産業翻訳における機械翻訳技術の有用性を評価する手法の構築

早川威士
株式会社アスカコーポレーション
takeshi.hayakawa@asca-co.com

1 はじめに

機械翻訳（MT）技術は、機械学習の導入を通じて着実に進歩を続けており、人工知能の応用技術として社会実装が期待されている。一方で、社会での翻訳ニーズは多様であり、他言語話者との意思疎通、母国語以外で記述された情報の理解や拡散、文書の多言語化など、様々な用途やシチュエーションで翻訳は行われる。この中で産業翻訳は主として、社会活動において作られる文書の翻訳をサービスとして提供する営為を指している。産業翻訳を特徴づける要素としては、(1) 一般に求められる品質が高く、その水準がユーザー（顧客）に依存すること、(2) 職能を有する翻訳者により行われること、(3) 専門性の高い技術文書や学術的文献、法律・行政文書が主な対象であること、(4) 商業的サービスとして資源（コスト）や時間（納期）の一定の制限のもとで行われること、などが挙げられる。

産業翻訳に MT を導入し活用するためには、単にテキスト変換の処理を MT に代替させれば事足りるものではなく、かかる産業翻訳の特性を理解し、MT が翻訳サービスの何を改善するかを明確にしなければならない。そのためには、MT 導入の効果を客観的で説明可能な有用性として評価し、MT を機能的なモジュールとして導入できるような翻訳プロセスの開発が欠かせない。しかし現状は、有用性の評価においては、この要請を満たすような画一的な評価指標は存在しない。BLEU に代表される自動評価指標は、システムとしての MT を評価する上では標準的な方法ではあるが、ベンチマークデータ（参照訳）との類似性に基づいた評価に過ぎないため、弾力的な運用は望みにくい。また、この評価手法の乏しさとも相まって、翻訳プロセスについて議論するための下地が整えられていないことも依然として課題である。MT とポストエディットの組み合わせが生産性の向上に必ずしも繋がっていないことは、その端的な例だと言える。

本稿では、評価手法の構築を主軸とし、いかに技術としての MT の有用性を可視化し、翻訳という営みのフレームワークを合理的に設計するかについて考察する。考察にあたっては、産業翻訳に求められる有用性とは何かを議論し、次いで評価指標を概観した上で著者が評価指標を構築したモデルケースを紹介する。

2 有用性の尺度

翻訳の評価を考えるとき、それは翻訳品質の評価と同義とみなされることもあるが、産業翻訳の評価においては必ずしもその限りではない。産業翻訳において翻訳はサービスであり、その文脈に即して何を評価すべきかを考える必要がある。国際標準化機構（ISO）が定める翻訳サービスの要求事項では、個別の翻訳プロジェクトが規格に適合するための仕様はユーザー（顧客）との合意に基づくべきとされている [1]。その仕様の例には、「accuracy and fluency（正確性と流暢性）」や「compliance with a style guide（スタイルガイドの遵守）」という翻訳品質に直接関わるものだけでなく、「project schedule and delivery date（スケジュールと納期）」や「quotation（費用見積り）」などの要素も含まれている。

すなわち、サービスの評価においては品質はサービスがもたらす価値のひとつに過ぎず、さらにその価値の実現のために必要な資源（コストと時間）との相対性の中で考えられなければならない。これまで、特に翻訳サービスで品質とコストおよび納期はそれぞれ相反するものとみなされてきた。MT がこの前提を覆しうるかどうかを考える上で、これらの要素を有用性の尺度として用いるのは妥当だと言えるだろう。

2.1 品質

翻訳の品質として中核的な位置を占める要素は情報の正確性と言語としての流暢性であり、この 2 軸を用いた評価は MT の評価にも長らく用いられてきた [2]。しかし、産業翻訳においてこれらは最も重要な要素ではあるものの、十分な品質の要件を満たすとは限らない。具体的には、顧客の指定するスタイルや、文書の使用目的に即した表現が適切になされているか、文書のドメインに一致した訳語が選択されているかなどが挙げられる。産業翻訳における品質の要件はプロジェクトごとに異なることが多く、この特異性をカバーできる評価が実用的である。

