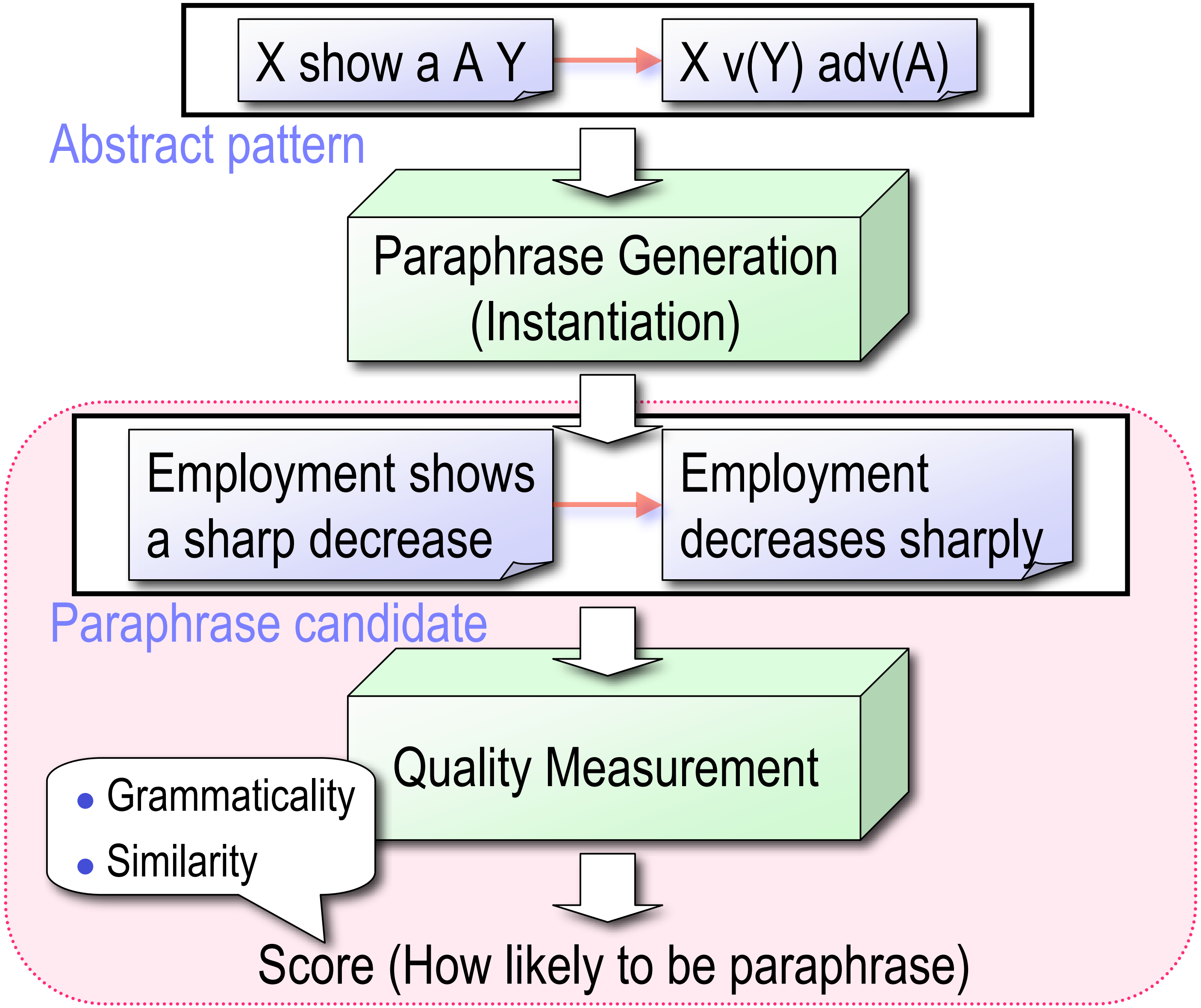
< COLING 2008, Aug. 19th, 2008 >

# A Probabilistic Model for Measuring Grammaticality and Similarity of Automatically Generated Paraphrases of Predicate Phrases

Atsushi FUJITA and Satoshi SATO

Nagoya Univ., Japan

# Overview


X show a A Y
X v(Y) adv(A)
Abstract pattern
Paraphrase Generation (Instantiation)
Employment shows a sharp decrease
Employment decreases sharply
Paraphrase candidate
Quality Measurement
• Grammaticality
• Similarity
Score (How likely to be paraphrase)

# Automatic Paraphrasing

- Fundamental in NLP
  - Recognition: IR, IE, QA, Summarization
  - Generation: MT, TTS, Authoring/Reading aids
- Paraphrase knowledge
  - Handcraft
    - Thesauri (of words) [Many work]
    - Transformation rules [Mel'cuk+, 87] [Dras, 99] [Jacquemin, 99]
  - Automatic acquisition
    - Anchor-based [Lin+, 01] [Szpektor+, 04]
    - Aligning comparable/bilingual corpora [Many work]

