Weakly-Supervised Statistical Segmentation of Japanese

Rie Kubota Ando and Lillian Lee
Cornell University
{kubotar,llee}@cs.cornell.edu

NAACL 2000

Knowledge-Based Approaches

Typical approach: morphological/grammatical analysis using handcrafted lexicons (dictionaries) and grammars.

May incorporate character-based heuristics: change in character type may indicate boundary.

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

Problem: unknown words (especially in technical domains)

Learning approaches

- Supervised approaches [Nagata 92,96; Papageorgiou 94; Kashioka et al. 98; Mori & Nagao 98]
  - ▷ Learn from pre-segmented training data (1000-190,000 sentences)
- Unsupervised approaches using knowledge bases [Yamron et al 93; Matsumoto & Nagao 94; Nagao & Mori 94; Takeuchi & Matsumoto 95; Nagata 97; Fuchi & Takagi 98]
  - ▷ Bootstrap from handcrafted rules and/or large lexicons

Both require substantial human effort.

Our Approach

We mostly rely on *simple statistics* from an *unsegmented* corpus; very small (5-50 sequence) pre-segmented training sets are used.

Advantages:

- 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

Focus: Long Kanji Sequences

These are important: contain domain terms and technical vocabulary, and make up a substantial portion of newswire text (e.g. 22% of 1993 NIKKEI).

But, these types of strings are also the most difficult to segment [Takeda & Fujisaki 87]:

- unknown word problem
- compound noun sequences (part-of-speech info doesn't help)
- character-type heuristics obviously don't help

Two unsupervised approaches [Teller & Batchelder 94, Tomokiyo & Ries 97] explicitly avoid kanji.

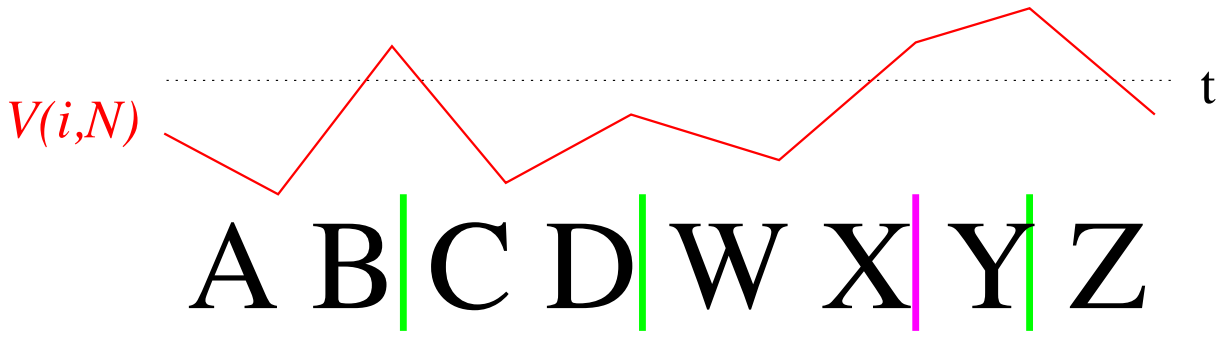
– new handwritten slide here: “algorithm” in generality

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]: frequency dips (never implemented)

[Sun et al. 98]: first and second local maxima in t-score differences and mutual information threshold

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.

Annotation Scheme

Two-level word and morpheme bracketing [Takeda & Fujisaki 87]

- Word = stem + affixes $\approx$ discourse entity
- Morpheme: affixes separated; $\approx$ smallest unit w/out meaning change

Three native speakers participated with high agreement rates (all disagreements at the morpheme level)

