制限言語執筆支援システムのユーザビリティ評価

宮田 玲† Anthony Hartley‡ 影浦 峡† Cécile Paris#
†東京大学大学院 ‡立教大学 #オーストラリア連邦科学産業研究機構

1 はじめに

原文の言語表現をコントロールすることで機械翻訳（MT）を実用的に活用する手法として、制限言語（Controlled Language: CL）や前編集の研究がこれまで進められており、一定の成果をあげている [1, 2]。例えば、Miyata *et al*.[3] は、自治体文書向けの日本語 CL ルールを作成し、複数の MT システムを用いて、その効果を検証している。評価実験の結果から、特定の MT システムに適したルールを選択することで、実用品質の MT 訳の割合が約 10–15%向上することが示された。

CL の運用の観点からは、数十のルールで構成される CL に従って原文を執筆する・書き換えることは、とりわけテクニカルライティングの訓練を受けていない一般的な執筆者にとって容易ではない、という課題がある。我々は現在、日本語 CL ルールに従った原文執筆を機械的に支援するシステムを開発している。具体的には、文章執筆中に、CL ルールに違反した言語表現パターンを検出し、インタラクティブに適切な書き換えを促す機能・インタフェースを実装した [4]。次のステップとして、この支援システムが、ユーザーにとって実際に有用であるか検証することが求められる。

国際標準化機構（ISO）[5][1]は、ユーザビリティを「ある製品が、指定されたユーザーによって、指定された利用の状況下で、指定された目標を達成するために用いられる際の、有効さ、効率及びユーザーの満足度の度合い」と定義している。また「有効さ」「効率」「満足度」は、以下のように定義される。

有効さ（effectiveness）：「ユーザーが、指定された目標を達成する上での正確さと完全さ。」

効率（efficiency）：「ユーザーが、目標を達成する際に正確さと完全さに費やした資源。」

満足度（satisfaction）：「不快さのないこと、及び製品使用に対しての肯定的な態度。」

ISO の定義に準拠しながら、ユーザビリティを評価する研究は、MT 出力結果を対象としたもの [7]、対話型 MT システムを対象としたもの [8] などがあるが、CL 執筆支援システムを対象とした研究はこれまで報告されていない。

そこで本研究では、ISO の定義に基づき、文章書き換えタスクを中心とした統制実験をデザイン・実施し、ユーザーのタスク遂行プロセス及びシステムへの満足度を定量的に計測することで、システムのユーザビリティを多面的に評価する。

2 CL ガイドライン

上述のように、CL の効果は特定の MT に適したルールを選択することで最大化される。本研究では、ルールベース MT の翻訳クラウド[2]及び統計的 MT の TexTra[3]の使用を想定して、2 種類の CL ガイドラインを作成した（それぞれ **CL-R** と **CL-S** とする[4]）。我々はこれまで、自治体文書向けの日本語 CL を作成し、各ルールの性能を評価しており [3]、その結果に基づき、各 MT に対して効果が期待できるルールを選んだ。**CL-R** は 30 ルール、**CL-S** は 31 ルールを含む（表 1）。

3 CL 執筆支援システム

システムの目的は、規定の CL ルール及び用語リストに準拠した原文執筆・書き換えの支援である。ユーザーとして文章作成を専門としない一般的な執筆者を想定する。原文執筆・書き換えの各段階におけるユーザーの意思決定を支援するため、ルール違反箇所の検出、書き換え候補の提示、半自動での違反箇所の修正、の各機能を実装した。システムは Web ブラウザから利用でき、主な使い方は以下の通りである（図 1 も参照）。

1. **入力ボックス**にテキストを入力する。
2. システムが自動的に各文を分析し、**CL 違反箇所**を赤字で、**用語の表記揺れ**を青字で表示する。
3. ユーザーは、必要に応じて**診断メッセージ**と**ルールの詳細**を参照しながら、原文を修正する。
4. 一部の色付けされた違反箇所については、クリックすると書き換え候補が提示される。
5. 提示された候補の 1 つをクリックすることで、入力ボックス内の該当箇所が自動的に書き換わる。

CL 違反箇所の検出機能は、表層の文字列・品詞パターンに基づく検索ルールを作成することで実現した（実装済みルールは、表 1 の最右列にチェックが付いている）。形態素解析には MeCab[5]を用いた。

