機械翻訳の現状、期待と今後の動向
－翻訳プロセスでのさらなる活用に向けて－

Current Status, Expectations and Future Directions for Machine Translation

名古屋大学大学院情報科学研究科特任教授 中岩　浩巳

PROFILE アジア太平洋機械翻訳協会（AAMT）会長、国際機械翻訳協会（IAMT）次期会長、言語処理学会前会長

nakaiwa@is.nagoya-u.ac.jp

1 はじめに

コンピュータが生まれた 1940 年代から、機械による翻訳に関する技術の研究開発は始まり、半世紀以上の時を刻んでいる。また、昨今の急激なグローバル化、特に、日本の急激な少子高齢化に伴う、人口減少による日本産業の海外依存度の増大が、ますます進んでいる。このような状況の変化に伴い、国際ビジネスを加速する機械翻訳への期待はますます高まっている。しかし、今までは様々な制約から、機械翻訳が広く活用される状況では必ずしもなかった。しかし、最近の機械翻訳を取り巻く環境、及び、翻訳を生業とする方々の機械翻訳への意識の変化に伴い、機械翻訳が活躍する場は急激に広がってきた。本稿では、最近の機械翻訳を取り巻く状況の変化を概説するとともに、特許翻訳という観点からその可能性と今後の展望について触れたい。

2 機械翻訳の歴史と最近の技術的変化

前述の通り機械翻訳はコンピュータが生まれた当初から研究が行われてきた。当時の翻訳技術は翻訳対象言語の単語を目的言語の単語に入れ替え、その後、目的言語にあう語順に並び替えるという単語直接方式がおもな手法であった。その後、チョムスキーらによる文法理論研究の発展に伴い、翻訳対象となる言語を文法に基づき構文解析し、対訳辞書や翻訳規則に基づき目的言語の構文構造に変換ののち、目的言語の文を生成する、変換（トランスファー）方式が主流となった。しかし、変換方式では、翻訳言語対ごとに対訳辞書や翻訳規則が必要となるため、多言語翻訳を実現するには言語対の組み合わせの数だけの膨大な知識を準備する必要がある。この問題を克服するために、架空の中間言語を設定し、翻訳対象言語を、構文構造などを介して中間言語表現に変換し、その中間言語表現を、構文構造などを介して目的言語に翻訳するという中間言語（ピボット）方式が提案された。これにより今まで言語対の組み合わせだけ必要であった辞書や翻訳規則が、翻訳対象言語→中間言語の辞書や規則、及び、中間言語→目的言語の辞書や規則だけを用意すればよくなり、例えば、十か国語の多言語翻訳を行う場合、90 種類の辞書や規則が必要であったところが、20 種類（双方向兼用の場合は 10 種類）の辞書や規則を用意するだけで済むようになった。

以上、当時の翻訳方式は、基本的に複数の言語が理解できる翻訳者や、言語の構造に明るい専門家が辞書や翻訳規則を人手で作成しているルールベース翻訳方式が採用されていた。よって、辞書や翻訳規則の構築に膨大な人的・時間的コストがかかっていた。また、翻訳対象分野や文体により、辞書や翻訳規則を一部修正しないと、十分な翻訳性能を達成することは困難であるため、そのためのコストも必要となっていた。

1980 年代の後半になると、過去に翻訳した対訳データを活用した翻訳技術が考案される。国際ビジネスを行っている企業は、出願特許やマニュアルなどをすでに大量に翻訳した対訳データの蓄積があり、それらの言語資源を再利用しようという動きである。その代表的な手法は統計翻訳技術である。統計翻訳では翻訳対象言語と

目的言語の文対を大量に用意し、単語がどの単語に翻訳されるか、また、語順はどう変化するか、共起する単語はどれか、などをこの対訳データから統計情報として計算し、確率の最も高い翻訳結果を出力する。当初は、単語単位での翻訳が主流であったが、上記ルールベース翻訳方式の技術的発展と同様に、フレーズ単位の翻訳、及び構文構造単位の翻訳など、より構造的な特徴を活用した統計翻訳にシフトしていった。また、最近ではディープニューラルネットワークを活用した翻訳方法も提案されている。

上記のように、近年機械翻訳技術へのパラダイムシフトが起こったが、統計翻訳方式がどの言語対においても優位という状況にはなかった。具体的には、英語とフランス語のように語順や語源などが近い言語対においては、統計翻訳方式がルールベース翻訳方式を凌駕する性能を達成しており、統計翻訳方式の利用が主流となっている。これに対し、日本語と英語のように語順も語源も大きく異なる言語対においては、翻訳過程において語順や構造的な大幅な変更が必要となるため、計算量上限られた語順操作しかできなかった従来型の統計翻訳方式では十分な翻訳性能が達成できず、ルールベース翻訳方式が主流となっていた。

