

# 企業の多言語情報発信を支援する取り組み 国際化をにらんだ産業文書の効率的作成へ向けて

† 井佐原均   立見みどり   ‡ 影浦峯   † Tony Hartley  
† 豊橋技術科学大学   ‡ 東京大学

## 1 はじめに

我が国の中核産業である自動車産業等は、国内拠点のみでなく、海外拠点においても、研究開発・生産・営業などの企業活動を積極的に進めている。これら産業の国際競争力の強化に向けた喫緊の課題の一つに、生産や営業に関わるさまざまなノウハウを的確に文書化し、さらには効率よく多言語化することがある。我々は自動車関連企業の協力を得て、情報通信技術を活用し、実務に必要な情報の多言語での発信を支援する環境の構築を目指している。これにより、海外での販売力の強化や、海外生産拠点の生産効率の向上が期待される。

本研究では制御言語を適切に拡張することにより制御言語と技術文書管理の狭間を埋める規格化日本語／英語の開発を行い、その規格に基づいて事後編集を含む翻訳フローにおいて機械翻訳システムを最適化することにより、文書執筆と多言語化の効率にブレークスルーをもたらすことを目指している。

本研究開発は、総務省の戦略的情報通信研究開発推進制度(SCOPE)の支援の下、地域 ICT 振興型研究開発「地域産業の国際競争力強化のための多言語情報発信支援の研究開発」として実施されている。

## 2 産業文書の多言語化に向けた課題

文書を効率よく多言語化するための取り組みの柱となるのは、機械翻訳システムの活用である。機械翻訳システムの性能や有用性については利用者ごとに議論の分かれるところであろうが、現時点で利用可能な機械翻訳システムの性能を最大限に生かして効率よく翻訳を行うためには、機械翻訳システム

の効果的な利用を支援するための環境を整備することが重要となる。

本研究では、機械翻訳システムを使って効率よく多言語化できるような産業文書の作成を支援するため、用語レベル、文章レベル、文書レベルの 3 方向から取り組む。以下、各レベルにおける課題を個別に解説する。

### 2.1 用語レベルの課題

機械翻訳の精度を向上させる手段としてユーザー辞書の整備は有効である。一般に、ユーザー辞書には各社固有の製品用語や固有名詞に加え、業界標準の技術用語や専門用語なども含めることが多い。業界標準語、事業分野別、製品別などの用語集を個別に作成し、翻訳する文書の分野や目的に応じて適用することで、コンテキストに合った用語を機械翻訳システムが選択できるようになる。

本研究ではさらに一歩進んで、あらゆる産業文書の作成に活用できるような共通語彙集の開発も視野に入れている。具体的には、技術用語で使用される語(一般動詞、形容詞、名詞などを含む)の意味を一意に定義し、英単語と一対一対応させた対訳用語集である。このような用語集を作成し、産業界全体で標準化することで、各語の意味を誰もが一意に認識し、あらゆる産業文書の曖昧性を排除することが可能となる。

### 2.2 文章レベルの課題

文書を国際化するには、さまざまな言語の読者を想定する必要がある。原文(本研究の場合は日本語)は、日本語を母語とする読者に加え、日本語を読解できるが日本語以外を母語とする読者、さらに

は専門知識を持たない外部の翻訳者にも読まれることになる。また、機械翻訳システムにとって処理しやすい文章である必要もある。英語に翻訳された文書は、英語を母語とする読者に加え、それ以上の数の、英語を母語としない読者が読む可能性がある。英語から多言語に翻訳すればさらに読者層は広がる。それらすべての読者に情報を正確に伝えるためには、原語の段階で、誤解を招くことのない、理解しやすい文章を執筆しなければならない。

このような文章を書くため、英語においては平易英語(Simplified English, Global English)や制御言語に関する研究が継続的に行われており(Adriaens 1994, Bernth & Gdaniec 2001, Kohl 2008)、機械翻訳システムによる訳文の品質向上にもある程度の効果があることが報告されている(O'Brien & Roturier 2007, Aikawa et al. 2007)。しかし日本ではまだこの分野での研究は始まったばかりで、2010年に「産業日本語研究会<sup>1</sup>」が設立され、特許文書をはじめとする産業文書に関する研究が進められている(渡邊 2010)。本研究では、より具体的に、産業文書の日英翻訳において効果の高い制御言語規則の発見とガイドラインの整備を目指す。

