Computation, Information, and Intelligence (ENGRI/CS/INFO/COGST 172), Spring 2007
4/18/07: **Lecture 35 aid — Word segmentation**

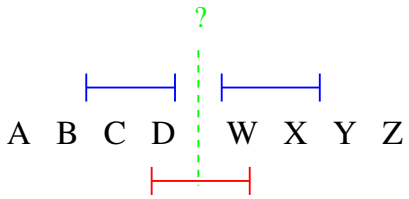
Topics: A mostly-unsupervised approach to the word segmentation problem, following R. K. Ando and L. Lee (2003). The question is whether simple statistics drawn from a large enough data-set can be used to accomplish a difficult language processing task.

I. Example sequence of Japanese kanji

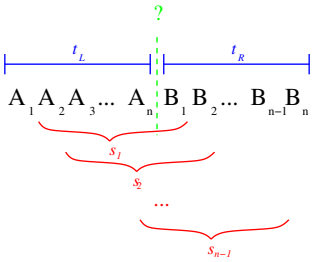
社長兼業務部長

II. N-gram evidence

(Character-level) bigram evidence considers the following situation:



The general n-gram situation looks like this:



for d := L, R
  for j := 1, 2, ..., n–1
    is #(t_d) > #(s_j)?

where we ask, for each choice of *tangent* n-gram t_L and t_R and for each choice of *straddling* n-gram $s_1, s_2, \ldots, s_{n-1}$, is $\#(t_d) > \#(s_j)$?

(OVER)

III. Evidence combination We use a "senatorial" system. Suppose we are looking at position i, and are only choosing block lengths from some fixed set N.

1. For each n in N, calculate the average number of "yes" votes among the $2 \times (n-1)$ n-gram comparisons.
2. The final vote $V(i, N)$ is the average of these averages.

IV. Making segmentation decisions

Draw a boundary if the evidence (plotted as a red line) for a location is either a *local maximum* (this induces the green boundaries) or, failing that, *above a threshold* (this induces the magenta boundary).

V. Evaluation metrics

- Precision: What percentage of what you thought were words were really words?
- Recall: What percentage of the real words did you mark as words?
- F: combines precision and recall: F = 2PR/(P+R)

VI. Word-level accuracy results Training data: 37 million characters worth of unsegmented kanji sequences from 1993 NIKKEI newswire, plus about 50 segmented sequences (representing roughly eight minutes of work); the latter is used for parameter setting (N and t).

The two algorithms on the left are two state-of-the-art (at the time) systems based on hand-crafted grammars and dictionaries containing 115,000 or 231,000 entries, respectively.