レポート

機械翻訳＋ポストエディットの実証研究：先行研究レビュー

立教大学大学院 異文化コミュニケーション研究科 博士後期課程
山田　優

実務翻訳の世界では、近年、機械翻訳(MT)を翻訳支援ツール(CAT=computer assisted translation)として利用できるか否かの議論は非常に活発化してきている。翻訳メモリ(TM)とMTを融合したツールが実用化されたりと(SDL Tradosの日本語MT対応、OmegaTのGoogle Translateとの連動など)、翻訳支援ツールとしてのMT活用が非常に注目されている。

MTをCATとして利用する方法としては、大きく分けて前編集(pre-edit)と後編集(post-edit)があるが、実用的には後編集、いわゆるポストエディット（以下PE）が主流だ。PEは説明するまでもなく、機械翻訳の訳出結果を後から人の手によって修正することである。厳密にはPEにも種類があり、原文の意味が分かればよい程度に目標言語に仕上げるためだけの簡易的なRapid post-editingから、出版や実務のレベルまでに品質を上げるFull post-editingがある(Allen, 2003を参照)。

さて、ここで実際問題になるのは、機械翻訳のPEは、TMやツールを使わない翻訳よりも、質・効率ともに優れているのかどうか、という事だろう。PEを前提として、機械翻訳の訳は使えるのか使えないのか、実務翻訳者の間でよくある議論である。しかし実務経験に基づく様々な意見や議論が交わされている一方で、客観的かつ実証的な研究は、ほとんど行われていない。少なくとも、日本においては、それが研究環境の現状であろう。そもそも「翻訳」というものがアカデミックの場で、研究の対象として認知されていないことに起因する。本稿の趣旨とずれるので詳述は避けるが、欧州では、翻訳学(translation studies)という分野が学問的に認知されている歴史があり、翻訳や通訳研究が大学等の機関で行われている。日本でもようやく学術的に翻訳研究を行う大学院や学会が増えてきているものの、まだまだ発展途上である。筆者は今、立教大学大学院の博士課程で翻訳研究を行っているが、このような研究ができる学術機関の数はまだまだ少ない。今後の発展と普及を願いたい。

以上、前置きが長くなったが、本レポートは2回の連載を予定しており、１回目の今回は、機械翻訳のポストエディット(MT+PE)を翻訳支援ツールとして活用した場合の、翻訳品質と作業効率に関する先行研究を概観する[注1]。海外での近年の研究結果を鑑みると、諸国語の組合せでは、MT+PEは実用レベルに達しているようである。次回のレポートでは、先行研究に基づき英語・日本語の組合せで、筆者が行った実験結果を掲載する。では、先行研究の詳細を見ることにしよう。

ALPAC報告書(1966)

MTの歴史において色々な意味で有名なALAPC報告書ではあるが、この報告書中にもポストエディット(PE)に関する記述があるので(Appendix 19)、その検証結果から見てみよう。ALPACは、23名の被験者（プロとアマチュア翻訳者が混在）に対し、英語・ロシア語の（技術文書）のPEについて実証検証を行っている。最初のアンケート調査では、翻訳者がPEを行った主観的な感想を、ツールを使わない普通の翻訳の場合との比較で調べている。感想の結果は、8名がPEの方が楽であったと回答した一方で、別の8名はPEよりも普通に翻訳をした方が楽だと回答した。また6名はPEも普通の翻訳も同じであると答えた。この結果自体が特に示唆することは無いのだが、興味深いのは、普通に翻訳をした方が楽だと回答した8名の翻訳者のうち6名が経験の長い翻訳者であり、逆にPEの方が楽だと答えた8名中6名は初心者の翻訳者であった点である。

実際の翻訳速度の結果と照らし合わせても、面白い結果が見られた。熟練翻訳者は、普通に翻訳を行う速度とPEの速度はあまり変わらず、初心者の翻訳者はPEを行うことにより普通に翻訳する速度の1.5倍から数倍程度の向上がみられた。つまり、熟練の翻訳者にとって機械翻訳のPEはあまりメリットがないが、初心者

にとっては翻訳効率の向上に非常に役に立つということになる。

しかし、PEの平均翻訳速度は被験者全体で均一化する傾向にあり、初心者がPEによって翻訳速度が上昇したとしても、熟練者が普通に翻訳する速度と大差がない。この結果のみから結論づけるならば、MT+PEは、プロの実務レベルではあまり役に立たなということになる。当時の機械翻訳の精度を考えれば、ある程度納得のいく結果ということだろうか。以下では、もう少し最近の研究を見てみることにしよう。

Krings (2001)

