< IWP 2005, Oct. 14th, 2005 >

A Class-oriented Approach to Building a Paraphrase Corpus

Atsushi FUJITA[1], Kentaro INUI[2]
[1] Kyoto University
[2] Nara Institute of Science and Technology

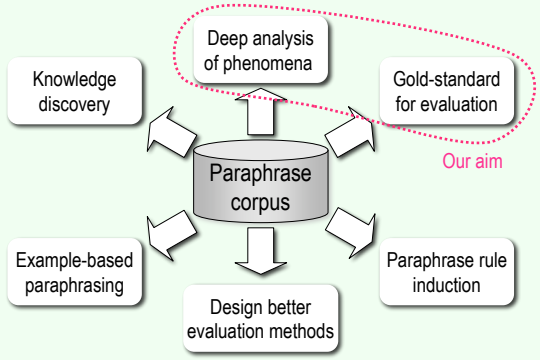
Requirements for handling paraphrases

- Transformation rules / patterns
  - Handcrafting
    $X_{verb} \Leftrightarrow S_0(X) + Oper_1(S_0(X))$
    [Iordansjaka et al., 1991] [Dras, 1999] [Sato et al., 1999] [Kondo et al., 1999] [Kondo et al., 2001] [Iida et al., 2001] etc.
  - Automatic acquisition
    [Barzilay et al., 2001] [Lin et al., 2001] [Shinyama et al., 2002] [Shimohata et al., 2002] [Pang et al., 2003] etc.
    burst into tears ⇔ cry
    X finds a solution to *Y* ⇔ *X* solves *Y*
- Paraphrase corpus (collection of paraphrase examples)
  - Few freely available resources [Dolan et al., 2004]

The leading indicators measure the economy...
The leading index measures the economy....

Purposes of paraphrase corpus

- Beneficial to activate the research field



Outline

1. Background
2. Issues and our class-oriented approach
3. Semi-automatic example collection
4. Preliminary trials
   1. Specification
   2. Discussion
5. Conclusion

Building paraphrase corpus

- Issues
  - to consider: variety, source, organization
  - to maximize: coverage, reliability, cost-efficiency
- Previous work

Manual production	Automatic acquisition
[Shirai et al., 2001]	[Barzilay et al., 2003]
[Kinjo et al., 2003]	[Shinyama et al., 2003]
[Shimohata et al., 2004]	[Dolan et al., 2004]

Reliability ⟶ Cost-efficiency

- Coverage is not ensured
- No focus on sorts/variety of paraphrases

Variety of paraphrases

Emma burst into tears and he tried to comfort her.
Emma cried and he tried to console her.
Lexical paraphrase ⇒ automatic acquisition

It was his best suit that John wore last night.
John wore his best suit last night.
Syntactic paraphrase ⇒ describable

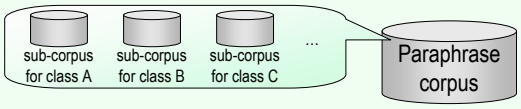
Steven made an attempt to stop playing Hearts.
Steven attempted to stop playing Hearts.

The breeze sways the trees.
The trees sway in the breeze.
Lexically compositional paraphrase
- syntactically regular
- semantically compositional

The room has already been warmed up.
The room is already warm.

A class-oriented approach

- Separately collect examples for each class



- Semi-automatic example collection
  - Automatic generation + human judgment
  - **Step 1**: Define a paraphrase class based on morpho-syntactic transformation patterns
  - **Step 2**: Collect all candidates using a paraphrase engine
  - **Step 3**: Judge candidate paraphrases in hand

