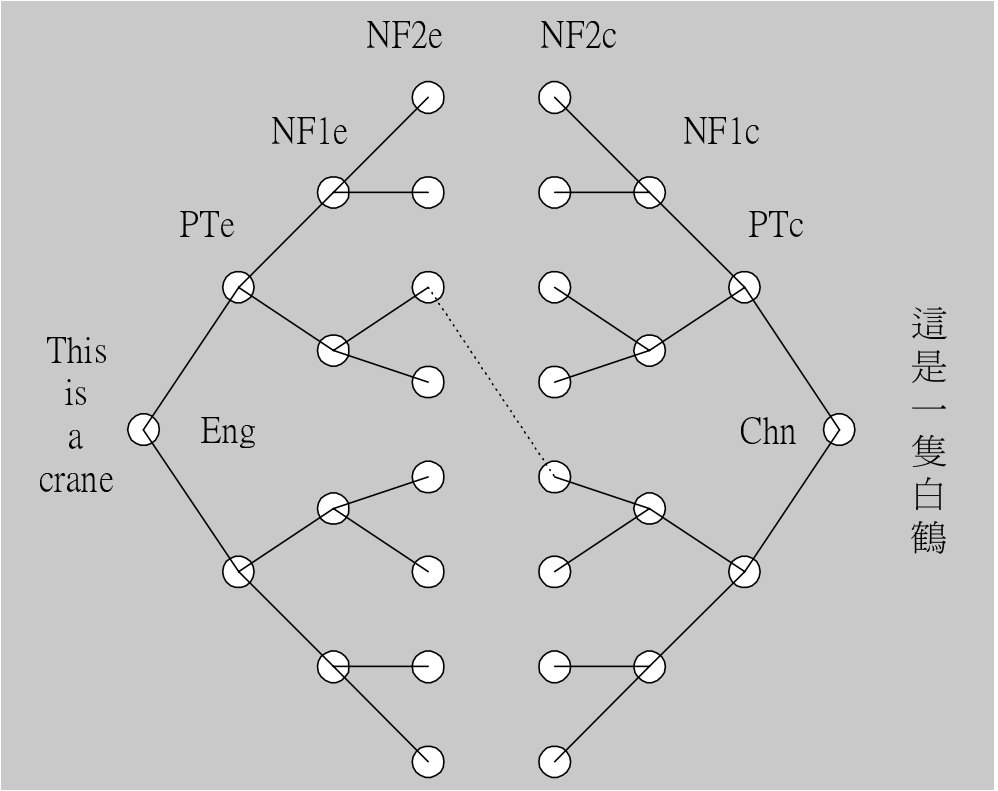
Embed Constraints through two-way learning (I)

- Architecture of a two-way training system



Embed Constraints through two-way learning (II)

- Example: the possible target words (such as, 白鶴) place constraints on possible senses of the source words “crane” through the given bi-lingual sentence pair



Embed Constraints through two-way learning (III)

- Original Optimizing Function (one way unsupervised learning)

$$\hat{\Lambda} = \arg\max_{\Lambda} \{ \max_{I} P(S_1^n, I | \Lambda) \}$$

- Modified Optimizing Function (two way unsupervised learning)

$$\overline{\Lambda} = \arg\max_{\tilde{\Lambda}} \{ \max_{\tilde{I}} P([S_i, T_i]_1^n, \tilde{I} \,|\, \tilde{\Lambda}) \},$$

$\overline{\Lambda}$ is in the subspace of that of $\tilde{\Lambda}$.

Adopting Appropriate Form (V):

- Eliminate the influence from those Non-discriminative parts by sharing the same sub-model
  - Select appropriate modeling unit
    - Use a larger context-sensitive unit, or adopt context-free compositional sub-units
    - Example: whole-parse-tree versus each production-rule, or syllable-model versus vowel-model in speech recognition
    - It is the issue of trading in the discrimination power for enhancing the robustness capability
  - Different non-discriminative sub-models can be tied together to form one sub-model; e.g., the vowel-part of the English E-set can be tied together
    - Models are tied by pooling their corresponding training data together

Adopting Appropriate Form (VI):

- Weight different knowledge sources (or parts) according to their contribution to the discrimination power
  - Different knowledge sources (such as lexical, syntactic, and semantic, etc.) have different dynamic ranges. They should be weighted according to their contribution to the discrimination power [Chiang 96]
  - Log-linear weighting model can be adopted
    - e.g., W_{lex} x Log P(Lex|Words) + W_{syn} x Log P(Syn|Lex, Words) ….
  - The weights could also be learned from the seed corpus as a set of parameters

Adopting Appropriate Form (VII):

