Computing Paraphrasability of Syntactic Variants Using Web Snippets

Atsushi FUJITA and Satoshi SATO
Nagoya Univ., Japan
Automatic Paraphrasing

- Fundamental in NLP
  - Recognition: IR, IE, QA, Multi-Doc.Summarization
  - Generation: MT, TTS, Authoring aids

- Resources required
  - Handcrafted knowledge
    - Thesauri [Many work]
    - Transformation rules [Mel’cuk+, 87] [Dras, 99] [Jacquemin, 99]
  - Automatic knowledge acquisition
    - Distributional similarity [Lin+, 01] [Szpektor+, 04]
    - Aligning comparable/bilingual corpora [Many work]
Paraphrase Knowledge

- Template-like representation
  - Lexical paraphrases
    - X wrote Y → X is the author of Y
    - X solves Y → X deals with Y
  - Morpho-syntactic paraphrases (syntactic variants)
    - X v Y → Y be v(Z)-PP by X
    - X show a A Y → X v(Y) adv(A)

- Lack of applicability conditions → incorrect results
Task Description

Computing *paraphrasability* between phrases

- **Input:** automatically generated paraphrase candidates
  - Pair of original and generated phrases \((s \text{ and } t)\)

- **Output:** paraphrasability score \([0,1]\)
  - Is \(t\) grammatical?
  - Does \(t\) hold if \(s\) holds? (semantic equivalence or inclusion)
  - Is \(t\) syntactically substitutable for \(s\) in some context?
Issues and Solutions

How to measure similarity between phrases?
- Contextual similarity: distributional similarity
  - Bag of words / Bag of dependency relations
- Constituent similarity: handling syntactic variants
  - Syntactic transformation + Lexical derivation

How to deal with data sparseness problem?
- Collect example sentences of phrases from Web snippets
  - Assessing grammaticality
Outline

1. Task Description
2. Paraphrases Handled
3. Proposed Method
4. Experiments
5. Discussion
6. Conclusion
Paraphrases of Predicate Phrases

- Symmetric vs. Asymmetric
  - Symmetric: $X$ change $Y$ $\leftrightarrow$ $X$ modify $Y$
  - Asymmetric: $X$ sprint $\rightarrow$ $X$ run

- Equivalent / Inclusion / Entailment vs. Inference
  - Equivalent / Inclusion / Entailment: $X$ change $Y$ $\leftrightarrow$ $X$ modify $Y$
  - Inference: $X$ married $Y$ $\rightarrow$ $X$ dated $Y$
  - $X$ sprint $\rightarrow$ $X$ run
  - $X$ snore $Y$ $\rightarrow$ $X$ sleep $Y$

- Lexical vs. Morpho-syntactic
  - Lexical: $X$ change $Y$ $\leftrightarrow$ $X$ modify $Y$
  - Morpho-syntactic: $X$ show a A $Y$ $\rightarrow$ $X$ v(Y) adv(A)
Paraphrases Handled

- Morpho-syntactic paraphrases (syntactic variants)
  - Syntactic transformation + Lexical derivation
    - Constituent similarity is guaranteed a little
    - e.g. Head-switching, Light-verb construction, Category-shift

Predicate phrases of Japanese

- Kakunin-o isogu
  - Checking-ACC to hurry
  - To hurry checking it

- Isoide kakunin-suru
  - In a hurry to check
  - To check it in a hurry
Syntactic Variant Generator for Japanese

kakunin : o : isogu
N  C  V

checking: ACC: to hurry

Trans. Pat.
N: C: V ⇒ adv(V): vp(N)

adv(isogu) : vp(kakunin)

Lex. Func.
adv(V)

Gen. Func.
vp(N)

isoide
{v(kakunin) : genVoice() : genTense()}

Lex. Func.
v(N)

Gen. Func.
genVoice()

