2017年3月8日
第8回産業日本語研究会・シンポジウム

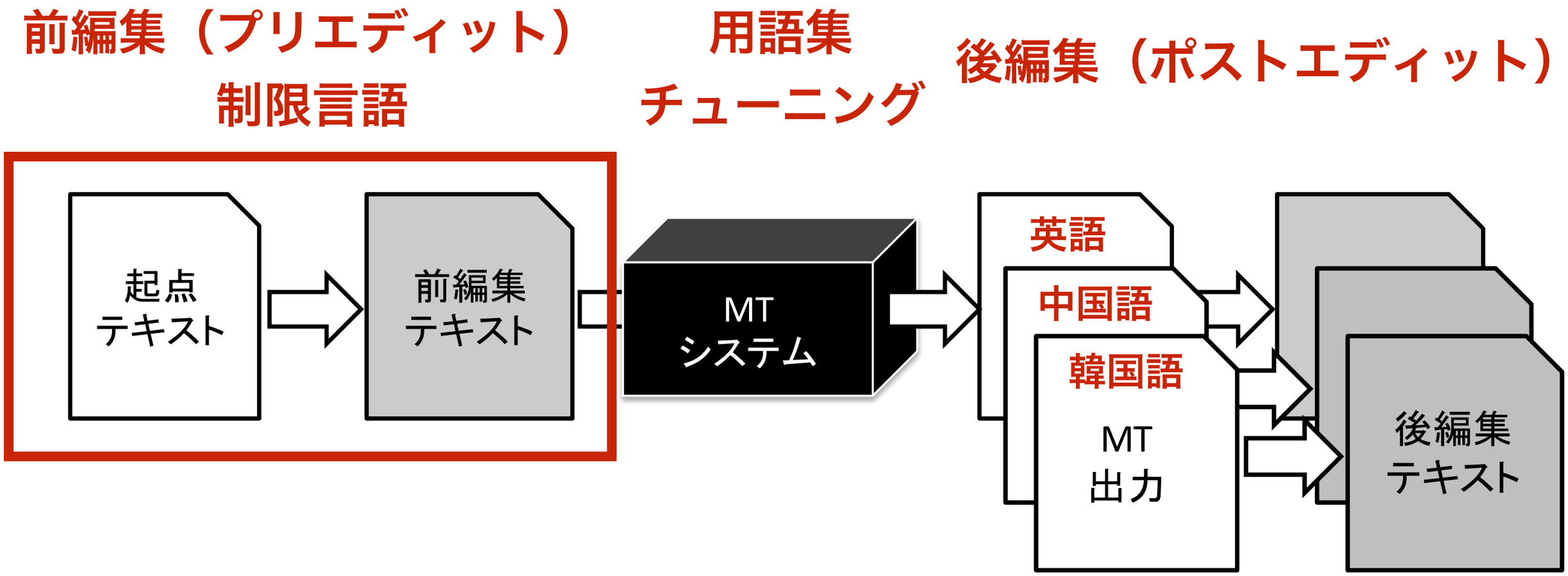
制限言語とオーサリング支援システム：機械翻訳を活用した文書の多言語展開に向けて

宮田玲†, Anthony Hartley‡, 影浦峡†, Cécile Paris#
†東京大学大学院教育学研究科, ‡立教大学, #CSIRO

機械翻訳（MT）を「うまく使う」には？

背景情報や文脈を活用することは難しい

→最初から曖昧・複雑な文を避ける



MuTUAL project

1. 自治体手続き文書
2. 日英方向
3. 統計的MT
   ルールベースMT

(Miyata et al. 2016a; 2016b)

制限言語

Controlled Language (CL):

> ‘A controlled natural language is a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics, while preserving most of its natural properties’.

