

A Category-oriented Approach to Paraphrase Corpus Construction

Computational Linguistics Lab.
Atsushi Fujita

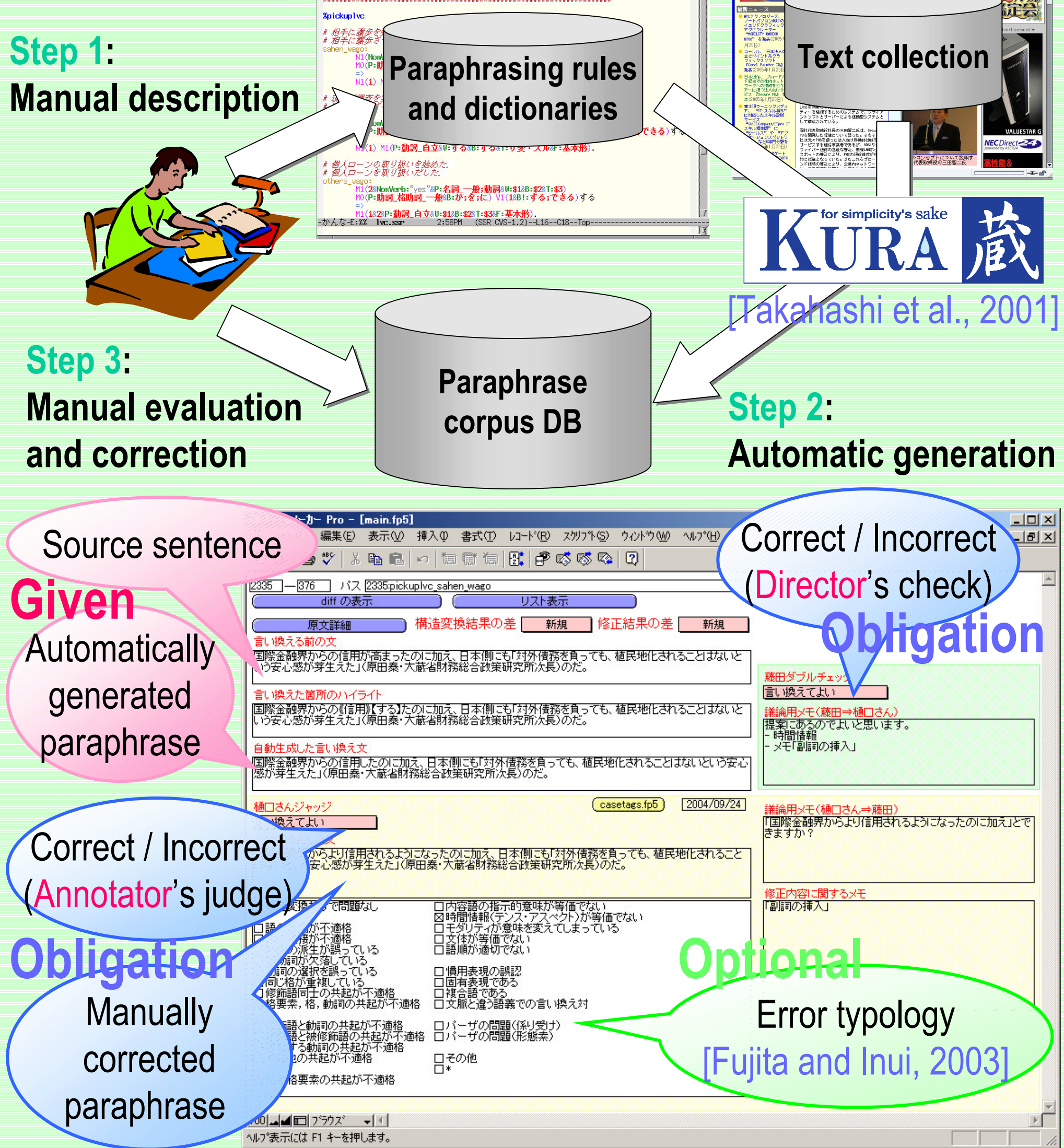
Background

- Towards intelligent treatment of text data on ubiquitous computing network
→ Processing contents written in **natural language** is important
- Handling **paraphrases** benefits various applications
 - Paraphrase generation → Reading assistance, Machine translation, etc.
 - Paraphrase recognition → Question answering, Multi-document summarization, etc.