Gen. Func.
genTense()

kakunin-suru
{φ, reru/rareru, seru/saseru}

isoide : {kakunin-suru : {φ, reru/rareru, seru/saseru} : {φ, ta/da}}

[Fujita+, 07]
Outline

1. Task Description
2. Paraphrases Handled
3. Proposed Method
4. Experiments
5. Discussion
6. Conclusion
Overview

1. Snippet Retrieval
2. Feature Extraction
3. Paraphrasability Computation

Candidate Generation

Snippets

Anchor

Features

Paraphrasability score $Par(s \Rightarrow t)$
Step 1. Snippet Retrieval

```
``Phrase search``
- Yahoo! JAPAN Web-search API
- 500 top snippets
Step 2. Feature Extraction

- **HITS**: # of pages Yahoo! API returns
  - Larger HITS $\Rightarrow t$ is more likely grammatical

- **BOW**: content words around the phrase in snippets
  - BOWs surrounding $s$ and $t$ have similar distribution
    $\Rightarrow s$ and $t$ are semantically similar

- **MOD**: modifiers and modifiees of the phrase in snippets
  - $s$ and $t$ share a number of modifiers and modifiees
    $\Rightarrow s$ and $t$ are syntactically substitutable
Extracting MOD Features

- Modifier / modifiee chunk (*bunsetsu*)
  - Relation types (Depend / Appositive / Parallel)
  - Base form of the head-word (content word)
  - Some types of functional words (if any)

Given phrase:
(I am) planning to verify the reproducibility of his experimental result in detail.
Step 2. Feature Extraction (Anc)

Source-focused feature extraction

1. Determine anchor $a$ which strongly associates with $s$
   - Noun which most frequently modifies $s$ (one of MOD features)

2. Retrieve snippets for $s$ AND $a$ and $t$ AND $a$

3. Extract BOW and MOD features from those snippets
Step 3. Paraphrasability Computation

- **Lin**: Lin’s measure [Lin+, 01]

\[
\text{Par}_{\text{Lin}}(s \Rightarrow t) = \frac{\sum_{f \in F_s \cap F_t} (w(s, f) + w(t, f))}{\sum_{f \in F_s} w(s, f) + \sum_{f \in F_t} w(t, f)}
\]

- **$F_s$, $F_t$**: Feature sets for $s$ and $t$
- **$w(x, f)$**: Weight of feature $f$ in $F_x$ (frequency in snippets)

- **skew**: $\alpha$-skew divergence [Lee, 99]

\[
\text{Par}_{\text{skew}}(s \Rightarrow t) = \exp(-d_{\text{skew}}(t, s))
\]

\[
d_{\text{skew}}(t, s) = D(P_s \parallel \alpha P_t + (1 - \alpha)P_s)
\]

- **$P_s = P(f \mid s)$, $P_t = P(f \mid t)$**
- **$\alpha$**: approximation degree of KL divergence [0,1]
Summary

Features: Contextual features of entire phrase
- c.f. Marginal features [Torisawa, 06] [Pantel+, 07]
- BOW, MOD

Weight of features: Frequency in snippets
- c.f. pair-wise MI [Lin+, 01] [Pantel+, 07]
- c.f. Relative Focus Feature [Geffet+, 05]

DS measures
- Lin’s measure (symmetric) [Lin+, 01]
- $\alpha$-skew divergence (asymmetric) [Lee, 99]
Outline

1. Task Description
2. Paraphrases Handled
3. Proposed Method
4. Experiments
5. Discussion
6. Conclusion
Setup: Candidate Generation

- 6 basic phrase types
- Most frequent 1,000+ phrases for each type
  - Mainichi newspaper corpus (1991-2005, 1.5GB)
  - Referring to dependency trees
- Syntactic variant generator for Japanese [Fujita+, 07]

<table>
<thead>
<tr>
<th>Phrase type</th>
<th># of tokens</th>
<th># of types</th>
<th>th. types</th>
<th>Cov. (%)</th>
<th>Output (Avg.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N: C: V$</td>
<td>20,200,041</td>
<td>4,323,756</td>
<td>1,001</td>
<td>10.7</td>
<td>1,536 (489)</td>
</tr>
<tr>
<td>$N_1 : N_2 : C : V$</td>
<td>3,796,351</td>
<td>2,013,682</td>
<td>107</td>
<td>6.3</td>
<td>88,040 (966)</td>
</tr>
<tr>
<td>$N : C : V_1 : V_2$</td>
<td>325,964</td>
<td>213,923</td>
<td>15</td>
<td>12.9</td>
<td>75,344 (982)</td>
</tr>
<tr>
<td>$N : C : Adv : V$</td>
