論文の内容の要旨

論文題目

Controlled Authoring for Document Multilingualisation Using Machine Translation

（機械翻訳を活用した多言語文書展開のための制限オーサリング）

氏　　名　　宮田　玲

1　自治体文書の多言語展開における課題と解決策

本論文の目的は、日本の自治体ウェブサイトの文書に焦点を当て、機械翻訳（Machine Translation: MT）を活用した多言語文書展開の枠組みの構築とオーサリング環境の開発を行うことである。

自治体は、外国人住民にも地域や生活に関する情報を正しく円滑に提供することが求められる。第1章では、多言語情報発信の手段としてMTの導入が進んでいる自治体ウェブサイトにおける課題を指摘した。具体的には、(1) 文書が十分に構造化されておらず必要な情報を的確に入手することが難しいという文書レベルの問題、(2) 外国語版のページでは品質の低いMT訳文が使われ、読み手に文意が正確に伝わらないという文レベルの問題を特定した。技術的には、従来難しいとされてきた日英・日中などのMT性能が向上してきており、自然言語の語彙・文法・スタイルに一定の制限を加えた制限言語（Controlled Language: CL）や分野に特化した対訳用語集と適切に組み合わせることで、MTを有効に活用できる可能性が高まっている。そこで課題の解決策として、制限オーサリングの枠組みから、原文書作成の段階で、効果的な情報の提示を可能とする文書構造の設計、読みやすく・機械翻訳しやすい言語表現の設計を行うことが有効であると提起した。さらに多くの自治体ではテクニカルライターや翻訳者を十分に確保できないという実状に鑑み、一般的な自治体職員の利用を想定した、制限オーサリング支援システムの設計・開発まで行うことにした。

第2章では、制限オーサリングと多言語化について、文書構造化、MT、CL、用語マネジメント、オーサリングシステムに関する先行研究を整理した。特にこれまで研究・開発されてきたCLは、文書を構成するテキストの機能的位置づけや伝達目標に応じて精緻化されていないということが明らかになった。言語構造に加えて文書内の言語形式の慣例にも違いがある日英などの言語間のさらなる翻訳改善のためには、文書構造とCLの対応づけによる、MT活用方法の高度化は重要な課題である。

