技術文書の多言語化を見据えた制限オーサリングと翻訳：基本方針と枠組み

宮田 玲† 柳 英夫‡ 影浦 峡♮ 萩原 秀章♯

†名古屋大学 ‡株式会社システートソリューションズ
♮東京大学 ♯トヨタ自動車株式会社

1 はじめに

産業翻訳において言語処理技術の導入が進んでいる。これまでも特にITや特許の分野において、翻訳メモリ（TM）などの翻訳支援ツールや機械翻訳（MT）とその後処理としてのポストエディット（PE）が利用されてきた。近年はニューラル機械翻訳（NMT）の急速な研究開発を背景に、MT (+PE) を活用しようという動きがますます高まっている。また産業翻訳においては、コスト削減のためだけでなく、これまで対応していなかった新しい言語方向への拡張のためにも、言語処理技術の活用が期待されている。しかし、どのような場面でどのような言語処理技術をどのように活用するか（あるいは、活用すべきでないか）についての知見は必ずしも十分整理されているわけではない[1]。汎用ツールとしてパッケージ化されたMTやTMをそのままの形で使うだけになってしまうケースも少なくない。産業翻訳の現場の細かいニーズに応じて、MTをはじめとした言語処理技術を効果的に活用するための枠組みと使いやすくカスタマイズ可能なツールが求められている。

このような背景の中、我々は現在、自動車関連の技術文書（修理書、解説書、オーナーズ・マニュアル[2]）を対象に、執筆と多言語化を促進するための枠組みの構築とそれに関連する各種言語処理ツールの開発に取り組んでいる。対象とする文書は、(1) 誤訳や訳し漏れは致命的な事故につながりうるため情報の十分性と正確性が強く求められる、(2) 製品のサイクルに応じて高い頻度で内容の類似したバージョンが作成される、(3) ほぼ同時にほぼ同じ内容が大量に多言語で展開される、という性格を持つ。また現時点では、執筆・翻訳は多くの制作会社・翻訳会社が関わりながら人手でなされており、時間・コストがかかるだけでなく、表現の一貫性を確保するのが難しいという課題がある。情報の十分性と正確性を担保できるのであれば、言語処理技術の活用による効率化が特に有効な文書ドメインであるといえる。

対象文書の表現に関する特徴としては、類似表現の反復が多いこと[3]と、各文は比較的小規模な構文セット・語彙と多様な用語（部品名等）から構成されることが観察された。このことは、あらかじめ用意した定型文とその翻訳版を適切に利用することで、原文執筆と翻訳作成の大部分をカバーできることを示唆している。したがって、我々はいわゆる「原文をその都度最初から翻訳する方法」を、人間翻訳・MTに関わらず、第一義的なものと考えない。特に現在のNMTは、訳し漏れ、低頻度語の誤訳、出力の不安定さが欠点として挙げられており、対象文書によってはデメリットがメリットを上回る。まずは、執筆工程において文書構造・言語表現・専門用語を適切に管理・統制した「制限オーサリング」とそれをベースにした半自動的な多言語文書作成に取り組む。その上で、カバーしきれない部分について、MT (+PE) や人間翻訳の役割を検討する。

本発表で報告する内容は、一つのケーススタディではあるが、技術文書の執筆・翻訳に広く当てはまるものであり、ある程度一般化可能な知見を提供するものである。以下では、我々のプロジェクトの基本的な方針を説明した上で、制限オーサリングと翻訳の枠組みを文書モデル、制限言語、用語管理の観点から概観し、執筆・翻訳の全体フロー、有用な言語処理技術のあり方、今後の研究開発方針を示す。

[1]複数の産業翻訳関係者から、手探りでMTを導入・運用している実態について聞いている。

[2]修理書は技術者向けに自動車の修理作業の基本概念や手順を示したもの、解説書は主にディーラー向けに自動車の機能や特徴を説明したもの、オーナーズ・マニュアルは利用者向けに自動車の機能や使い方を説明したものである。

[3]新規文書は既存の文書をなるべく踏襲して作成されるためである。なお大きく、同一車種の文書内で共通して使われる表現（たとえば、ある車種の部品の「取り付け」と「取り外し」に関する文書は似ている）、異なる車種（同一車種の異なるモデルも含む）の文書間で共通して使われる表現（たとえば、異なる車種間でも部品の「取り付け」に関する文書は似ている）に分けて考えることができる。

2　基本方針

プロジェクトの基本方針は大きく以下の 4 点である。

1. 執筆工程と翻訳工程を合わせて最適化すること
2. 文書の単位を考慮して言語表現を設計すること
3. 誰もが参照可能な形でルールを明文化すること
4. ルールを活用するためのツールを提供すること