<td>1,209,265</td>
<td>923,475</td>
<td>211</td>
<td>3.9</td>
<td>8,281 (523)</td>
</tr>
<tr>
<td>$Adv : N : C : V$</td>
<td>378,617</td>
<td>233,952</td>
<td>201</td>
<td>14.1</td>
<td>128 (50)</td>
</tr>
<tr>
<td>$N : C : Adv$</td>
<td>788,038</td>
<td>203,845</td>
<td>861</td>
<td>31.4</td>
<td>3,212 (992)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>26,698,276</strong></td>
<td><strong>7,912,633</strong></td>
<td>6,190</td>
<td></td>
<td><strong>176,541 (4,002)</strong></td>
</tr>
</tbody>
</table>
Examples of Phrases

<table>
<thead>
<tr>
<th>Structure</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>N:C:V</td>
<td><strong>kakunin-o</strong> isogu</td>
<td><strong>kentou-o</strong> sarani susumeru</td>
</tr>
<tr>
<td></td>
<td>checking-ACC to hurry</td>
<td>consideration-ACC further to go ahead</td>
</tr>
<tr>
<td></td>
<td>to hurry checking it</td>
<td>to take consideration further</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structure</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>N₁:N₂:C:V</td>
<td><strong>songai-baisho-o</strong> motomeru</td>
<td><strong>nodo-ga</strong> itai</td>
</tr>
<tr>
<td></td>
<td>damage-reparation-ACC to demand</td>
<td>throat-NOM be painful</td>
</tr>
<tr>
<td></td>
<td>to demand reparation for damage</td>
<td>to have a sore throat</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structure</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>N:C:V₁:V₂</td>
<td><strong>toukei-o</strong> tori-hajimeru</td>
<td><strong>takai hyouka-o</strong> ukeru</td>
</tr>
<tr>
<td></td>
<td>statistics-ACC to take-to start</td>
<td>high assessment-ACC to receive</td>
</tr>
<tr>
<td></td>
<td>to start collect statistics</td>
<td>to be rated high</td>
</tr>
</tbody>
</table>
Setup: Computing Paraphrasability Scores

15 measures:

- Proposed: \{HITS,\{BOW,MOD,HAR\}\} × \{Lin, skew\} × \{Nor, Anc\}
- BL (Mainichi): HITS using 1.5GB newspaper corpus

<table>
<thead>
<tr>
<th>Phrase type</th>
<th>Nor.HITS Output</th>
<th>Nor.HITS Ave.</th>
<th>Nor.BOW.* Output</th>
<th>Nor.BOW.* Ave.</th>
<th>Nor.MOD.* Output</th>
<th>Nor.MOD.* Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>N : C : V</td>
<td>1,405 (489) 2.9</td>
<td></td>
<td>1,402 (488) 2.9</td>
<td></td>
<td>1,396 (488) 2.9</td>
<td></td>
</tr>
<tr>
<td>N_1 : N_2 : C : V</td>
<td>9,544 (964) 9.9</td>
<td></td>
<td>9,249 (922) 10.0</td>
<td></td>
<td>8,652 (921) 9.4</td>
<td></td>
</tr>
<tr>
<td>N : C : V_1 : V_2</td>
<td>3,769 (876) 4.3</td>
<td></td>
<td>3,406 (774) 4.4</td>
<td></td>
<td>3,109 (762) 4.1</td>
<td></td>
</tr>
<tr>
<td>N : C : Adv : V</td>
<td>690 (359) 1.9</td>
<td></td>
<td>506 (247) 2.0</td>
<td></td>
<td>475 (233) 2.0</td>
<td></td>
</tr>
<tr>
<td>Adv : N : C : V</td>
<td>45 (20) 2.3</td>
<td></td>
<td>45 (20) 2.3</td>
<td></td>
<td>42 (17) 2.5</td>
<td></td>
</tr>
<tr>
<td>N : C : Adv</td>
<td>1,459 (885) 1.6</td>
<td></td>
<td>1,459 (885) 1.6</td>
<td></td>
<td>1,399 (864) 1.6</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16,912 (3,593) 4.7</td>
<td></td>
<td>16,067 (3,336) 4.8</td>
<td></td>
<td>15,073 (3,285) 4.6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1,368 (488) 2.8</td>
<td></td>
<td>1,366 (487) 2.8</td>
<td></td>
<td>1,360 (487) 2.8</td>
<td></td>
<td>1,103 (457) 2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7,437 (897) 8.3</td>
<td></td>
<td>7,424 (894) 8.3</td>
<td></td>
<td>6,795 (891) 7.6</td>
<td></td>
<td>3,041 (948) 3.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,517 (697) 3.6</td>
<td></td>
<td>2,497 (690) 3.6</td>
<td></td>
<td>2,258 (679) 3.3</td>
