Weakly-Supervised Statistical Segmentation of Japanese

Rie Kubota Ando and Lillian Lee

NAACL 2000 version (8 pages):

http://www.cs.cornell.edu/home/llee/papers/segmentnaacl.home.html

Natural Language Engineering 2003 version ($\sim$ 20 pages):

http://www.cs.cornell.edu/home/llee/papers/segmentjnle.home.html

Segmentation

Japanese, Chinese, Thai, ...: no spaces between words

社長兼業務部長

We would like to segment character sequences into words.

▷ IR, NLP applications.

Japanese Character Types

1. Katakana: transliterations of borrowed words
2. Hiragana: closed-class words, markers, etc.
3. Kanji: proper nouns, domain terms, technical vocabulary

Problem: unknown words, particularly in technical domains

Japanese Word Segmentation

We concentrate on long sequences of *kanji* (technical vocabulary). Cf. [Teller & Batchelder 94, Tomokiyo & Ries 97].

Standard approaches:

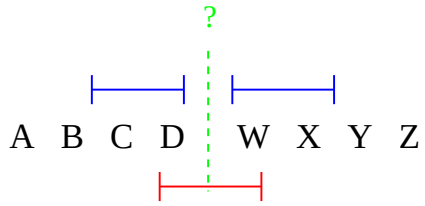
- Utilize dictionaries [Yamron et al 93; Matsumoto & Nagao 94; Nagao & Mori 94; Takeuchi & Matsumoto 95; Nagata 97; Fuchi & Takagi 98]
- Infer rules from *pre-segmented* training data (1000-190,000 sentences) [Nagata 92,96; Papageorgiou 94; Kashioka et al. 98; Mori & Nagao 98]

Alternative approach: rely mostly on simple statistics in *unsegmented* data.

- Little human effort involved $\Rightarrow$ more data can be used
- No lexicon $\Rightarrow$ "unknown" words not a (special) problem
- No Japanese-specific heuristics $\Rightarrow$ greater portability

Segmentation Method

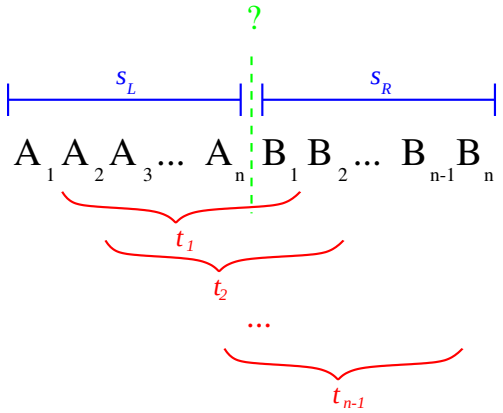
Consider evidence from n-character sequences.



Is $\#(\text{C D}) > \#(\text{D W})$ in the unsegmented data?

Is $\#(\text{W X}) > \#(\text{D W})$ in the unsegmented data?

Each "yes": vote for splitting



for d := L, R
  for j := 1, 2, ..., n-1
    is #(s_d) > #(t_j)?

Combining Evidence

Issue: more $(n+1)$-gram questions than n-gram questions.

"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.

The TANGO algorithm

Let $V(i,N)$ denote support for a word boundary at position i.

Place boundary at i if $V(i,N)$ a local max, or greater than threshold t
(Thresholds ANd maximums for NGrams that Overlap):

cf. [Nagao-Mori 94], [Sun et al. 98]

[Saffran, Aslin, Newport 1996] on segmentation of syllable sequences into wo

A small amount of annotated data is needed to learn parameter values.

Experimental Framework

Data: 37M kanji characters from '93 NIKKEI newswire.

Five train-development-test splits:

- Test: 450 long sequences, hand-segmented
- Development: ≤ 50 long sequences, hand-segmented, disjoint from test
- Training: remainder, unsegmented

Parameter training: all combinations of 2-grams through 6-grams, grid search on threshold.

Baseline Algorithms

State-of-the-art morphological segmenters with hand-crafted grammars.

Chasen 1.0 [Matsumoto et al. 97]: 115K lexicon entries

Juman 3.61 [Kurohashi and Nagao et al. 98]: 231K lexicon entries

("Training": add words in segmented data to lexicon)

Precision and Recall

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)

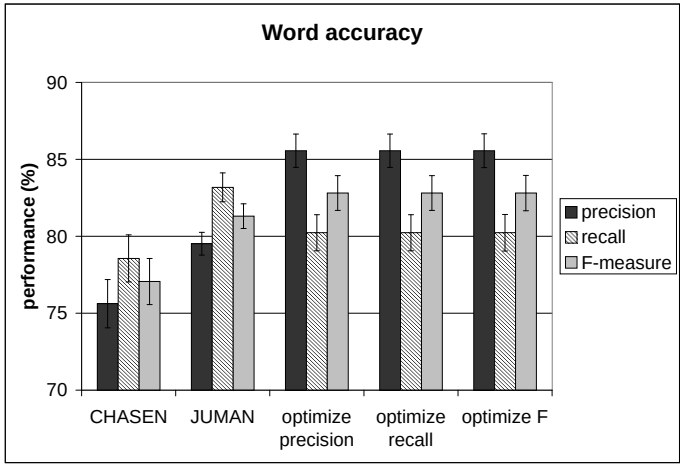
Example:

True: |this|is|a|sentence|

Output: |this|i|s|a|sen|tence|

Precision: 2/6 = 33%

Recall: 2/4 = 50%



Contribution of Segmentation Conditions

Do we need both the local maximum (M) *and* the threshold condition (T)?

Yes!

	optimize precision	optimize recall	optimize F-measure
word	M	M & T	M
morpheme	M & T	T	T

Other Related Work

- High-frequency character n-grams: Nagao & Mori 94, Ito & Kohda 95.
- Thresholds and maxima of complex count-difference statistics: Sun, Shen, & Tsou 98.
- More sophisticated models and learning methods: e.g., Kazama, Miyao, & Tsujii 01

Much more work on Chinese.

Summary

We have presented a mostly-unsupervised algorithm for segmentation.

- Results rival morphological analyzers
- Very little annotated data is needed
- Simple statistics can be effective