# Representation of Paraphrase Knowledge

Fully-abstracted

[Harris, 1957]

- X V Y → X's V-ing of Y — Nominalization
- X V Y → Y be V-PP by X — Passivization
- X show a A Y → X v(Y) adv(A) — Removing light-verb

[Lin+, 2001]

- X wrote Y → X is the author of Y
- X solves Y → X deals with Y

[Barzilay+, 2001]

- burst into tears → cried
- comfort → console

Fully-lexicalized

# Instantiating Phrasal Paraphrases

- Over-generation leads to spurious instances
  - cf. filling arguments [Pantel+, 07]
  - cf. applying to contexts [Szpektor+, 08]

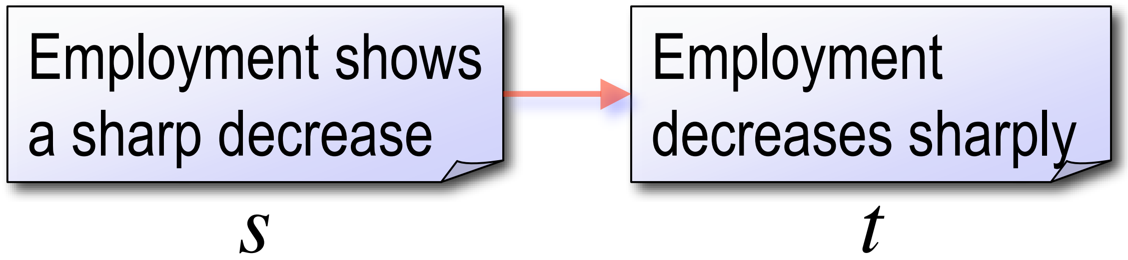
X show a A Y → X v(Y) adv(A)

| | | |
|---|---|---|
| Employment shows a sharp decrease → | Employment decreases sharply | OK |
| Statistics show a gradual decline → | Statistics decline gradually | Not equivalent |
| The data show a specific distribution → | The data distribute specifically | Not grammatical |

# Task Description

- Measuring the quality of paraphrase candidate

  **Input**: Automatically generated phrasal paraphrases




  **Output**: Quality score [0,1]

# Quality as Paraphrases

- Three conditions to be satisfied
  1. Semantically equivalent
  2. Substitutable in some context
  3. Grammatical
- Approaches
  - Acquisition of instances
    - 1 and 2 are measured, assuming 3
  - Instantiation of abstract pattern (our focus)
    - 1 and 2 are weakly ensured
    - 3 is measured, and 1 and 2 are reexamined

# Outline

1. Task Description
2. **Proposed Model**
3. Experiments
4. Conclusion

# Proposed Model

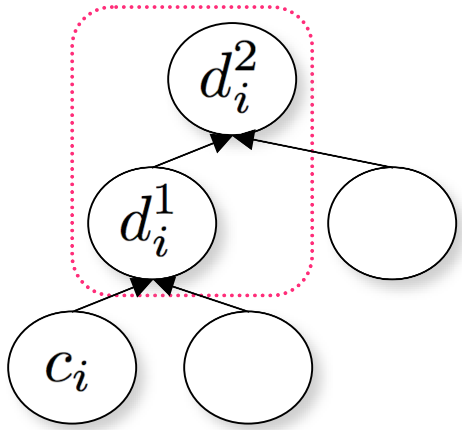
- Assumptions
  - $s$ is given and grammatical
  - $s$ and $t$ do not co-occur
- Formulation with a conditional probability

$$
\begin{aligned}
P(t|s) &= \sum_{f \in F} P(t|f)P(f|s) \\
&= \sum_{f \in F} \frac{P(f|t)P(t)}{P(f)} P(f|s) \\
&= \underbrace{P(t)}_{\text{Grammaticality}} \underbrace{\sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)}}_{\text{Similarity}}
\end{aligned}
$$

# Grammaticality Factor

- Statistical Language Model
  - Structured *N*-gram LM
  - Normalized with length