1 は作業プロセスのあり方に関わるもので、我々は翻訳を視野にいれた執筆を試みる。対象文書の予備調査により、(i) 同一文書内に重複した内容の記述があり翻訳量が増大していること、(ii) 類似した伝達内容に対して異なる構文・語彙・表記が使われており TM マッチ率が減少していることが確認された。特に (ii) の問題は、後工程における MT などの言語処理技術に対しても問題となりうる。既にローカリゼーション等の産業翻訳においては、文書の多言語展開を想定して、ロケールの変数化等の処理を施した国際化版を作る試みがなされている [1]。また MT を視野に入れた原文編集手法であるプリエディットや読みやすさ・翻訳しやすさを担保するための制限言語の研究が進められている [2, 3]。我々はこれらの知見を参照しつつ、なるべく上流工程から文書を統制し[4]、それに応じて翻訳工程を最適化することで、原文品質・翻訳品質の向上とプロセス全体の効率改善を目指す。

2 は情報の配置と言語表現の位置付けに関する方針である。執筆・翻訳は文書を扱うのであって、文を扱うのではない[5]。たとえば、見出し内では体言止めを使う一方で、本文の箇条書き内では常体を使うなど、文書内の位置・役割に応じて、求められる表現形式が変わりうる。この事実は誰もが知っていながらも、文書の要素としてどのような体系を準備し、言語表現と対応付けるかは、具体的に明らかになっていない。我々は、まず伝達される情報のタイプごとに文書の構造と内容を調査し、基本的なモデルを設計した上で、言語表現との対応関係を規定する。

3 は作業を促進するためのメタレベルの方針であり、我々は文書構造から言語表現、用語までを広くカバーするルールの構築を目指す。これまで執筆・翻訳は、個々人の経験やスキルに依存しがちであり、成果物に対して客観的基準に照らして評価するということも十分になされてこなかった。特に複数の制作会社・翻訳会社に執筆・翻訳を依頼する場合、成果物の品質を揃えるためにも、明文化されたルールが不可欠である。また、たとえば作業者とチェッカーの間で意見が割れた場合でも「○○というルールに準拠すると、××の方が適切である」といった形の議論が可能になり、成果物の一貫性・品質が担保されるだけでなく、作業プロセスの円滑化が期待できる。なお本プロジェクトでは、既存の執筆・翻訳生産物に基づきながらも、それに留まるわけではなく、規範的な枠組みの策定を行う。従来の執筆・翻訳方法との齟齬が生まれる可能性もあるため、これまでテクニカルライティングや翻訳論の分野でも取り組まれてきた執筆・翻訳に関する技術を参考にしつつ、できるだけ「なぜそうすべきか」の説明を含めてルール化する。

4 は具体的な言語処理技術の設計と実装に関わるものである。3 で構築するルールは、文書構造、言語表現、用語の各領域にわたり、使いこなせるようになるまでには、一定の練習が必要であるし、熟練のテクニカルライターや翻訳者であっても漏れなく正確にルールに準拠することは容易ではない。我々は、本プロジェクトの執筆・翻訳プロセスに最適化した言語処理ツールの開発を行う一方で、他の産業翻訳場面にも移転可能な形で基盤技術を整備することを目指している。

このような基本方針のもと、まずは概念実証フェーズとして、日本語・英語の修理書を対象に、文書モデルの作成、制限言語の構築、用語の管理と利用、の 3 つの枠組みから研究を進めている。3〜5 節ではそれぞれの枠組みについて説明する。

3　文書モデルの作成

文書モデルとは、暫定的に「特定の伝達目標を達成するために何をどのような配置で書くべきかを抽象化し形式的に示したもの」とする。我々は規範性を担保するために、技術文書の標準規格として実績のある DITA (Darwin Information Typing Architecture) で定義される情報型[6]をベースとし、自動車関連の技術文書の特徴に応じてカスタマイズする方針をとる。DITA はあくまで汎用的な文書規格であり、「何を書くか」（内容要素とする）を詳細に指定したものではない。そこで、我々は DITA に基づく文書構造に内容要素を割り当てる。具体的には、既存の文書を対象として、節を最小単位としたテキストのスパンに、何について書かれているかの指示的な要約を与えながら、ボトムアップに内容要素の類型を作成し、それを DITA 構造と対応させる。