Krings (2001)は、機械翻訳のPE作業プロセスを検証した近年では最も意欲的な研究である。Think Aloud Protocolを用いて、翻訳者のPEプロセスと機械翻訳を使わない翻訳プロセスとの比較検証を詳細に行った。使用した機械翻訳はルールベース型(SystranおよびMetal)で、言語の組合せは英語から仏語であった。

相対的な作業効率(時間)では、普通の翻訳よりもMT+PEの方が20%程度上昇した。興味深いのは、機械翻訳の訳出と(raw MT output)とPE後のテキストとの類似度(similarity level)を比較すると、4割弱程であったという報告である。つまりMT訳の6割近くが、PEにおいて修正されたことになる。Kringsの実験が行われた10年前の機械翻訳の精度はALPAC報告書当時よりも向上していると思われるが、それでも6割という修正量は、実用化のレベルとして考えるには、少し多いようにも見受けられる。そうであるとしても、MT+PEによって翻訳に要した時間が2割減少したのは、むしろ驚くべき結果であろう。英語・日本語で2割の効率アップが達成できるとすれば、実務では歓迎されるものかもしれない。

Bowker & Ehgoetz (2007)

翻訳の品質の評価は難しい問題だ。品質の定義の仕方しだいでは、作業効率も変わってくるからだ。Bowker & Ehgoetz(2007)の研究では、この品質をユニークかつ実践的に扱い、機械翻訳の検証を行った。

翻訳の品質は、Skopos（翻訳の目的）によっても左右されると考えられるが(Vermeer, 1989/2000等参照)、Chesterman & Wagner (2002:80)は、翻訳を「サービス（業）」と捉え、その品質を測るには顧客の満足度を調べるのも一つの方法になりうる、と提案している。これは、受容者評価(recipient evaluation)と言われ(Trujillo, 1999)、Bowker & Ehgoetzは、実務で重きの置かれる3要素＝CQD（コスト、品質、納期）と関連づけて翻訳の品質を評価した。

大学の事務関連業務で発生する文書の翻訳を検証対象とした。大学や企業のように予算と時間の限られた状況では、翻訳の需要があっても、その全てを外注できない。そこで安価でスピーディーなMT+PEを利用できないか、調査するのがこの研究の目的であった。

同じ原文に対して3種類の訳出物を用意し、翻訳のユーザーとなる大学教授に対してアンケートを実施した。3種類の翻訳とは、(1)翻訳者がゼロから翻訳したもの、(2)MT+PEしたもの、(3)機械翻訳のみを行ったもの、である。品質の順位は、当然、(1)>(2)>(3)の順になる。しかし、これにコストと納期の条件を加える。(1)が一番高価で納期も長い。これに対して(2)は、(1)の5～10分の1程度。(3)は更に少なく100分の1程に設定した。数字の割合は実務の予想工数に基づいている。この条件において、ユーザーはどれを選択するのかが焦点だ。

結果は、(1)を選んだのが全体の32.3%、(2)が67.7%、(3)を選んだ人は誰もいなかった。つまり、使用目的が限定された翻訳であれば、7割弱の人はMT+PEの品質レベルで満足できるという。逆に、機械翻訳そのままでは、いくら安価かつ短納期であっても、実用レベルに達しないということである。また、(1)を選んだ3割の人は、人間の翻訳者による翻訳を必要としていたわけだが、この中には文学部や外国部学部など言語に関わる学部の教授らが多く含まれていたこともあり、言葉・言語に対する意識の違いが結果に反映されていたと、Bowkerらは分析する。

この調査で実施されたPEはRapid post-editingなので、通常のFull post-editingよりも品質は落ちていたにもかかわらず、条件次第では実用化レベルになるというのは、非常に興味深い結果である。

Fiederer & O' Brien (2009)

では、コストや時間的の制限を設けずにポストエディットを行った場合の品質は、普通の翻訳に比べてどうなのだろうか。Fiederer & O'Brien (2009)は、翻訳品質のみに焦点を当てた評価を行った。Hutchins & Somers (1992, p.163)によると、翻訳品質には３側面ある。Accuracy（正確性）、clarity（理解しやすさ）、style(スタイル)である。これら3点から、O'Brienらは

MT+PE後の翻訳の品質と普通の翻訳とを11人の評価者の評価結果に基づき調査した。

まず、clarity、すなわち、訳文が読みやすいかどうか等の基準では、PEも普通の翻訳もほぼ同等の評価であった。翻訳者が訳文の読みやすさに注意を払うのは当然であるが、PEでも普通の翻訳同様にclarityは修正されるようである。