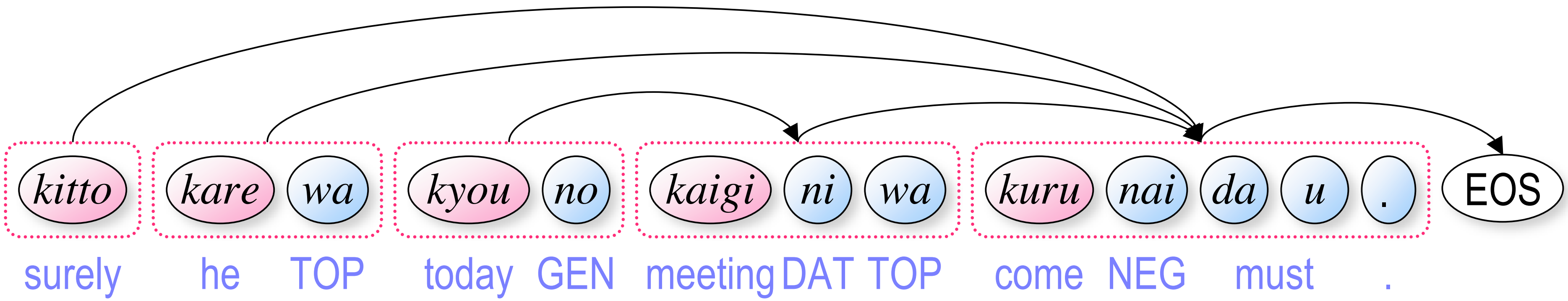
History of $c_i$




$$P(t) = \left[ \prod_{i=1\ldots|T(t)|} P_d(c_i|d_i^1, d_i^2, \ldots, d_i^{N-1}) \right]^{1/|T(t)|}$$

# Grammaticality Factor: Definition of Nodes

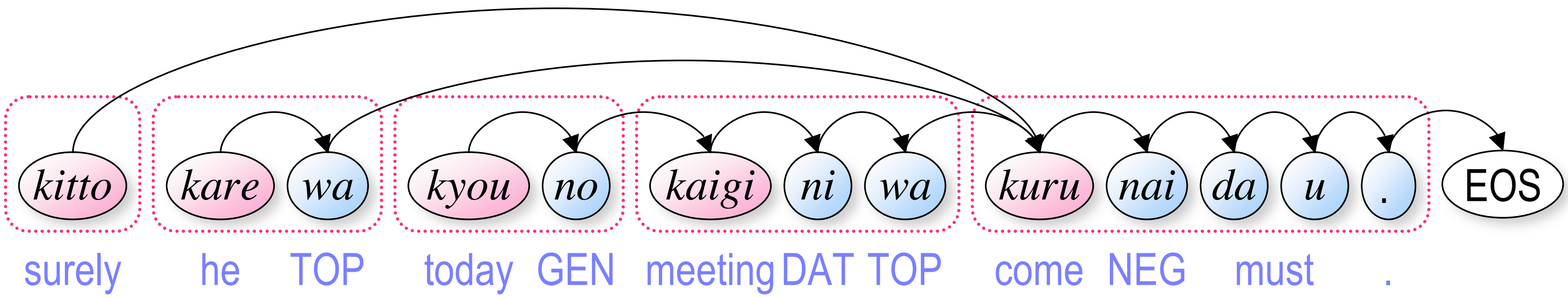
- For Japanese
  - What present dependency parsers determine
    - *Bunsetsu*: {Content word} + {Function word} *
    - *Bunsetsu* dependencies
  - *Bunsetsu* can be quite long (so not appropriate)




(He will surely not come to today’s meeting.)

# Grammaticality Factor: MDS

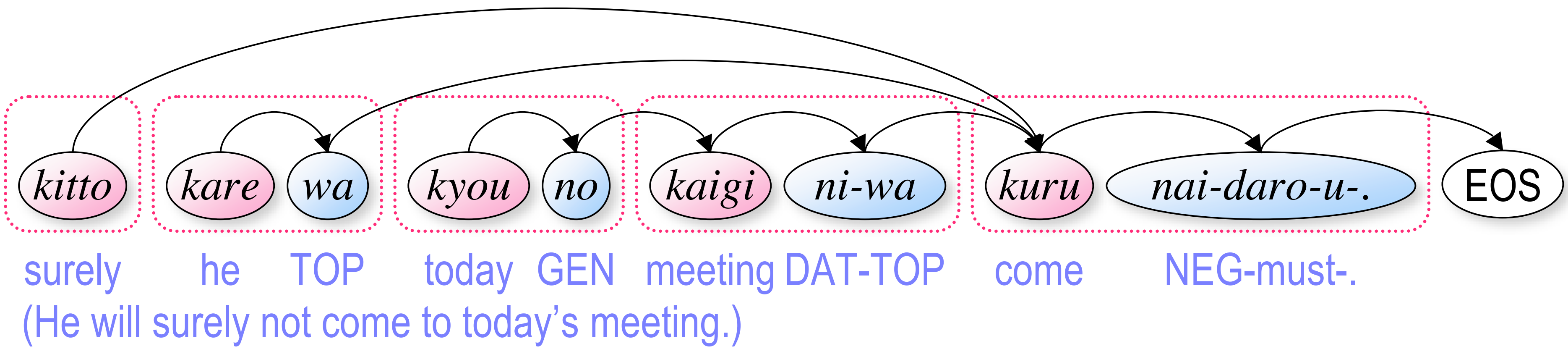
- Morpheme-based Dependency Structure [KURA, 01]
  - Node: Morpheme
  - Edge:
    - Rightmost node → Head-word of its mother *bunsetsu*
    - Other nodes → Succeeding node




(He will surely not come to today's meeting.)