本研究では、英語において確立されている制御言語規則や日本語技術ライティングのガイドラインを元を選び出した制御言語規則に基づいて日本語を書き換え、複数の機械翻訳システムで翻訳してその品質を評価する。また、書き換え後の日本語も、日本語を母語とする読者にとって不自然な文になってはならない。そのため、書き換え前後の日本語の品質を評価するためのテストも行う。これらについては現在実験を行っている段階であるため詳細は省くが、実際の対象読者から生の評価を収集するため、英文和文それぞれのテキストの評価者として、協力企業の日本本社および海外拠点の社員に実験に参加してもらった。

## 2.3 文書レベルの課題

産業文書の作成および機械翻訳システムでの翻訳に適した文書構造については十分な議論や研究

がなく、標準化され浸透している仕様はいまだ存在しない。しかし特にユーザーマニュアルなどで広まりつつある DITA (Darwin Information Type Architecture)などに見られるように、xml ベースの形式が今後も主流になると思われる。本研究では、産業文書内のテキストの機能要素を、文書、段落、文、句のレベルで同定し、文書構造にマッピングすることを目指す。これにより、各要素に異なるタグを付けて役割を示すことで、たとえば、「文書を印刷する」という日本語の表現を、タイトルでは“Printing a Document”に、手順の記述では“Print the document”に機械翻訳システム側で訳し分けることも可能になる。機能要素を同定して構造化することにより、機械翻訳システムのパフォーマンス向上だけでなく、人手による文書執筆の効率の向上にも寄与することを目指す。

## 3 現状と当面の取り組み

多言語化に向けたこれらの課題に実践環境で取り組むべく、まずは協力企業における文書作成の現状を調査した。協力企業では、日本国内で蓄積されてきた業務のノウハウを海外拠点でも活用すべく、現在各部署において業務マニュアルの執筆を進めており、近い将来それらの翻訳を計画している。しかしこの調査では、本研究で目指すところと、企業での文書作成の現状の間に大きなギャップがあることがわかった。以降、その現状と、解決に向けた当面の取り組みについて紹介する。

### 3.1 用語レベルの現状と当面の取り組み

社内で使用する業務マニュアルという性格上、IT製品マニュアルなどと異なり、用語を統一し徹底する必要性は十分に認識されていない。マニュアルごとに巻末に用語集が添付されているが、同じ概念が複数の言葉で表されている(「育児休暇」と「育児休業」)、複数の日本語の用語に同じ英語訳が割り当てられている(「正規従業員」と「固定人員」の英訳が両方とも“regular employee”)、マニュアル本文で使われていない用語が掲載されている、などの問題が見つかった。また、「競合他社」と「同業他社」といった、ニュアンスの異なる語が明確な使い分けの基

<sup>1</sup> <http://www.tech-jpn.jp>

準なしに使用されている例もあった。日本語だけで情報をやり取りしているときには、状況に応じて無意識に解釈を調整し、当事者間で曖昧さを吸収できているかもしれないが、日本語から英語へ、英語から多言語へと翻訳する過程で、その曖昧さは増大され、コミュニケーションに多大な支障をもたらすことになるだろう。また用語の不統一は、機械翻訳の際に訳文の品質低下の原因となる。

理想的には全社的な取り組みとして用語集を作成し、一元管理して担当者が定期的にメンテナンスすることが望ましい。しかし社内文書の作成のために用語管理専任のスタッフを確保することは難しい。長期的な方向性を踏まえつつ現在の資源で取り組める課題として、まずは業務マニュアル単位での正確な用語集の作成を提案した。その手順は次のとおりである。(1) マニュアル執筆後に、自動用語抽出システム(「言選 Web <sup>2)</sup>」など)を使ってシステムティックに用語を抽出する。(2) 生成されたリストを元に、意味が重複した用語について再検討し、同じ意味の語はひとつにまとめる。(3) 原語と訳語を一対一対応にする。

マニュアルごとの用語集が整備された後、マニュアル間で用語を揃える必要があるが、これには、意味を識別した上で類似の語を抽出するシステムが必要となり、具体的な手法は現在検討中である。

