2nd November, 2015
MT Summit 2015: MT Researchers' Track

# Japanese Controlled Language Rules to Improve Machine Translatability of Municipal Documents

**Rei Miyata**†, Anthony Hartley‡, Cecile Paris#,
Midori Tatsumi‡, Kyo Kageura†

†U of Tokyo, ‡Rikkyo U, #CSIRO

# Outline

1. Background and objectives
2. CL formulation
3. CL evaluation
4. Results and discussions
5. Conclusion and future plan

# Outline

1. Background and objectives
2. CL formulation
3. CL evaluation
4. Results and discussions
5. Conclusion and future plan

# Background

**Social**

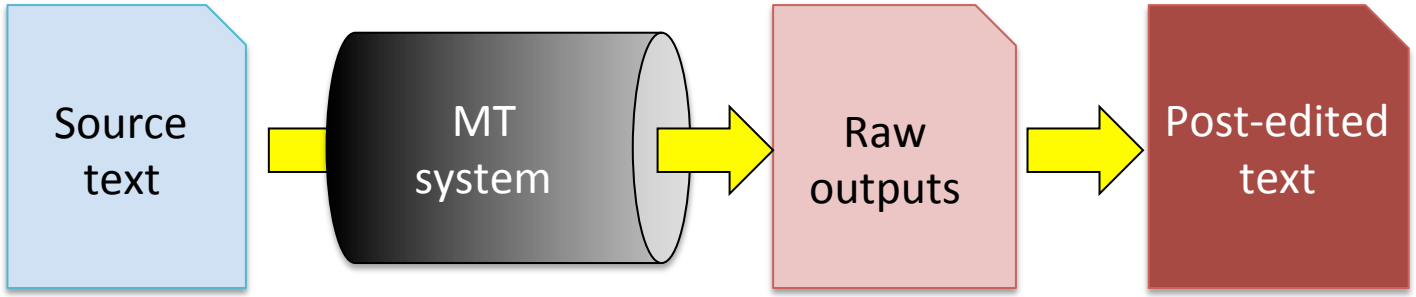
- ✓ Multilingualisation of municipal websites in Japan
- ✓ Widespread use of MT

**Technical**

- ✓ Difficulties in Ja-En MT
- ✓ 'Black box' use setting

**Controlled Language (CL)**

ex) AECMA, Caterpillar Technical English, KANT
ex) Technical Japanese, Simplified Technical Japanese



# Problems

- Applicability of CL rules (O'Brien 2006)
  - To what extent can CL rules be effectively generalised across MT systems and language pairs?
- Compatibility of source text (ST) and target text (TT) quality (Hartley et al. 2012; Tatsumi et al. 2013)
  - Conflict between ST and TT quality

# Objectives and scope

- Objectives

1. **[CL formulation]**
   Propose and implement a protocol to formulate CL rules
2. **[CL evaluation]**

   Investigate to what extent we can improve MT quality through CL rules, without degrading ST readability.

   **Applicability**

   **Compatibility**

- Scope
  - Municipal website documents
  - Japanese to English MT (4 systems)

# Outline

1. Background and objectives
2. CL formulation
3. CL evaluation
4. Results and discussions
5. Conclusion and future plan

# Requirements

Our CL rules should:

1. help to raise the quality of MT outputs (TT quality)
2. not degrade the quality of the Japanese source texts (ST quality)

# CL formulation protocol

- Principle

Comparing original source texts and more machine translatable ones rewritten by humans

- Trial and error protocol
  1. Rewrite a source text aiming at a better quality of MT output
  2. Record how the text was changed and assess the quality of the output
  3. Repeat steps 1 and 2, until achieving satisfactory quality of the MT output

# Implementation of the protocol

| ST | MT | Change |
|---|---|---|
| 電力会社に連絡、使用開始手続き完了後、ブレーカーのスイッチを入れます | You can turn on the contact, the procedures after completion of the electric power companies | [Original sentence] |
| 電力会社に連絡します。使用開始の手続きが完了した後、ブレーカーのスイッチを入れます。 | Will contact the electric power company. Procedures for activation is complete, you turn on the breaker. | Split sentence<br>Add “します”<br>Add “の”<br>“完了後”→“完了した後” |
| 電力会社に連絡してください。使用開始の手続きが完了した後に、ブレーカーのスイッチを入れてください。 | Please contact the electric power company. Please turn on the breaker after the procedure of the activation is completed. | “します”→“してください”<br>Add “に”<br>“ます”→“てください” |

