機械翻訳の実用的利用に向けた取り組み

豊橋技術科学大学情報メディア基盤センター教授 **井佐原　均**

PROFILE 京都大学大学院工学研究科電気工学専攻修士課程修了。京都大学博士（工学）。通商産業省工業技術院電子技術総合研究所、郵政省通信総合研究所（現・独立行政法人情報通信研究機構）を経て、2010年1月より現職。

1 はじめに

インターネット上に多くの言語情報が存在するようになり、情報検索が実用化され、情報抽出の研究開発も進んでいる。機械翻訳の分野においても、統計翻訳（SMT）や用例翻訳（EBMT）といった大規模なデータに基づくコーパスベースの機械翻訳の開発が進み、ビジネスへの応用も行われている。このような翻訳システムは大量のデータに支えられて、言語的に近い言語対の間の翻訳では実用性が高まっている。しかしながら、日本語と英語といった異なったタイプの言語間での翻訳では、機械翻訳はまだまだ改良の余地があるという意見も多い。10年前に我々は情報検索で得られた英語の文を機械翻訳し、対訳表示で見た場合の有効性をTOEICのテストを用いて検証し、ほとんどの日本人にとっては当時の英日翻訳システムであっても有益であることを示したが[1]、これは必要な情報が分かれば良いという情報受信型の翻訳であり、日本から海外への情報提供といった情報発信型のサービスでの日英翻訳の有効性は検証されていなかった。

本稿では、情報発信型の翻訳での機械翻訳の有効性向上の取り組みについて述べる（図1）。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（破線）。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貢献する。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掲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ろう。

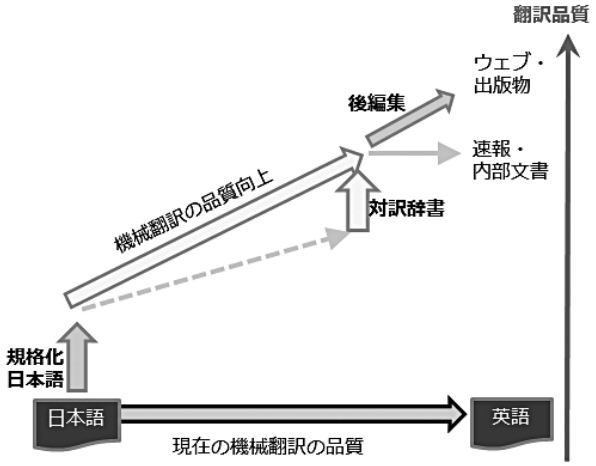

図1　機械翻訳の精度向上

2 規格化日本語[2、3]

ここでは、規格化日本語への具体的な取り組みとして、我々が中京地区の企業と協力して進めている取り組みを

概説する[1]。

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。

規格化日本語の開発においては、定めた規格が実際に使われ、有効であることが最重要である。このため、企業の協力を得て、実際のマニュアルを参照し、意見交換をしつつ、詳細な検討を行い、機械翻訳等の機械処理に適した文書の特徴を定めた。この結果を基に、規格化日本語の第１版の開発を進めるとともに、すぐに使える（現場に受け入れられる・平易な）文章作成基準を作成した。この基準の一部を図２に示す。図２の「15. 文中で記号を多用しない」を適用した例を以下に示す（図３）。この

1	1 つの文には 1 つの事柄を書く
2	1 文を 50 文字以内に収める
3	箇条書きを利用して簡潔に書く
4	助詞を省略しない
5	必要に応じて主語を明示する
6	主部と述部を正しく対応させる
7	主語以外に安易に「は」を付けない
8	目的格の助詞には「が」より「を」を使う
9	リスト内の項目のスタイルを統一する
10	具体的な表現や直接的な表現を使う
11	重複表現を省いて簡潔に書く
12	正しい文法に沿って書く
13	漢字で書くことが標準となっている言葉は漢字で書く
14	誤字をなくす
15	文中で記号を多用しない

図２　文章作成の基準（一部）

既存マニュアルの文

一般的には、標準的な条件の下で生産を行った場合の原価 =「標準原価」と言われるが、社内では、「当期首時点の実力原価 = 基準原価」と言う

その機械翻訳出力

It is called prime cost = when generally, production it does under standard condition “standard prime cost”, but inside the company, you call “capability prime cost = standard prime cost of this term neck point in time”

書き換えた文

一般的には、標準的な条件の下で生産を行った場合の原価を「標準原価」と呼ぶが、社内では、当期首時点の実力原価を基準原価と呼ぶ。

書き換えた文の機械翻訳出力

Prime cost when generally, production it does under standard condition is called “standard prime cost”, but inside the company, capability prime cost of this term neck point in time is called standard prime cost.

