翻訳の品質と効率: 実社会におけるニーズと工学的実現可能性

藤田 篤　　山田 優
情報通信研究機構　関西大学

1 はじめに

産業のグローバル化や機械翻訳 (MT) 技術の発展に伴い，翻訳産業をとりまく状況は急速に変化しつつある．例えば，大規模な対訳データや深層学習に基づく手法などによって MT の性能が向上しており，翻訳会社 (LSP) やクラウド翻訳サービスなどへの応用の期待が高まっている．また，実務翻訳や翻訳教育，翻訳研究においては，翻訳の品質評価やデータの分析などの観点で，自然言語処理 (NLP) 技術，およびより広範な各種情報技術に対する潜在的なニーズがある．すなわち，翻訳ワークフローの洗練・進化が求められている．

このような状況を鑑み，我々は，本学会第 22 回年次大会において『文理・産学を越えた翻訳関連研究』と題したテーマセッション [1] を企画し，文理・産学の垣根を越えて翻訳に関連するニーズとシーズを共有することを図った．このセッションには，翻訳産業，翻訳プロセス，翻訳教育などを対象とした発表が 12 件集まり，学際的な研究や協働の可能性について議論がなされた．発表者の立場はおおまかには下記の 3 種類に区分できる（図 1 も参照されたい）．

a) 言語処理・翻訳研究: 情報処理・分析の研究，MT や翻訳関連技術に関わる研究者・技術者
b) 翻訳産業: LSP，翻訳者などの実務者
c) 翻訳者の養成: 大学のコース，職業訓練校の教育者

これらの研究発表に基づいて，我々は各 2 群間の研究課題を整理した [14]．また，a) 言語処理・翻訳研究の成果が b) 翻訳産業において求められる人材やスキルに変化をもたらしているものの，そのような人材・スキルをどのように育成するかに関する b) 翻訳産業と c) 翻訳者の養成の 2 群間の連携は希薄であることをふまえ，両群の関係者が集う日本通訳翻訳学会においても議論を始めた [13]．

一方で，図 1 の 3 群の連携は翻訳にまつわる課題の一部にすぎない．そこで，前回のテーマセッションで共有されたシーズとニーズを前提とし，今回のテーマセッションでは，実社会の様々なシーンにおいて（プロダクトとしての）翻訳に求められる要件，および**翻訳の品質と効率**を制御するための翻訳ワークフローデザインやエコシステムなどのアイディアを共有したい．

[1] http://anlp.jp/nlp2016/#ts3

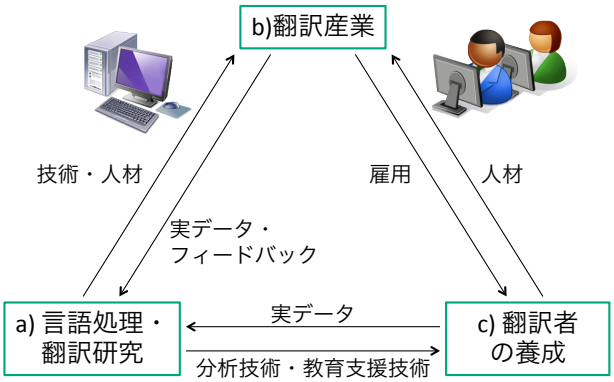

図 1: 前回のテーマセッションを通じて見られた 3 つのプレイヤー群とそれらの間の関係．

本稿では，そのような議論の足場とすべく，翻訳の品質と効率に関する既存の議論をまとめておく．

2 種々の翻訳戦略と具体例

Hutchins ら [6][2] は，一定の品質を保証することを前提とした翻訳の戦略を，人間が介在する程度を軸に整理し，次の 4 種類の典型を示した．

(a) 人間による翻訳
(Traditional Human Translation; HT)
(b) 機械に支援された HT
(Machine-aided Human Translation)
(c) 人間に支援された MT
(Human-aided Machine Translation)
(d) 高品質な MT
(Fully Automatic High Quality Translation)