しかし、近年、日本語と英語のような大きく語順が異なる言語対に対しても高い性能を達成する統計翻訳手法が考案されている。例えば、英語を日本語に翻訳する際に、英語を日本語の語順に入れ替えたのちに、日本語語順の英語を日本語に逐語訳する手法が提案されている。この手法では、特許分野の英日翻訳において、歴史上初めてルールベース翻訳方式を凌駕する翻訳品質を達成した。

3 翻訳業界からの機械翻訳への期待

上記のような、機械翻訳の技術的進歩は、機械翻訳の最も有望な利用者の 1 つである翻訳業界および翻訳を生業とする(一部の)方々の意識を変化させることとなった。

翻訳業務は、ここ近年翻訳ツールの翻訳プロセスへの導入に伴い、急激に変化してきた。具体的には、過去に翻訳した大量のデータの有効活用である。前述の通り、過去に翻訳した分野と類似の文を翻訳する場合には、過去に翻訳した履歴が参考になる。翻訳ツールでは、これから翻訳しようとする文と、過去に翻訳した文との類似性を計算し、ある閾値以上に類似した翻訳文が見つかった場合、その翻訳文対の一部を流用・参照しながら新たな文を翻訳することで、翻訳の効率化と、表現の統一を図る。この過去の翻訳履歴のことを翻訳メモリーという。このような翻訳プロセスへの移行は、特に国際企業におけるマニュアルなど、過去に大量の翻訳履歴をすでに持っている場合には、翻訳を発注する企業側から、この対訳データを提供し、翻訳メモリーによる翻訳をするように指定することが頻繁に行われるなどの、翻訳の発注形態の変化を引き起こしている。よって、受注する翻訳業者、及び、実際に翻訳を行う翻訳者は、否応にも翻訳メモリーを活用した翻訳ツールの活用しなければならない状況にある。実際に翻訳メモリーを活用したほうが、すべて一から翻訳するよりも効率的に翻訳が行われることが広く知られており、これが翻訳発注側及び翻訳業者側の翻訳メモリーの積極的導入につながっている。

しかし、上記翻訳メモリーで一定以上の類似度のある対訳データが見つからなかった場合には、従来通り一から翻訳する必要があった。よって、これらの文に対しても翻訳の効率化を行うことを目的に、機械翻訳技術の活用が注目されてきた。前述の通り、翻訳メモリーがすでに準備されている状況においては、この過去の翻訳履歴を、統計翻訳に活用することで、過去の翻訳データの翻訳傾向を反映させた翻訳結果を得ることができる。また、最近の統計翻訳技術の品質改善は、それをさらに後押ししている。

このような機械翻訳技術の翻訳プロセスへの導入への期待の高まりから、翻訳業界での機械翻訳技術に関する講演やパネルディスカッションなど、情報提供する機会が大幅に増えている。例えば、日本最大の翻訳業界団体である日本翻訳連盟では、年 1 回開催される翻訳祭において、機械翻訳に関する複数のセッションが設置され、会場が溢れるほどの聴衆を集めた。また、同連盟のセミナーにおいても機械翻訳がテーマに選ばれ、多くの聴衆を集めている。また、製品やサービスの使用説明を扱う専門家の団体であるテクニカルコミュニケーター協会で

は、同会主催の TC シンポジウムにおいて、機械翻訳に関する活用法などに関するセッションが複数開催されている。さらに、機械翻訳に関する団体であるアジア太平洋機械翻訳協会主催の機械翻訳フェアにおいても、日本翻訳連盟、テクニカルコミュニケーター協会、及び、機械翻訳協会それぞれが参加するパネルディスカッションが開催されている。

また、海外においては、機械翻訳の上手な活用法や、活用ツール、また、翻訳データを共有するための会員制組織 TAUS が設立され、主要な国際企業が会員メリットを享受している状況にある。

以上の通り、機械翻訳に対する注目は以前に増して高まっている。

4 特許翻訳における課題と今後の方向性

特許文書の翻訳需要としては、大きく、出願された特許の内容を確認するための翻訳（インバウンド）と、外国に特許を出願するための翻訳（アウトバウンド）に分れる。インバウンド翻訳では、その主な目的が、特許の出願内容の把握であるため、多少の翻訳ミスがあっても内容が分かれば許容される。その意味で、インバウンド用途においては、機械翻訳は十分活用できるレベルに達しているといえる。これに対して、アウトバウンド翻訳では、翻訳された文章が、そのまま特許の審査対象となるため、翻訳内容の正確性や、特許文としての表現の適切性など、インバウンド翻訳より高い翻訳品質が求められる。よって、機械翻訳の利用においては、現在の技術レベルを考えると、機械翻訳前後の人間による編集作業（前編集及び後編集）が必須となる。