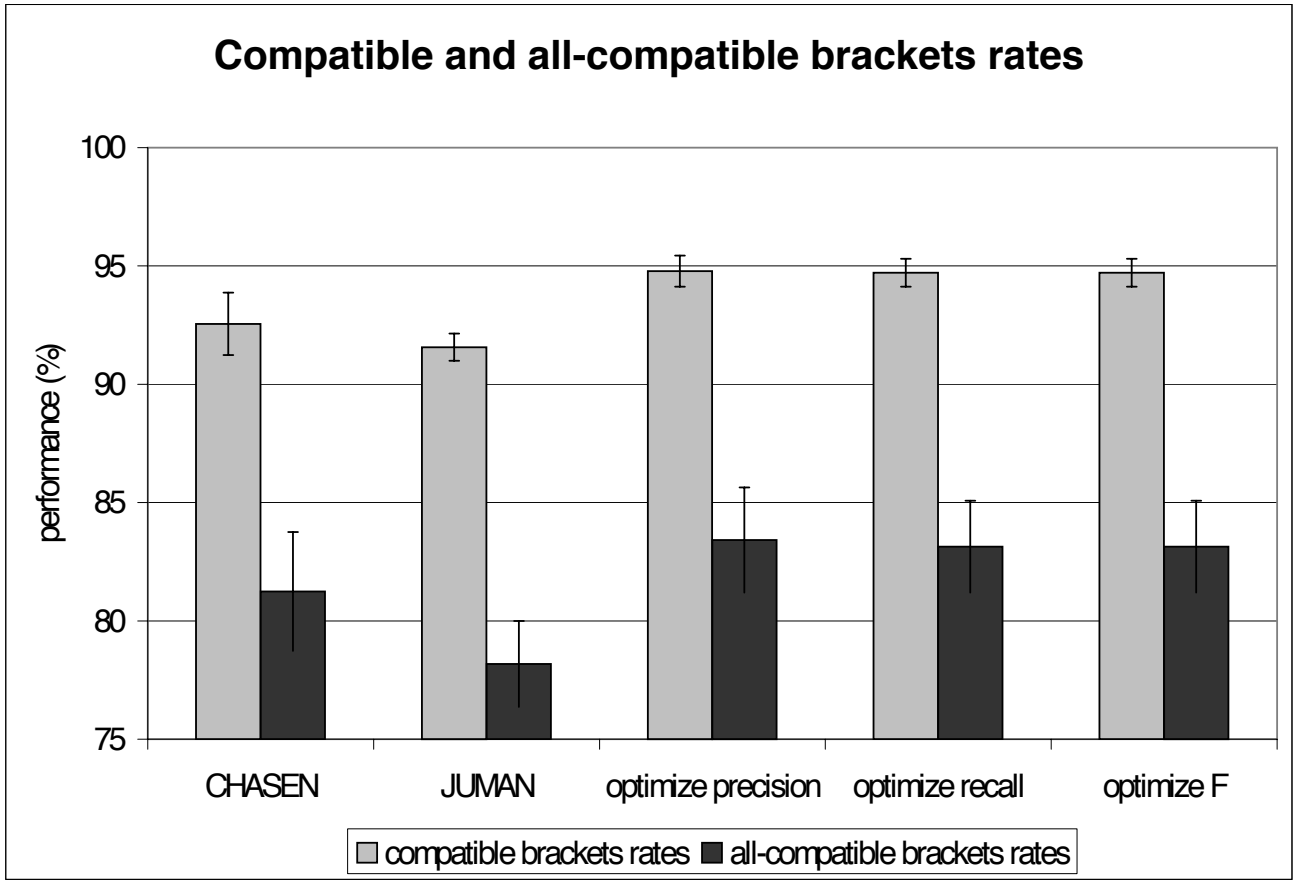
Compatible Brackets

Counts only egregious errors to compensate for different segmentation policies:

- any traditional crossing bracket (xb)
- any morpheme-dividing bracket (md)

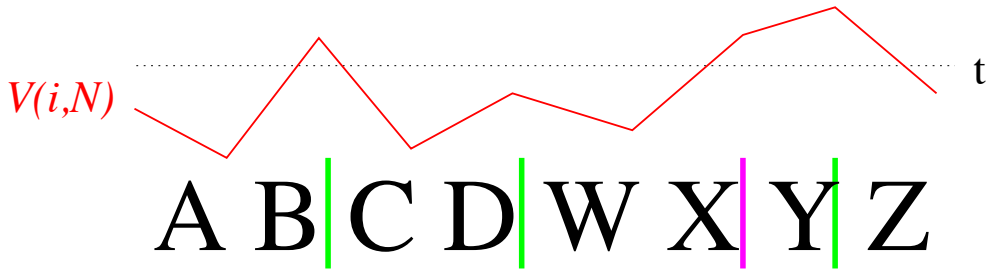
	Word errs (prec.,rec.)	Morpheme errs (prec.,rec.)	Comp.-bracket errs: (crossing,morph.-div'g)
[[data][base]][system]	-	-	-
ldatabaselsysteml	0,0	1,2	0, 0
ldatalbaselsysteml	2,1	0,0	0, 0
ldatalbasesysteml	2,2	1,2	1, 0
ldatabaselsyslteml	2,1	3,3	0, 2

Compatible Brackets Performance



Note: cannot optimize for compatible brackets directly.