図３：文の修正による翻訳精度の向上例

1　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。

基準に沿って、既存のマニュアルを書き換えることにより、機械翻訳の精度が向上することを実証した。具体的には、この平易な基準に沿って作成されたマニュアルとその機械翻訳出力が、日本語および英語として適切な読みやすいものになっているかどうかを被験者実験で確かめた。「既存マニュアルの文」と「書き換えた文」を日本人 20 名に提示し、書き換え後の文が日本語のマニュアルとして、より適切な（読みやすい）文になっていることを確認した。また、「既存マニュアルの文の機械翻訳出力」と「書き換えた文の機械翻訳出力」を外国人８名に提示し、書き換えた文の機械翻訳出力がより良い英文になっていることを確認した。

このような規格を執筆手引の形にまとめ、またマニュアル作成のひな型を Word 文書のマクロで提供し、実際の現場で数十人規模でのマニュアル執筆を行った。今後、このデータを検討し、日本語規格の充実を図る予定である。

3 対訳表現の整備 [4]

実際の産業文書を高精度で翻訳する環境を実現するためには、その文書に出現する用語の辞書を整備することが必要である。このような用語は分野や企業に特化したものであり、かつ新規の語彙が常に作成されるという特徴がある。このため人手による作成は速度と経費の点で困難であり、文書から自動的に用語を取り出す技術が必要となる。

翻訳を対象とする場合、取り出すべき対象は単語だけでは不十分であり、頻出する言い回しなど、意味のあるひとまとまりの句を取り出し、その全体に対して対訳（となる語句）を与える必要がある。我々が開発したシステムは、文書中の単語の連接情報を用いることにより、このような語句の抽出を可能としている。語句抽出は「候補の選定」と「用語の推定」の２段階で行なわれる。まず、文書集合中の一定長までの形態素列（単語や単語の活用部分の並び）のすべてを対象として、多くの文書に使われていて、かつ、いくつかの文書の中では繰り返し使われている形態素列を、統計的指標を用いて候補として選定する。この方法で、人間が知らない用語を認識し、理解する感覚を計算機に持たせることができる。次に、候補を構成する形態素間の連接の強さを測ることで、その候補が用語かどうかを推定する。たとえば、「お／台／場」が用語であるかどうかを推定する場合、文書集合中の「お／台／場」に隣接する形態素の種類は、形態素「場」を削った「お／台」や形態素「お」を削った「台／場」に続く形態素の種類よりも多いという仮説を統計的指標で検証することで、「お台場」を用語と推定できる。この方法で、人間が一部を聞いただけで、残りの内容を予測するような感覚を計算機に持たせている。本手法では、対象として、名詞や複合名詞に限らず、全ての形態素列を対象とするため、テキスト集合中で特徴的な動詞や助詞を含む長い名詞句も獲得できる。

この手法の実証として、自動車や楽器マニュアルの日英対訳の提供を受け、実験を行った。図４、５に自動車のマニュアルから取り出した「意味のある語句」の例を示す。このように単語や複合語だけではなく、意味のあるひとまとまりの語句も取り出せることが本手法の特徴である。翻訳作業においては、語が適切に訳されている

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（SAB）
シート エアバッグ インライン 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
シート カバー
シート クッション の 助手席 乗員 検出 センサ

図４　マニュアルから取り出した「意味のある語句」の例（日本語）

the rear wiper motor output shaft
the A/C pressure transducer harness connector
the blend door actuators
disconnect the body wire harness connector (3) from the
brake lamp switch output circuit
checking camshaft position sensor signal with a lab scope

図５　マニュアルから取り出した「意味のある語句」の例（英語）

だけではなく、定まった言い回しが適切に訳されていることが必要である。したがって、語だけでなく句も取り出せる我々の手法が有効である。

4 クラウドソーシング後編集

絶えず更新される情報をプロの翻訳者に依頼して後編集するには膨大なコストが必要となる。コストを抑えるためにはプロの介入を最小限に抑える事が重要である。そこで我々は、プロの翻訳者ではないが、対象文書の内容についての知識がある人のボランティアベースによる後編集（集合知による後編集、あるいは Crowd Sourcing Post-editing）を提案した。