**1 rewriter, 100 sentences in municipal domain, 3 MT systems**

# 38 features to be regulated

1. multiple verbs in a sentence
2. lack of subject
3. lack of object
4. connection
5. particle Ga (が) for object
6. enumeration A-Mo, B-Mo (Aも、Bも)
7. Te-kuru (てくる) / Te-iku (ていく)
8. inserted adverbial clause
9. ending clause with Noun
10. Sahen-noun + Desu (です)
11. attributive use of Shika - Nai (しか～ない)
12. verb + You (よう)
13. A or not
14. Sahen-noun + Wo (を) + Suru (する)
15. Sahen-noun+ Sareru（される）
16. particle Nado (など・等)
17. giving and receiving verb
18. redundant word
19. compound word
20. omission
21. suffix
22. particle Made
23. particle De (で)
24. particle No (の) to mean "by" or "from"
25. per A
26. particle Te (て)
27. if particle To (と)
28. particle He-Ha (へは)
29. particle Ni-Ha (には)
30. particle No-Ka (のか)
31. demonstrative pronoun (ko-so-a-do)
32. particle Ni (に)
33. Japanese Kana / Chinese Kanji / number
34. bullet mark
35. unit
36. punctuation (sentence separation)
37. square bracket
38. wave dash (～)

# CL rules (examples)

Rule 1. Avoid using multiple verbs in a sentence.

Rule 2. Avoid omitting subject.

Rule 9. Do not end clause with noun.

Rule 11. Avoid using attributive use of Shika-Nai (しか-ない).

Rule 14. Avoid using Sahen-noun + Wo (を) + Suru (する).

Rule 20. Do not omit parts of words or sentences.

Rule 24. Avoid using particle No (の) to mean “by” or “from”.

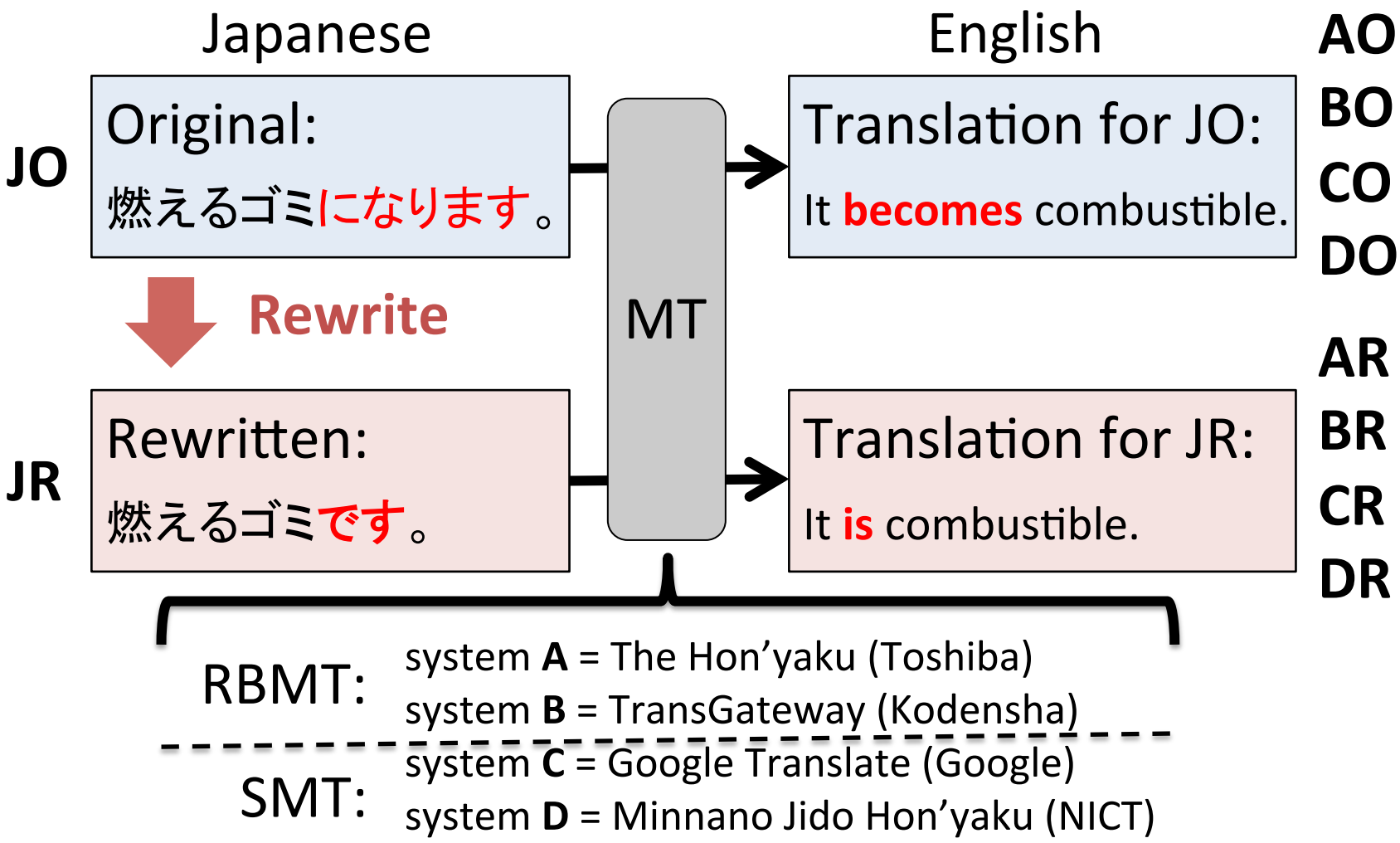
Rule 36. Use punctuations properly to separate sentences.