- Adopt Class-Based Modeling (see appendix), if necessary, to avoid over-fitting
  - Class-based modeling also trade in the the discrimination power for the robustness
  - Better to be non-uniformly adopted
    - Use the detailed model when its corresponding training data is sufficient, and adopt the class-based model when its corresponding training data is not enough

Adopting Appropriate Form (VIII):

- Avoid Over-Fitting:
  - Using Cross-Validation Set to Select Appropriate Model Complexity
    - Compromise between the Size of the Training Corpus and the Model Complexity
    - For example, in the case of deciding the number of word-classes we should divide: the larger number of word-classes you choose, the bigger the maximum likelihood value you can obtain from the training set; however, it might even deteriorate the performance in the testing set
  - Using Non-uniform Model Complexity
    - Use a mixture of coarse (e.g., class-based) and refined models (e.g., lexicon-based)
    - Use refined models when the corresponding training data is sufficient, and adopt coarse models when we don't have enough training data

Educating Initial Guess

- Heuristically Guessing Initial Parameters
  - EM/Viterbi can be easily trapped at a local maximum that is not matching human preference
  - Try to provide a good starting point for leading the searching process to converge to the desired local maximum point
    - Using problem (or domain) knowledge to make heuristically guessing for the initial parameter set
    - Example: (1) PP-Attachment prefer minimum attachment and right association (=> assign heuristic guessing on left/right association, say as [0.3, 0.7], according to prior knowledge).
    - Example: (2) non-uniform initial state segmentation in E-Set Speech Recognition (assigning more states to the consonant part)

Educating Initial Guess (cont.)

- Bootstrapping with a Seed Corpus (see Appendix)
  - Adopting a seed corpus (annotated) to guide the learning process (unveil the human preference)
  - Compromise between Corpus Annotation Cost & Performance
  - Obtain hints of human preference, thus has a better chance to converge toward desired local maximum
  - Bootstrap in incremental stages
    - Avoid the effect of seed corpus to be overriden by the fluctuation resulted from the guessing of the large corpus to be mixed

Enhancing Discrimination Power

- Executing Adaptive Learning on Seed Corpus
  - However, avoid over-tuning
  - Multiple Modules Learning: [Chiang *et. al*, 96]
    - Independent Learning: parameters of different modules are learned independently to optimize individual criteria of module performance
    - Incremental Learning: modules are trained on-by-one; parameters of current module are trained to optimize the performance criteria of the system
    - Joint Learning: parameters of all modules are trained simultaneously to optimize the system performance criteria

Enhancing Discrimination Power (cont.)

