クラウドソーシング後編集
―機械翻訳の実用的利用に向けた取り組み―

Crowd-sourcing post-editing

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: 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、独立行政法人情報通信研究機構を経て、現職。産業日本語研究会世話人代表。

1 はじめに

インターネットの普及により多言語での情報の受発信は、近年ますます重要になっている。機械翻訳システムの性能は徐々に向上しており、翻訳支援や情報獲得支援には利用可能な状況になっている。しかし、未だに精度は完璧ではなく、前処理や後編集などが必要とされている。

昨年度の JAPIO YEARBOOK において、「機械翻訳の実用的利用に向けた取り組み」として、情報発信型翻訳の精度向上の可能性として、「規格化日本語」「対訳表現の整備」「クラウドソーシング後編集」の３つの可能性を示した。本稿では、この中で特に「クラウドソーシング後編集」について、その後の進展について述べる。

2 クラウドソーシング後編集

情報検索によって得られた英語ページの意味を把握するために英日機械翻訳システムの出力をそのまま利用することが可能であるが、ビジネスにおける文書を日本語から英語に翻訳して、印刷物とするといった情報発信の場面においては、まだまだ後編集が必要となる。しかしながら、絶えず更新される情報をプロの翻訳者に依頼し、後編集するには膨大なコストが必要になる為、誰もが利用できるわけではない。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。そこで我々はボランティアベースによる後編集の効果を実証することにした。この手法をクラウドソーシング後編集と呼ぶ。

本稿では機械翻訳システムの出力を、母語話者ではあるが翻訳のプロではない人が複数人で後編集することを提案する。後編集においては、分野の知識を持つ母語話者の集合知によって、分野の知識のないプロの翻訳者と同等の理解可能な翻訳文を低コストで得ることが可能であるとの仮説の下、その実証を試みた。

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。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。場合によっては後編集の対象となる文章に対して、ある程度の知識を持っている人が参加することでプロが行う後編集より適切な文が得られると考えられる（図１）。

3 豊橋技術科学大学の英語ホームページでの実証

豊橋技術科学大学の英語版ホームページには機械翻訳システム（Microsoft Translator）が設置されており、約 40 か国語に翻訳することが可能となっている。これを用いて、クラウドソーシング後編集の実証実験を行った。英語版ホームページの現状を表１に示す。

本学の英語版ホームページの翻訳結果に対し、大学の留学生が自分の母語に訳された結果を後編集した。後編集を行った言語は、アラビア語・インドネシア語・スペイン語・ドイツ語・フランス語・ベトナム語・ポルトガ