ただし、産業翻訳では MT の出力の評価をもってサービスの評価とはみなせないケースの方が多いだろう。それは、産業翻訳のサービスが品質管理を前提としており、誤りの起こりうる MT の後工程としてポストエディットの実施を想定しているためである。そのため、MT だけでなくポストエデ

ィットも含めたプロセス全体における品質を、いかに評価するかも評価フレームワークを設計する上では重要になる。

2.2 時間

翻訳者による翻訳（人手翻訳）の特徴的なデメリットとして、スケーラビリティが低い、すなわち翻訳のボリュームが増えればそれがそのまま時間的負荷になってしまうことが挙げられる。MT は人手翻訳よりもはるかに高速で処理を行うことが可能であり、MT を用いることで翻訳にかかる時間を短縮したいというのは合理的な希求である。しかし、翻訳プロセスの全体を考えたとき、MT の導入が直ちに時間短縮に繋がるかどうかは自明ではない。MT そのものより、ポストエディットなど前後プロセスに要する時間をどの程度短縮できるかを考えることは必須と言って良いだろう。また、間接的な要因として、MT を適用するために使用するツールの習熟度や、個別のプロジェクトを始めるための準備期間なども時間に影響する要素として挙げることができる。

2.3 コスト

金銭的コストは市場によって決められる側面もあるが、翻訳を持続性のあるサービスとして維持していくにはそのコストを検証的に測定する必要があるだろう。MT を導入するには、一般に設備投資が必要であり、エンジン構築などにかかるイニシャルコストと、サーバー使用料や電気料金からなるランニングコストを負担することになる。さらに MT の性能を高めていくには自然言語処理技術の専門家によるメンテナンスもコストの要因になる。

しかし、MT を導入することで単純にこれらのコストが増加するとなれば、メリットを可視化しにくく社会実装の推進は困難になる。したがって従来のコスト構造の見直しが求められることもあるだろう。その中で、ポストエディットのコスト設定、特に人手翻訳をポストエディットにスイッチするときの作業レートを決めるためには、その根拠を客観的に示すことができるような指標を用いる必要がある。その意味では、前項の時間もコストに関わってくる要素である。

2.4 その他の要素

一般に翻訳サービスはテキスト変換のみだけでなく、文書のフォーマッティングやデザインも含むことが多い。これらは必ずしも MT やポストエディットと独立して行われるわけではなく、例えばウェブドキュメントの翻訳では翻訳と並行して HTML タグの処理が必要になる。その制約も MT の有用性に影響するだろう。

また、翻訳プロセスそのものとは関連しないが、秘密保持やデータセンターの設置条件も MT の導入に際して考慮すべき要件である。機械学習は運用のために大量のデータを必要とするため、サービス利用の条件としてデータの提供を求めることがある。この条件がビジネス的観点から許容できるかどうかは、そのシステムの採否において決定的な要因となる可能性がある。

3 評価手法の構築

MT の評価の手法は、人手評価と自動評価に大別される。いずれも MT の評価としては確立した手法であるが、前項に記した個別の有用性を測定するには、それぞれの指標が何を表現しているのかを把握し、目的に応じて使い分ける必要がある。

3.1 人手評価

人手評価は目的に応じた柔軟な評価指標の設計が可能である。ただし、評価者個人に基づくバイアスや主観性を最小化するための方策が欠かせない。最も一般的な手法のひとつは、複数名の評価者をアサインし、評価の一致度を測定するという方法である [3]。評価の一致度は評価作業の信頼性を表すと考えられ、一致度を高めるためには作業前のキックオフミーティングや、パイロットスタディ、作業指示書のバリデーションなどを行うことが望ましい。

厳密な評価結果を得るには、人手評価は非常にコストのかかる手法である。金銭的なコストだけでなく、翻訳の高度なスキルを有する同水準の評価者を多数確保することは、資源的、時間的制約を伴う。このため、人手評価を実施するには最小限の評価データから目的とする結果が得られるよう、期待する効果やサンプルサイズを適切に設定することが重要である。