Goal

- To create a large-scale paraphrase corpus (collection of examples)
- To analyze how we paraphrase
- To improve paraphrase generation and recognition systems
 - by inducing paraphrase patterns from corpus
 - by identifying useful linguistic knowledge for paraphrase generation
- As a standard test-set for evaluating paraphrase generation systems

Method



Advantages and future work

- Advantages of the proposed method
 - it allows amateur to engage in judge
 - it leads to rapid corpus construction (three-times faster than [Shirai et al., 2001])
 - it collects incorrect examples (for analysis)
- We will create corpora for other categories
- Discussion on required expertise
- To collect reliable and diverse data

Reading assistance

Simplifying text

tell one of my thinking

... verges with the worst ratio

voice a piece of my conviction

... is getting worse and worse

Multi-document summarization

Recognizing and summarizing same news

intranets of enterprises

to keep security

to ensure security

networks within enterprises

Product

Collection of paraphrase examples

Features

Category-oriented

Use rules and dictionaries for particular category of paraphrase phenomena

Semi-automated

Use an existing paraphrasing system to collect all possible paraphrases, reducing human labor and bias

Trials

- To confirm the feasibility of the proposed approach through creating several types of corpus (paraphrase corpora)

Paraphrasing of light-verb constructions (LVC)

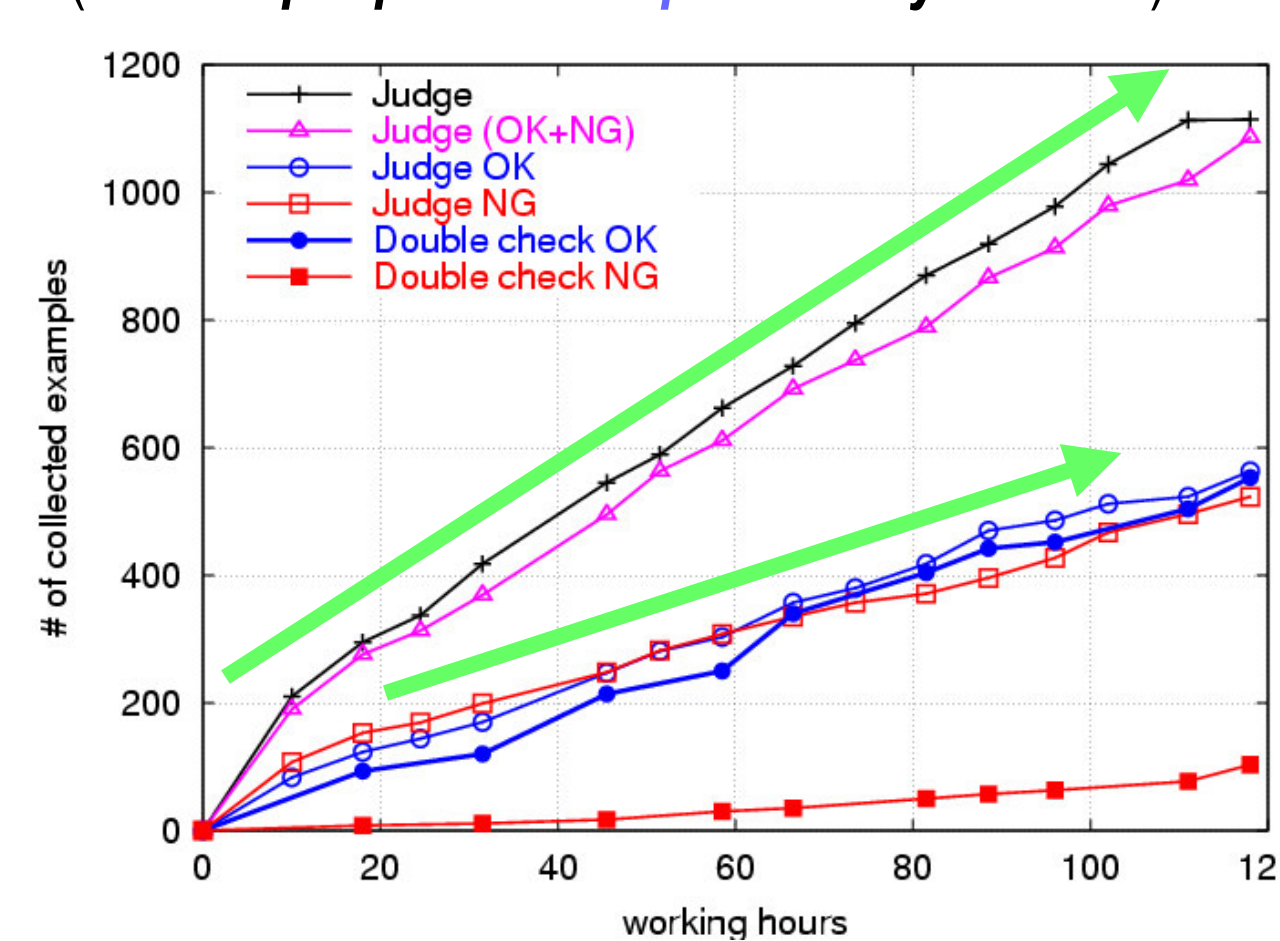
Transitivity alternation (Trans.alt)

多くの人が彼の演説に感動を受けた。
(A lot of people **received** a good impression by his talk.)