Rule 38. Avoid using wave dash (〜).

# Outline

1. Background and objectives
2. CL formulation
3. CL evaluation
4. Results and discussions
5. Conclusion and future plan

# Evaluation framework



- Data

4 sentences * 38 rules = 152 sentences

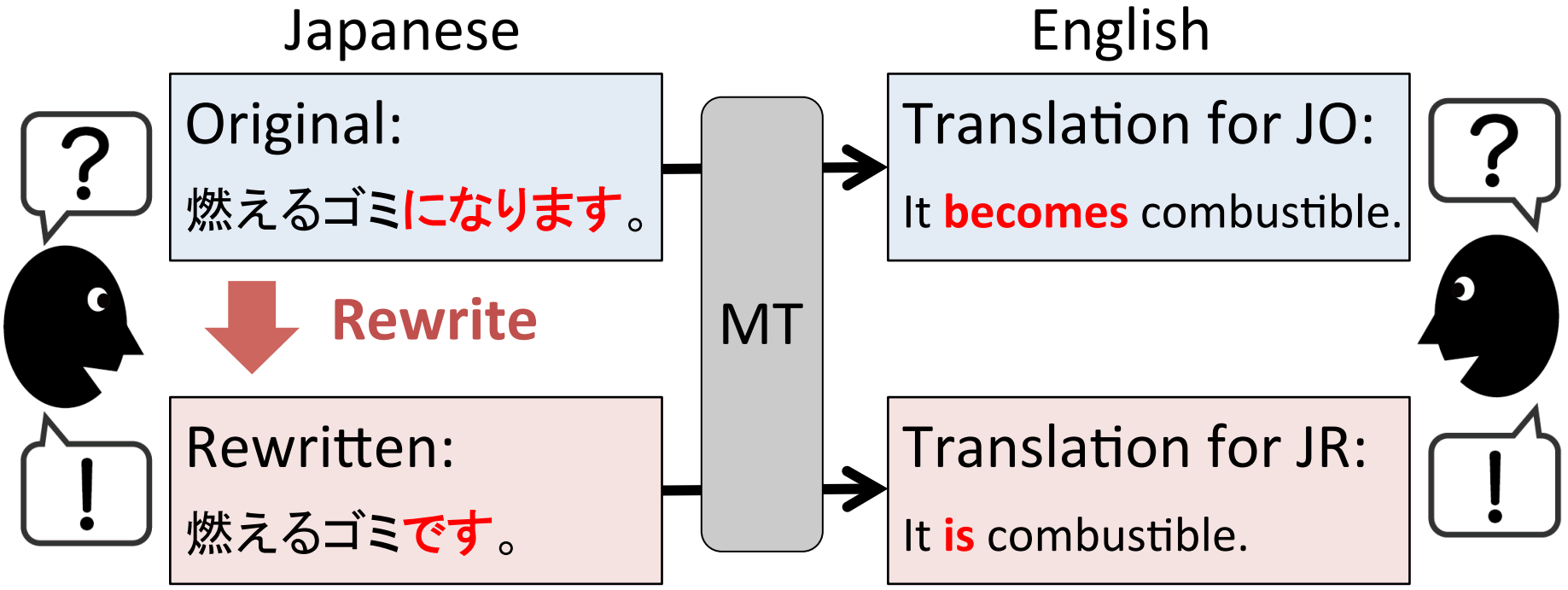
  - 152 **J**apanese **O**riginal sentences (**JO**)
  - 152 **J**apanese **R**ewritten sentences (**JR**)
  - 304 **H**uman translations (**HO**/**HR**)
  - 1216 MT sentences
    - 304 outputs of system **A** (**AO**/**AR**)
    - 304 outputs of system **B** (**BO**/**BR**)