(Kuhn 2014, p.123)

ex) ASD-STE100, Caterpillar Technical English, SMART, PLAIN, Basic English

ex) 特許ライティングマニュアル, Simplified Technical Japanese, やさしい日本語

自治体文書向け規制表現パターン (Miyata et al. 2015)

構文レベル	
1	一文に複数の動詞
2	主語の欠如
3	目的語の欠如
4	読点を用いた並列要素の列挙
5	目的語への助詞「が」
6	並列表現「Aも、Bも」
7	「てくる」／「ていく」
8	文中の副詞句の挿入
9	体言止め
10	サ変名詞＋「です」
11	「しか～ない」
12	動詞＋「ように」
13	「かどうか」
14	サ変名詞＋「をする」
15	サ変名詞＋「される」
語彙レベル	
16	「など」／「等」
17	授受動詞
18	冗長語
19	複合語
20	列挙項目中の要素の省略
21	接尾辞
22	助詞「まで」（宛先の用法）
23	助詞「で」
24	助詞「の」（「による」「から」の意
25	単位表現「につき」の省略
26	助詞「て」
27	助詞「と」（条件用法）
28	助詞「へは」
29	助詞「には」
30	助詞「のか」
31	指示代名詞（こそあど）
32	助詞「に」
表記レベル	
33	ひらがな表記
34	箇条書き記号
35	機種依存文字
36	読点
37	強調のカギ括弧
38	波ダッシュ

例

- 列挙項目中の要素の省略を避ける

ST 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。

MT 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.

⇩

ST 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。

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

例

- 尊敬用法の「れる・られる」を避ける

ST 託児を利用される場合は、１０日前までにファミリー・サポート・センター事務局へ予約をお願いします。

MT Used daycare, I'd like to make a reservation to a Family Support Center office 10 days in advance.

⇩

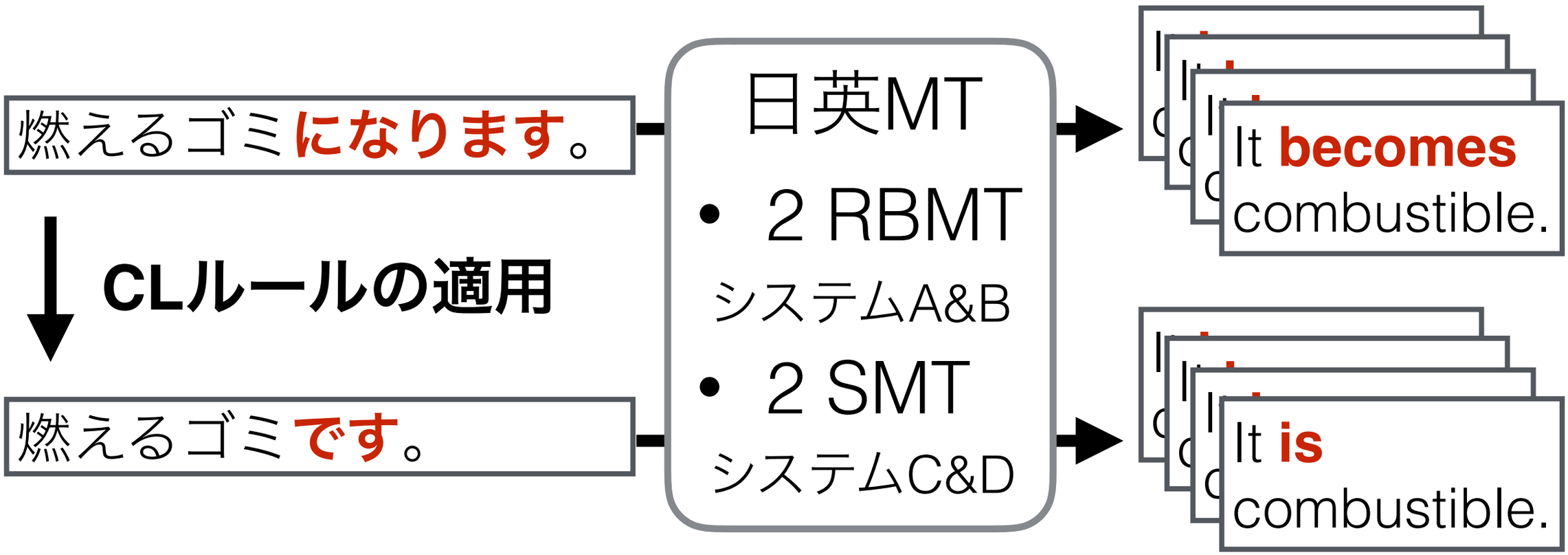
ST 託児を利用する場合は、１０日前までにファミリー・サポート・センター事務局へ予約をお願いします。