特許文章の翻訳を機械翻訳という観点からとらえると、利点･欠点が存在する。利点はその大量な多言語データの存在である。前述の通り、最近広く使われている統計翻訳技術では、翻訳したい分野の対訳データをどれだけ事前に集められるかによって、その分野のテキストの翻訳に対する翻訳品質が変わってくる。この点、特許文書は同じ特許を複数の国に出願する場合、母国語で作成された特許明細は翻訳され、さらに、その出願された特許は、ある一定期間後には公開される。他の国に出願された同じ特許は、パテントファミリーの情報により、検索が可能である。以上のことから、他国に出願された膨大な特許を検索し、学習用の対訳データとすることにより、特許文書向けの統計翻訳システムの構築が容易できるのである。また、特許の文章は、成立させることを目的としているため、比較的、限定的で明確な文章構造や表現を持っている。これは、表現のバリエーションが存在すると、翻訳品質の維持が難しいという、機械翻訳の特性を考えると、有利な点である。また、特許文には限らないが、別の観点として、訳語の統一がある。特許文では、提案手法が明確に記述されていることが必要となるため、その文中に現れる用語も同じものを指し示すのであればまったく同じ表現で訳されることが望ましい。この点、機械翻訳では、同じ語に対して同じ訳が用いられる傾向が強いため､用語統一が容易という利点もある。

反面、特許の文は、一般的に長い文が多いという傾向がある。その典型的な例が請求項の文である。ここまで長いと 1 文をそのまま機械翻訳にかけるだけでは、活用できる品質の翻訳結果を得ることは困難であるため、文を分割するなどの様々な前処理、後処理が行われるのが普通である。また、日英翻訳のように、語順や構造が大きく異なる言語間の翻訳においては、文が長いと語順の変換操作が困難となり、翻訳品質が低下してしまう。この問題に関しては、前述の通り、語順を翻訳の前処理として並び替える統計翻訳技術が提案され、品質が改善されつつある。また、特許の性質上新しい技術に関する記述が主となるため、新たな語が頻繁に出現するという傾向にある。よって未知語を事前に発見しその訳語を登録するなどの処理が必要となる。ただし、前述の通り、特許は対訳データの宝庫であるため、大量の対訳データに対し対訳単語対を発見する手法を適用することで、ある程度は事前に訳語を登録することも可能である。さらに、複数国に出願された特許は確かに同じ内容の文書が複数言語で翻訳されるため、統計翻訳では学習データとして活用できるという利点はあるが、現在の統計翻訳技術では、学習データとして文対応関係が付与されている必要があり、実際に複数言語で出願された特許は、文対応情報が付与されているわけではないため、事前に文対応付けが必要となる。また、出願特許の構造は、出願国に応じて異なる場合が多く、出願国に応じた構造的対応

関係を認定する必要がある。これに対しては、特許文を対象とした対訳文対の自動抽出手法がいくつか提案されている。

翻訳技術の主流が、ルールベース翻訳方式から、統計翻訳方式、またその両者のハイブリッド方式と変わっていくにつれて、大量の対訳データが収集・活用できる特許翻訳はその実用性も含めて、機械翻訳にとって極めて有望な分野であるといえる。また、過去の翻訳データを活用できることは、翻訳メモリーを活用した翻訳支援など、様々な翻訳支援ツールとの連携の可能性が高いことを示している。欧米では、前述の通り、翻訳言語間の構造の近さから、英日翻訳のような構造的に遠い言語間よりの機械翻訳の品質が高いため、機械翻訳を翻訳プロセスの中に組み込むことが一般的に行われている。英日においては、昨今の急速な翻訳性能向上に伴い、翻訳プロセスへの機械翻訳の活用が始まってきている。しかし日英は、まだまだ十分な翻訳性能を達成できていないため、実務的な翻訳プロセスに組み込むためにはさらなる技術革新が必要となる。また訳文品質が十分でなくても、機械翻訳結果をうまく活用するために、業務フローの最適化を検討することも現時点では必要である。

5 まとめ

以上、本稿では、機械翻訳の歴史を振り返るとともに、機械翻訳を取り巻く状況の変化を、技術面と、翻訳業界の意識面、また海外における機械翻訳の利用状況の観点から概説した。最近の機械翻訳に対する期待の増大を受けて、機械翻訳技術を提供する側と、機械翻訳を利用する側がより一層連携し、海外の事例も参考にしながらソリューションを提供できるよう、アジア太平洋機械翻訳協会（AAMT）会長として、今後も努力していきたい。