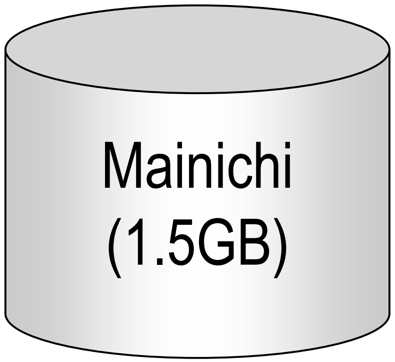
## Grammaticality Factor: CFDS

- Content-Function-based Dependency Structure
  - Node: Sequence of content words or of function words
  - Edge:
    - Rightmost node → Head-word of its mother *bunsetsu*
    - Other nodes → Succeeding node



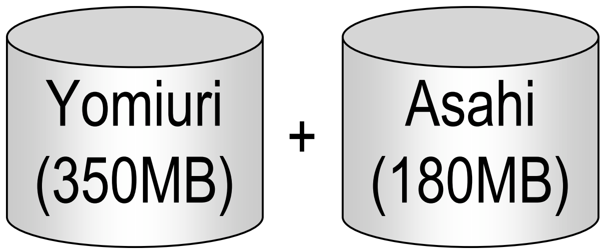
# Grammaticality Factor: Parameter Estimation

- MLE for 1, 2, and 3-gram models




| Node Type | # of alphabets |
|---|---:|
| MDS | 320,394 |
| CFDS | 14,625,384 |
| *Bunsetsu* | 19,507,402 |

- Linear interpolation of 3 models
  - Mixture weights were determined via an EM



# Similarity Factor

- A kind of distributional similarity measure

$$\sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)}$$

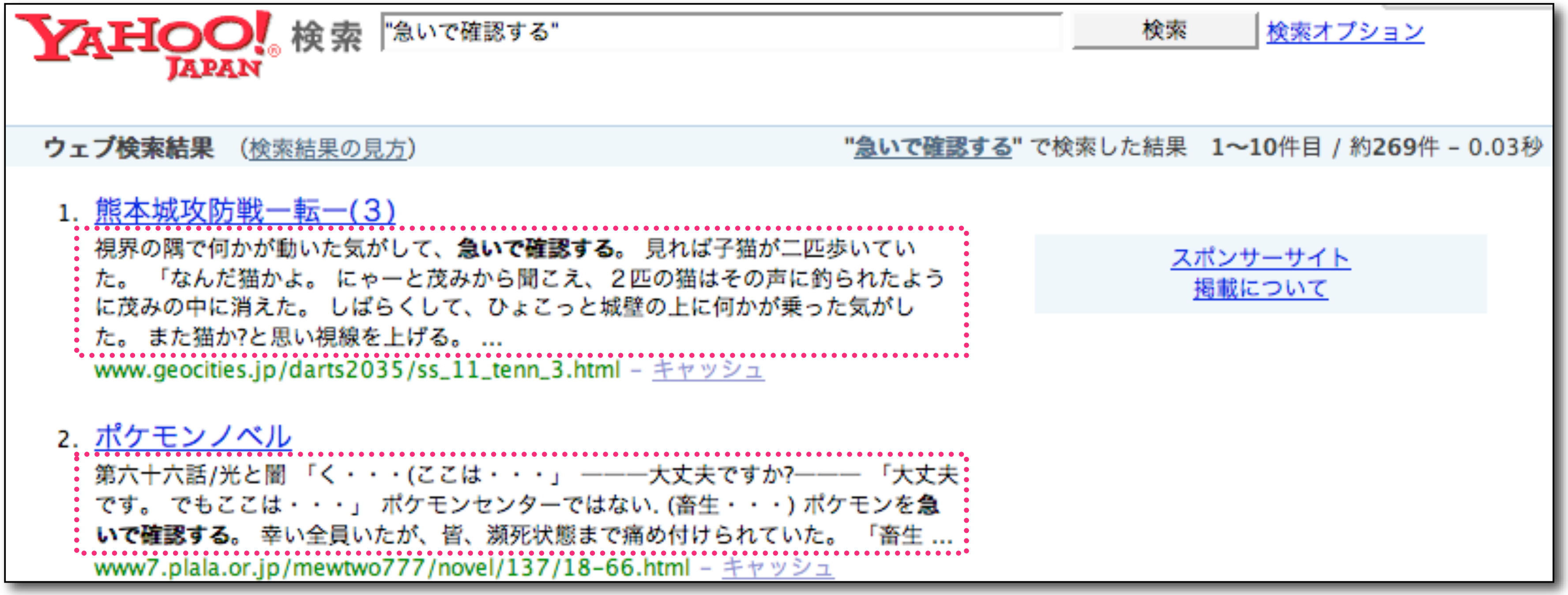
- Contextual feature set ($F$)