これらのうち戦略 (a)〜(c) は少なからず人間が介在するものであり，当該作業者の能力および誠実さが担保されれば，一定の品質の保証が可能である．戦略 (b) の具体例としては，翻訳メモリ (TM)，用語集管理，辞書引き機能，対訳コンコーダンサなどを用いた翻訳が挙げられる．日本においても，そのような機能を備えた翻訳支援ツール（翻訳作業環境）[3] が一般的に用いられており，現代の翻訳産業におけるベースライン的な戦略とみなせるだろう．

戦略 (c) は，戦略 (b) よりも機械的な処理の比重が大きい．言い換えれば，翻訳作業の根幹をなす部分について機械的な処理の結果をある程度信頼できる場合の

[2] http://www.hutchinsweb.me.uk/IntroMT-TOC.htm
[3] Trados，Memsource，MemoQ など．

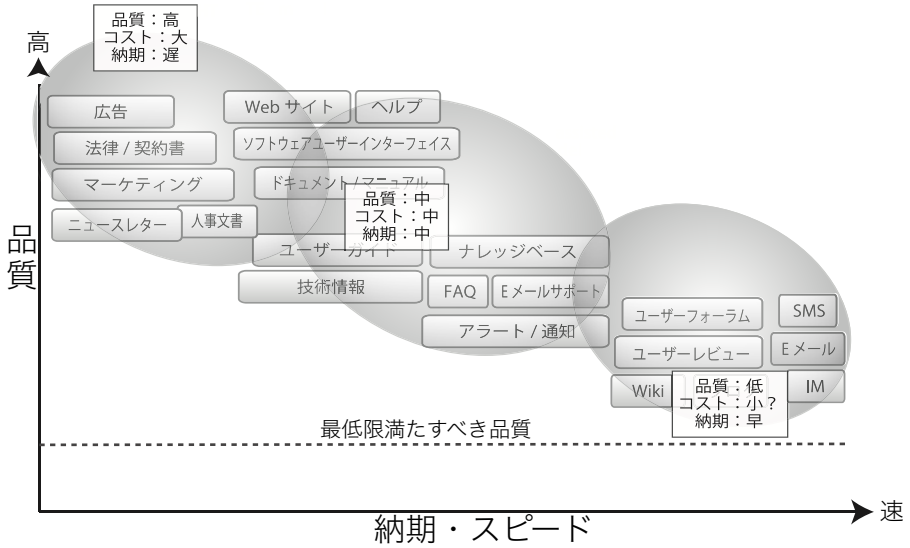

図 2: 品質 (Quality) と納期・スピード (Delivery) に基づく翻訳事例の分布.

戦略である．最たる具体例は，まず MT 結果の後編集（ポストエディット; PEMT）である．欧州言語の一部の言語間では早くから MT の精度がある程度高かったため，この手法がよく用いられている．他にも，インタラクティブ MT [5] と呼ばれる手法がこの戦略に該当する．これは，テキスト入力時の予測入力と同様に，人間が翻訳を入力する際に MT によって後続の表現を予測して提示するものである．

戦略 (a)〜(c) には人間の作業を含むため，短期間に大量の翻訳を行うことは困難であるし，新たな言語対や新たな専門分野に対応する人材をまかなう際のコストも大きい．LSP および翻訳者がより大きな収益をあげるためには，品質を保ちながら効率・生産性を上げることが大きな課題である．戦略 (a)〜(c) の中では最も高い効率が期待できるのは戦略 (c) であるが，その多寡は MT の性能に依存する．効率を担保する手段として，品質推定 (Quality Estimation; QE) [9] と呼ばれる，参照訳なしに MT 結果を評価する技術に期待が寄せられている [3]．例えば，文レベルの QE 技術は，PEMT と，より人間が主体となる戦略 (a)，(b) との選択的切り替えに資する．また，MT 結果の個々の語の品質を推定できれば，PEMT の対象を可視化したり，インタラクティブ MT において提示する候補を制御するなどして，効率を改善できる．