3.2 自動評価

自動評価は、MT の評価結果を何らかのスコアとして出力し、そのスコアの大小により MT の性能を表現する。自動評価指標の信頼性はこれまで人手評価との相関の高さによって検証されてきた。しかし、翻訳のどの要素を重視して開発されたかは指標によって異なるため、これを利用して、単に信頼できる指標かどうかにとどまらず、指標の意味するところを踏まえた評価が可能になると考える。

BLEU　MT の評価において最も頻用されている指標である [4]。BLEU は MT による訳文中の連続する単語が参照訳と適合する率を計算しており、2.1 項で述べた正確性と流暢性を表現している。BLEU 自体がこうした multi-modal な評価を志向していることから、この指標は総合的な評価に用いるのが妥当だと言える。ただし、BLEU には計算手法が複数存在するため [5, 6]、他の研究データと直接比較する際には注意が必要である。

編集距離　Levenshtein 距離とも呼ばれ [7]、MT の

訳文と参照訳の比較を編集（挿入、置換、削除）の工数で表し、参照訳との類似度を示している。発展形として、単語単位の距離を測定する WER[8]、並び替えの概念を導入した TER[9] がある。シンプルな指標ではあるが、転移学習などによりドメインに特化させた MT の評価など語彙の一致度が重要な状況では有用な指標となる可能性がある。

hTER　計算自体は前述の TER[9] と同一であるが、参照訳ではなくポストエディットした修正文との距離を測定したものを特に hTER と呼ぶ。同じ計算でも解釈は大きく異なり、TER が参照訳との類似性を表しているのに対し、この hTER は編集工数、すなわちポストエディットの労力を直接的に測定する指標である。2.2 項、2.3 項で述べたとおり、ポストエディットの労力は産業翻訳の生産性を大きく左右する要素であると言えるため、今後重要な意義を持つ指標になると考える。

3.3　その他の評価手法

前項までに記載した評価の多くは MT の品質を測定あるいは推定することにフォーカスしているため、有用性の一部をなす時間やコストを評価するには、MT 出力以外の要素を評価するのも有効な方法だろう。例えば、ポストエディット作業の負荷を評価するために、所要時間を直接測定したり [10]、視線移動を記録する手法 [11] などがこれまで提案されている。

4　評価手法構築のモデルケース

本項では、著者が実際に評価を設計した 2 つの事例を記述する。これらの事例では、学習データに多様な対訳コーパスを含む汎用の MT システム（汎用 MT）と、医学分野の対訳コーパスを用いて転移学習でドメイン適応させた MT システム（ドメイン MT）の出力を比較するという状況を想定した。

4.1　エラー分類の評価

背景　MT は over-generation（湧出し）や under-generation（訳抜け）など特徴的なエラーを生むことが知られており [12]、これらは情報の正確性を損ねるため翻訳の品質において重要な課題である。しかし、既存の評価指標でこのエラーを抽出できる手法はないため、人手評価によりエラーの分類と抽出を行った。また、これらのエラーがいかなる条件下でより頻度が高くなるかを評価したデータも存在しないため、2 つの MT システムを用意し比較を行った。

方法　英 → 日翻訳を対象に、汎用 MT と、ドメイン MT のエラー出現率を比較した。評価は学習データとは別の医学分野のデータ約 200 文を対象に行い、評価者 3 名が翻訳のエラー抽出と分類をアノテーションした。エラーは誤訳、構文構造の誤り、不自然な表現、不適切な専門用語、訳抜け、湧出し、訳語不統一の 7 種類に分類した。

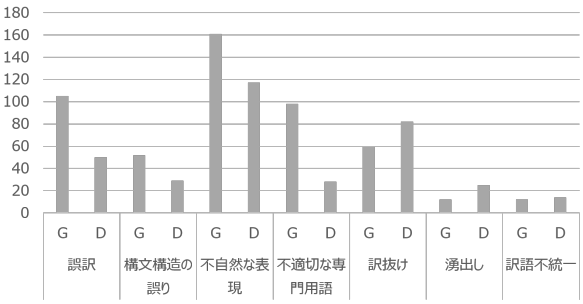

Figure 1: MT 訳文のエラー分類ごとの出現頻度。G＝汎用 MT からの出力、D ＝ドメイン MT の出力

結果　抽出されたエラーの総数は汎用 MT よりもドメイン MT の方が少なかった（500 対 345）。エラー分類に基づいてみると、誤訳や不適切な専門用語のエラーはドメイン MT の方が顕著に減少したが、訳抜け、湧出しのエラーはむしろドメイン MT の方が増加した（図 1）。