次に、accuracyであるが、これはPEの方に軍配が上がった。Accuracyは訳抜けや原文への忠実性ということになるので、機械翻訳を使うと普通の翻訳よりもaccuracyについては良い結果になることが分かった。普通の翻訳では、意外と訳抜けや誤訳が多いということの証明とも言える。

最後のstyleは、目標言語の言語使用域などに適合した訳文となっているかという基準となる。Clarityとの違いが微妙ではあるが、styleの方は、短文評価でなく文章（テキスト）としての一貫性や、特定分野での言い回しに準じているかという要素が評価対象になる。結果は、普通に翻訳をした方がPEよりも優れていた。

以上、品質3要素の比較をまとめると、clarityはPEも普通の翻訳も同等、accuracyはPEが有利、styleでは普通に翻訳をしたほうが有利ということであった。品質の詳細分析では、PEも普通の翻訳もどちらも甲乙つけがたい結果であった。

しかし興味深いのは、総合的に判断してPEと普通の翻訳のどちらの品質が良かったか、と質問したところ、ほぼ全員の評価者が「普通の翻訳」が良かったと答えている。これは、詳細分析結果とは裏腹に、MT+PEの品質に対する悲観な結果とも解釈できる。しかしO'Brienらは、この理由として、styleの点数差が大きかった点を指摘し、評価者はstyleの要素を過大評価している可能性があることを示している。つまり、翻訳の品質評価を行う場合には、style的要素が重視されてしまい、それによりaccuracyやclarityが軽視されうるということである。逆に言えば、翻訳者自身も翻訳中にstyleに多大な注意を払っているということでもあり、PEの作業とは少し異なるのかもしれない。この点は、筆者が行った実験結果とも絡んでくるので、翻訳品質の評価方法とポストエディットのやり方を考える上で非常に重要となる要素であろう。

O'Brien (2006a)

機械翻訳にかける前に、原文に含まれる文法的曖昧性などを取り除いておけば、機械翻訳後の訳出精度が上がり、結果としてPEに要する労力が低減し効率アップにつながると、予想できる。O'Brien(2006a)は、この原文の前編集(pre-edit)に制限言語(CL)を使用することによって、その後に生成される機械翻訳の結果をPEすることにより、どの程度の効率化が図れるかを調べた。

O'Brienは、この「PE効率」を、時間的(temporal)、技術的(technical)、認知的(cognitive)側面から検証した。IBMのWebsphere(ルールベース型)を使用して、制限言語で書き直した原文（前編集有り）と書き直さない原文（前編集無し）とを機械翻訳にかけ、それぞれのPEの作業効率を調査した。

時間的な処理速度(総ワード数÷所要時間)の比較では、予想通り、前編集した機械翻訳結果をPEしたほうが速かった。ただ、分節(segment)を個別に見た場合、前編集をした方が全ての分節で速かったかと言えば、そうでない箇所も観察された。O'Brienはこの理由を次のように説明する。PEを行う場合は、単語の位置などを変えるだけで良いことがある。この操作を行うために「カット&ペースト」機能を使えば効率が上がるが、翻訳者（後編集者）の多くは、新たにキーボードから文字入力をしていた。入力作業は、認知的に負荷がからないからなのかもしれないが、このような冗長な技術的作業は、時間的な効率性からは無駄である。全ての分節で時間が短縮できなかった理由を、このような技術的操作が関与していたとした。しかし、原文に対応する訳語をキーボード入力するというプロセスは、ひょっとすると翻訳という基本行為となんらかの関係があるのかもしれないと、筆者は考えている。

さて、認知的な負荷の問題であるが、通常、翻訳者が翻訳の問題に直面すると、入力の手を止めて考えたり、調べ物をしたりと、訳出作業が一時中断する。つまり、一時中断(ポーズ)の割合が多ければその分だけ、翻訳者が難問に直面する割合が高くなり、認知的負荷も高くなると言われている。O'Brienは両方のPEのケースについて、ポーズの割合を調査したが、違いは全く見られなかった。実験参加者の実験後のコメントの中に、「PEは、普通に翻訳をするより疲れる」という感想が散見された。Kringsの調査でも指摘されていたことだが、PEは、原文と訳文を行き来する回数が増えるために、直線的な作業になりづらいらしい。つまり、前編集により機械翻訳の下訳の精度が上がったとして

も、「PE」という作業の性質上、原文と訳文と照らし合わせるための認知負荷は、さほど変わらないのかもしれない。

いずれにしても、目に見える結果として、前編集とPEを組み合わせれば、時間的な作業効率が向上することは、この実験で実証されたといえる。

O'Brien (2006b)

