A Compositional Approach toward Dynamic Phrasal Thesaurus

Atsushi FUJITA, Shuhei KATO, Naoki KATO, Satoshi SATO
Nagoya Univ., Japan
Computing Semantic Equivalence (SE)

- Fundamental in NLP
  - Recognition: IR, IE, QA
  - Generation: MT, TTS, Summarization

- Previous attempts used ...
  - Thesauri   [So many work]
  - Tree kernels   [Collins+, 01] [Takahashi, 05]
  - Statistical translation models   [Barzilay+, 03] [Brockett+, 05]
  - Distributional similarity   [Harris, 64] [Lin+, 01] [Weeds+, 05]
  - Syntactic patterns   [Mel’cuk+, 87] [Dras, 99] [Jacquemin, 99]
Computing Semantic Equivalence (SE)

- Fundamental in NLP
  - Recognition: IR, IE, QA
  - Generation: MT, TTS, Summarization

- Previous attempts used ...
  - Thesauri
  - Tree kernels
    - Words are not necessarily the unit of meaning
      - (polysemous words, meaning of construction)
    - Cannot generate paraphrases
  - Statistical translation models
    - Corpus is not almighty
      - (data sparseness, cost)
  - Distributional similarity
  - Syntactic patterns
    - No thorough list
Our Proposal

- Phrasal Thesaurus
  - A mechanism for directly computing SE between phrases

- be in our favor
- be favorable for us
- its reproducibility
- if it is reproducible
- show a sharp decrease
- decrease sharply
- investigate why there was a fire
- investigate what started a fire
- make an investigation into the cause of a fire
- investigate the cause of a fire
Aim

- Implement tools and resources
  - Application-independent module
  - Human aids: writing / reading texts
- Confirm phrase is appropriate unit for computing SE
  - Ambiguity of words $\gg$ Ambiguity of phrases (more suitable to handle)

This is a preliminary progress report (w/o concrete evaluation)
Outline

1. Motivation & Aim
2. Range of phenomena
3. System & implementation
4. Discussion
5. Conclusion
Towards Phrasal Thesaurus

- What sorts of phrases?
- How to handle a variety of expressions?

- be in our favor ↔ be favorable for us
- its reproducibility ↔ if it is reproducible
- show a sharp decrease ↔ decrease sharply
- investigate the cause of a fire ↔ investigate why there was a fire
- investigate what started a fire ↔ make an investigation into the cause of a fire
Range of phrases

- Predicate phrase (cf. various exps. in RTE)
  - Reliably captured using recent technologies
  - Approx. corresponds to single event
    
    [Chklovski and Pantel, 2004] [Torisawa, 2006]

- Our target language: Japanese

  - **noun phrase + case marker + predicate**
    - Various noun phrases
    - Various predicates
    - Case markers indicate grammatical roles of noun phrases
Classification of noun phrases in Japanese

- **noun phrase**
  - **content**
    - **formal noun**
      - "koto"
      - "mono"
      - "no"
    - **single word**
    - **compound**
      - $N_1 N_2$
      - $N + \text{suffixes}$
      - $N_1 + \text{"no"} + N_2$
    - **modified**
      - $\text{Adj} + N$
      - $\text{Adjectival verb} + N$
      - $\text{clause} + N$

- **common noun nominalization**
Classification of predicates in Japanese

predicate

verb phrase

Adj
Adjectival verb + “da”
Adv + “da”
Copula

single word

original verb
Sino-Japanese verb
lexical compound
light verb
Adv + “suru”
original + original
Sino + original
Sino + Sino
N + Sino

compound

single word compound
Range of phrases

- Our target language: Japanese
  - noun phrase + case marker + predicate

- Variation of phrases => Variation of words
  - Various combinations of open-class words
Range of phenomena

- Variation of paraphrases of phrases
  - Variation of paraphrases of words
    - Difficult (hard?) to statically enumerate
    - No previous work explicitly collected:
      - “All verbs that can be passivized”
      - “All noun-verb pairs that compose light-verb constructions”
    - How to handle them?
Paraphrases of predicate phrases