<td></td>
<td>1,156 (548) 2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>342 (174) 2.0</td>
<td></td>
<td>339 (173) 2.0</td>
<td></td>
<td>322 (168) 1.9</td>
<td></td>
<td>215 (167) 1.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41 (18) 2.3</td>
<td></td>
<td>41 (18) 2.3</td>
<td></td>
<td>39 (16) 2.4</td>
<td></td>
<td>14 (7) 2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,235 (809) 1.5</td>
<td></td>
<td>1,235 (809) 1.5</td>
<td></td>
<td>1,161 (779) 1.5</td>
<td></td>
<td>559 (459) 1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12,940 (3,083) 4.2</td>
<td></td>
<td>12,902 (3,071) 4.2</td>
<td></td>
<td>11,935 (3,020) 4.0</td>
<td></td>
<td>6,088 (2,586) 2.4</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation 1: Ev.Gen

Question
- Can a correct paraphrase have the highest score among candidates for a source phrase?

Judgment (2 assessors)
- For 200 input, the best candidates of 15 models
Results 1: Ev.Gen

- Mainichi << *.HITS ≧ *.BOW.* < *.MOD.* ≧ *.HAR.*
  - Web enables us to compute paraphrasability accurately
- Candidates with higher scores are more likely correct
  - e.g. Lenient Prec. over 93% (th=0.5)
- Nor.X.* ≧ Anc.X.* (discuss later)
Evaluation 2: Ev.Rec

Question

- How is the method useful for collecting paraphrase instances?

Judgment (2 assessors)

- 200 best candidates for each of 15 models
Results 2: Ev.Rec

- Mainichi $\approx \ast$.HITS $\ll \ast$.BOW.$\ast \approx \ast$.MOD.$\ast \approx \ast$.HAR.$\ast$
  - DS measures outperformed HITS
  - Lenient Prec. almost reach a ceiling

- Nor.X.$\ast \approx$ Anc.X.$\ast$ again
  - Anchor selection might be inappropriate
  - 2 or more content words make $\ast$ rarely ambiguous
Results 2: Ev.Rec

Remaining problems

- Dropping $N_1$ from $N_1:N_2:C:V$
  - Typically functions as generalization
    - $yagai-konsato$ to $konsato$
      - outdoors-concert to concert
  - $N_1$ sometimes plays as the semantic head of $N_1:N_2$
    - $shukketsu-taryou-de$ to $shibou-suru$
      - blood loss-plenty-ABL to die due to heavy blood loss
    - $taryou-de$ to $shibou-suru$
      - loss-plenty-ABL to die due to plenty

- Solutions:
  - Semantic parsing, Phrase boundary detection, etc.
Outline

1. Task Description
2. Paraphrases Handled
3. Proposed Method
4. Experiments
5. Discussion
6. Conclusion
Discussion: Issues Addressed

- Measurement of paraphrasability between phrases
  - Reasonably nice (Ev.Gen: over 65%, Ev.Rec: over 96%)
    - Combining constituent and contextual similarities
  - Room for improvement
    - Feature selection [Hagiwara+, 08]
    - Feature weighting [Lin+, 01] [Geffet+, 05]

- Data sparseness problem
  - Not perfectly solved
  - TSUBAKI offers larger number of snippets [Shinzato+, 08]
Discussion: Technical Issues

- **Coverage**
  - For 50% input, no candidate is output
  - More robust generation system
    - To generate a wider range of paraphrases
    - To handle other types of phrases with less human-labor

- **Portability**
  - 90% of candidates are filtered out due to 0 HITS
  - Use SLMs to prune incorrect candidates before querying
Conclusion

Computing *paraphrasability* between phrases

- Input: paraphrase candidates
  - Automatically generated
  - Syntactic variants
  - Predicate phrases in Japanese

- Output: paraphrasability score [0,1]
  - Is $t$ grammatical?
  - Does $t$ hold if $s$ holds? (semantic equivalence or inclusion)
  - Is $t$ syntactically substitutable for $s$ in some context?

Proposed method achieved reasonable results

- Ev.Gen: over 65% (over 93% w/ th=0.5), Ev.Rec: over 96%