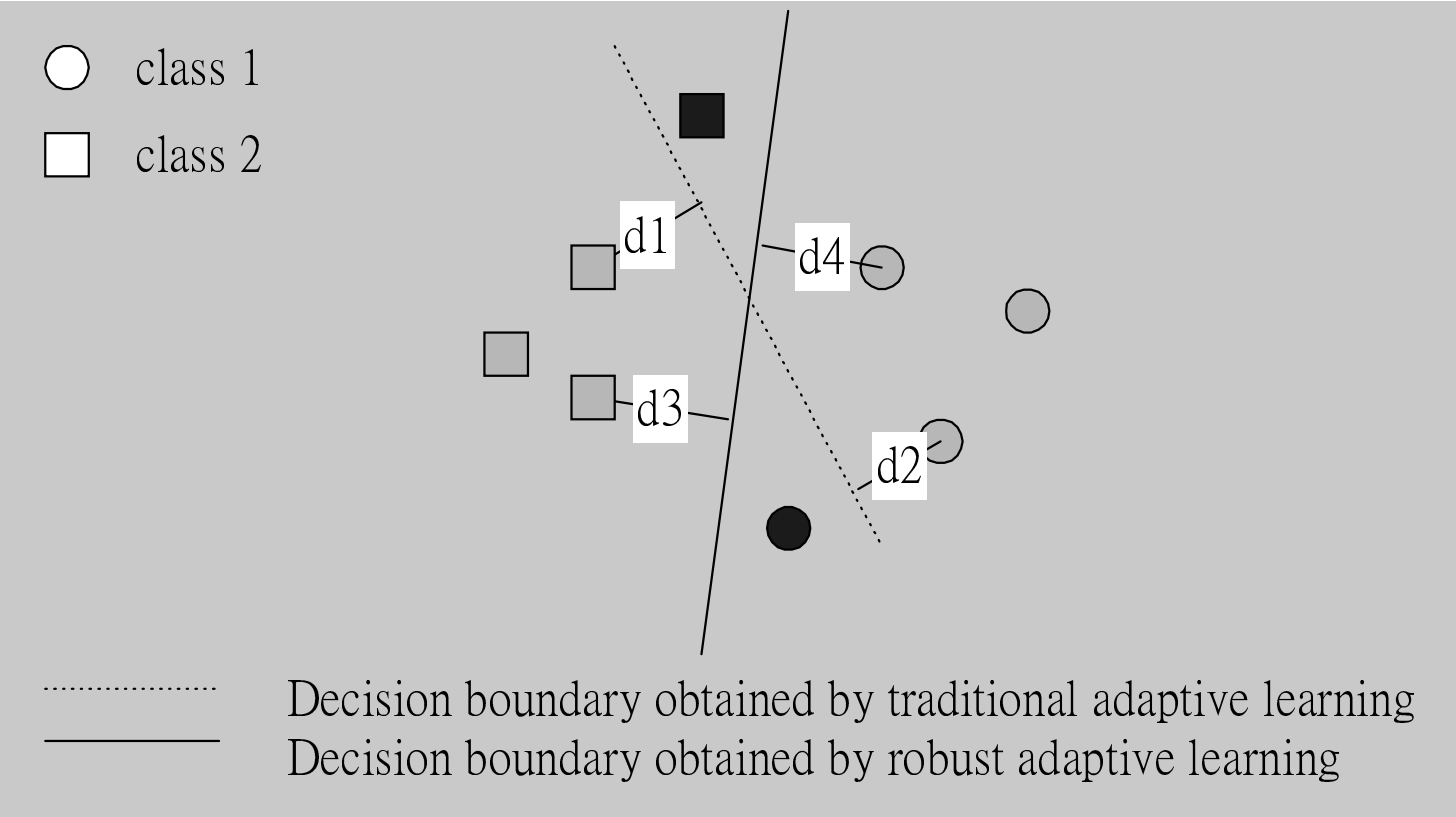
- Enhancing Learning Efficiency: Parameter Tying
  - Tie those rare and highly correlated events into a new class
  - Enlarge the training procedure coverage scope:
    - such rare parameters will be trained (instead of being ignored from training) when their correlated events are trained
  - Training Efficiency will be higher, as more percentages of the parameters will be better trained

Enhancing Robustness

- Smoothing
  - Managing those unseen/un-reliable/under-trained parameters
  - Back-off smoothing, interpolation (from multiple sources), and tying
- Enlarging Tolerance Margins between Correct Label and its corresponding Most-Competitive Candidate during the Adaptive Learning Procedure
  - Attempt to achieve the maximum separation between different classes
  - Provide the safety zone for tolerating possible statistical variation and data scatting over in the testing set

Achieve Maximum Separation

- Maximum Separation Classification [Su and Lee 1994]
  - Green: Training Set; Red: Testing Set
  - Traditional -- Training Set is separated BUT testing set is NOT
  - Robust -- Testing set is better improved by Maximum Separation



Suggested Unsupervised Learning Steps (I)

1. Develop Models that mostly Reflect Human Inference Behavior, Embed Constrains and Fit Training Data

   - Select Discriminative Features based on which human make preference
     - Must be jointly considered with the adopted form
   - Select Appropriate Form
     - Determine appropriate Feature Dependency
     - Decide suitable Model Complexity with a Cross-Validation Set
   - Determining desired Feature Space and Form is an Iterative Design Process

2. Initial Guess

   - Adopting Annotated Seed Corpus for Initial Model Parameters
   - Smoothing Parameters for Unseen Events (with respective to seed corpus) in Training Set before processing those un-annotated corpus

Suggested Unsupervised Learning Steps (II)

3. Re-generating Prediction According to New Model Parameters

- EM: Re-calculating the Expectation of Sufficient Statistic
- Viterbi: Re-labeling Corpus according to new Model Parameters

4. Re-Estimation of Model Parameters via MLE

- EM: Using the expectation (which implies that every possibility is considered)
- Viterbi: Using the guessed labels (only one possibility is considered)

5. Repeat the above Prediction and Estimation Steps until joint likelihood value of the training corpus converge

- to maximize the likelihood value

Suggested Unsupervised Learning Steps (III)

6. Conduct Discriminative/Robust Learning in Seed Corpus (Tying Parameters)
   - to compensate for criteria mismatch

7. Bootstrap Seed Corpus Incrementally Stage by Stage
   - training data is increased incrementally to avoid overriding the Seed

8. Using the Cross-Validation Set to Check the Effectiveness of Each Step
   - check whether the performance is starting to degrade

9. Iterate the above design procedures until you are satisfied
   - You should end up with a model that better reflects human preference with well-trained parameters