各文を複数名で後編集する場合、二人目以降は、原文と、機械翻訳システムによる翻訳出力と、それまでの後編集結果を参考にして、更に良い文を作ることができる。最終的に得られる後編集結果はプロの翻訳家が翻訳した（あるいは後編集した）文に近い品質となると考えられる。他の人の後編集結果を参考にする事ができるので、翻訳技術の乏しい人でも参加する事ができる。また、修正に自信がある文だけを後編集することができる。

豊橋技術科学大学では、多言語での情報発信を実現するため、英語版ホームページに Microsoft Translator を設置している（図６）
（http://www.tut.ac.jp/english/introduction/）。

図６　大学ホームページの機械翻訳と後編集
（この画面では、英語からベトナム語への翻訳出力に対して後編集を行っている）

まず最初に、集合知による後編集の省力化の有効性を検証する為に、Microsoft Translator CTF（Collaborative Translation Framework）を用いて本学の留学生（９か国語）に対し、母語とする言語の翻訳結果を後編集するように依頼した。母語が同じ留学生を一つのグループとし、英語版ホームページを母語に翻訳した結果に対し後編集を行った。現在、本学の英語版ホームページ上の約 2,500 文が Microsoft Translator で翻訳可能である。留学生はプロの翻訳者（後編集者）ではないが、対象言語の母語話者であること、今回の対象が大学のホームページの記述であり、大学の実体についての知識を持っていることから、プロの翻訳者と同程度の後編集が可能になると考えている。

ボランティアの集合知による後編集が有効であるかどうかを示すために、各グループが後編集した結果の品質を人手評価と自動評価とで評価している。人手評価では、筆者らは、翻訳先言語を理解しないため、機械翻訳の結果と後編集結果の比較や、後編集結果の品質評価を各言語の母語話者に依頼する。具体的には、後編集結果を参照し、適切な後編集結果がある場合には、それを指定した。既存の後編集結果では満足できなかった場合には、さらに後編集を行う。この場合は、既存の後編集結果の何処が何故、問題であったかを事後に確認する。現在インドネシア語、スペイン語、中国語、ベトナム語の４カ国語について実験中である。また、自動評価では TER（Text Error Rate）を使って評価を行う予定である。

この実験では誰がどのような修正を行ったかを記録する事ができなかった。その点を踏まえ、厳密に統制した実験により集合知による省力化が有効かを評価する。現在、本学の日本人学生４人を被験者に英日方向の同様の実験を行っている。ここでは、ある文に対し、どの学生が何番目に後編集を行い、どのような編集を行ったかを記録している。この実験によって前回までの後編集結果を参照する事の有効性や、後編集に必要な最低限の人数を見極める。

5 おわりに

本稿では精度向上の進んだ現在の機械翻訳システムを実際の翻訳サービスの場面で、より有効に活用するための関連技術について述べた。各技術は実際の文書を対象に開発・評価を行っており、学術的にも実用的にも有効な成果を出しうると考えている。

参考文献

[1] Fuji, M. et al. (2001). Evaluation Method for Determining Groups of Users Who Find MT Useful. In Proceedings of the Machine Translation Summit VIII.

[2] Tatsumi, M., et al. (2012). Building Translation Awareness in Occasional Authors: A User Case from Japan. In Proceedings of EAMT2012.

[3] Hartley, A. et al. (2012) Readability and Translatability Judgments for “Controlled Japa-nese”. In Proceedings of EAMT2012.

[4] Yamamoto, E. et al. (2008). Extraction of Informative Expressions from Domain Specific Documents. In Proceedings of LREC 2008.

[5] Aikawa, T. et al. (2012). The Impact of Crowdsourcing Post-editing with the Collaborative Translation Framework. In Proceedings of JapTAL2012.

[6] 相川孝子、井佐原均（2011）。ホームページの多言語化に向けた機械翻訳とコミュニティによる後編集の活用、言語処理学会第 17 回年次大会発表論文集。

[7] 山本健太郎、相川孝子、井佐原均（2012）。機械翻訳出力の後編集の集合知による省力化、言語処理学会第 18 回年次大会発表論文集。