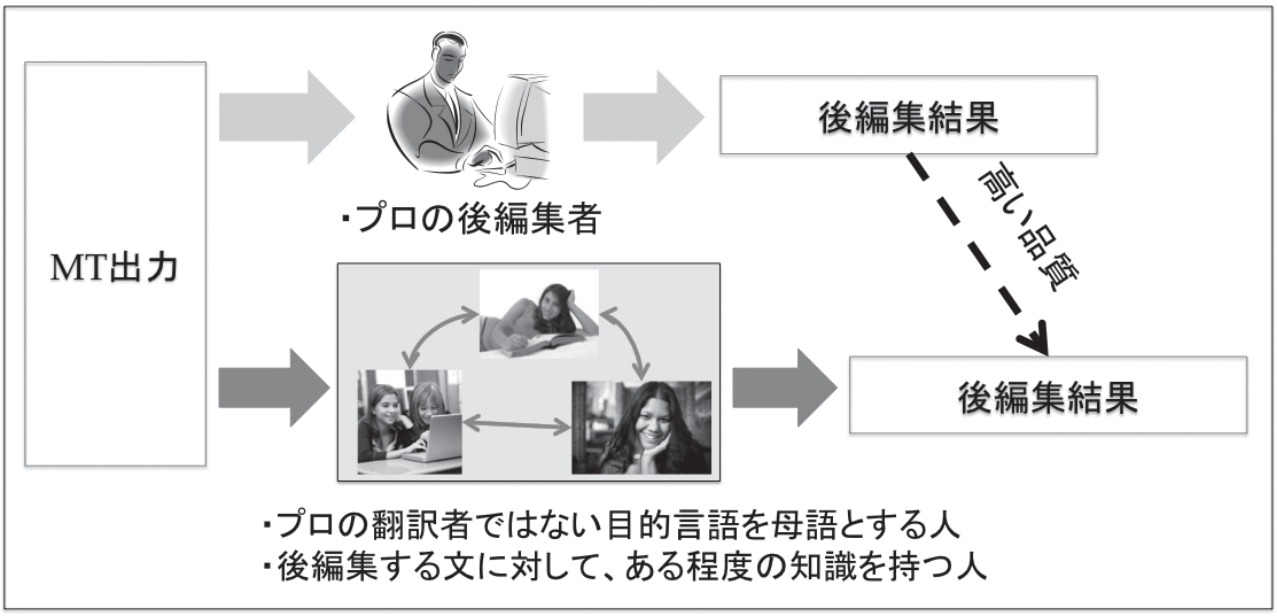

図1　プロによる後編集と集合知（クラウドソーシング）後編集

表1　ホームページのサイズと機械翻訳システム

対象となるページ数（ページ）	64
対象となる文数（文）	約 2,500
翻訳エンジン	Microsoft Translator

ル語・中国語・韓国語の 9 か国語である。また、詳細な統制実験として、日本人学生 4 名による日本語の後編集実験も行った。後編集終了後に、プロの翻訳者によって、どの後編集が適切かを判定した。適切な後編集結果が無い場合は、プロの翻訳者が後編集を行った。これらの処理ののちは、対象とした言語のホームページは誤りのない的確な翻訳となる。

本実験に参加した留学生の内訳と得られたデータ数を表 2 に示す。9 か国語の 22 人が参加した。留学生は母語に翻訳された本学のホームページを参照し、Microsoft Translator に付属した CTF（Collaborative Translation Framework）の機能を用いて、後編集を行う。その時、英語原文、機械翻訳システムの出力、（もしあれば）現在までに行われた後編集結果を参照し、より適切な翻訳が可能と考えた場合に、それを入力する。

各留学生には 30 時間を目処に作業をするように指示をした。また、後編集に際しては、意味が正しく伝わるレベルにまで編集するように指示し、文体の良し悪しについては考慮しなく良い旨、指示を行った。後編集作業は約 2 か月で行われた。

表2　留学生の内訳と得られたデータ数

目的言語	人数	後編集の対象となった文	総後編集数
アラビア語	2	397	723
インドネシア語	2	1,285	1,559
スペイン語	4	1,841	3,643
ドイツ語	1	147	192
フランス語	2	512	647
ベトナム語	2	1,341	1,929
ポルトガル語	1	204	308
中国語	6	1,637	2,269
韓国語	2	598	707

4 人手評価

総後編集数の上位 4 カ国語（中国語、インドネシア語、スペイン語、ベトナム語）を対象に人手評価を行った。

評価はプロの翻訳者に入力文に対しての機械翻訳結果と後編集結果をランダムな並びで提示し、最も良いものを選択する形式とした。もし候補の中に十分な翻訳が存在しなかった場合はプロの翻訳者（後編集者）が自ら修正を行った。

表３　後編集量と選択結果

	中国語	インドネシア語	スペイン語	ベトナム語
センテンス数（文）	1,638	1,287	1,849	1,344
有効センテンス数（文）	1,627	1,286	1,848	1,343
LCPE が選択された数（文）	1,088	875	894	783
LCPE 以外の CPE が選択された数（文）	171	37	329	116
MT が選択された数（文）	250	127	291	134
プロの翻訳者によって修正された数（文）	118	247	334	310
LCPE が選択された（%）	67%	68%	48%	58%
LCPE 以外の CPE が選択された（%）	11%	3%	18%	9%
MT が選択された（%）	15%	10%	16%	10%
プロの翻訳者によって修正された（%）	7%	19%	18%	23%