- X is charged by Y
- Y announced the arrest of X
- X married Y
- X dated Y
- X prevent Y
- X lower the risk of Y
- X change Y
- X modify Y
- X change Y
- X alter Y
- X buy Y
- X acquire Y
- X solve Y
- Y is solved by X
- X gives Y a fright
- Y is frightened of X
- X is in our favor
- X is favorable for us
- X show a sharp decrease
- X decrease sharply
- X get the sack
- X be dismissed from employment
- X see the light
- X realize the truth
Paraphrases of predicate phrases

- X is charged by Y → Y announced the arrest of X
- X married Y → X dated Y → X prevent Y → X lower the risk of Y
- X change Y → X modify Y → X change Y → X alter Y
- X buy Y → X acquire Y → X solve Y → Y is solved by X
- X gives Y a fright → Y is frightened of X
- X is in our favor → X is favorable for us
- X show a sharp decrease → X decrease sharply
- X get the sack → X be dismissed from employment
- X see the light → X realize the truth
Paraphrases of predicate phrases

- X is charged by Y → Y announced the arrest of X
- X married Y → X dated Y → X prevent Y → X lower the risk of Y
- X change Y → X modify Y → X change Y → X alter Y
- X buy Y → X acquire Y → X solve Y → Y is solved by X
- X gives Y a fright → Y is frightened of X
- X is in our favor → X is favorable for us
- X show a sharp decrease → X decrease sharply
- X get the sack → X be dismissed from employment
- X see the light → X realize the truth
Paraphrases of predicate phrases

- X is charged by Y
- Y announced the arrest of X
- X married Y
- X dated Y
- X prevent Y
- X lower the risk of Y
- X change Y
- X modify Y
- X change Y
- X alter Y
- X buy Y
- X acquire Y
- X solve Y
- Y is solved by X
- X gives Y a fright
- Y is frightened of X
- X is in our favor
- X is favorable for us
- X show a sharp decrease
- X decrease sharply
- X get the sack
- X be dismissed from employment
- X see the light
- X realize the truth
Compositional paraphrases (syntactic variants)

- Syntactic transformation + Lexical derivation
  \[ \Rightarrow \text{Dynamic generation (Dynamic Phrasal Thesaurus)} \]

- **X solve Y**  \[\rightarrow\]  **Y is solved by X**  \[\rightarrow\]  **X V Y**  \[\rightarrow\]  **Y be v(Z)-PP by X**

- **X gives Y a fright**  \[\rightarrow\]  **Y is frightened of X**

- **X is in our favor**  \[\rightarrow\]  **X is favorable for us**  \[\rightarrow\]  **X be in Z's Y**  \[\rightarrow\]  **X be adj(Y) for Z**

- **X show a sharp decrease**  \[\rightarrow\]  **X decrease sharply**  \[\rightarrow\]  **X show a A Y**  \[\rightarrow\]  **X v(Y) adv(A)**
Compositional paraphrases (syntactic variants)

- Syntactic transformation + Lexical derivation
  \[\Rightarrow\] Dynamic generation (Dynamic Phrasal Thesaurus)

- Our target language: Japanese

- Trivial? No.
  - Not exhaustively explored
  - Beneficial [Dolan+, 04] [Romano+, 06]
Outline

1. Motivation & Aim
2. Range of phenomena
3. System & implementation
4. Discussion
5. Conclusion
System overview

- **Input**: Phrase (string)
- **Output**: List of paraphrases

Morphological analysis

Syntactic transformation

Surface generation

SLM-based filtering

isoide : kakunin-sa : re : ta
isoide : kakunin-shi : ta
isoide : kakunin-suru
isoide : kakunin-sa : reru

(We) hurry checking it.