MT+PEの作業が、実際に翻訳者の認知負荷にどのくらい影響しているのかを、技術的なキーボード入力のポーズの割合だけからでは観察不可能であることが先の実験から判明した。そこで、O'Brien(2006b)では、人間の瞳孔の動きと開き具合を測定できるアイトラッキング装置を用いて、PE作業に要する作業者の認知負荷を測定した。実験は、もともと翻訳メモリにおけるファジーマッチ(Fuzzy Match)のマッチ率と瞳孔の開き具合との相関を調査する目的で行われたのだが、メモリ内にMTの訳文も混ぜて行ったのが、この研究のユニークな点であった。

結果は、大方の予想通り、翻訳メモリの70%～100%マッチ前後までは、マッチ率に従って瞳孔拡張は減少し続けた。またノーマッチ(No Match)で瞳孔拡張は最大になった。つまり、ゼロからの翻訳（ノーマッチ状態）では翻訳者の認知負荷が最も大きくなり、近似箇所を修正するだけの作業（ファジーマッチ状態）では、認知負荷も小さくなることが証明された。

この結果は想定内なのだが、特筆すべきは、機械翻訳のPEの作業における認知負荷が、予想以上に低かったという結果である。驚くことに、機械翻訳の修正作業（PE）で、瞳孔拡張は、85～90%ファジーマッチとほぼ同等だったのだ。翻訳メモリを使ったことのある方なら想像できるだろうが、85%マッチの場合は、大抵、1つか2つの単語を入替える程度の修正作業でしかない。非常に単純な作業なので、認知的負荷が低いのは頷ける。これが機械翻訳のPEでも同じだということだ。つまり機械翻訳の訳出精度を高いということの裏付にもなる。この実験で使用された言語ペアは、英語→仏語／独語であった。英語と日本語の組合せならば、まだこのレベルにはならないだろう。

Guerberof(2009)

O'Brien(2006b)の結果を受けてGuerberof(2009)は、統計的機械翻訳(SMT)を使った英語→西語での、PEの作業時間と品質に関する追試を行っている。彼女の実験も、翻訳メモリのファジーマッチとSMTの訳文をメモリ内に混在させて比較検証を行った。結果は、機械翻訳を修正する場合の方が、翻訳メモリのファジーマッチを修正するよりも、時間と品質ともに優位であった。

この理由として、翻訳メモリの修正の場合は、（人間の）翻訳者の訳文が近似文（下訳）として表示されるため、文章がこなれていて自然であるために、差分箇所を見つけ出すのに時間が掛かってしまうというものであった。また品質（この場合は、誤訳や訳漏れがないという基準を用いた）についても、翻訳メモリの文章がこなれているために、訳抜けがあったとしても、見逃してしまうことがあると指摘された。これに対して機械翻訳の訳文は、ぎこちない直訳が多いので、原文と訳文の一対一対応が比較的容易になり、品質的にも有利になるというものであった。

MT+PEと、ゼロからの翻訳（ノーマッチ）との品質の比較では、僅かながらゼロからの翻訳が優勢であったものの、所要時間とのバランスを考慮した総合的評価では、MT+PEに軍配が上がる。つまり、英語→西語での翻訳は（分野が制限されるという条件はつくものの）もはやゼロから翻訳するよりも、そして翻訳メモリを使うよりも、MT+PEが一番良いということが言えるのだ。

実は、Guerberofの研究の動機は、翻訳者へのワード単価をいくらに設定すべきか悩んでいたことに端を発する。というのも、すでに彼女が働く翻訳会社では機械翻訳を導入しており、この実験のようにPEの作業だけを翻訳者に発注していたからだ。もしも、このGuerberof研究結果とO'Brien(2006b)の結果が採用されることになれば、MT+PEの作業は、翻訳メモリ85～90%マッチと同じ単価、すなわち、通常のノーマッチの4分の1程の値段になってしまう。実際の効率はここまで向上はしていないので、この数字が額面通り使われることはないとしても、翻訳業界の横行する単価の値崩れは、この方面からも押し寄せていることを、改めて実感させられた研究結果である。

Garcia (2010)

これまでの実証研究は、英語とヨーロッパ言語の組合せであったが、Garcia(2010)は英語→中国語での検証を行った。時間と品質について、MT+PEとゼロから

の翻訳とを比較した。品質基準にはNAATIの試験基準を使用した。

結果は、MT+PEもゼロからの翻訳も、どちらも時間、品質ともにほとんど変わりがなかった。これはまだアジア言語での機械翻訳の精度が、ヨーロッパ言語との組合せよりは、劣っているということを暗示しているのかもしれないが、仮にそうだとしても、機械翻訳を使うことが決してマイナスに働くことのないレベルまでは近づいているとも解釈することができる。