    } RBMT

    - 304 outputs of system **C** (**CO**/**CR**)
    - 304 outputs of system **D** (**DO**/**DR**)

    } SMT

# Evaluation framework




**(2) Japanese readability for ST readers**

**(1) MT quality for TT readers**

# (1) MT quality evaluation

- Two-step evaluation method (Tatsumi et al. 2013)
  1. [Understandability]
     How much they understood and how much effort was required
  2. [Accuracy]
     How close the meaning of human translation was to their understanding of MT output
- 24 adult English speakers with little Japanese knowledge
  - Each sentence was evaluated by 3 judges

# (2) Japanese readability evaluation

- Method (Hartley et al. 2012)
  - Present both JO and JR
  - Ask judges to evaluate the readability of each sentence on a four-point scale
- 3 Japanese native speakers
  - Each sentence was evaluated by 3 judges

# Outline

1. Background and objectives
2. CL formulation
3. CL evaluation
4. Results and discussions
5. Conclusion and future plan

# (1) MT quality evaluation

- 4 categories of MT quality

| **Category** | Understandability | Accuracy |
|---|---|---|
| MT useful | ○ | ○ |
| MT inaccurate | ○ | × |
| MT unintelligible | × | — |
| HT unintelligible | × | — |

(even human translation was not understandable)

# Overall results

| | MT useful | MT inaccurate | MT unintelligible | HT unintelligible |
|---|---|---|---|---|
| AO | 27.4% | 4.6% | 62.1% | 5.9% |
| AR | 30.9% | 5.5% | 58.8% | 4.8% |
| BO | 23.2% | 5.0% | 66.0% | 5.7% |
| BR | 27.2% | 5.7% | 63.4% | 3.7% |
| CO | 26.5% | 3.9% | 64.7% | 4.8% |
| CR | 30.0% | 6.8% | 58.3% | 4.8% |
| DO | 27.0% | 6.4% | 61.0% | 5.7% |
| DR | 26.3% | 6.8% | 60.1% | 6.8% |

**3–4% increase in system A, B and C**

# MT useful case (rule1–13)

(%)

| | Rule | Feature | A | B | C | D |
|---|---|---|---|---|---|---|
| | 1 | multiple verbs in a sentence | −25.0 | 8.3 | −8.3 | −16.7 |
| ✔ | 2 | lack of subject | 25.0 | 25.0 | 25.0 | −16.7 |
| | 3 | lack of object | 8.3 | 33.3 | 0.0 | −8.3 |
| ✔ | 4 | connection | 0.0 | 16.7 | 8.3 | 16.7 |
| | 5 | particle Ga (が) for object | 0.0 | 8.3 | −8.3 | −16.7 |
| | 6 | enumeration A-Mo, B-Mo (Aも、Bも) | 8.3 | −16.7 | 0.0 | −16.7 |
| | 7 | Te-kuru (てくる) / Te-iku (ていく) | −16.7 | 0.0 | 8.3 | 25.0 |
| ✔ | 8 | inserted adverbial clause | 33.3 | −16.7 | 8.3 | 16.7 |
| | 9 | ending clause with Noun | −8.3 | 0.0 | 0.0 | 25.0 |
| ✔ | 10 | Sahen-Noun + Desu (です) | 16.7 | 16.7 | 25.0 | 16.7 |
| | 11 | attributive use of Shika – Nai (しか～ない) | −8.3 | 0.0 | 25.0 | 8.3 |
| ✔ | 12 | verb + You (よう) | 8.3 | −8.3 | 8.3 | 16.7 |
| ✔ | 13 | A or not | 16.7 | 25.0 | 8.3 | 8.3 |