[4] 5 節で触れる用語の管理については、執筆工程よりもさらに上流にある製品の設計工程にも関わるものである。

[5] 正確には、文書の単位を前提として、また明示的に考慮して、個別の文（言語表現）を扱う。

[6] Concept, General Task, Machinery Task, Reference などの型が定義されている。

DITA に基づく手続き型の文書構造				対応する内容要素
attention				禁止事項, 徹底事項, 推奨事項, 注意事項の確認
info				モノ, 補足
steps	step	cmd		動作, 使用用品
	conditionalstep	case	condition	実施条件
			cmd	動作, 使用用品
		attention		禁止事項, 徹底事項, 推奨事項
		tutorialinfo		方法
		info		モノ, 補足
postreq				作業後の必要作業

表 1: 文書構造と内容要素の対応

これまで、自動車の修理書の中でも比較的構造が明確な手続き型の文書を対象に、DITA の Machinery Task 型に基づいた文書構造とそれに対応した内容要素を定義した（表 1）[7]。

このような文書モデルは、新規文書の執筆や既存文書の診断のガイドラインに使えるだけでなく、次節に示す制限言語の各モジュールと組み合わせることで、文書を考慮した言語処理を可能とする（詳細は 6 節）。

4 制限言語の構築

上記の文書モデルは、基本的には言語に依存しない。言語表現レベルで具体的にどのように執筆・翻訳するかについては、制限言語の枠組みを用いる。制限言語と一言に言っても、言語表現のクラス（語彙、構文、表記）、制限の強度（使うべき表現を指定するか、使ってはいけない表現を指定するか）、形式化の度合い（メタ言語的に指示するか、表現そのものを指定するか）に応じて、様々な形がありうる。それらを踏まえつつ、以下の 3 つのモジュールを開発する。

制限語彙　一般語[8]を対象に、使用可能な語を指定したリストのことである。予備調査では、手順の箇条書きで使われる主に動作を表す文末動詞は、述べ 820259 語、異なり 1012 語であった。また「移動する」「移動させる」「移動を行う」といった表層的な表記の揺れを統合したところ、604 語となった。さらに、「メモする」「記録する」といった文書中でほぼ同じ意味・用法を持つ語を統合すると、動作を表す基本的な動詞は、500 語程度に収まると予想している。これらの語を分類し内容要素と対応させた上で、定義や用法を明文化し、制限語彙としてまとめる。

スタイルガイド　構文、表記レベルで望ましい（あるいは望ましくない）表現パターンを明文化したルール集のことである。『日本語スタイルガイド』[6] や ASD-STE100 [7] 等のガイドラインを参考にしながら、複雑な構文や曖昧な表現を規制するルールを日英それぞれ数十種類ずつ作成予定である。読みやすさと（機械）翻訳しやすさを両立させること、文書構造を考慮してルールを定義することが重要である。

表現テンプレート集　主に部品名や数値を変数化した定型表現のテンプレートを一定のカテゴリーごとにまとめたものである。たとえば「コネクター X をエアバッグ Y から切り離す」のような定型的な表現は、修理書において頻出するため、事前に部品名や数字等の可変要素を変数化して「[A] を [B] から切り離す」（[A], [B] には部品名が入る）のようにテンプレート化しておく[9]。表現テンプレートは、既存のテキストから可変要素を自動で判別したものを人手で修正した上で、制限語彙およびスタイルガイドに準拠するように書き換えることで作成する。さらに「[A] を [B] から切り離す」と「Disconnect [A] from [B]」など言語間で表現テンプレートの対応をとっておく。

5 用語の管理と利用

自動車関連の技術文書には、部品名などの用語が大量に含まれる。制限オーサリングの原則として、一つの対象（部品等）には各言語で一つの名前を一貫して使用することが重要である。これまでも自動車部品の名称は品名コードとともに日英対訳で管理されているが、部品以外にも、工具、テスト手法、ソフトウェア機能など自動車分野の専門用語は多い。これらを一定のカテゴリーに分類した上で、言語間で対応させて管理しておくことが重要である。

また 4 節で述べた表現テンプレートの変数部分に、よく使われる用語や使われ得る用語カテゴリーを対応させることで、用語の柔軟な検索・利用が可能となる。

[7]別途 [4] にて、文書モデルの詳細と文書診断への応用について発表予定である。表 1 は [4] からの引用である。

[8]現時点では、動詞、名詞、形容詞、副詞を対象としている。

[9][B] の部分はさらに「[B1] および [B2]」のように展開できるようにしておくことで、「コネクター X をエアバッグ Y およびコンピュータ Z から切り離す」等のバリエーションをカバーできる。