この研究の特徴は、実験にGoogle翻訳者ツールキットという環境を使った点にあった。Garciaが指摘するように、翻訳支援ツールの歴史は翻訳メモリの単体使用から機械翻訳との融合というように変容してきた。Googleが用意する翻訳者ツールキットでは、機械翻訳に主眼をおき、翻訳メモリは副次的にしか機能しない。また、このようなツールを使うことが、翻訳者の翻訳への考え方にも影響を与えている。

実験参加者に対して行った調査では、「Google翻訳者ツールキットを使った翻訳のほうが、使わないよりも翻訳しやすい」という意見が、実験後には増えていた。興味深いのは、そういった意見と実際のデータとの関係である。ツールを好むと述べた翻訳者の訳出物の品質は、ツールを使わなかった時の品質よりも優れている場合が多かった。逆に品質が悪くなったケースもあることはある。また「ツールを使わないほうが良い」と回答した翻訳者は、己を良く理解してか、その翻訳品質は、ツールを使うと確かに悪くなっていた。

テクノロジーに対する向き不向きはあるにせよ、全体としては、機械翻訳の活用を前向きに受け入れる翻訳者の数が上回っており、機械翻訳に敵対心を抱いていないというのは、翻訳の未来を考えるうえで何かヒントを与えてくれそうな結果だと思う。

まとめ

以上、まばらではあるが、MT+PEに関する文献を見てきた。このように欧州言語ペア間では、分野や使用目的をすれば、ほぼ実用レベルに達していることが実証データからも分かった。特にO'Brien(2006b)で示されたようにPEの認知負荷がTMの85%マッチ相当というデータは非常に衝撃的である。また、Garcia(2010)が言うように、若い世代ではPEという作業を積極的に受け入れる傾向があるものも面白い。ALPACの悲劇以前のようにMTに対し盲目的に期待を寄せるのではなく、PEという作業を介して、MTを見ていくことが、より現実的かもしれない、ということだろうか。

ということで、次回は、英語→日本語でのPEの実験結果を書きます。

【註】

（注1）本稿は、2010年9月1日発行の『翻訳通信』（発行人：山岡洋一氏）に掲載した記事に加筆および修正をしたものである。

【参考文献】

Allen, J. (2003) Post-editing, In H. Somers (Ed.) *Computers and translation: A translator's guide* (pp. 297-317). Amsterdam/Philadelphia: John Benjamins,

ALPAC. (1966). *Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council*. Washington, D.C.: National Academy of Sciences, National Research Council. Retrieved from http://www.nap.edu/openbook.php?record_id=9547&page=R1

Bowker, L, and Ehgoetz, M. (2007). Exploring user acceptance of machine translation output: A recipient evaluation. In D. Kenny and K. Ryou (Eds.) *Across Bound- aries: International Perspectives on Translation* (pp. 209–224). Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Bowker, L. (2005). Productivity vs quality: A pilot study on the impact of translation memory systems. *Localisation Focus 4(1)*. 13-20.

Chesterman, A. and E. Wagner. (2002). Can Theory Help Translators? A Dialogue Between the Ivory Tower and the Wordface. Manchester: St. Jerome Publishing.

Fiederer, R. and O'Brien, S. (2009). Quality and machine translation: A realistic objective? *The journal of specialised translation*, 11. 52-74.

Garcia, I. (2010). Is machine translation ready yet? *Target, 22(1)*. 7-21.

Guerberof, A. (2009.) Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus, 7(1)*. 11–21.

Hutchins, J. and Somers, H. (1992). *An introduction to machine translation*. London: Academic Press Limited.

Krings, H. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes* (G. S. Koby, Ed). Ohio: Kent State University Press.

O'Brien, S. (2008). Processing fuzzy matches in translation memory tools: an eye-tracking analysis. In S. Gopferich, A. Jakobsen, & I. Mees (Eds.), *Looking at eyes: Eye-tracking studies of reading and translation process*. (pp. 79-102). Copenhagen studies in language: Copenhagen business school.

O'Brien, S. (2006a). Controlled language and post-editing. *MultiLingual*, October/November 17-19. Retrieved from https://216.18.156.115/multiligual/downloads/screenSupp83.pdf

O'Brien, S. (2006b). Eye-tracking and translation memory matches. *Perspectives: Studies in translatology*, 14 (3), 185-205.

Trujillo, A. (1999) Translation Engines: Techniques for Machine Translation, London: Springer.

Vermeer, H. J. (1989/2000). Skopos and commission in translational action. In L. Venuti (Ed.), Translation studies reader (pp. 227-38). London: Routledge.