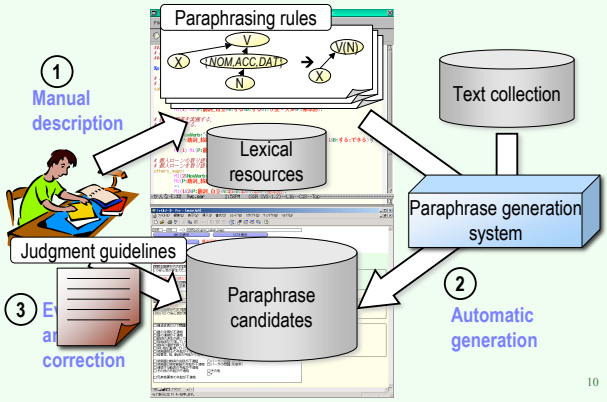
Aim of this study

- Confirm the feasibility of the method through practice
  - Given
    - A paraphrase class
    - A text collection
  - Collect paraphrase examples belonging to the class
    - At a minimal human labor cost
    - As exhaustively as possible from the text collection
    - As reliable as humanly possible

Outline

1. Overview
2. Issues and our class-oriented approach
3. Semi-automatic example collection
4. Preliminary trials
   1. Specification
   2. Discussion
5. Conclusion

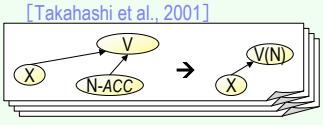
Semi-automatic example collection



Step 1: Pattern description

- Morpho-syntactic paraphrasing patterns
  - Pairs of dependency trees
  - Implemented on a paraphrase generation system

[Takahashi et al., 2001]



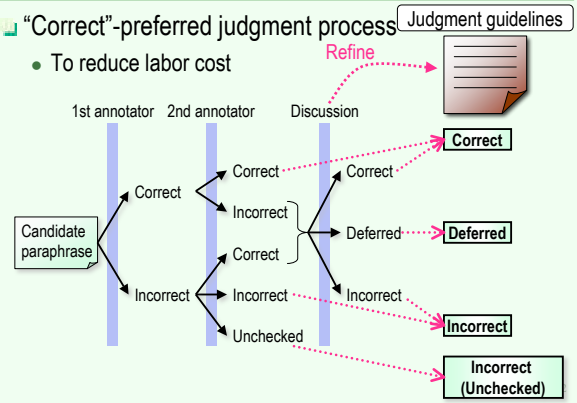
X: variable for any word
N: variable for a noun
V: variable for a verb
V(N): verbalized form of N

映画-が 彼-に 感動-を 与える
film-NOM him-DAT impression-ACC to give-ACTIVE
(The film made an impression on him.)

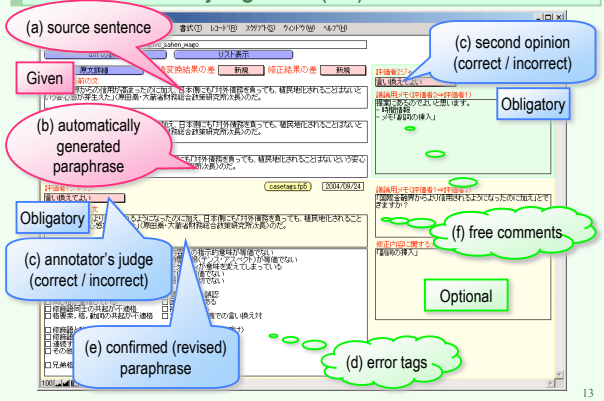
映画-が 彼-を 感動-させる
film-NOM him-ACC to be impressed-CAUSATIVE
(The film impressed him.)

Step 3: Manual judgment (mutual judgment)

- "Correct"-preferred judgment process
  - To reduce labor cost



Step 3: Manual judgment (I/F)



Outline

1. Overview
2. Issues and our class-oriented approach
3. Semi-automatic example collection
4. Preliminary trials
   1. Specification
   2. Discussion
5. Conclusion

Target classes

- Paraphrases of light-verb constructions (LVC)

映画-が 彼-に 感動-を 与える
film-NOM him-DAT impression-ACC to give-ACTIVE
(The film made an impression on him.)

映画-が 彼-を 感動-させる
film-NOM him-ACC to be impressed-CAUSATIVE
(The film impressed him.)

- Transitivity alternation (TransAlt)

そよ風-が 木々-を 揺らす
breeze-NOM tree-ACC to sway-Transitive
(The breeze sways the trees.)