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

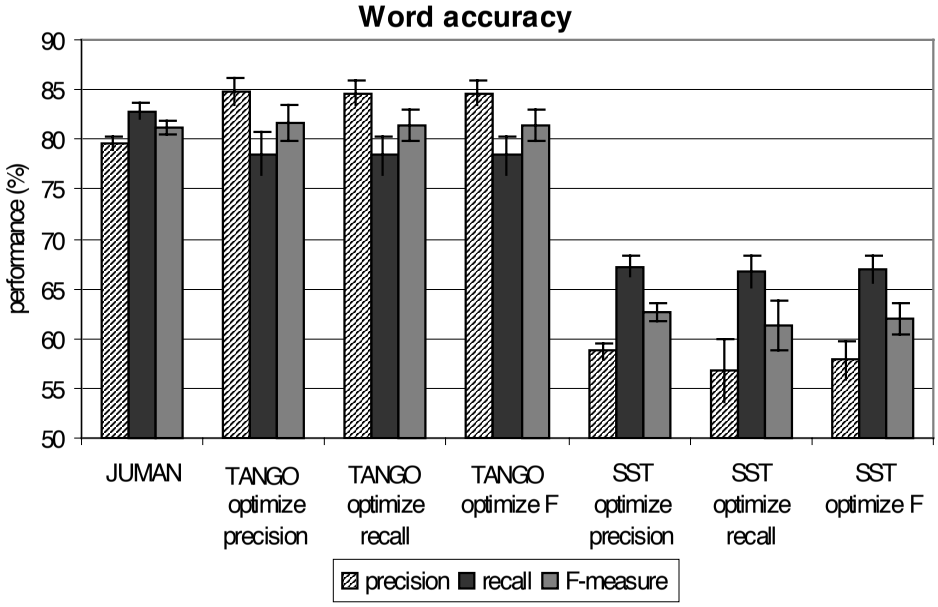
Other Related Work

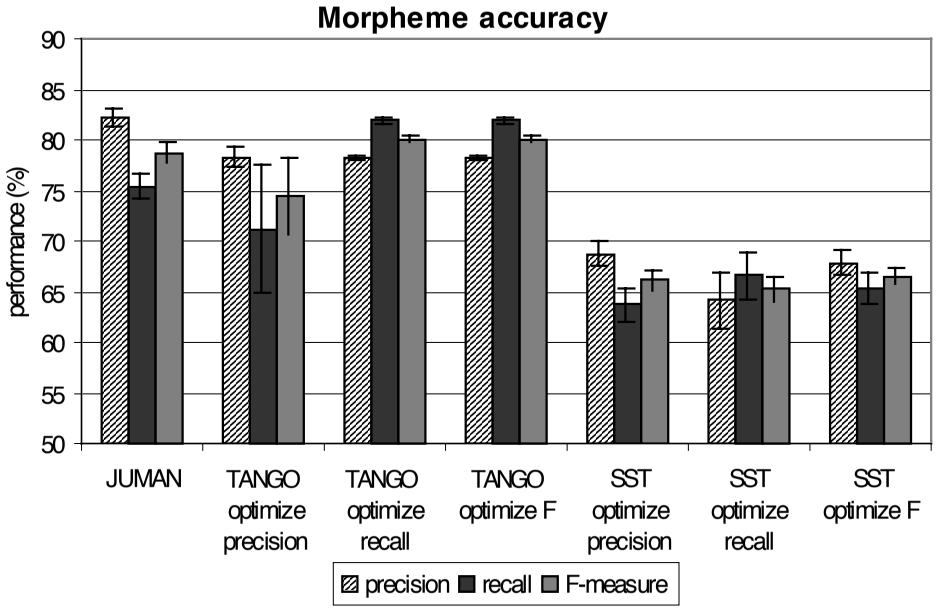
High-frequency character n-grams: Nagao & Mori 94, Ito & Kohda 95.

Thresholds and maxima of count statistics: Sun, Shen, & Tsou 98.

- More complex statistics incorporating numerical differences

Comparison with Sun et al.





Summary

We have presented a mostly-unsupervised algorithm for segmentation.

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