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（112306003）

Research and Development on Support Mechanism of Multilingual Information Outbound for Intensification of the International Competitiveness of Regional Industries

研究代表者
井佐原均　豊橋技術科学大学
Hitoshi Isahara, Toyohashi University of Technology

研究分担者
トニー　ハートレー†　神崎 亨子†
Tony Hartley†　Kyoko Kanzaki†
†豊橋技術科学大学
†Toyohashi University of Technology

研究期間　平成 23 年度～平成 24 年度

概要
本研究開発は、ノウハウの文書化を進める地域の自動車産業および、そのような文書の翻訳を受注している翻訳会社と協力して進めた。具体的には、(1)制限文法とテクニカルライティングの両面から、日本語文書の作成を効率化し可読性を向上させる無理のない規格化日本語を開発するとともに、(2) 機械翻訳のチューニングと最適化および翻訳支援環境を構築し、(3)これらが実際のノウハウ文書の作成と翻訳に有効であることを実証した。

１．まえがき

東海地域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。本プロジェクトでは、実際に企業と連携しつつ、ICT 技術を用いて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的とする。

機械翻訳を用いた翻訳過程で精度を向上するポイントとしては、前編集、機械翻訳、後編集の３つが考えられる。前編集においては、日本語を人間にとっても、機械にとっても分かりやすく記述するための規格を導入し、入力文を制約することにより、機械翻訳結果の品質向上が可能となる。機械翻訳に関しては、翻訳エンジンそのものの精度向上はもちろん大切であるが、既存のシステムの活用においては、対訳辞書の整備が翻訳精度の向上に直結する。情報発信型の翻訳が、情報受信型の翻訳と異なる点の一つは入力をコントロールできることである。入力文を規格化し、対象分野の対訳辞書を整備することにより、機械翻訳出力の精度が向上し、速報性が重視される文書や、内部での利用のための文書の翻訳には十分な精度となることが期待される。出版物やウェブ上で一定期間提示される文書などの場合には、さらに翻訳の精度を高めるための後編集作業が必要となる（図 1）。

機械翻訳をはじめとする自然言語処理技術を産業文書に適用する場合には、各企業や分野に対してシステムを適合させる必要がある。機械翻訳システムそのものを適合させる場合、用例翻訳や統計翻訳といったコーパスベースの翻訳システムであれば、学習用の対訳文書を準備することによって、機械翻訳システムの精度を向上することが可能となる。個々の企業で大量のデータを集積することには困難が伴うが、企業文書のデータの共有によって大量データを利用可能とする試みもある。

本研究開発では機械翻訳のアルゴリズムを改良するのではなく、既存の機械翻訳システムの利用を前提に、入力文書をコントロールし、対訳専門用語辞書を構築することによって、翻訳精度を向上し、必要な場合には後編集によって求める翻訳精度を実現する試みである。このような取り組みによって機械翻訳をはじめとする自然言語処理技術の活用の幅が広がることを期待している。入力文の制約においては、規格化日本語という考え方により、企業や自治体文書を翻訳しやすい文書に変更することを試みている。対訳専門用語辞書の構築においては、マニュアル文書を対象に、そこに出現する「意味のあるひとまとまりの語句」を半自動で取り出すことを目指している。

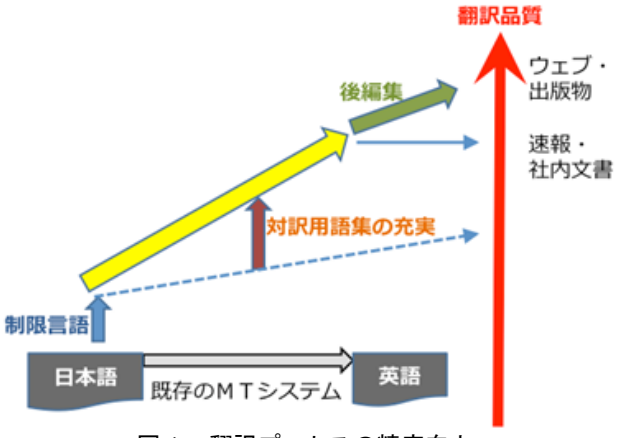

図１　翻訳プロセスの精度向上

２．研究開発内容及び成果

本研究開発においては、「規格化日本語の開発」、「MT のチューニングと最適化および翻訳支援」、「これらの技術の実証」の研究開発を実施した。

「規格化日本語の開発」においては、自動車関連企業において、本研究開発で作成した文書規格に基づく文書作成を行い、機械翻訳システムを用いて、その結果を検証した。初年度となる平成 23 年度においては、開発期間が 6 か月程度と限られていたため、産業文書用の日本語の規格の開発に注力することとしていた。協力企業の実文書を詳細に

検討し、日本語の規格を提案するとともに、その規格を用いて書かれた日本語文書と、その翻訳（英訳）文書が、日本語原文の読みやすさを損なわず、かつ英語訳文の理解容易度を向上させる、ということを示した。また、協力企業との協同作業として、同社の 100 名を超える社員に対し、日本語規格を用いた文作成についての教育を行った。平成 24 年度においては、上記の教育に基づいて作成した文書を対象に機械翻訳実験を行い、訳文の品質向上を示した。また、協力企業だけではなく、地方自治体のホームページも実験対象に加えた。また、成果を企業や関連協会と共有するとともに、標準化に向けての活動を行った。ホームページを対象としたガイドラインの一部を以下に示す。