  **BOW**: Words surrounding $s$ and $t$ have similar distribution
  $\Rightarrow$ $s$ and $t$ are semantically similar

  **MOD**: $s$ and $t$ share a number of modifiers and modifiees
  $\Rightarrow$ $s$ and $t$ are substitutable

# Similarity Factor: Parameter Estimation

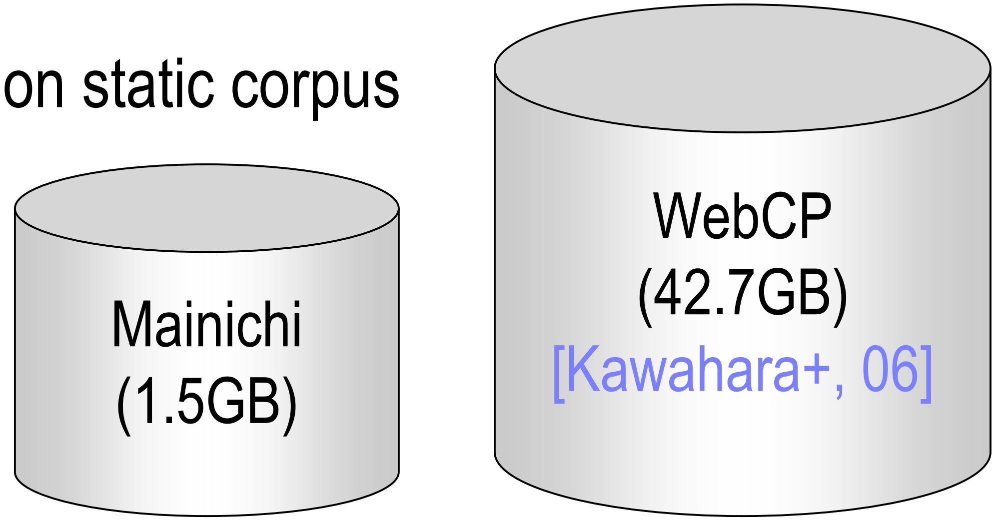
- Employ Web snippets as an example collection
  - To obtain sufficient amount of feature info.
  - Yahoo! JAPAN Web-search API
    - “Phrase search”
    - 1,000 snippets (as much as possible)



# Similarity Factor: Parameter Estimation (cont'd)

- MLE
  - $P(f|p)$
    - Based on snippets


  - $P(f)$
    - Based on static corpus



# Summary

- What is taken into account
  - Grammaticality of $t$
  - Similarity between $s$ and $t$
- You do not need to enumerate all the phrases
  - cf. $P(ph \mid f)$, $pmi(ph, f)$
- Options

$$P(t|s) \quad = \quad \underbrace{P(t)}_{\text{Grammaticality}} \underbrace{\sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)}}_{\text{Similarity}}$$