繰り返しになるが，上記は一定の品質を保証するための戦略を分類したものである．MT 技術は，参照訳を用いた自動評価尺度 [4] の出現により大きく改善したが，それ単体では，現時点で品質に関する前提を満たしていない [5]．それでは，MT は，どのような状況下で，どの程度の品質を担保することを目指すべきだろうか? まずは，目的に応じて担保すべき「一定の品質」，すなわち犠牲にできることと譲れないこと，を具体化する必要がある．今回のテーマセッションをこのような議論の端緒とするため，次節では実社会における様々な翻訳事例を挙げる．

3 実社会におけるニーズ

翻訳産業に求められる翻訳の品質とは，それがビジネスである限りにおいて，QCD の 3 指標（Quality: 品質，Cost: コスト，Delivery: 納期・スピード）で評価される．これらの指標は一般にトレードオフの関係にある．実社会における様々な翻訳事例を，要求される品質 (Q) と納期・スピード (D) の 2 軸で整理したものを図 2 に表す．情報の発信や拡散 (dissemination) の用途は主に図の左上に，情報の収集 (assimilation) の用途は主に右下に配置される．例えば，情報発信の事例である出版翻訳の場合，高い品質を担保するために高いコストと長い時間をかけざるをえない．また，企業などのウェブサイトを翻訳する場合も，ブランドを損なわないために，これに準ずる QCD が当てはまる．英日翻訳の人間翻訳 (HT) では，原文 1 ワード単価 10 円程度が相場であり，スループットは上級者でも 1 日あたり 2000 ワードである．そのコストを払っても満足できる品質が，求められる品質であるともいえる．

[4] BLEU，NIST，WER，TER，METEOR，RIBES など．

[5] Bar-Hillel [2] は，戦略 (d) を「研究目標として非現実的であるだけでなく原理的に不可能である」と評した．

一方で，図 2 の右下に示すように，情報収集の用途では品質を多少犠牲にしても高スピードで翻訳を行う必要がある．2 節で紹介した戦略 (b)，(c) は QCD の 3 指標に柔軟性を与え，このような翻訳のニーズへの対応を可能にした．ただし一般に，スピードを求めるにつれて到達可能な品質は下がる．このような背景から，最低限満たすべき品質 (図 2 の下部) というものも意識されるようになってきた．

3.1 翻訳の品質

3.1.1 品質に関わる概念

翻訳そのものの品質 (intrinsic quality) を評価する指標としてよく用いられる適切さ (adequacy) と流暢さ (fluency) という概念は，表裏一体で切り離して考えることができないことが多い．適切さというと，原文と訳文が完全に対応しているかの幻想を生むが，必ずしもそうとは限らない．例えば，ニューラル機械翻訳 (NMT) で，自動車の「ベンツ」が「BMW」と誤訳されたことが最近取り沙汰されているが，いずれの語も「ドイツ車」という上位概念を共有している点は見逃せない．このようなエラーは実は HT でも頻繁に起こりうる．すなわち NMT が人間らしいエラーを生じていると言えるのかもしれない．例えば，字幕翻訳など字数制限のある現場では,「カリフォルニア」を「西海岸」と訳す場合もある．適切さを評価する際には，このような訳出の可能性をも考慮する必要があろう．

流暢さを決定している要因も様々である．まず，文法的な正しさと，ネイティブならばどのように表現するか (native-like selection)，ということの間には大きな乖離がある．数ある要因のひとつに有標・無標 (un/markedness) の違いが挙げられる．また，“Length” を日本語では「長さ」と言うが,「短さ」とは言わない．この類の判断は，大規模なコーパスから，ある程度統計的に実現できるかもしれない．流暢さの下位要素として，読みやすさ (clarity) も考慮する必要がある．これには結束性 (cohesiveness) や情報構造 (theme-rheme progression) の問題が関係する．例えば次の文章を見てみよう．