入力文には専門用語や固有名詞が出現するものがあり、プロの翻訳者が正否を評価できないものも有った。評価できなかったものを除いた結果を有効センテンスとした。この結果を表３に示す。ここで CPE は Crowd(-sourcing) Post Edit の略であり、LCPE は Last Crowd Post Edit、すなわち最後に行われた修正結果を示す。各言語とも最後に行われた修正（LCPE）が約半数以上選ばれた。さらにプロの翻訳者が訂正したのは各言語とも約２割以下であった。このことからプロが介入しなくてもある程度の品質の文が得られたと考えられる。

たとえば、中国語についてみれば、クラウドソーシング後編集を行わない場合、プロ翻訳者はＭＴ出力を正しい（そのまま後編集せずに利用可能）とした 15%以外の文（つまり 85%の文）を修正しなくてはならないが、クラウドソーシングの結果、（プロの翻訳者によって LCPE が適切であると判断された）67% が修正不要となり、プロの翻訳者による後編集は残りの 18%で済む。（もし、LCPE 以外の後編集結果も参照して適切な訳文を選択するならば、新たに編集入力を行う修正は 7%で済む。）

クラウドソーシング後編集を経て、なお後編集が必要な部分について、機械翻訳の出力を後編集する場合と、LCPE を後編集する場合の労力を比較するために、機械翻訳出力と LCPE のそれぞれと、プロの翻訳者が入力した訳文との間の編集距離を計算した。この結果を表４に示す。例えば中国語においては、MT 出力を修正する場合の TER が 67.822 であるのに対し、後編集結果を修正する場合の TER が 35.829 と約半分となり、18%残った修正必要部分においても、修正の作業量はるかに少なくなる。

表４　TER（Translation Err Rate）による評価

言語	MT 出力とプロによる修正文間	LCPE とプロによる修正文間
インドネシア語	38.160	20.238
スペイン語	35.634	25.046
ベトナム語	47.287	18.620
中国語	67.822	35.829

以上をまとめると、ある実例において機械翻訳システムの出力はその 85%の文が後編集が必要であったが、クラウドソーシング後編集によって、後編集が必要な出力は 18%に減り、その 18%においても修正量は約半分になる。非常に乱暴に言えば、85 の労力が９（18%の半分）になることになり、プロの翻訳者による後編集のコストを大幅に減少できる可能性がある。

5　まとめ

コストを抑え、質の高い後編集結果を得ることを目的にボランティアベースのクラウドソーシング後編集を提案した。クラウドソーシング後編集ではプロの翻訳者の介入を従来の後編集の半分以下に抑えるという結果が得られた。

クラウドソーシング後編集を行うことで、約半数以上の文がプロの翻訳者の後編集結果と同等の品質であることを実証できた。このことからプロの介入を従来の半分以下に抑えられると言える。さらに後編集の対象となる文章に対し、ある程度の知識を持つ人が参加することで専門用語を正しく選択でき、プロの翻訳者より質の高い後編集が行える。以上のことからクラウドソーシング後編集はコストを抑え、品質の後編集結果を得るために有効な手法だと言える。

参考文献

[1] Midori Tatsumi, Takako Aikawa, Kentaro Yamamoto and Hitoshi Isahara (2012), "How Good Is Crowd Post-Editing? Its Potential and Limitations", AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012), pp. 69-77.

[2] Takako Aikawa, Kentaro Yamamoto and Hitoshi Isahara (2012), "The Impact of Crowdsourcing Post-editing with the Collaborative Translation Framework", JapTAL 2012, LNAI 7614, pp. 1-10, Springer-Verlag Berlin Heidelberg.

[3] Hitoshi Isahara (2012), "Toward Practical Use of Machine Translation", JapTAL 2012, LNAI 7614, pp. 23-27, Springer-Verlag Berlin Heidelberg.