MDS / CFDS

BOW / MOD

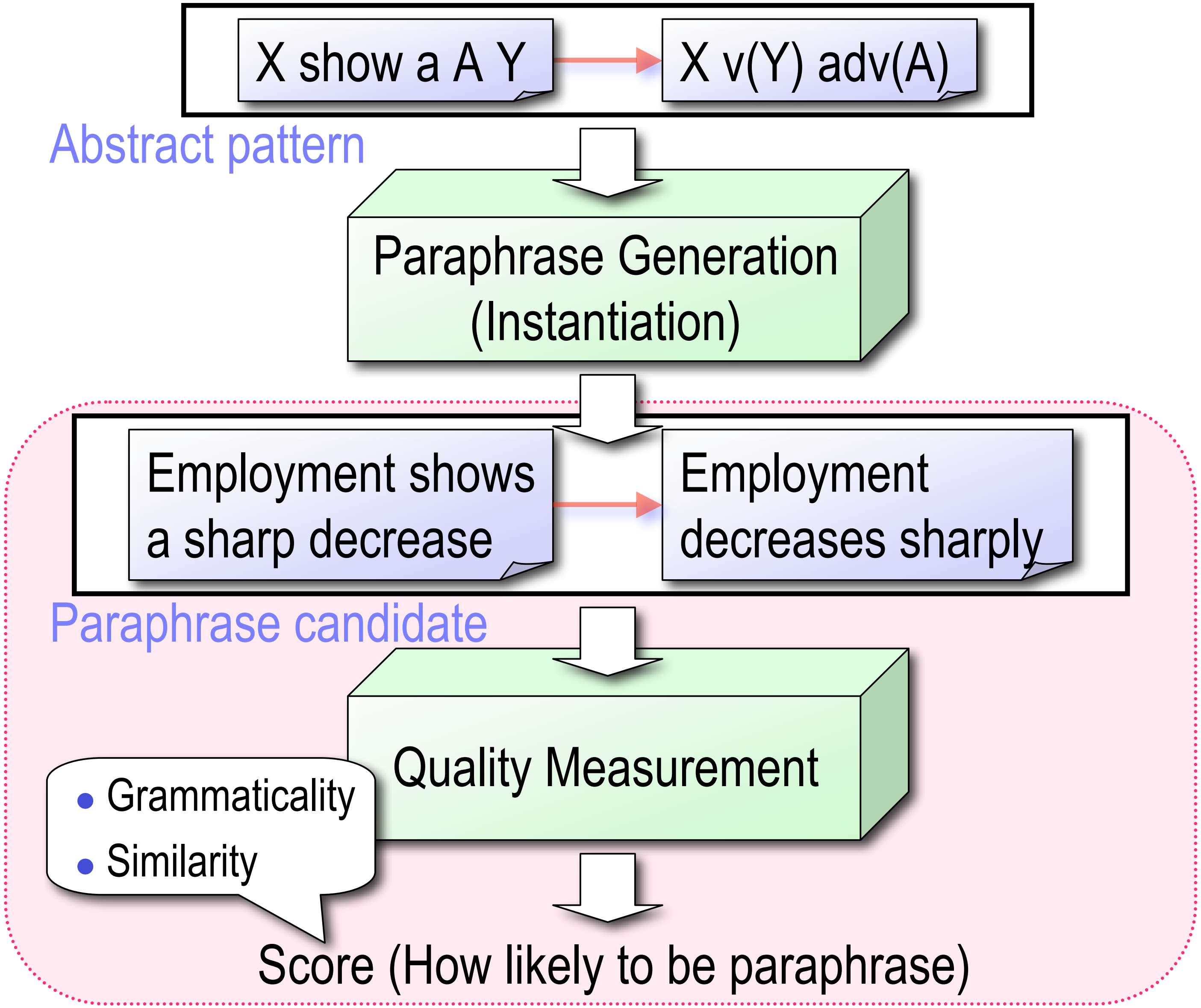
Mainichi / WebCP

max # of snippets (1,000 / 500)

# Outline

1. Task Description
2. Proposed Model
3. Experiments
4. Conclusion

# Overview


X show a A Y
X v(Y) adv(A)
Abstract pattern
Paraphrase Generation (Instantiation)
Employment shows a sharp decrease
Employment decreases sharply
Paraphrase candidate
Quality Measurement
Grammaticality
Similarity
Score (How likely to be paraphrase)

# Test Data

- Extract input phrases
  - 1,000+ phrases × 6 basic phrase types
  - Mainichi (1.5GB)
  - Referring to structure