MT If you use day care, I'd like to make a reservation to a Family Support Center office 10 days in advance.

考慮すべき点

- 目的：MT訳の品質向上／原文の可読性向上…
- 言語方向：日英／日中／日韓／英仏…
- 文書ドメイン：産業文書／特許文書／自治体文書…
- 使用するMT：RBMT／EBMT／SMT／NMT

ルールの性能評価実験の枠組み



原文の品質:

「読みやすさ」

(Hartley et al. 2012)

MT訳の品質:

1.「理解度」

2.「正確性」

評価結果

MT訳の品質：向上したか？　原文の品質：低下していないか？

Rule	MTシステム A	B	C	D	原文
1		✔			
2	✔	✔	✔		✔
3	✔	✔			✔
4		✔	✔	✔	
5		✔			✔
6	✔				
7			✔	✔	✔
8	✔		✔	✔	✔
9				✔	
10	✔	✔	✔	✔	✔
11			✔	✔	
12	✔		✔	✔	✔
13	✔	✔	✔	✔	
14			✔		✔
15		✔		✔	✔
16			✔	✔	✔
17		✔			✔
18			✔	✔	✔
19		✔		✔	

Rule	MTシステム A	B	C	D	原文
20	✔	✔	✔		
21	✔		✔		
22	✔		✔		
23		✔			
24	✔				✔
25	✔	✔	✔	✔	✔
26		✔			✔
27		✔			✔
28	✔		✔	✔	✔
29		✔			✔
30		✔			
31			✔		✔
32					
33		✔			✔
34	✔		✔		✔
35	✔	✔	✔	✔	
36	✔				✔
37	✔	✔		✔	✔
38	✔				

複数のMTに有効なルール

- 主語を省略しない

家庭や地域は、子どもが多くの時間を日常的に過ごす場所であり、
[子どもは] 生活の中で様々なことを学んでいきます。

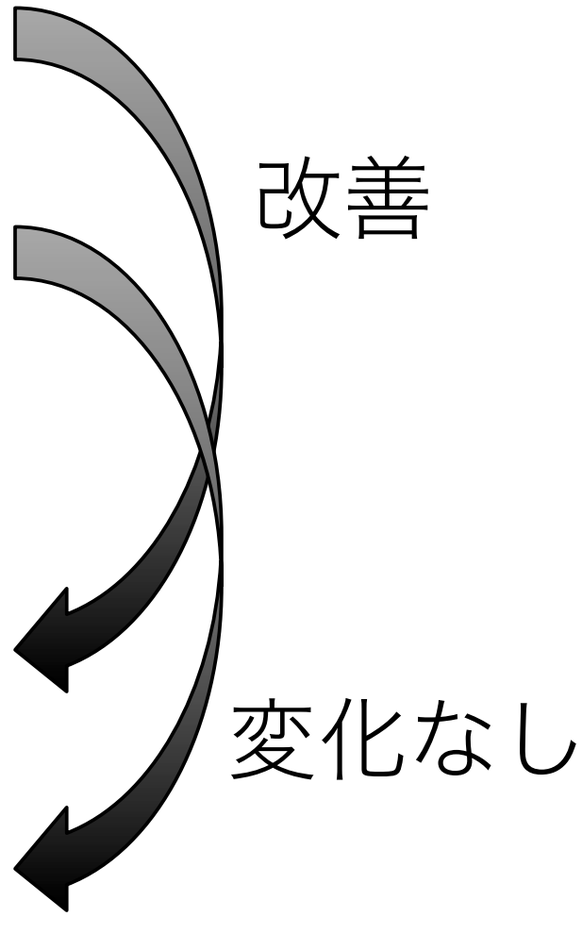
システム	CL	翻訳結果
A	前	A home and the community are places where a child spends much time daily, and **study** that it is various in a life.
A	後	A home and the community are places where a child spends much time daily, and **a child studies** that it is various in a life.
B	前	A house and an area are the place where a child spends much time daily, and **various things will be learned** in the life.
B	後	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.
C	前	Home and regions, children are routinely spend place a lot of time, **you will learn** a variety of things in life.
C	後	Home and regions, children are routinely spend place a lot of time, **children will learn** a variety of things in life.

