Part VI: An Example for Unsupervised Learning - Extracting Chinese New Words [Chang 97a, 97b]

- Task Definition
  - Generate potential word list from the given corpus
  - Optimization Criteria: improve precision and recall simultaneously
- System Architecture
  - Word segmentation: with contextual constraints
  - Cohesion judge: ranking module according to the likelihood values (two classes model with association features)
  - Two-stage iterative approaches to improve recall, in addition to improving precision
- Consideration in designing Unsupervised Learning
- Prospective Improvement for Unsupervised Learning

Task Definition: Extracting Chinese New Words

- Task: extract new words from the given un-segmented text corpus
  - **Input:**
    - An un-segmented Chinese Text Corpus, and a system dictionary of known words
  - **Output:**
    - Potential New Words in the Text Corpus (that were not in the system dictionary)
  - **Criteria:**
    - Improve joint precision-recall performance

Example of Compound Word Extraction

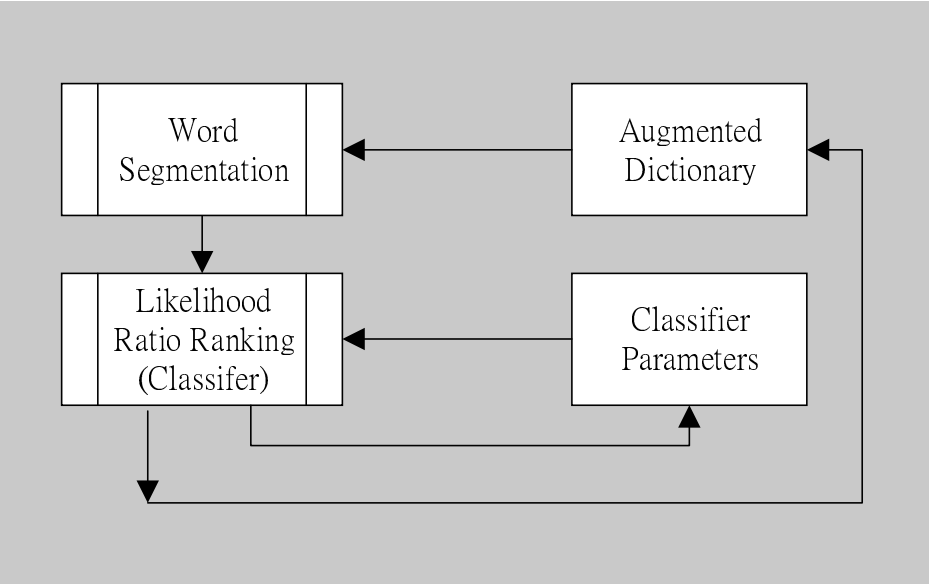
- China Times 1997/7/26:
  - **[台經院]**指出，隨著股市**[活絡]**與景氣**[回溫]**，第一季車輛及零件營業額成長十六.八一%，顯示民間需求**[回升]**。再加上爲加入ＷＴＯ，開放進口已是時勢所趨，也將帶動消費成長。 **[台經院]**預測今年民間消費全年成長率可提昇至六.七四%。
  - 在投資方面，第一季國內投資出現**[回升]**走勢， **[固定資本]**形成實質增加六.五六%，其中民間投資實質增加八.九五%。在持續有民間大型投資計畫進行、國內**[房市] [回溫]** 、與政府開放投資、加速執行公共工程等多項因素下，預測今年全年民間投資將成長十一.八%。
  - **[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。
- New words: proper names, jargons, lexicalized compounds, ...