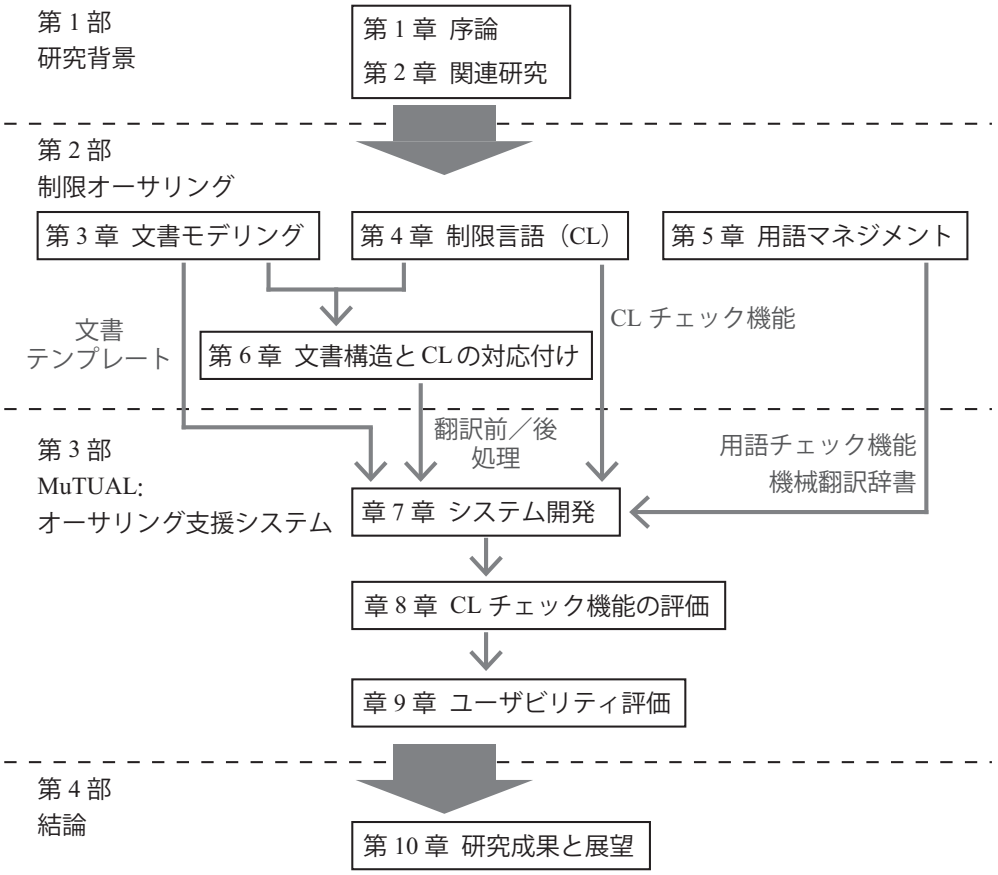

図 1　本論文の構成

以上を踏まえ、第 3〜6 章で自治体文書の日英翻訳を想定した制限オーサリングの枠組みの構築と検証を、第 7〜9 章でオーサリング支援システムの開発と評価を行った（図 1）。

2　制限オーサリングの枠組みの構築と検証

第 3 章では、文書レベルの研究として、自治体の手続き型文書（例えば、住民登録の仕方や転出届の出し方を説明する文書などを指す）の構造を定式化した。ジャンル分析の手法を参考にしながら、自治体ウェブサイトから収集した 123 の手続き型文書を対象に、文書の機能的な要素を網羅的に抽出・類型化した。さらにそれらの要素を用いて、技術文書用 XML 標準規格 Darwin Information Typing Architecture (DITA) であらかじめ定義される「タスク型」の文書構造を詳細化することで、自治体手続き型文書の構造を定式化した（図 2）。これは、正確かつ漏れのない自治体手続き型文書を作成するための指針となるだけでなく、第 6 章で文書構造に応じた CL ルールを定義するための基盤となる。

初期DITA	詳細化DITA（一部抜粋）
事前条件 （prereq）	個人条件 イベント条件 アイテム条件
背景情報 （context）	説明（概要、目的、効力、 罰則、関連概念）
手順 （steps）	1. 必要なものを持参する 2. 申請場所に行く 3. 様式を提出する 4. 手数料を払う
期待結果 （results）	得られる結果（所要期間、 交付物、連絡）
タスク完了後 の操作 （postreq）	関連手続きへの誘導

図 2　自治体手続き型文書の構造

第 4 章では、文（言語）レベルの研究として、日本語 CL ルールの作成と評価を行った。まず以下の 2 つの手続きで、自治体文書を対象とした CL ルールを 60 種類作成した。

1. テクニカルライティングや作文技術に関する既存のルール・ガイドラインを収集・整理する
2. MT の出力結果を分析し、翻訳精度に関わると考えられる言語表現を類型化する

続いて、自治体ウェブサイトから構築したテキストデータを用いて、作成したルールの効果を「日本語文の読みやすさ」と「翻訳英文の理解度・正確性」の観点から評価した。4 つの異なる MT システムの評価結果を検証することで、少なくとも 3 つのシステムに有効な汎用ルールを 18 種類同定した。さらに個別の MT システムに最適なルールを選択することで、実用品質の訳文が最大 14% 増加することが明らかになった。また 42/60 ルールにおいて、原文品質の向上ないし維持が確認できた。

第 5 章では、文（用語）レベルの研究として、自治体分野に特化した日英対訳制限用語集の構築と評価を行った。まずは自治体生活情報の日英対訳コーパス（15391 文対）から、人手により 3741 の対訳用語対を収集した。さらに、用語のバリエーションを類型化した上で、頻度情報などを参照しながらバリエーションを統制し、約 2800 の制限用語を定義した。構築した用語集の十分性を、統計的な手法により推定した結果、自治体分野の用語が、日本語では約 55–65%、英語では約 45–60% カバーされていることが示された。自治体分野の用語集はこれまでほとんど整備されていないことを踏まえると、実用の観点から良好な結果だといえる。