(1) あるところにおじいさん が いました．
おじいさん は 山へ芝刈りに行きました．

名詞句に付随する新情報マーカーの「が」が旧情報「は」へと格下げされることにより，2 つの文の結束性が高まっている．このように結束性を担保するには，現状の機械翻訳のように個々の文 (sentence) を独立に翻訳するのではなく，文と文がつながる文章 (text) や文脈 (context) を考慮した解釈と処理が必要になる．

文章を超えたレベルの品質 (extrinsic quality)，例えば，言語表現が果たしうる語用論的な役割についても考慮する必要がある．文脈によっては「敵は来ない」と「もう大丈夫だ」は同義である．これは発語内行為 (illocutionary act) などとも関係する．また発話の受容側の理論では，Nida [8] が，文化的差異を乗り越える方略として「動的等価」を提唱している．翻訳の等価は，翻訳の受容者の反応[6] の等価と捉える．聖書における「子羊」は，その動物が存在しないアイスランド人向けに翻訳をするならば「アザラシ」に置き換えることも可能であると説く．Nida は，こうすることで原文と訳文に対する読者の「反応」が等しくなると考えた．翻訳の現場においても，動的等価に近い手法（一般に，意訳と言われるもの）が頻繁に用いられる．

3.1.2 品質の評価方法

翻訳の品質に関する種々の指標のうちどこまでを担保すべきかは翻訳（プロダクト）の用途に依存する．また，品質評価を行う際の評価基準も，評価の目的に応じて異なるであろう[7]．Translation Automation User Society (TAUS) はこのことをふまえ，Dynamic Quality Framework (DQF) という評価の枠組みを提唱した．これは，訳文中の適切ではないスパンを認定し，各々に対して「綴り誤り」や「固有表現の訳出誤り」などのカテゴリを付与するための，すなわちミクロな評価のためのツールである．現在では，欧州の Quality Translation LaunchPad で開発された Multidimensional Quality Metrics (MQM)[8] と誤りの体系を共有しており[9]，MQM における決定木 [7] を用いたカテゴリ分類が可能になっている．DQF や MQM は，2 節で述べた戦略 (a)〜(c) の出力，すなわち翻訳者が産出する翻訳や PEMT の出力に対する包括的なエラー分析を目的として設計されている．

一方，MT 結果の品質を直接評価する場合，上記以外の問題点も多く含まれる．また，そもそも現時点の MT は個々の文を独立に翻訳する方式が主流であり，文献 [11] のような試みはあるものの，多くの場合は，文章レベルおよび文章を超えたレベルの品質を担保しようとしていない．実際に，翻訳に関する評価型ワークショップ[10] における人間による主観評価でも，文を単

[6] Austin [1] の言うところの発話媒介行為 (perlocutionary act) と考えてもよい．

[7] 例えば豊島ら [10] は翻訳の初学者に対するフィードバックを想定した評価基準・評価方法について検討している．

[8] http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics

[9] https://www.taus.net/evaluate/qt21-project

[10] WMT，WAT，IWSLT など．

位とした，マクロな評価が行われている．またその際，適切さと流暢さを考慮した絶対評価が行われる場合もある一方で，異なる MT システムの結果どうしの比較に基づく相対評価が行われる場合もある．

3.2 翻訳の効率・生産性

翻訳の納期・スピードは図 2 の右側の翻訳事例で特に重要視される．このようなニーズに応える方法としては MT システムをそのまま用いることが考えられるが，図の中央の用途においては，より高い品質を保証するために PEMT や TM が必要である．2 節で述べたように，HT と比べると PEMT や TM の方が効率・生産性が高い．このため，日本の翻訳産業において，図の左上の用途向けにも PEMT や TM が採用され始めている．