[1]以下の定義文は、日本工業規格による翻訳 [6, p.2] から引用。

[2]高電社, http://www.kodensha.jp/platform/cloud/

[3]NICT, https://mt-auto-minhon-mlt.ucri.jgn-x.jp

[4]R と S はそれぞれ、RBMT（ルールベース MT）と SMT（統計的 MT）の先頭文字を示す。

[5]MeCab, http://taku910.github.io/mecab/

No	ルール	CL-R	CL-S	システムへの実装
1	一文はできる限り 50 文字以内におさめてください。	✓	✓	✓
2	箇条書きで書くときは、列挙項目の前後の文を完結させてください。	✓	✓	
3	修飾語と被修飾語の関係を明確にしてください。	✓	✓	
4	「が」を使って文をつなげるのは、「しかし」の意味を持つ場合（逆接用法）だけにしてください。	✓	✓	✓
5	「ので」の意味で「ため」を使わないでください。	✓		✓
6	1 文に複数の否定表現を使わないでください。	✓		✓
7	可能や尊敬の意味で「〜れる」「〜られる」を使わないでください。	✓	✓	✓
8	複数の意味に解釈できる言葉ではなく、なるべく明確な意味を持つ言葉を使ってください。	✓	✓	
9	「〜という」表現はなるべく省いてください。	✓		✓
10	「思われる」「考えられる」は必要なとき以外は省いてください。	✓		✓
11	「たり」は単独で使わず、「〜たり〜たり」と並列で使ってください。		✓	✓
12	なるべく標準的な和英辞典に載っている語を使ってください。	✓	✓	
13	複合サ変名詞＋「する」の形を使わないでください。		✓	✓
14	誤字、脱字がないように注意してください。また、同音異義語や助詞の抜けにも注意してください。	✓	✓	
15	主語はなるべく省略しないでください。	✓	✓	✓
16	目的語はなるべく省略しないでください。	✓	✓	
17	並列要素を読点でつなげないでください。	✓	✓	✓
18	目的語につける助詞は、「〜が」ではなく「〜を」を使ってください。	✓	✓	✓
19	「〜てくる」「〜ていく」を使わないでください。		✓	✓
20	副詞句を主語と述語の途中になるべく挿入しないでください。		✓	
21	体言止めは使わないでください。	✓	✓	
22	サ変名詞＋「です」の形をなるべく使わないでください。	✓	✓	✓
23	「〜しか〜ない」の形を使わないでください。	✓	✓	✓
24	動詞＋「ように」の形を使わないでください。		✓	✓
25	「など」「等」をなるべく使わないでください。		✓	✓
26	授受動詞「〜あげる」「〜くれる」は使わないでください。	✓	✓	✓
27	冗長な表現をなるべく使わないでください。	✓	✓	✓
28	長い複合名詞をなるべく使わないでください。	✓	✓	✓
29	項目を列挙する際、項目の一部を省略しないでください。	✓	✓	✓
30	単位表現には「〜につき」「〜あたり」を使ってください。	✓	✓	✓
31	節の区切りに「〜て」をなるべく使わないでください。	✓	✓	✓
32	条件 if の用法で「〜すると」を使わないでください。	✓		✓
33	動詞はなるべく平仮名ではなく漢字で表記してください。	✓	✓	✓
34	文頭の記号は削除してください。	✓	✓	✓
35	機種依存文字を使わないでください。	✓	✓	✓
36	カギ括弧を使って単語を強調しないでください。	✓	✓	✓
用語	正しい表記を使ってください。	✓	✓	✓

表 1: CL ルールとシステムへの実装の有無

4 実験枠組み

ISO の定義に従い、システムのユーザビリティを次の 3 つの側面から評価する : (1) システムの利用により CL 違反箇所は減少するか（有効さ）; (2) システムの利用により書き換え作業時間が減少するか（効率）; (3) システムはユーザーに肯定的に受け入れられるか（満足度）。これら 3 つを定量的に計測するために、実験参加者をシステムの利用あり／なしの 2 つのグループに分けて、CL 違反箇所・用語の表記揺れを含む日本語原文を書き換える作業を遂行してもらい、(1) 違反箇所の修正数、(2) 修正に要した時間、(3) システムへの主観的満足度、を測定する。