(It) was checked in a hurry. (x)
(We) checked it in a hurry. (x)
(We) check it in a hurry. (o)
(It) is checked in a hurry. (x)
1. Morphological analysis

- **Input**: Phrase (string)
- **Output**: Array of morphemes w/ POS-tag
  - Using MeCab-0.91, a state-of-the-art morphological analyzer

(We) hurry checking it.

```
kakuninoisogu
```

```
kakunin : o : isogu
N  C    V
```

- checking    ACC    to hurry

MeCab + post-process

- **N**: noun
- **V**: verb
- **Adj**: adjective
- **An**: adjectival verb
- **Adv**: adverb
- **C**: case marker
  - etc.
2. Syntactic transformation: knowledge used

- **Transformation pattern**
  - Generates skeletons of syntactic variants

- **Generation function**
  - Enumerates expressions made of the given set of words

- **Lexical function**
  - Generates different lexical items in certain relation
2. Syntactic transformation: example

\[ \text{checking: ACC to hurry} \]

\[ \text{kakunin} : o : isogu \]

\[ N \quad C \quad V \]

\text{Trans. Pat.}

\[ N: C: V \Rightarrow \text{adv}(V):vp(N) \]

\[ \text{adv(isogu)} : \text{vp(kakunin)} \]

\text{Lex. Func.}

\[ \text{adv}(V) \]

\text{Gen. Func.}

\[ \text{vp}(N) \]

\[ \text{isoide} \]

\{ \text{v(kakunin)} : \text{genVoice()} : \text{genTense()} \} \]

\text{Lex. Func.}

\[ \text{v}(N) \]

\text{Gen. Func.}

\[ \text{genVoice()} \]

\text{Gen. Func.}

\[ \text{genTense()} \]

\text{Gen. Func.}

\[ \text{v}(N) \]

\[ \text{genVoice()} \]

\[ \text{genTense()} \]

\[ \text{kakunin-suru} \]

\{ \phi, \text{reru/rareru, seru/saseru} \}

\text{isoido}

\{ \text{kakunin-suru} : \{ \phi, \text{reru/rareru, seru/saseru} \} : \{ \phi, \text{ta/da} \} \}
3. Surface generation

- **Input**: Bunch of candidate phrases
- **Output**: List of candidate phrases
  - 1. Unfolding
  - 2. Lexical choice (exclusively used auxiliaries)
  - 3. Conjugation

\[
\text{Surface generation}
\]

\[
\begin{align*}
\text{isoide} &: \{\text{kakunin-suru} : \{\phi, \text{reru/rareru, seru/saseru}\} : \{\phi, \text{ta/da}\}\}
\end{align*}
\]
4. SLM-based filtering

- **Input**: List of candidate phrases
- **Output**: List of grammatical phrases
  - Grammaticality assessment
    - Initial model: if occur in Mainichi 1999-2005 (1.8GB)

```
isoide : kakunin-suru,  isoide : kakunin-shi : ta,  
isoide : kakunin-sa : reru,  isoide : kakunin-sa : re : ta,  
isoide : kakunin-sa : seru,  isoide : kakunin-sa : se : ta
```

(SL) was checked in a hurry.  (x)
(We) checked it in a hurry.  (x)
(We) check it in a hurry.  (o)
(SL) is checked in a hurry.  (x)
Knowledge development

- Paraphrase phenomena ⇒ Create patterns
  - Not necessarily from examples
  - Same manner as
    - MTT [Mel'cuk+, 1987]
    - STAG [Dras, 1999]
    - FASTR [Jacquemin, 1999]
    - KURA [Takahashi+, 2001]

  - Frame ⇒ Register various expressions
Comparison w/ previous work

**MTT** [Mel’cuk+, 1987]
- Paraphrasing rules at 7 levels
- More than 60 Lexical functions

**FASTR** [Jacquemin, 1999]
- Structural transformations (**Syntagma**)
- Semantic links (**Paradigm**)

**Ours**
- Transformation at SSynt level only (cf. MTT)
- Predicate phrase, not technical term (cf. FASTR)
- One-to-N generation by Gen.Func.