簡潔な文を書く	(a)一文はできる限り 70 文字以内におさめてください。それ以上になる場合でも、100 文字以内にはおさめてください。 (b)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。 (c)文の中に、括弧書きで長い説明を入れないでください。
語句の関係をはっきり示す	(d)主語と述語の関係を明確にしてください。 (e)修飾語と被修飾語の関係を明確にしてください。

「MT のチューニングと最適化および翻訳支援」においては、用語自動抽出システムを改良し、実際の文書（自動車マニュアル）を対象に、再現率の減少を抑えつつ（83%⇒72%）、適合率の向上（約 10 倍）を実現した。また、他の文書（楽器マニュアル）でも同様の効果があることを示した。

我々のシステムは、文書中の用語の連接情報（統計情報）を用いることにより、意味のあるひとまとまりの語句の抽出を可能としている。本研究の目標は、全自動の辞書構築ではなく翻訳前の英語テキストから自動的に抽出した用語句候補を人手で眺めて、分野特有の用語集を取り出すことである。このためには人手で調査できる程度の数の、また正解を十分に含んだ、用語句候補を取り出す必要がある。

処理の手順は「統計的に用語（句）を抽出」⇒「用語の特徴を用いて間引き」⇒「人手での選定」となる。

用語自動抽出システムにより、103,828 文からなる英語版マニュアルから 65,712 個の用語候補を抽出した。抽出した用語の中に抽出したい 100 語句の内の 83 語句を確認した。この 83 語句を部分形態素列として含む語句は 1,767 語句が抽出されていた。再現率でみれば、83%ということになるが、適合率は極めて低く、抽出した文字列は膨大な量であり、ノイズを多く含んでいる。この 6 万強の語句を人間が見て、100 個弱の語句を選び出すのは現実的ではない。このため、正解データの取りこぼしを増やさず、かつ正解でない語句を排除するような手法が必要となる。

このため、取り出した語句の品詞情報・字種情報や連接情報を利用しての間引きを行った。具体的には、以下に示すような形態素に関する条件、統計的な条件、字種による条件を利用した。この結果、マニュアル原文から取り出した専門用語候補 65,712 個を 4,947 個に絞ることができた。この中で正解語句を含む文字列は 175 文あり、正解語句としては間引き前に 83 個含まれていたものが、72 個維持されている。5000 程度の候補から正解を目視で選ぶことは十分実用的であると思われる。取り出される用語の例を以下に示す。

シート アッセンブリ ハーネス コネクタ
シート エアバッグ スクイブ 回路 の 短絡 の 点検
the rear wiper motor output shaft
the A/C pressure transducer harness connector
Control Change and Channel Mode

「実証」については、上記の研究開発をそれぞれ実際の文書や企業での文書作成場面、翻訳会社の作成する用語集を対象に行った。また、本研究開発の成果を ISO の国際標準とする目途が立った。さらに、産業日本語研究会、テクニカルコミュニケーター協会、システム開発文書品質研究会をはじめ、産業界との交流を進めた。

３．今後の研究開発成果の展開及び波及効果創出への取り組み

平成 25 年 3 月には、企業・翻訳会社等から約 30 名の参加者を得て、日本語の規格化と多言語情報発信についての意見交換会を行った。本成果の今後の活用に大きな期待が寄せられた。

本研究開発によって得られた成果は、自然言語処理をはじめとする情報処理技術の国際競争力の強化に直結するとともに、他分野の活動の情報発信に寄与するという点で、間接的には広い分野の国際競争力強化に寄与する可能性を持つ。また、国際標準化を行うことにより、広い波及効果の創出を目指している。

４．むすび

今回、研究開発の対象とした文書は、自動車関連企業の社内文書、地方自治体のウェブページ、自動車マニュアル、楽器マニュアル、である。これら以外にもすでに複数の会社からデータの提供を受けており、今後はこれらの文書も対象に手法の有効性を実証していく。また、データを提供していただいた企業はもちろん、他の企業にも本研究開発の成果の利用を呼び掛ける。本研究開発の中で、産業日本語研究会、テクニカルコミュニケーター協会、システム開発文書品質研究会との交流を進めた。これらの活動とも連携していきたい。

【誌上発表リスト】

[1] Midori Tatsumi、Tony Hartley、Hitoshi Isahara、Kyo Kageura、T. Okamoto、K. Shimizu、“Building Translation Awareness in Occasional Authors: A User Case from Japan”、EAMT2012（平成 24 年 5 月 28 日）
[2] Tony Hartley、Midori Tatsumi、Hitoshi Isahara、Kyo Kageura 、Rei Miyata 、“ Readability and Translatability Judgments for ‘Controlled Japanese’”、EAMT2012（平成 24 年 5 月 29 日）
[3] Hitoshi Isahara、“Toward Practical Use of Machine Translation ”、 8th International Conference on Natural Language Processing (JapTAL2012)（平成 24 年 10 月 24 日）

【国際標準提案リスト】

[1] ISO・TC37、24620-1、Language resource management - Simplified natural language -- Part 1: Basic concepts and general principles、Working Draft 平成 25 年 9 月予定、Draft International Standard 平成 26 年 2 月予定

【参加国際標準会議リスト】

[1] ISO・TC37 SCs meetings、Madrid、平成 24 年 6 月 24 日～29 日