解釈　ドメイン適応により語彙選択にまつわるエラーは減少したが、言語生成に関わるエラー頻度に改善は見られなかった。しかしこの傾向を意識することで、ドメイン MT のポストエディット効率改善に寄与するかもしれない。

4.2　固有表現の再現率

背景　Neural MT は柔軟な表現が可能だとされているが、裏を返せば必ずしも学習データに忠実な出力を行うわけではない。しかし、産業翻訳では文書によって特定の訳語が決められている定型的な固有表現があり、これらに対する MT の訳文が多様であると却って品質を損ねることになる。そこで、ドメイン適応を行ったときに転移学習の学習データに含まれる固有表現がどの程度 MT 出力で再現されるかを調べた。

方法　ドメイン適応の学習に用いた日本語を原文とする対訳コーパス約 10000 文の中から、出現頻度 10 回以上の単語またはフレーズを固有表現として 100 個選択した。固有表現を含む文を 1 文ずつ（計 100 文）抽出し、それらは学習データからは除外して検証データとした。検証データの日本語文を汎用 MT とドメイン MT で英訳し、固有表現の訳出が対応する英語のフレーズに一致する割合を算出した。

結果　固有表現再現率は汎用 MT で 67.5%、ドメイン MT で 83.0% であり、ドメイン MT の方が再現率は優れていた。固有表現の訳出例を Table 1に示す。

解釈　ドメイン適応により固有表現を適切に訳出できることが示されたが、それでも不一致率は 15% 以上にのぼる。この改善がポストエディットの効率化につながるかどうかは、時間計測の評価などのアプローチが必要だろう。

固有表現原文	参照訳（正例）	ドメイン MT	汎用 MT
装置	Apparatus	Apparatus	Equipment
質量偏差試験	Mass Variation	Mass deviation test	Mass deviation test
比活性	Specific activity	Specific activity	Specific activity
医薬品各条	the monograph	the monograph	each drug article
紫外可視吸光度測定法	Ultraviolet-visible Spectrophotometry	Ultraviolet-visible Spectrophotometry	UV-visible absorbance measurement method

Table 1: 2 種類の MT における固有表現の訳出例

5 おわりに

本稿では、MT を産業翻訳に導入するための評価の考え方と、モデルケースを紹介した。現時点では MT の有用性を評価できる万能の指標は存在しないが、各指標の意図を理解して用いることでそこに意義を持たせることはできる。人手評価と自動評価を組み合わせたり、独自に評価指標を設計することでより評価の表現力は増し、さらに評価をプロセス設計にフィードバックすることで、MT を用いた翻訳プロセスのメリットを強調し、課題を解決していくことができると考える。

References

[1] ISO 17100:2015 Translation services - Requirements for translaton services. Standard, International Organization for Standardization, Geneva, CH, 2015.

[2] Linguistic Data Annotation Specification: Assessment of Adequacy and Fluency in Translations. Revision 1.5. Technical report. Technical report, Linguistic Data Consortium and others, 2005.

[3] Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23, 1981.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[5] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics, 2004.

[6] Jianfeng Gao and Xiaodong He. Training MRF-based phrase translation models using gradient ascent. 2013.

[7] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[8] Sonja Nießen, Franz Josef Och, Gregor Leusch, Hermann Ney, et al. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *LREC*, 2000.

[9] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200, 2006.

[10] Maarit Koponen, Wilker Aziz, Luciana Ramos, and Lucia Specia. Post-editing time as a measure of cognitive effort. *Proceedings of WPTP*, pages 11–20, 2012.

[11] Michael Carl, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. The process of post-editing: a pilot study. *Copenhagen Studies in Language*, 41:131–142, 2011.

[12] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany, August 2016. Association for Computational Linguistics.