効率・生産性の指標として，ここまでは主にスループットのみを考えてきたが，実際にはそれ以外の要因も考慮する必要がある．現時点では，PEMT には翻訳者が従事することが多い．しかし，実際にポストエディターに求められる資質は HT のそれとは異なる．すなわち，従来の翻訳教育・翻訳者訓練において優秀な学習者が，必ずしも MT のエラーの検出能力や修正能力など，PEMT に対する適性が高いとは限らない [12]．ISO においてもポストエディターの要件の整理が進められている．また，HT に比べると PEMT はむしろ苦痛であると語る経験者も多い [4]．このような心的負荷の要因として，MT の流暢さが向上したことによって適切さの問題を検出しづらくなっていることが挙げられる．この傾向は，NMT の台頭によってさらに強まっている．

他の要因としては，PEMT をする品質に達していない MT 結果が混在したり，MT 品質のばらつきが多く，PEMT できるものとできないものが混在していることなども考えられる．PEMT によって実社会における翻訳の効率・生産性を高めるには，MT の結果が下訳として使える品質に達していること，およびポストエディターがそのような判断を容易にできなければならない．このような課題に対して，2 節で述べたインタラクティブ MT や MT の品質推定 (Quality Estimation; QE) 技術への期待が高まっている．

翻訳産業においては，今後もテクノロジーと駆使した翻訳手法が使われ続ける．そして翻訳者は，このような状況の変化に追従せざるを得ないだろう．

4 おわりに

本稿では，翻訳の品質と効率をめぐる既存の議論および戦略について概観した．今回のテーマセッションにおいて，言語処理研究者，翻訳研究者，翻訳産業に関わる方，実務翻訳者など，異なる背景を持つ参加者間での議論の足場となれば幸いである．

参考文献

[1] J. L. Austin. *How to Do Things with Words*. Oxford University Press, 1962.

[2] Y. Bar-Hillel. The present state of research on mechanical translation. *American Documentation*, Vol. 2, No. 4, pp. 229–237, 1951.

[3] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the 1st Conference on Machine Translation* (*WMT*), pp. 131–198, 2016.

[4] R. Fiederer and S. O'Brien. Quality and machine translation: A realistic objective? *The Journal of Specialised Translation*, Vol. 11, pp. 52–74, 2009.

[5] S. Green, S. I. Wang, J. Chuang, J. Heer, S. Schuster, and C. D. Manning. Human effort and machine learnability in computer aided translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 1225–1236, 2014.

[6] W. J. Hutchins and H. L. Somers. *An Introduction to Machine Translation*. Academic Press, 1992.

[7] A. Lommel, A. Görög, A. Melby, H. Uszkoreit, A. Burchardt, and M. Popović. QT21 deliverable 3.1: Harmonised metric, 2015. http://www.qt21.eu/wp-content/uploads/2015/11/QT21-D3-1.pdf.

[8] E. A. Nida. *Toward a Science of Translating*. Brill, 1964.

[9] L. Specia, D. Raj, and M. Turchi. Machine translation evaluation versus quality estimation. *Machine Translation*, Vol. 24, No. 1, pp. 39–50, 2010.

[10] 豊島知穂, 藤田篤, 田辺希久子, 影浦峡, A. Hartley. 校閲カテゴリ体系に基づく翻訳学習者の誤り傾向の分析. 通訳翻訳研究への招待, pp. 47–65, 2016.

[11] F. Ture, D. W. Oard, and P. Resnik. Encouraging consistent translation choices. In *Proceedings of Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pp. 417–426, 2012.

[12] M. Yamada. Can college students be post-editors? an investigation into employing language learners in machine translation plus post-editing settings. *Machine Translation*, Vol. 29, No. 1, pp. 49–67, 2015.

[13] 山田優, 藤田篤, 影浦峡, 武田珂代子, 立見みどり. 文理・産学を越えた翻訳関連研究: 端緒の議論と今後の展望. 日本通訳翻訳学会第 17 回年次大会予稿集, p. 34, 2016.

[14] 山田優, 藤田篤, 影浦峡. 言語処理学会テーマセッション『文理・産学を越えた翻訳関連研究』開催報告. *AAMT Journal*, Vol. 63, pp. 36–40, 2016.