Basic Language Models and System Architecture

- Two Modules
  - Word Segmentation Module:
    - A Viterbi Training module: to get best word segments
    - According to an augmented dictionary, i.e., the union of system dictionary plus high frequency character n-grams
  - Likelihood Ratio Test Module:
    - A two-class classification module: used to rank word candidates (in best segments) by likelihood ratio
    - Can also be used to determine whether an n-gram is a word, but not used in this manner

Basic Language Models and System Architecture

- Integration of the Modules
  - Iteratively apply word segmentation and use the relative rank information of the segments to improve the augmented dictionary for segmentation
    - improve the segmentation parameters and classifier parameters as well



Language Model for Word Segmentation (Viterbi Training)

- Segmentation Stage: Find the best segmentation pattern S*

$$S^*(V(t)) = \arg\max_{S_j} P(S_j = w_{j,1}^{j,m(j)} | c_1^n, V(t))$$

- which maximizes the following likelihood function of the input corpus

$$P(S_j = w_{j,1}^{j,m(j)} | c_1^n, V(t)) \approx \prod_{i=1,m(j)} P(w_{j,i} | V(t))$$

- c_1^n : input characters c_1, c_1, ..., c_n

- S_j : j-th segmentation pattern, consisting of { $w_{j,1}$, $w_{j,2}$, ..., $w_{j,m(j)}$ }

  - V(t): vocabulary (n-grams in the augmented dictionary) used for segmentation at the t-th iteration

  - S*(V): the best segmentation (is a function of V)

Language Model for Word Segmentation (cont.) (Viterbi Training)

- Re-estimation Stage: Estimate the word probabilities which maximize the likelihood of the input text:
  - Initial Estimation:

$$P(w_{j,i}|V) = \frac{Number(w_{j,i})\ in\ corpus}{Number\ of\ \ all\ w_{j,i}\ in\ corpus}$$

  - Reestimation:

$$P(w_{j,i}|V) = \frac{Number(w_{j,i})\ in\ best\ segmentation}{Number\ of\ \ all\ w_{j,i}\ in\ best\ segmentation}$$

Language Model for Two-Class Classifier (Log-Likelihood Ratio Ranking Module)

- **Input:** n-grams in the given un-segmented 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)

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

- **Decision Rules:**

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

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

- **NOTE:** We don't really use it for assigning class label when joining in unsupervised learning. Instead, the associated LLR's are used for sorting to identify **relative ranking order** of character n-grams, and hence it works as a ranking module.

Integration of Knowledge Sources

- Conventional System Schemes:
  - Segmentation (with known words) + Merge adjacent characters + Qualification with a filter
- Characteristics:
  - Independent knowledge sources, one-pass, non-iterative
    - Word Segmentation: Use contextual constraints (or contextual probabilities) to find the best segmentation
    - Filter: Use word association features (e.g., mutual information, dice) to filter out unlikely compound words
      - many filtering approaches filter out unlikely candidates in a feature-by-feature filtering manner, one feature one filtering step
    - No information sharing between the two modules

Problems with Segment-Merge-Filtering Schemes

- Merge-type errors cannot be recovered:
  - Types of errors: over-segmentation, under-segmentation (mis-merging)
  - **New words** may be merged with neighbors into **known words** in a system dictionary, and thus will not be extracted
    - Example: known word:土地公 & new word:公有
    - [土地公有政策] => [土地公][有][政策]
- **Simple filtering will *never* improve recall**
  - Successful filtering ⇒ precision improved, recall **unchanged**
  - Unsuccessful filtering ⇒ both precision and recall **degraded**

Problems with Segment-Merge-Filtering Schemes

- Association features not used jointly; instead, used independently
  - Worse than jointly considering all association features
- Information cannot be shared between word segmentation and filtering
  - Inherent contextual constraints cannot be used by filter
  - Word association features do not help select candidate word for segmentation module
- Model parameters are not improved iteratively
  - Performance of segmentation and filtering is unlikely to be perfect in only one pass with unsupervised mode

Strategies for Extracting Chinese New Words

- Strategies
  - Use augmented dictionary (system dictionary+high frequency n-grams)
    - to prevent from pre-mature rejection of new words by using only known words for segmentation
    - new words have the chance to compete with known words during segmentation
  - Iterative Approach to provide a chance for improving **recall**:
    - Word Segmentation → Qualification (→ Re-estimate Parameters) → Segmentation → Qualification (→ Re-estimate Parameters) …
    - Why: (See Next Slide)
  - Use a two-class classifier which jointly considering all features: likelihood ratio test
  - Use ranks of likelihood ratio to identify very likely or very unlikely candidates, instead of using the value for filtering out candidates with non-positive values
  - ☆ Filter => Likelihood Ratio Ranking Module (aka LRRM)