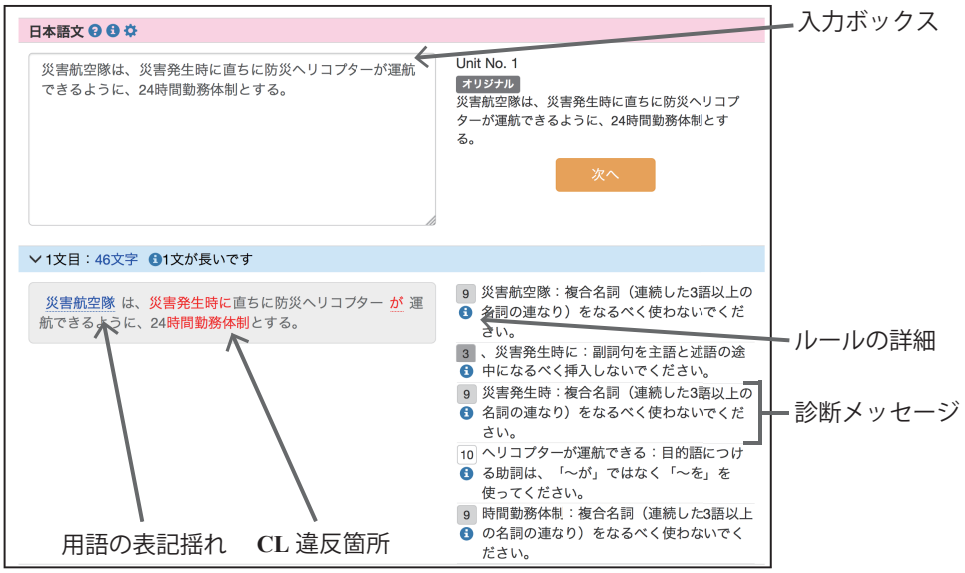

図 1: CL 執筆支援システムのインタフェース（実験作業環境）

4.1 実験デザイン

データ： 表 1 の 36 種類の CL ルールについて、少なくとも 1 つの違反箇所が含まれるように、自治体ウェブサイトから日本語文を 30 文抽出した。違反箇所の修正数を計測するために、データ中の違反箇所を人手でアノテーションした。さらに、筆者自らがデータ中の用語 2 つに手を加え、表記揺れを作成した。最終的なデータセット（30 文）は、ガイドライン **CL-R** の違反 67 箇所、**CL-S** の違反 76 箇所を含む。

条件： 対照実験を行うため、システムの利用あり（Treatment）／利用なし（Control）[6]の 2 条件を設けた。また、使用するガイドラインは、**CL-R** と **CL-S** の 2 種類である。すなわち、本実験には以下の 4 条件がある：**[TR] T**reatment 条件／ CL-**R** を使用；**[TS] T**reatment 条件／ CL-**S** を使用；**[CR]** Control 条件／ CL-**R** を使用；**[CS]** Control 条件／ CL-**S** を使用。

手順： 参加者は、図 1 の入力ボックス内に一文ずつ表示されるセンテンスを、内容は保持したまま、CL ガイドライン・用語リストに従って修正する。「次へ」ボタンを押すと、修正結果及び修正に要した時間が自動的に記録され、次のセンテンスが表示される。

質問紙： システムに対する満足度を測定するために、ユーザビリティ研究で標準的に使用される質問紙 System Usability Scale (SUS) [9] を用いる。SUS は表 2 に示す 10 の質問からなり[7]、各質問は「1: とてもそうは思わない」から「5: とてもそう思う」の 5 件法である。

[6]Control 条件の場合、紙媒体の CL ガイドライン一覧と用語リスト（自治体分野の用語 100 語）のみを使用する。
[7]オリジナルの英語版を、筆者らが日本語に翻訳した。

1	このシステムをしばしば使いたいと思う。
2	このシステムは、必要以上に複雑であると感じた。
3	このシステムは簡単に使えると思った。
4	このシステムを使えるようになるには、専門家のサポートが必要だと感じる。
5	このシステムにあるさまざまな機能やガイドは統一感があると感じた。
6	このシステムは一貫性に欠けるところが多いと思った。
7	たいていの人は、このシステムをすぐに使えるようになるだろう。
8	このシステムはとても使いづらいと感じた。
9	このシステムを安心して使えると感じた。
10	このシステムを使い始める前に、多くのことを学ぶ必要があった。

表 2: System Usability Scale (SUS) の質問項目

4.2 実験の実施

実験参加者として、大学生 12 名（日本語母語話者）を集め、ランダムに 3 名ずつ 4 条件（TR, TS, CR, CS）に割り振った。参加者はまず、CL ガイドラインと用語リストに一通り目を通した上で、練習課題として 5 文を書き換えた。書き換え実験の本番では、参加者は 30 文全て（但し、順番はランダム）を書き換えた。なお過剰な作業負荷がかからないように、30 文を 10 文ずつの 3 セットに分け、各セットの間に短い休憩を挟んだ。全ての書き換えが終了後、システムを利用した参加者（Treatment）を対象に、質問紙 SUS を実施した。また全員を対象に、事後インタビューを行い、作業やシステムの問題点・改善点などを尋ねた。

	Treatment			Control		
	TR	TS	平均	CR	CS	平均
違反修正数	55.0	62.7	58.8	49.7	55.7	52.7
違反見逃し数	11.0	13.3	12.2	16.3	20.3	18.3
修正率（%）	83.3	82.5	82.9	75.3	73.3	74.3

表 3: 実験結果：有効さ（CL 違反箇所の修正率）

	Treatment			Control		
	TR	TS	平均	CR	CS	平均
総時間	2405	2206	2306	3744	2844	3294
1 文毎	80.2	73.5	76.9	124.8	94.8	109.8
1 修正毎	43.7	35.2	39.5	75.3	51.1	63.2