### 3.2 文章レベルの現状と当面の取り組み

業務マニュアルの書き手は多くの場合、各業務の担当者であり、専門のライターではない。このことは、特に機械翻訳を前提とした原文を執筆する上で障壁となる可能性が高い。現状では、用字用語や文末スタイルなどの統一がとられておらず、また 100 文字を超える長い文章も多い。さらに、「(教育を)展開する」「(システムを)回す」といった、曖昧な動詞表現も多々見られる。これらは機械翻訳の際に精度を低下させる原因となることが多いため、ある程度の執筆規則やガイドラインの導入は免れない。しかしプロのライターでない執筆者にとって、細かい規則やガイドラインは執筆を支援するより返って負担を増やすことになりかねない。このため、本プロジェクトでは、前述

のセクション 2.2 で説明した取り組みの結果を基に、少数の、最も高い効果が期待できる規則のみを採用することで、執筆者の負担を抑えながら機械翻訳精度を少しでも向上させることを目指す。

### 3.3 文書レベルの現状と当面の取り組み

IT マニュアル制作の世界では、SGML、xml、XLIFF、DocBook、DITA など、マークアップ言語による文書作成が主流となって久しいが、企業の社内文書では事情が異なる。MS Office が事実上の標準となっているが、その中でも、本研究の協力企業においては、文書作成の主要なソフトウェアとして、Word ではなく Excel が使われている。このことが、機械翻訳の利用において大きな障害となった。

Excel のセル内に埋め込まれた図表が数多くあり、それらは機械翻訳システムで直接扱うことができないばかりか、テキスト形式に変換することすらできないものが多かった。また、自動改行を使用せず、1 文が複数行のセルにまたがって書かれていることも多々あり、その場合は機械翻訳システムに読み込まれたとしても、セルごとに入れられた句単位でしか認識されず、1 つの文として翻訳されない。今回詳しく調査した 3 つの業務マニュアルでは、10～20%のテキストが図表内にあり、10～45%のテキストが、複数のセルに分断された文で構成されていた。この現状を改善しないかぎり、用語レベルや文章レベルで機械翻訳の精度向上に取り組んでも、実際に生成される訳文にその効果が反映されない。

そこで、とりあえずは Excel から Word に移行してもらうよう、標準的な文書構造を定めた Word テンプレートを作成した。テンプレートでは、タイトル、概念説明文、手順説明文、参考文献などの構成要素ごとに個別のスタイルを設定した。スタイルを使用して書式設定した場合、Word ドキュメントを xml 形式に変換すると、そのスタイル名がタグとなってテキストに埋め込まれる。これにより、将来何らかのマークアップ言語形式に移植することになった際にも、タグ設定の手間を省くことができる。ただし業務マニュアルは業務によって記述すべき内容が大きく異なるため、標準として定めた文書構造をどの程度共通して使用できるかは未知数である。

<sup>2)</sup> <http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html>

## 4 おわりに

以上、本研究の目的と課題、および実践環境での実行に当たっての現状と当面の解決策について、用語レベル、文章レベル、文書レベルごとに述べた。企業の国際競争力を高めるには、情報発信力の強化が不可欠であり、そのために克服すべき課題は多い。しかし現状はその課題に取り組む以前に整備しなければならない問題も多いことがわかった。今後の長期的な目標の達成に向けて、地道な努力が必要である。

### 【参考文献】

- Adriaens, G. (1994). Simplified english grammar and style correction in an MT framework: The LRE SECC project. *Translating and the Computer 16*, London, UK. pp. 78-88.
- Aikawa, T., Schwartz, L., King, R., Mo, C. O., & Lozano, C. (2007). Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. Paper presented at the *MT Summit XI*, Copenhagen, Denmark. pp. 1-7.
- Bernth, A., & Gdaniec, C. (2001). MTranslatability. *Machine Translation*, 16(3), 175-218.
- Kohl, J. R. (2008). *The global english style guide: Writing clear, translatable documentation for a global market*. Cary, NC: SAS Institute Inc.
- O'Brien, S., & Roturier, J. (2007). How portable are controlled language rules?: A comparison of two empirical MT studies. Paper presented at the *MT Summit XI*, Copenhagen, Denmark. pp. 345-352.
- 渡邊 (2010).イノベーションインフラとしての産業日本語: 特許版産業日本語と産業日本語プラットフォームの開発について *Japio 2010 YEARBOOK*, 160-165.