木々-が そよ風-に 揺れる
tree-NOM breeze-DAT to sway-Intransitive
(The trees sway in the breeze.)

Resources

- LVC
  - 4 paraphrasing patterns (e.g. (7) in paper)
  - 20,155 pairs of 〈*deverbal noun*, *verb*〉
    - 〈感動(*impression*), 感動する(*to be impressed*)〉
    - 〈誘い(*invitation*), 誘う(*to invite*)〉
- TransAlt
  - 8 paraphrasing patterns (e.g. (10) in paper)
  - 212 pairs of 〈*intransitive verb*, *transitive verb*〉
    - 〈揺れる(*to sway-Intransitive*), 揺らす(*to sway-Transitive*)〉
    - 〈壊れる(*to break-Intransitive*), 壊す(*to break-Transitive*)〉

Results of trials

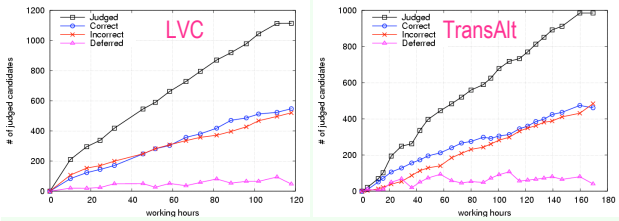
	Paraphrase class	LVC	TransAlt
Step 1	# of paraphrasing patterns	4	8
	Size of dictionary	20,155	212
Step 2	# of source sentences	10,000	25,000
	# of generated candidates	2,566	985
Step 3	# of judged candidates	1,067	964
	# of incorrect candidates	520	503
	# of correct candidates	547	461
	# of paraphrase examples	591	484
	Working hours	118	169.5

Working hours: 2 annotators' working time for (1) Judgment, (2) Discussion, and (3) Re-judgment after refining guidelines

Aim of this study (reminder)

- Confirm the feasibility of the method through practice
  - Given
    - A paraphrase class
    - A text collection
  - Collect paraphrase examples belonging to the class
    - At a minimal human labor cost
    - As exhaustively as possible from the text collection
    - As reliable as humanly possible

Cost-efficiency



- Not so wasteful human labor cost
  - 7.1 candidates / man-hour
  - 3.7 paraphrase examples / man-hour
  - TransAlt is 1.75 times more difficult than LVC due to test

Exhaustiveness

- The initial resource is not necessarily optimal
  - Paraphrasing patterns
  - Derivation pairs
- How are they optimal?
  - Estimated coverage: 77% (158/(158+48))
    - 158 paraphrases for 750 excerpted sentences
    - Manual examination obtained another 48 paraphrases
      - 47 misses can be salvaged by resource enhancement
      - Errors of shallow parsers hurt only once
  - Use of patterns is realistic approach
  - Manual examination ensures coverage

Reliability

- Strategy
  - Classification bases on guideline & linguistic intuition
  - Inter-annotator discussion refined judgment guidelines
- Agreement ratio increased (in case of LVC)
  - 74% (3rd day) → 77% (6th day)
    → 88% (9th day) → 93% (11th day)
  - It's still not easy to explain "why this is correct / incorrect"
- Future plan
  - Involve an expert to make sure of judgment guidelines
  - Involve the 3rd annotator for judgment

Conclusion

- Feasibility of a semi-automatic example collection
  - Class-oriented example collection
  - Employing a paraphrase generation system
- Promising results
  - Reasonable human labor cost, but need reduction
  - Moderately exhaustive at initial stage
  - Typically reliable, but some marginal cases
- Paraphrase sub-corpora consist of
  - LVC: 1067 candidates / 591 examples
  - TransAlt: 964 candidates / 484 examples

Future work

- Discussion on required expertise
  - It is not easy to explain "why this is correct / incorrect"
  - Involve an expert to make sure of judgment guidelines
- Build sub-corpora for other paraphrase classes
- Extrinsic evaluation through case studies
  - Resultant provides both correct and incorrect examples
  - Immediately available for analysis and system evaluation
- Publicly open the resource
  - Paraphrase corpus, Lexical resources, Judgment guidelines