Extracting Chinese New Words

- Why Iterative ?
  - **Recall Improvement:** Truncated candidates could be replaced by other more likely segments (judged by contextual probability) at later segmentation iterations, thus extracting likely new words
    - ⇨ **Recall** could be improved, in addition to improving precision (by filtering)
    - ⇨ Joint improvement of precision-recall becomes possible
  - **Information Sharing:** Contextual probability used by Word Segmentation and association features used by filter help each other in improving the model parameters
    - WS: producing better segments iteration by iteration, highly probable new words are moved to the word-class, thus refine two-class classifier model
    - Filter: provide correct candidate ranking for truncating unlikely n-grams, thus improve the dictionary used by the word segmentation module
    - ⇨ Contextual information and Association features are iteratively integrated

Unsupervised Training for New Word Extraction

- **Initialization:**
  - Initial augmented dictionary = {system dictionary + high frequency n-grams in text (frequency count >=5)}
  - Initial word segmentation probability = relative frequency in text corpus
  - Initial two-class classifier parameters: divide n-grams into word & non-word according to system dictionary & estimate feature distribution for the two classes
- **Jointly train & improve two modules:** Word Segmentation+Ranking Module
  - LRRM: a two-class classifier, using likelihood ratio between word-class and non-word class to rank possibility of an n-gram being a word

Unsupervised Training for New Word Extraction (cont.)

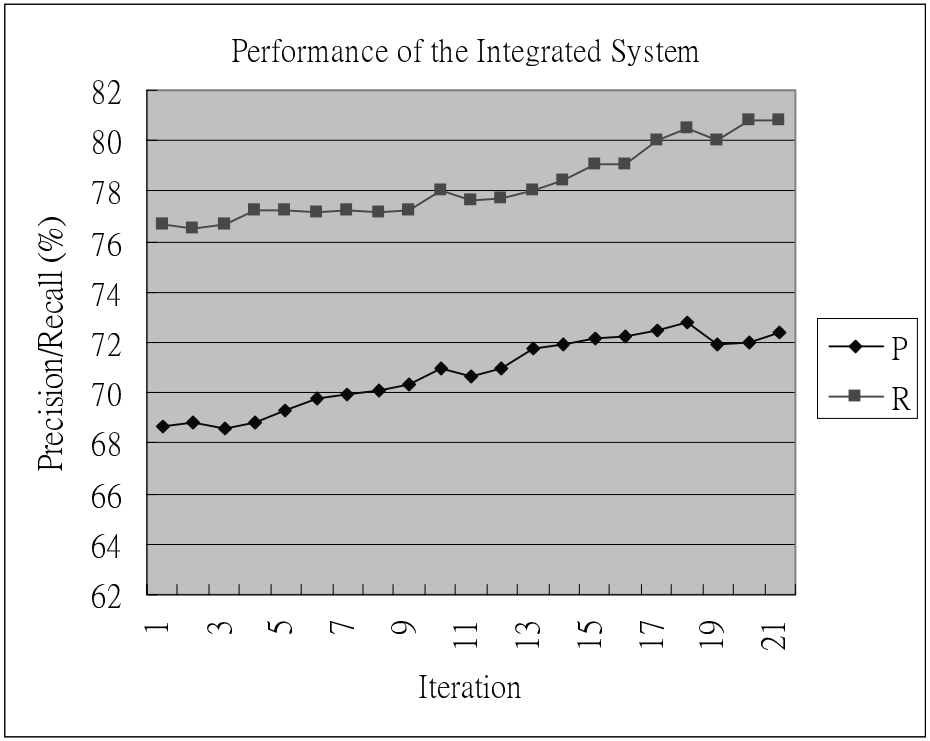
- **Jointly train & improve two modules (cont.)**
  - **Viterbi Training:** for Training Word Segmentation Module:
    - Use initial probabilities for finding the best word segments
    - Re-estimate word probabilities from best segments
    - **Repeat:** until converge or running a specified iterations
  - Sort word list in Word Segmentation results by Likelihood Ratio
  - Delete unlikely words (not in system dictionary) from augmented dictionary
  - **Update word/non-word class parameters of LRRM:** with highly likely new words (change the estimates to the word-class)
- **Repeat:** Joint Training to Iteratively improve the Viterbi-Training and LRRM modules

Extracting Chinese New Words (I)

- System Architecture:
  - See Figure 1 & 2 at the end of this part.