第 6 章では、文書構造と対応づけた CL の作成を進めた。第 3 章で定めた DITA の各文書要素に適した言語表現パターンを、起点言語（日本語）と目標言語（英語）の両方において定義することで、文書要素に特化した CL ルールを 4 種類作成した。例えば、DITA の「手順」要素においては、日本語では「A を持参します」のように文末表現に「〜します」を用い、英語では「Bring A」のように命令形を用いると定めた。しかし、ここで原文「A を持参します」をそのまま MT を用いて訳すと、「To bring A」のように必ずしも目標言語として適した結果にならないことがある。そこで、例えば「A を持参します」を「A を持参しろ」とあらかじめ命令形に変換する、といった翻訳前処理工程の導入が有効であると考えた。4 種類の CL ルールの内、MT を用いたときに起点言語と目標言語でずれが生じる 2 ルールを同定し、翻訳前処理用の変換ルールを定義した。翻訳前処理の適用前後の MT 訳文を分析した結果、(1) 翻訳前処理が確かに有効であること、(2) MT システムによってはさらに翻訳後処理が必要であることが明らかになった。

3　制限オーサリング支援システムの開発と評価

以上第 3 章から第 6 章で提案した枠組みを踏まえ、第 7 章では、制限オーサリング支援システム MuTUAL の全体設計を行い、プロトタイプを開発した（図 3 参照）。穴埋め式に文章を構成する**文書テンプレート**は、第 3 章で定式化した文書構造に準拠している。**CL 執筆アシスタント**は、執筆者が統制された文章を効率的かつ正確に作成できるよう、CL ルール（第 4 章で作成）と制限用語集（第 5 章で構築）に違反した言語表現をリアルタイムに検出し、書き換え候補を提示する。**MT システム**は、ユーザー辞書に制限用語集が登録されており、自治体用語を一貫して正しく翻訳できる。第 6 章で提案した**翻訳前処理／後処理**の機能もシステム内部に組み込まれている。MuTUAL の最大の特徴は、文書要素に応じて最適化された、CL 執筆アシスタント・翻訳前処理／後処理が起動することである。これにより、文書要素に応じた MT による訳し分けが実現できる。

図 3　制限オーサリングシステムのインターフェイス

システムの評価として、中核モジュールである CL 執筆アシスタントの性能評価とユーザビリティ評価を行った。まず第 8 章では、CL チェック機能の性能を、テストセットを用いて、精度と再現率の観点から評価した。実装した 30 のルールの内、20 のルールは精度・再現率ともに 0.7 以上（さらにその内 13 のルールは最大値 1）と実用での検証に向けて十分な値を達成した。第 9 章では、システムのユーザビリティを、ISO の定義に従い、効果、効率、満足度の 3 つの観点から評価した。実験協力者 12 人（日本語を母語とする大学生）を、システムのサポートあり／なしの 6 人ずつのグループに分け、CL 及び制限用語集に違反した箇所を含む文章を書き換える実験を行った。システムに対する満足度は、質問紙・インタビューにより定量的・定性的に評価した。また書き換え前後の原文とその MT 訳文の品質を、人手により評価した。評価の結果、システムを用いることで、(1) 違反箇所の書き換え成功率が約 9% 向上したこと（効果）、(2) 書き換えに要する時間が 30% 以上減少したこと（効率）、さらに (3) システムの機能・インターフェイスは概ね好意的に受け容れられたこと（満足度）、が明らかになった。またユーザーは全ての機能を有効に活用できたわけではなかったこともあり、原文・翻訳文品質の向上へのシステムの寄与は限定的であったが、CL ルールと制限用語集に従うことで、翻訳文品質が向上することが示された。

4　制限オーサリングの枠組みと支援システムの有効性

以上をまとめると、本論文で提案した制限オーサリングの枠組みは、原文品質を下げることなく、機械翻訳しやすい構造化文書を作成する上で有効であることが示された。また本研究で作成した CL ルールや制限用語集は、専門的な文章技術を有しない人にも十分使いこなせるものであることが示唆された。ユーザビリティ評価の結果、執筆効率の大幅な向上が見られたことからも、本論文で提示した制限オーサリングの枠組みと支援システムは自治体における運用・実証実験が可能な段階にあると考えられる。