Trans. Pat.
$N{:}C{:}V \Rightarrow adv(V){:}vp(N)$

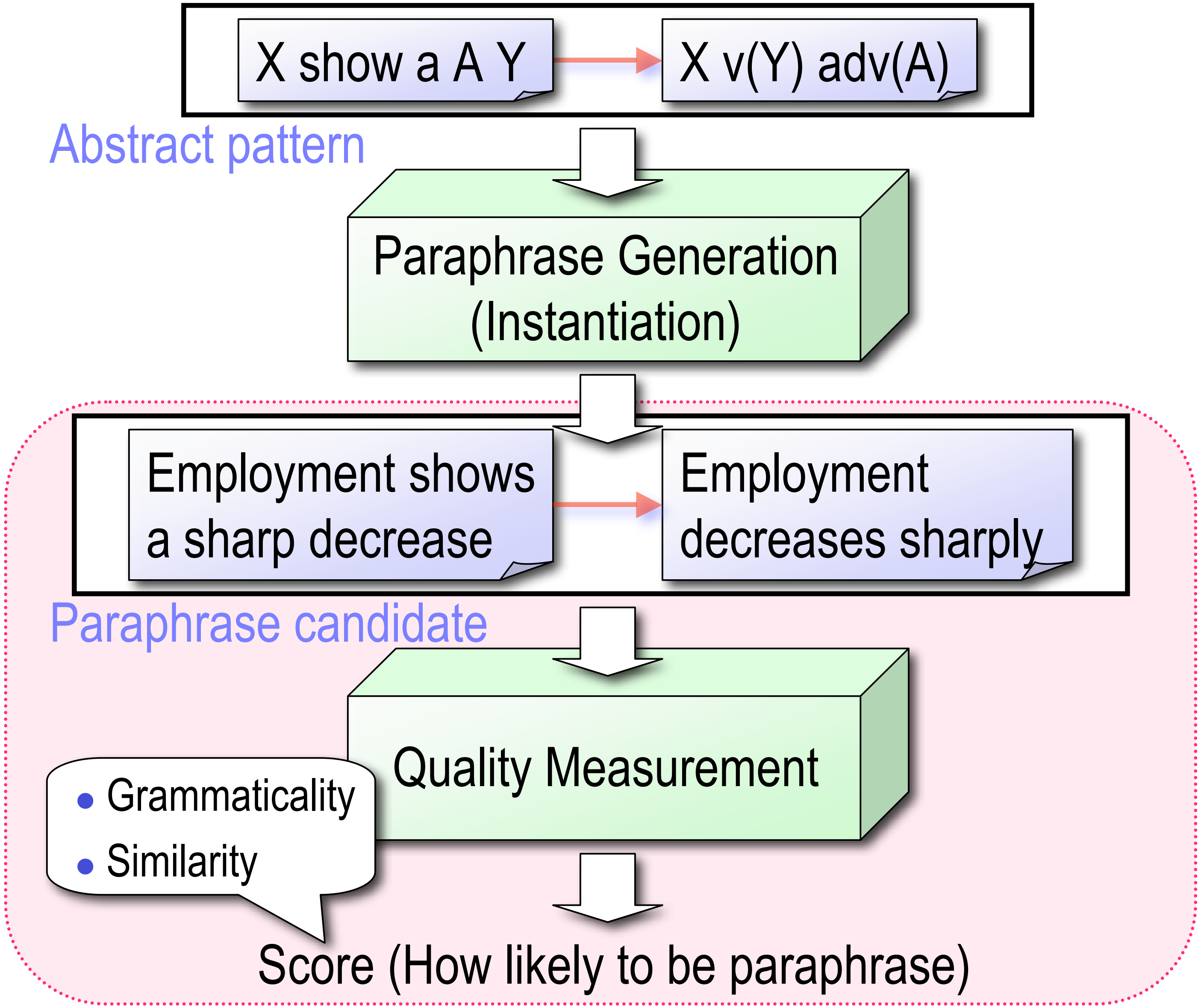
Gen. Func.
$vp(N)$

Lex. Func.
$adv(V)$

- Paraphrase generation [Fujita+, 07]
  - 176,541 candidates for 4,002 phrases
- Sampling
  - Candidates for 200 phrases
  - Diverse cases (see column Y)

| | All | Sampled | | |
|---|---|---|---|---|
| Phrase type | $s$ | $s$ | $\langle s, t \rangle$ | Y |
| $N{:}C{:}V$ | 489 | 18 | 57 | 3.2 |
| $N_1{:}N_2{:}C{:}V$ | 966 | 57 | 4,596 | 80.6 |
| $N{:}C{:}V_1{:}V_2$ | 982 | 54 | 4,767 | 88.3 |
| $N{:}C{:}Adv{:}V$ | 523 | 16 | 51 | 3.2 |
| $Adj{:}N{:}C{:}V$ | 50 | 2 | 8 | 4.0 |
| $N{:}C{:}Adj$ | 992 | 53 | 173 | 3.3 |
| Total | 4,002 | 200 | 9,652 | 48.3 |

# Overview


X show a A Y
X v(Y) adv(A)
Abstract pattern
Paraphrase Generation (Instantiation)
Employment shows a sharp decrease
Employment decreases sharply
Paraphrase candidate
Quality Measurement
Grammaticality
Similarity
Score (How likely to be paraphrase)

# Viewpoint

- How well a system can rank a correct candidate first?
- Models evaluated
  - Proposed model
    - All combination of options
    - P(t) × P(f) × Feature set × max # of snippet
      2 2 2+1 2

      HAR: harmonic mean of BOW and MOD scores
  - Baselines
    - Lin's measure [Lin+, 01]
    - $\alpha$-skew divergence [Lee, 99]
    - HITS

**Similarity** only

**Grammaticality** only

# Results (max 1,000 snippets)

- # of cases that gained positive judgments
  - Models except CFDS+Mainichi << the best models