表 4: 実験結果：効率（書き換え作業時間；単位は秒）

Q.	TR1	TR2	TR3	TS1	TS2	TS3	平均
1	4	3	5	3	5	4	4.0
2	1	2	1	2	2	2	1.7
3	5	3	5	2	4	4	3.8
4	3	4	1	4	4	3	3.2
5	2	4	5	2	5	4	3.7
6	4	2	1	2	1	2	2.0
7	4	4	5	2	4	4	3.8
8	1	2	1	3	2	2	1.8
9	3	3	5	3	5	4	3.8
10	4	3	1	4	4	2	3.0
SUS 値	70	68	100	54	80	78	75.0

表 5: 評価結果：満足度（質問紙 SUS）

5 結果と考察

5.1 有効さ（CL 違反箇所の修正率）

表 3 にシステムの「有効さ」の結果を示す。「修正率」は、30 文中の全ての違反箇所の内、CL ガイドライン・用語リストに従って正しく修正できた箇所の割合を示す。全体として、Treatment グループは、Control グループよりも修正率が約 9%高い（対応のない t 検定で有意：$t = -2.878, df = 10, p = .016$）。個別のルールの結果を分析すると、4 ルール（表 1 の No. 12, 14, 16, 29）の修正率は、Treatment グループの方が、Control グループより低い値だった。この内、No. 12, 14, 16 はシステムに実装されていない。この結果は、利用者はシステムに頼る傾向があり、システムが検出できない違反箇所を見落としがちであることを示唆している。

5.2 効率（書き換え作業時間）

表 4 にシステムの「効率」の結果を示す。「1 修正毎」は、違反箇所を 1 つ修正するのに要した時間（秒）を示しており、この結果から、Treatment グループは、Control グループより 30%以上高い効率で書き換え作業を遂行できたと分かる（対応のない t 検定で有意：$t = 2.826, df = 10, p = .018$）。

5.3 満足度（システムに対する主観的評価）

最後にシステムの満足度に関する質問紙 SUS の回答結果を表 5 に示す。最下行の「SUS 値」[8]は、100 点満点でシステムの満足度を示す。SUS 値の平均は 75.0 と、改善の余地はあるが、比較的高い値だと言えよう。

表 5 から特に、多くの参加者は、システムの「学習の難しさ」に関係する Q. 4 と Q. 10 に同意していることが分かる。事後インタビューでも、システムで提供されている書き換え支援機能（書き換え候補の表示など）の全てを使いこなすことは難しいという指摘があった。

6 おわりに

本研究では、CL 執筆支援システムのユーザビリティを、ISO で定義される「有効さ」「効率」「満足度」の 3 つの側面から評価した。実験参加者の原文修正作業の遂行プロセスを、システムの利用あり／なしの条件で比較した結果、(1) システムの利用により、違反箇所の修正率が約 9%向上し、(2) 修正効率が 30%以上改善され、(3) システムの満足度は概ね高いことが示され、本システムの有用性が実証された。また、システムの「学習の難しさ」に関する課題も明らかになったため、今後、機能・インタフェースを改善していきたい。

謝辞 科研費（16J11185）、KDDI 財団の調査研究助成「自治体文書の多言語化支援システムの開発」の支援を受けた。

参 考 文 献

[1] Aikawa, T. *et al.* Impact of Controlled Language on Translation Quality and Post-Editing in a Statistical Machine Translation Environment. *MT Summit XI*, 1–7, 2007.

[2] Kuhn, T. A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1): 121–170, 2014.

[3] Miyata, R. *et al.* Japanese Controlled Language Rules to Improve Machine Translatability of Municipal Documents. *MT Summit XV*, 90–103, 2015.

[4] Miyata, R. *et al.* Evaluating and Implementing a Controlled Language Checker. *CLAW 2016*, 30–35, 2016.

[5] ISO. Human-Centred Design Processes for Interactive Systems. 1999.

[6] JIS. 人間工学—インタラクティブシステムの人間中心設計プロセス. 2000.

[7] Doherty, S & O'Brien, S. Assessing the Usability of Raw Machine Translated Output: A User-Centered Study Using Eye Tracking. *International Journal of Human Computer Interaction*, 30(1): 40–51, 2013.

[8] Alabau, V. *et al.* User Evaluation of Interactive Machine Translation Systems. *EAMT 2012*, 20–23, 2012.

[9] Brooke, J. SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*. Taylor & Francis, 189–194, 1996.

[8]否定的な尋ね方をしている偶数番号の質問の評価値を 1–5 から 5–1 に反転させた上で、10 項目全ての評価値を足し合わせて、合計値を 2 倍した値である。