システム依存のルール

- 「～しか～ない」を避ける

この店では、現金しか使えません。

A. In this shop, you can not use only cash.
B. In this store, we do not use only cash.
C. Only cash can be used at this store.
D. In this shop, only cash is usable.



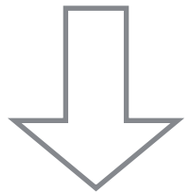
この店では、現金のみ使えます。

A. In this shop, you can use only cash.
B. In this shop, I use only cash.
C. Only cash can be used at this store.
D. In this shop, only cash is usable.

読みやすさが低下することも・・・

- 並列要素を読点でつなげない

本園では飼育実習、学芸員実習の受け入れを行っております。



本園では飼育実習及び学芸員実習の受け入れを行っております。

前半のまとめ

- 目的・言語方向・文書ドメイン・MTシステムに応じて、制限言語（CL）をデザインし、効果を検証することが重要

  - システムAに有効なルールが、システムBに有効とは限らない

  ➔ 特定のMTにあわせてルールを選択する

- MT訳の品質向上に寄与するルールが、原文の読みやすさを低下させることもある

  ➔ CLの用途・目的に応じて、バランスをとる

CLオーサリングの難しさ

表記ゆれ

ルール 5: 目的語への「が」

災害航空隊は、災害発生時に直ちに防災ヘリコプターが
運航できるように、24時間勤務体制とする。

ルール 8:長い副詞句の挿入

ルール 19: 複合名詞

何が難しいのか？

1. CLの**理解**が難しい

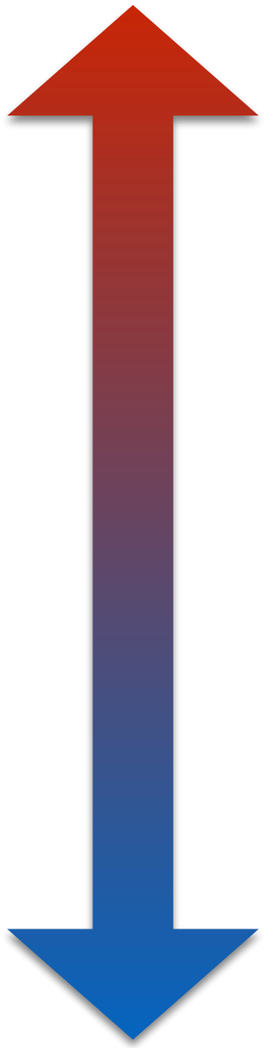
    - ガイドラインが意味不明
    - 文法用語が分からない（e.g. サ変名詞ってなに？）

2. CLの**使用**が難しい

    - ルールに違反した箇所を見逃してしまった
    - 修正すべき箇所は見つけたけど、どう直せばいいか分からない
    - 語彙・用語をいちいちチェックするのは大変

オーサリング支援

CLの実運用における執筆・書き換え支援の重要性



自動書き換え

(Mitamura & Nyberg, 2001; Shirai et al., 1998)

機械的な執筆・書き換え支援

→ CLチェッカー

ex) EN (Mitamura et al., 2003), DE (Rascu, 2006),
EL (Karkaletsis et al., 2001), JA (Nagao et al., 1984)