たとえば「[A] を使用して、[B] を研磨する」というテンプレートでは [A] に「工具」カテゴリーを、[B] に「部品」カテゴリーを対応させる[10]。

6 全体フローと言語処理技術

以上の枠組みを用いた執筆・翻訳の全体フローとして、まず執筆者は、伝達情報のタイプに応じて適切な文書モデルを選択し、それに準拠しながら文書構造と内容要素に関する基本設計を行う。各内容要素の具体的なテキスト化にあたっては、なるべく表現テンプレートを用いる（起点言語は日本語とする）。これにより、制限語彙とスタイルガイドへの準拠は確実となる上、対応する英語のテンプレートが準備してあれば、ほぼ自動的に英語への変換が可能となる。表現テンプレートを使用せずに執筆するテキストについても、制限語彙とスタイルガイドに準拠することで、一貫性が高まり、TM マッチ率や翻訳効率の向上が期待できる。

関連する言語技術は大きく、表現の作成と制御に関する技術と言語変換・翻訳に関する技術に分けられる。

前者については、まず表現テンプレートの柔軟な検索・候補提示・入力支援が重要である。表現テンプレートは、文書（特に内容要素）と対応付けられているので、検索対象のテンプレートの候補を限定することができる。さらに、表現テンプレートの変数部分に入る用語は、用語の分類体系から順次選択できるようにする他、異表記や類義語を含めて曖昧検索できるようにすることで執筆者の入力を支援する。

また起点言語・目標言語両方において、制限語彙、スタイルガイド、用語集に沿ってテキスト中の表現をチェックし、適切な言い換え候補を提示するツールの開発が有効である。特にスタイルガイドは文書構造・内容要素と対応付けられており、たとえば「禁止事項」を執筆中は文末が「しないでください」に統一されているかチェックする、といった詳細な設定が可能である。

後者については、記入済の表現テンプレートを多言語に変換するツールが有用である。基本的には、テンプレート自体と変数部分を多言語間で対応させておけばよいが、たとえば英語における動詞の単数形・複数形の処理など、言語ごとの文法を踏まえた調整機構が必要である。また表現テンプレートの言語間対応が事前に取れていない場合もあるため、従来型の TM 機能も併用する[11]。

[10] さらに絞り込んで、[A] に入りうる具体的な工具名の集合を決めておくことができれば、より効率的な用語検索が可能となる。

[11] なお言語間対応の取れた表現テンプレートは、変数部分付き TM とみなすことができる。

制限語彙や用語集を適切に用いることで、MT の導入も現実的になるが、特に NMT を用いる場合は、登録した語句の確実な翻訳を行うための工夫が課題となる。制限オーサリングにより表現の統制された対訳テキストは、TM や MT のデータとしても有効活用できるだろう。将来、データが十分に集まれば、対象文書に特化した翻訳モデルの作成も可能である。

7 おわりに

本稿では、自動車関連の技術文書を対象に、多言語化を見据えた制限オーサリングと翻訳の基本方針と枠組みを示した。文書モデル、制限言語、用語管理の観点から、最適な執筆・翻訳フローを検討し、必要となる言語処理技術について整理した。

現在は概念実証フェーズとして、日英の自動車修理書の一部を対象として、枠組みの具体化を行っている。今後は、評価実験により枠組みの有効性を検証することが必要である。執筆・翻訳の成果物については、文書・表現の特徴（内容の重複度合い、表現揺れの数、既存テキストの再利用率）や読み手の理解度・読解速度の観点から評価し、作業プロセスについては、所要時間や作業負荷の観点から評価する予定である。

日英以外の言語への拡張に際しては、英語をピボット言語とした MT と PE の活用も検討している。またイラストなどの非言語要素の効果的な利用についても取り組む予定である。

謝辞 研究データはトヨタ自動車株式会社からご提供いただいた。本研究の一部は科研費（17H06733）の支援を受けた。

参 考 文 献

[1] Pym, A. Exploring Translation Theories (2nd ed.). New York: Routledge, 2014.

[2] 宮田玲, 藤田篤. 機械翻訳向けプリエディットの有効性と多様性の調査. 通訳翻訳研究への招待, No. 18, pp .53–72, 2017.

[3] Kuhn, T. A Survey and Classification of Controlled Natural Languages. Computational Linguistics, Vol. 40, No. 1, pp. 121–170, 2014.

[4] 杉野峰大, 宮田玲, 松崎拓也, 佐藤理史. 文書モデルの作成と文書診断に向けた予備的考察：自動車の修理書を対象に. 言語処理学会第 25 回年次大会, 2019 (発表予定).

[5] OASIS. Darwin Information Typing Architecture (DITA) Version 1.3. http://docs.oasis-open.org/dita/dita/v1.3/dita-v1.3-part3-all-inclusive.html

[6] 一般財団法人テクニカルコミュニケーター協会. 日本語スタイルガイド第 3 版. テクニカルコミュニケーター協会出版事業部, 2016.

[7] ASD. Simplified Technical English, ASD-STE100, Issue 7. http://www.asd-ste100.org