Extracting Chinese New Words (II)

- Results: precision and recall both increase almost monotonically without sacrificing one for another
  - bigram new word precision (P) & recall (R):



Comparison between Example & Suggested Steps (I)

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

- Select Discriminative Features based on which human make preference
  - Segmentation: uses character N-grams (could be integrated with POS tags)
  - Classifier: uses (mutual information, entropy) vector jointly, instead of using them as individual association measures for filtering candidates

Comparison between Example & Suggested Steps (II)

- 1. Develop Models (cont.)
  - Select Appropriate Form
    - Determine appropriate Feature Dependency:
      - segmentation model is known to be conditional on potential candidate list, thus it motivates us to design procedures for refining augmented dictionary, which is used for segmentation, iteratively
      - classifier was based on likelihood ratio test for minimum error rate
    - Decide suitable Model Complexity with Cross-Validation Set: Not applied to the two-class ranking module (since there are only two features)
      - Feature selection could be conducted as feature number increases
    - Integrate the two different knowledge source in an iteratively improved manner

Comparison between Example & Suggested Steps (III)

2. Initial Guess

- Adopting Annotated Seed Corpus for Initial Model Parameters: segmented seed corpus is currently not available; will be a plus if available (was applied in another task [Chang 95])
- Using a System Dictionary and high frequency n-grams as possible anchor points for word segmentation, and estimating segmentation parameters as relative frequency in un-segmented input
- Using the System Dictionary for dividing n-grams into two classes (word/non-word) for estimating initial classifier parameters
- Smoothing Parameters for Unseen Events (with respective to seed corpus) in Training Set: N/A

Comparison between Example & Suggested Steps (IV)

3. Re-generating Prediction According to New Model Parameters

- Viterbi-type labeling for word segmentation

4. Re-Estimation of Model Parameters via MLE

- Viterbi Training within word segmentation module

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

- within each joint training iteration of two module, the likelihood associated with the word segmentation module is maximized

Comparison between Example & Suggested Steps (V)

6. Conduct Discriminative/Robust Learning in Seed Corpus (Tying Parameters)

- currently not applied; using a segmented seed for adjusting segmentation parameters and the two-class parameters would be helpful

7. Bootstrap Incrementally Stage by Stage

- not applied in this task; may better utilize the seed if applied

Comparison between Example & Suggested Steps (VI)

8. Using the Cross-Validation Set to Check the Effectiveness of Each Step

   - not applied

9. Iterate the above design procedures until you are satisfied

   - using a two-stage iterative approach to integrate the two modules did provide a system that meets our expectation: improving precision and recall simultaneously without trading one for another

Refinement to the Unsupervised Learning Procedure

- Possible Future Refinement
  - Feature set: joint more useful association measures in the classifier (ranking module), including feature selection mechanism for the best subset; using tagging information to help segmentation, etc.
  - Initial Guess: will be a plus with segmented seed corpus; applying smoothing to the initial parameters could be helpful as well
  - Discriminative/Adaptive Learning on Seed: could be applied to adjust the parameters to get maximize precision and recall (in terms of weighting sum or F-measure) as we did for English compound extraction [Chang 97]
  - Bootstrapping: Incrementally enlarge the training size with the seed fixed would possibly leads us to a better set of parameters

Open Discussion

- Other factors affecting unsupervised learning performance ?
- Other methods to improve performance ?
- Other methods to mining implicit constraints, anchor points, … ?
- Other automatic or semi-automatic methods to reduce annotation costs & provide useful seed ?