✔ : positive effects on 4 MT systems
✔ : positive effects on 3 MT systems
} Generally applicable CL rules

# Example

## Rule 2. Avoid omitting subject

**[AO]** A home and the community are places where a child spends much time daily, and study that it is various in a life.

**[AR]** A home and the community are places where a child spends much time daily, and a child studies that it is various in a life.

**[BO]** A house and an area are the place where a child spends much time daily, and various things will be learned in the life.

**[BR]** A house and an area are the place where a child spends much time daily, and a child will learn various things in the life.

**[CO]** Home and regions, children are routinely spend place a lot of time, you will learn a variety of things in life.

**[CR]** Home and regions, children are routinely spend place a lot of time, children will learn a variety of things in life.

# Applicability of the rules

- Generally applicable CL rules
  - 11 rules have positive effects on at least three MT systems
- MT-dependent CL rules
  - Effectiveness of the CL rule is varied depending on the systems
  - RBMT (system A, B) versus SMT (system C, D)?

  → Different "reactions" do not correlate to the different "architectures"

# Optimal rule set

*18 rules for system A, 19 for B, 19 for C and 16 for D

| | MT useful | MT inaccurate | MT unintelligible | HT unintelligible |
|---|---|---|---|---|
| AO | 20.6% | 3.9% | 68.9% | 6.6% |
| AR | 37.3% | 3.9% | 55.7% | 3.1% |
| BO | 20.6% | 4.4% | 70.6% | 4.4% |
| BR | 38.2% | 3.5% | 54.4% | 3.9% |
| CO | 21.1% | 4.8% | 70.2% | 3.9% |
| CR | 36.4% | 5.3% | 53.1% | 5.3% |
| DO | 17.7% | 7.8% | 69.3% | 5.2% |
| DR | 34.4% | 6.8% | 53.1% | 5.7% |

**More than 15% increase for all systems**

# (2) Japanese readability evaluation

- 2 categories of JA readability

| Category | Option |
|---|---|
| Acceptable | 1. Easy to read |
| | 2. Fairly easy to read |
| Unacceptable | 3. Fairly difficult to read |
| | 4. Difficult to read |

# Japanese readability (rule1–13)

| Rule | Feature | Improvement (%) |
|---|---|---|
| 1 | multiple verbs in a sentence | −16.7 |
| 2 | lack of subject | 41.7 |
| 3 | lack of object | 33.3 |
| 4 | connection | −25.0 |
| 5 | particle Ga (が) for object | 33.3 |
| 6 | enumeration A-Mo, B-Mo (Aも、Bも) | −41.7 |
| 7 | Te-kuru (てくる) / Te-iku (ていく) | 16.7 |
| 8 | inserted adverbial clause | 0.0 |
| 9 | ending clause with Noun | −16.7 |
| 10 | Sahen-Noun + Desu (です) | 33.3 |
| 11 | attributive use of Shika – Nai (しか～ない) | −8.3 |
| 12 | verb + You (よう) | 8.3 |
| 13 | A or not | −16.7 |

# Compatibility of ST and TT quality

| Rule | A | B | C | D | ST |
|---|---|---|---|---|---|
| 1 | | ✔ | | | |
| 2 | ✔ | ✔ | ✔ | | ✔ |
| 3 | ✔ | ✔ | | | ✔ |
| 4 | | ✔ | ✔ | ✔ | |
| 5 | | ✔ | | | ✔ |
| 6 | ✔ | | | | |
| 7 | | | ✔ | ✔ | ✔ |
| 8 | ✔ | | ✔ | ✔ | ✔ |
| 9 | | | | ✔ | |
| 10 | ✔ | ✔ | ✔ | ✔ | ✔ |
| 11 | | | ✔ | ✔ | |
| 12 | ✔ | | ✔ | ✔ | ✔ |
| 13 | ✔ | ✔ | ✔ | ✔ | |
| 14 | | | ✔ | | ✔ |
| 15 | | ✔ | | ✔ | ✔ |
| 16 | | | ✔ | ✔ | ✔ |
| 17 | | ✔ | | | ✔ |
| 18 | | | ✔ | ✔ | ✔ |
| 19 | | ✔ | | ✔ | |