| Trans. Pat. | $N:C:V \Rightarrow adv(V):vp(N)$ |
| Lex. Func. | $adv(V)$ |

| Trans. Pat. | $N:C:V \Rightarrow adv(V):vp(N)$ |
| Lex. Func. | $adv(V)$ |

| Gen. Func. | $vp(N)$ |
Current scale of knowledge

- **Transformation pattern**
  - Starting from N:C:V
    - N₁:N₂:C:V, N:C:V₁:V₂, ... : 37 patterns

- **Generation function**
  - As a by-product of generalizing transformation patterns
    - Content phrases (5): NPs, VPs
    - Functional expressions (4): Case markers, Auxiliaries

- **Lexical function**
  - Lexical derivation (10 dics, totally 6,322 word pairs)
  - Noun-to-interrogative (1)
To ensure coverage

1. Enumerate Trans. Pat. for N:C:V

2. Extend them for more complex types of phrases
The body of Lex. Func.

**IPADIC-2.7.0 + Mainichi 1999-2005 (1.8GB)**

<table>
<thead>
<tr>
<th>POS-pair</th>
<th>$D$</th>
<th>$C$</th>
<th>$D\cup C$</th>
<th>$J$</th>
<th>cleaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun - verb</td>
<td>3,431</td>
<td>-</td>
<td>3,431</td>
<td>3,431</td>
<td></td>
</tr>
<tr>
<td>noun - adjective</td>
<td>308</td>
<td>667</td>
<td>906</td>
<td>475</td>
<td>done</td>
</tr>
<tr>
<td>noun - adjectival verb</td>
<td>1,579</td>
<td>-</td>
<td>1,579</td>
<td>1,579</td>
<td></td>
</tr>
<tr>
<td>noun - adverb</td>
<td>271</td>
<td>-</td>
<td>271</td>
<td>271</td>
<td></td>
</tr>
<tr>
<td>verb - adjective</td>
<td>252</td>
<td>-</td>
<td>252</td>
<td>192</td>
<td>done</td>
</tr>
<tr>
<td>verb - adjectival verb</td>
<td>74</td>
<td>-</td>
<td>74</td>
<td>68</td>
<td>done</td>
</tr>
<tr>
<td>verb - adverb</td>
<td>74</td>
<td>-</td>
<td>74</td>
<td>64</td>
<td>done</td>
</tr>
<tr>
<td>adjective - adjectival verb</td>
<td>66</td>
<td>95</td>
<td>159</td>
<td>146</td>
<td>done</td>
</tr>
<tr>
<td>adjective - adverb</td>
<td>33</td>
<td>-</td>
<td>33</td>
<td>26</td>
<td>done</td>
</tr>
<tr>
<td>adjectival verb - adverb</td>
<td>70</td>
<td>-</td>
<td>70</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6,158</td>
<td>762</td>
<td>6,849</td>
<td>6,322</td>
<td></td>
</tr>
</tbody>
</table>
Outline

1. Motivation & Aim
2. Range of phenomena
3. System & implementation
4. Discussion
5. Conclusion
Discussion (.future work)

- Sufficient condition
  - Patterns does not ensure paraphrasability perfectly
  - Extensional definition of selectional preferences [Pantel+, 2007]

- Structured transformation
  - For flexible and accurate matching
  - Less impact due to short phrase

- Methodology of resource development
  - Modularization of Gen. Func. is inconsistent
  - Requires linguistic expertise
  - Simple KBs are preferable (cf. MTT)
Conclusion & Future work

Notion of Phrasal Thesaurus is introduced

- Compositional paraphrases of predicate phrases
- Preliminary progress report of resource development

Future work

- Development
  - Resources
  - SLM (Structured, Web, etc.)
  - Applicability conditions
- Intrinsic / extrinsic evaluation

Trans. Pat.
N:C:V \implies adv(V):vp(N)

Gen. Func.
vp(N)

Lex. Func.
adv(V)