※その他商用ソフト：Acrolinx, MAXitなど

手作業の執筆・書き換え

介入のフェーズと深度

- **フェーズ：**いつサポートするのか？
  - 執筆段階（authoring）
  - 書き換え段階（rewriting）
- **深度：**どこまでサポートするのか？
  - 違反箇所の指摘（detection）
  - 書き換え候補の提示（suggestion）
  - 自動修正／修正支援（correction）

日本語CLオーサリング支援ツール

Rewriting task　JA　EN

Japanese Sentence

災害航空隊は、災害発生時に直ちに防災ヘリコプターが運航できるように、24時間勤務体制とする。

1文目：46文字　一文がやや長いです

災害航空隊 は、災害発生時に直ちに防災ヘリコプター が 運航できるように、24時間勤務体制とする。

9 災害航空隊：複合名詞（連続した3語以上の名詞の連なり）をなるべく使わないでください。

9 災害発生時：複合名詞（連続した3語以上の名詞の連なり）をなるべく使わないでください。

8 ヘリコプターが運航できる：目的語につける助詞は、「〜が」ではなく「〜を」を使ってください。

9 時間勤務体制：複合名詞（連続した3語以上の名詞の連なり）をなるべく使わないでください。

実用に関する問い

- どのくらいの精度・再現率で、ルールに違反した言語表現を自動的に検出できるか？
- どのくらいの精度であれば、人間のユーザー（執筆者）はストレスなくツールを使うことができるか？

	違反箇所	違反箇所ではない
検出した	**[A]** 正しく検出	**[B]** 間違って検出
検出しなかった	**[C]** 検出漏れ	**[D]** 検出せずにOK

$$\textbf{精度} = \frac{[A]}{[A] + [B]}$$

$$\textbf{再現率} = \frac{[A]}{[A] + [C]}$$

CL違反箇所検出性能の評価結果

(Miyata et al. 2016c)

#V: データ中の違反箇所の数

P: 精度　R: 再現率　F: F値

No	検出表現	#V	P	R	F
2	主語の欠如	26	0.630	0.654	0.642
3	目的語の欠如	15	0.333	0.667	0.444
4	読点を用いた並列要素の列挙	20	0.740	1.000	0.851
5	目的語への助詞「が」	5	1.000	0.800	0.889
7	「てくる」／「ていく」	6	1.000	1.000	1.000
8	文中の副詞句の挿入	7	0.286	0.571	0.381
9	体言止め	4	0.111	0.750	0.194
10	サ変名詞＋「です」	6	1.000	1.000	1.000
11	「しか～ない」	4	1.000	1.000	1.000
12	動詞＋「ように」	5	1.000	1.000	1.000
15	サ変名詞＋「される」	4	0.500	1.000	0.667
16	「など」／「等」	22	1.000	1.000	1.000
17	授受動詞	4	1.000	1.000	1.000
18	冗長語	5	1.000	1.000	1.000
19	複合語	35	0.897	1.000	0.946
20	列挙項目中の要素の省略	5	0.429	0.600	0.500
25	単位表現「につき」の省略	5	1.000	1.000	1.000
26	助詞「て」	14	1.000	0.857	0.923
27	助詞「と」（条件用法）	5	1.000	0.800	0.889
33	ひらがな表記	4	0.364	1.000	0.533
34	箇条書き記号	9	1.000	0.889	0.941
35	機種依存文字	7	1.000	1.000	1.000
37	強調のカギ括弧	6	0.500	0.500	0.500
	Total	223	0.676	0.870	0.761