2 judges' OK

1 or 2 judges' OK

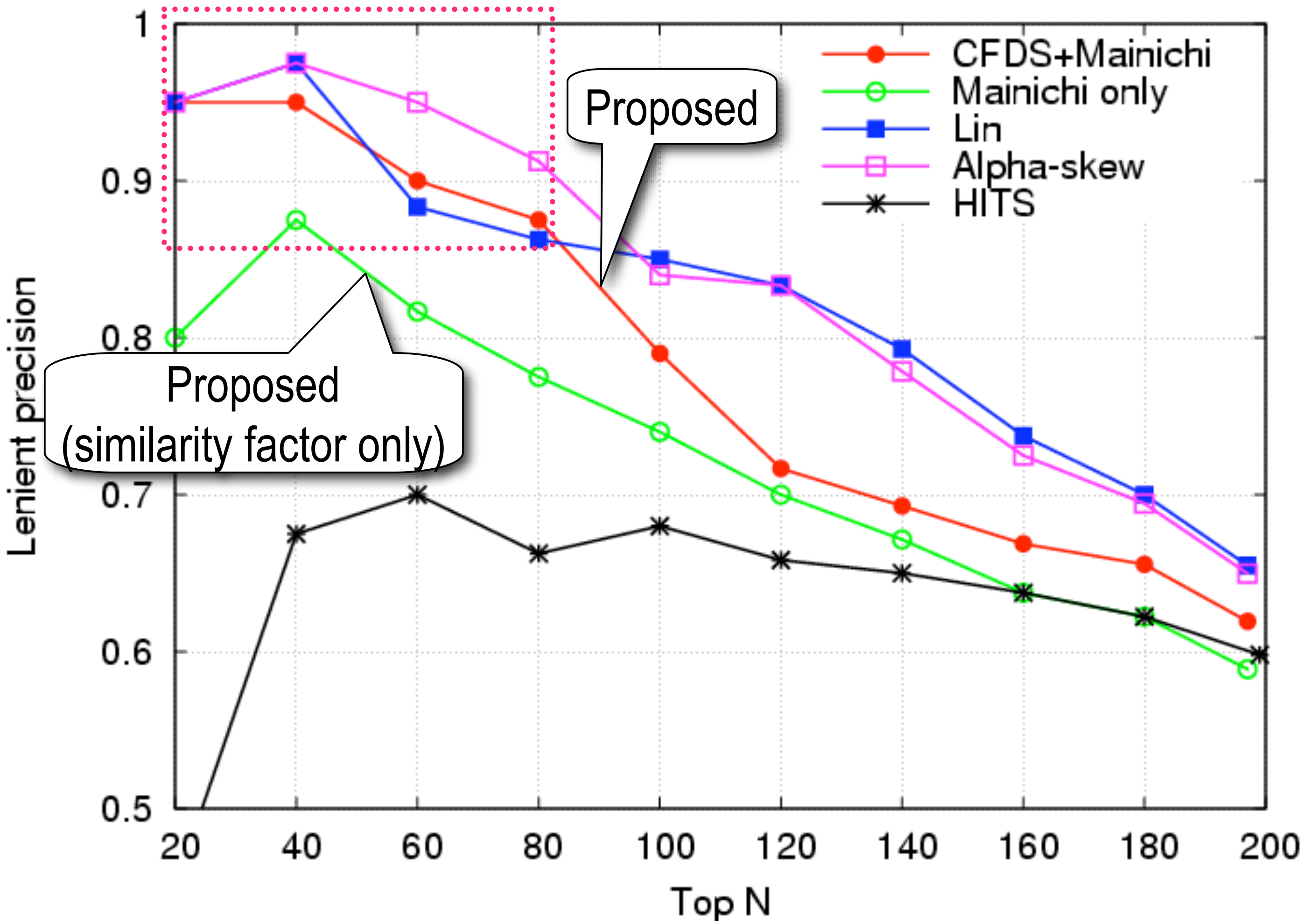
| Model \ Feature | Strict | | | Lenient | | |
|---|---|---|---|---|---|---|
| | BOW | MOD | HAR | BOW | MOD | HAR |
| CFDS+Mainichi | 79 | 82 | 83 | 121 | 121 | 122 |
| Lin | 79 | 88 | 88 | 116 | 128 | **129** |
| $\alpha$-skew | 84 | **89** | **89** | 121 | 128 | 128 |
| HITS | 84 | | | 119 | | |

**XXX**: best

XXX: significantly worse than the best (McNemer's test, p<0.05)

## Results (max 1,000 snippets, HAR)

- Lenient precision and score
  - Best candidate ∧ Relatively high score ⇒ High precision



# Considerations

- Harnessing the Web led to accurate baselines
  1. Looking up the Web … Feature retrieval
     + Grammaticality check
  2. Comparing feature distributions … Similarity check
- Two distinct viewpoints of similarity are combined

  **Constituent similarity**:
  - Syntactic transformation + Lexical derivation [Fujita+, 07]

  **Contextual similarity**:
  - Bag of words / Bag of modifiers

Trans. Pat.
$N{:}C{:}V \Rightarrow adv(V){:}vp(N)$

Gen. Func.
$vp(N)$

Lex. Func.
$adv(V)$

# Diagnosis shows the room of improvement

$$P(t|s) = \overbrace{P(t)}^{\text{Grammaticality}} \overbrace{\sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)}}^{\text{Similarity}}$$

max # of snippets
(1,000 / 500 / 200 / 100)

MDS < CFDS

Mainichi > WebCP

BOW < MOD ≒ HAR

A2: MDS cannot capture collocation of content words

A5: No significant difference (Even Web is not sufficient?)

A3: Combining with P(t) dismisses the advantage

A4: Linguistic tools are trained on newspaper articles

# Conclusion & Future work

- Measuring the quality of paraphrase candidates
  - **Input**: Automatically generated phrasal paraphrases
  - **Output**: Quality score [0,1]
    - Semantically equivalent
    - Substitutable in some context
    - Grammatical

  **Similarity** (semantically equivalent; substitutable in some context)
  **Grammaticality** (grammatical)

  - Overall: 54-62% (cf. Lin/skew: 58-65%, HITS: 60%)
  - Top 50: 80-92% (cf. Lin/skew: 90-98%, HITS: 70%)
- Future work
  - Feature engineering (including parameter tuning)
  - Application to non-productive paraphrases