多くの人が彼の演説に感動した。
(A lot of people **were impressed** by his talk.)

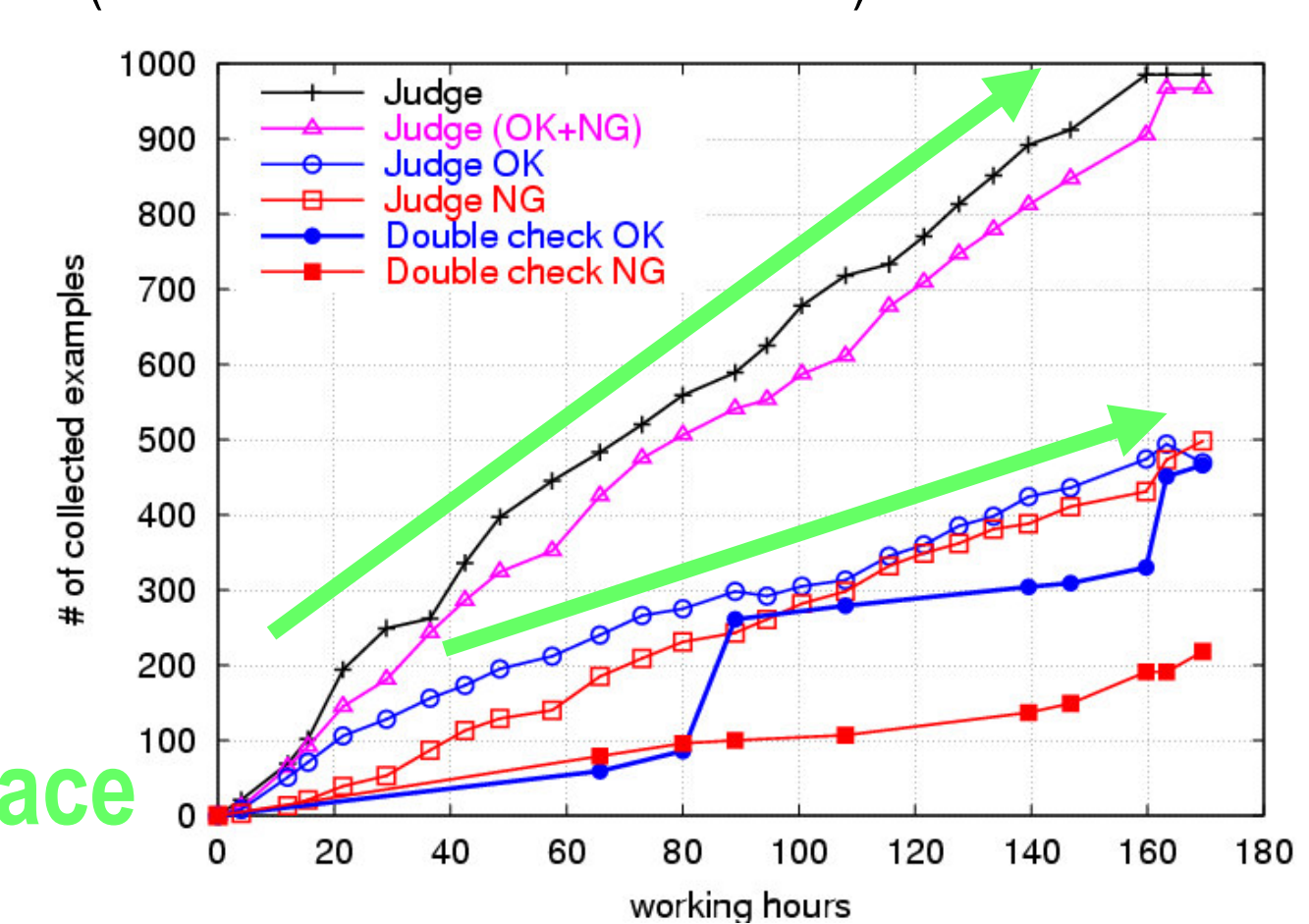
彼はハンマーで壁を壊した。
(He **broke the wall** with a hammer.)

ハンマーで壁が壊れた。
(The wall **broke** with a hammer.)



4.3 correct examples / 1 hour

Stable working pace



2.8 correct examples / 1 hour