例

「しか～ない」

「しか」＋「ません」のパターンの検索

自生地には観察会の2日間**しか**入れ**ません**

サ変名詞＋「される」（尊敬用法）

尊敬用法の同定

すでに**請求された**方は対象になりません

受身用法の同定

在留期間が3か月を超えて適法に在留する外国人の方も、住民票に**記載される**ようになります

後半のまとめ

- まずは、執筆者＝人間にとって、何が難しいかの見極めが肝心

- システムでどこまでサポートできるのか、サポートすべきか？

- 比較的簡単な文字列パターンマッチングで、（ある程度）CL違反箇所を検出できそう

- 再現率／精度のバランス

今後の課題と展望

- ニューラルMTと制限言語の相性
  - 翻訳結果の制御可能性：RBMT > SMT > NMT
- 違反箇所の検出性能の向上
- システム・ユーザビリティの実証評価
  - 機能とインタフェースの改善
  - CLオーサリングの訓練

参考文献

Adriaens, G. and Schreurs, D. (1992). From Cogram to Alcogram: Toward a controlled English grammar checker. COLING 1992, 595–601.

AECMA (1995). A guide for the preparation of aircraft maintenance documents in the aerospace maintenance language AECMA Simplified English. AECMA Document, PSC-85-16598.

Kuhn, T. (2014). A survey and classification of controlled natural languages. Computational Linguistics, 40(1): 121–170.

Hartley, A., Tatsumi, M., Isahara, H., Kageura, K., and Miyata, R. (2012). Readability and translatability judgments for 'Controlled Japanese'. EAMT 2012, 237–244.

Karkaletsis, V., Samaritakis, G., Petasis, G., Farmakiotou, D., Androutsopoulos, I., Markantonatou, S., and Spyropoulos, C. D. (2001). A controlled language checker based on the Ellogon text engineering platform. NAACL 2001, Software Demonstrations, 90–103.

Mitamura, T., Baker, K., Nyberg, E., and Svoboda, D. (2003). Diagnostics for interactive controlled language checking. EAMT/CLAW 2003, 237–244.

Mitamura, T. and Nyberg, E. (2001). Automatic rewriting for controlled language translation. NLPRS 2001 Workshop on Automatic Paraphrasing: Theories and Applications, 1–12.

Miyata, R., Hartley, A., Kageura, K., and Paris, C. (2016a). 'Garbage Let's Take Away': Producing understandable and translatable government documents: A case study from Japan. Social Media for Government Services, 367–393. Springer, Basel.

Miyata, R., Hartley, A., Kageura, K., Paris, C., Utiyama, M., and Sumita, E. (2016b). MuTUAL: A controlled authoring support system enabling contextual machine translation. COLING 2016, System Demonstrations, 35–39.

参考文献

宮田玲, Hartley, A., 影浦峡, Paris, C. (2017). 制限言語執筆支援システムのユーザビリティ評価. 言語処理学会第23回年次大会 (発表予定).

Miyata, R., Hartley, A., Paris, C., and Kageura, K. (2016c). Evaluating and implementing a controlled language checker. CLAW 2016, 30–35.

Miyata, R., Hartley, A., Paris, C., Tatsumi, M., and Kageura, K. (2015). Japanese controlled language rules to improve machine translatability of municipal documents. MT Summit XV, 90–103.

長尾真, 田中伸佳, 辻井潤一. (1984). 制限文法にもとづく文章作成援助システム. 情報処理学会研究報告, NL(44): 33–40.

Nyberg, E. and Mitamura, T. (2000). The KANTOO machine translation environment. AMTA 2000, 192–195.

O'Brien, S. (2003). Controlling controlled English: An analysis of several controlled language rule sets. EAMT/CLAW 2003, 105–114.

O'Brien, S. (2006). Controlled language and post-editing. Multilingual, 17(7): 17–19.

小倉英里, 工藤真代, 柳英夫. (2010). シンプリファイド・テクニカル・ジャパニーズ：英訳を視野に入れて日本語を作る. 情報処理学会研究報告, 2010-DD-78(5): 1–8.

Rascu, E. (2006). A controlled language approach to text optimization in technical documentation. KONVENS 2006, 107–114.

Shirai, S., Ikehara, S., Yokoo, A., and Ooyama, Y. (1998). Automatic rewriting method for internal expressions in Japanese to English MT and its effects. CLAW 1998, 62–75.