| Rule | A | B | C | D | ST |
|---|---|---|---|---|---|
| 20 | ✔ | ✔ | ✔ | | |
| 21 | ✔ | | ✔ | | |
| 22 | ✔ | | ✔ | | |
| 23 | | ✔ | | | |
| 24 | ✔ | | | | ✔ |
| 25 | ✔ | ✔ | ✔ | ✔ | ✔ |
| 26 | | ✔ | | | ✔ |
| 27 | | ✔ | | | ✔ |
| 28 | ✔ | | ✔ | ✔ | ✔ |
| 29 | | ✔ | | | ✔ |
| 30 | | ✔ | | | |
| 31 | | | ✔ | | ✔ |
| 32 | | | | | |
| 33 | | ✔ | | | ✔ |
| 34 | ✔ | | ✔ | | ✔ |
| 35 | ✔ | ✔ | ✔ | ✔ | |
| 36 | ✔ | | | | ✔ |
| 37 | ✔ | ✔ | | ✔ | ✔ |
| 38 | ✔ | | | | |

# Example (adverse effect)

## Rule 20. Avoid omission

**[JO]** 月・水・金曜日の午前9時から午後4時まで開設しており、3月末まで開設しています。

**[BO]** It's established from a month and 9:00am of water and Friday to 4:00pm and it's established until the end of March.

↓

**[JR]** 月曜日・水曜日・金曜日の午前9時から午後4時まで開設しており、3月末まで開設しています。

**[BR]** It's established from 9:00am of Monday, Wednesday and Friday to 4:00pm and it's established until the end of March.

# Outline

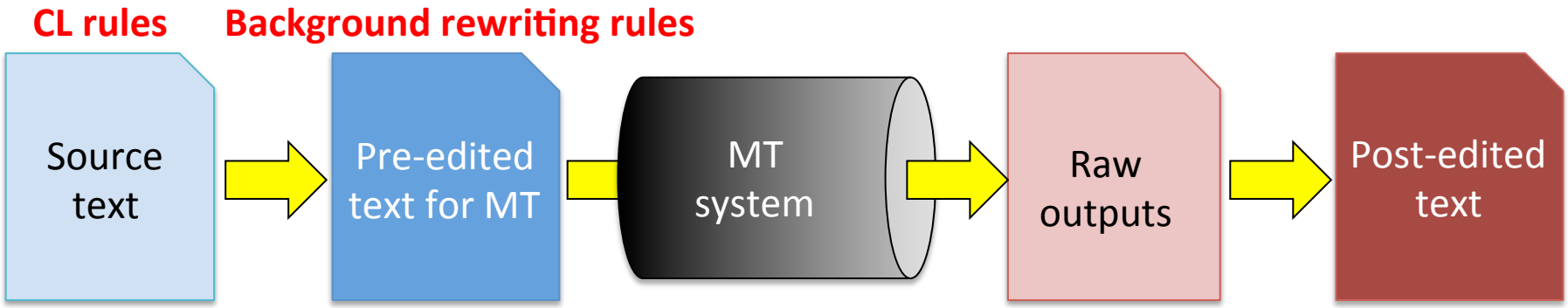
1. Background and objectives
2. CL formulation
3. CL evaluation
4. Results and discussions
5. Conclusion and future plan

# Conclusion

- Applicability
  - 11 general rules effective for at least three MT systems
  - **More than 15% improvement if we compile optimal rule sets**
- Compatibility
  - About two thirds of the CL rules improved or retained Japanese source readability
  - Degradations in readability for humans often correlate with redundancy generated by the rules

# Future plan

- CL rules that degrade ST quality become background (automatic) rewriting rules




- Implement our CL formulation protocol to:
  - other language pairs
  - other text domains

Thank you!

mutual-project.com