機械翻訳出力の後編集の集合知による省力化

†山本 健太郎　　‡相川 孝子　　†井佐原 均
†豊橋技術科学大学
‡マイクロソフトリサーチ

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。そのため、どこでどんな間違いが起こるかわからない機械翻訳の導入は、危険が高すぎると懸念する組織も多いと思われる。

近年では完璧ではない機械翻訳を活用するために後編集の研究が行われている。しかし、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。

本研究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。この集合知により、プロの翻訳者の後編集によって得られるような理解可能な翻訳文を低コストで得ることが可能である事を示すことが本研究の目的である。

2　後編集の集合知による省力化

絶えず更新される情報をプロの翻訳者に依頼して、後編集するには膨大なコストが必要となる。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を提案した。

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。最終的に得られる後編集結果はプロの翻訳家が翻訳した (あるいは後編集した) 文に近い品質となると考えられる。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。

後編集を行う人数は多ければ良いという訳ではなく、ある人数以降は後編集結果に改善が見られなくなると思われる。無駄を省くために人数の見極めが重要になってくる。

3　Microsoft Translator CTF (Collaborative Translation Framework)

Microsoft は Microsoft Translator を開発する過程で、翻訳システムの精度向上、対応言語の拡大等に注力すると同時に、「どうやったら人間と機械が共同して翻訳の質を高め、情報の多言語化につとめる事ができるのか」 という観点からの検討を進め、共同翻訳フレームワーク (CTF) 機能を開発した。

CTF 機能には 「編集機能」 がある。これは文字通り機械翻訳結果を人間が確認し、訂正・編集を加えることができるという機能で、編集された結果は、Microsoft Translator のデータベースに送られ、以降の同一ページの翻訳に利用されると共に、翻訳精度の向上に利用される。この CTF は Web 上で Widget として走らせ、その上にユーザーからのフィードバックを受け入れるユーザーインターフェースを付加したものである (図 1)。

図 1: CTF 機能による編集結果のリスト

CTF 機能による編集は全ユーザーが行えることから、悪意のある編集が行われる可能性がある。これを防ぐために「権威ユーザー指定機能」が実装されている。この機能は Web マスターが特定のユーザーを選び、この選ばれた権威ユーザーによって編集された文を、「信頼できる翻訳」として、このサイトで優先して使う事ができるという機能である。

4 検証実験

豊橋技術科学大学では、多言語での情報発信を実現するため、英語版ホームページに Microsoft Translator を設置している (図 2)。今回、集合知による後編集の省力化の有効性を検証する為に、本学の留学生に対し、母語とする言語の翻訳結果を後編集するように依頼した。

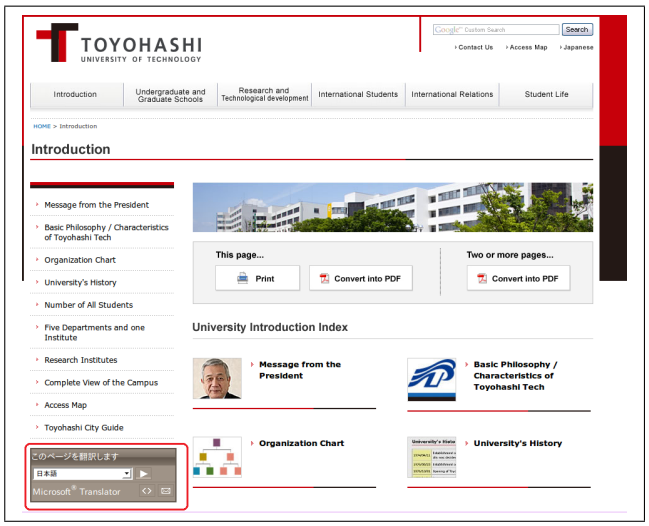

図 2: 豊橋技術科学大学の英語版ホームページ

4.1 実験内容

母語が同じ留学生を一つのグループとし、英語版ホームページを母語に翻訳した結果に対し後編集を行う。現在、本学の英語版ホームページ上の約 2,500 文が Microsoft Translator で翻訳可能（したがって、後編集可能）である。表 1 に後編集の結果を示す。

表 1 で「後編集を行った文」とは、編集可能な約 2,500 文のうちで各グループが実際に後編集を行った文の数である[1]。「後編集結果」とは各グループが「後編集を行った文」に対して行った修正の総数である[2]。

[1] この数と 2500 文との差は、後編集が必要でないか、時間が無くて後編集ができなかったかのいずれかである。本研究では、いずれであるかは把握していない。

[2] 2 人目以降の後修正者は、自分より前の後編集者が見落とした個所に新たに修正を加えるか、自分より前の後編集者の修正（あるいは不修正）に満足して何もしないか、自分より前の後編集者の修正を再度修正するか、のいずれかである。

表 1: 留学生の内訳と後編集結果

目的言語	人数 (人)	後編集を行った文 (文)	後編集結果 (文)
アラビア語	2	397	723
インドネシア語	2	1,285	1,559
ポルトガル語	1	204	308
スペイン語	4	1,841	3,643
中国語	6	1,637	2,269
ベトナム語	2	1,341	1,929
フランス語	2	512	647
ドイツ語	1	147	192
韓国語	2	598	707

4.2 検証方法

ボランティアの集合知による後編集が有効であるかどうかを示すために、各グループが後編集した結果の品質を人手評価と自動評価とで評価する。人手評価では、筆者らは、表 1 に示した言語を理解しないため、機械翻訳の結果と後編集結果の比較や、後編集結果の品質評価を各言語の母語話者に依頼する。具体的には、後編集結果を参照し、適切な後編集結果がある場合には、それを指定する。既存の後編集結果では満足できなかった場合には、さらに後編集を行う。この場合は、既存の後編集結果の何処が何故、問題であったかを事後に確認する。また、自動評価では TER を使って評価を行う予定である。

5 おわりに

機械翻訳は、検索した英語文書を日本語化して読むといった、情報受信のためには現状でも一定の有効性を持っている。一方、印刷文書やホームページといった情報発信型においては、そのままでは不十分な場合が多い。しかし、多くのボランティアによって、低コストで理解可能な後編集結果を得られることで、機械翻訳の利用は益々活発になるだろう。

ボランティアの可能性については、ウィキペディアに示されるような知的満足からくる物もあろうし、留学生が母国の後輩のために、また PTA が子供達のために、と様々な可能性が考えられるよう。

我々は引き続き後編集の集合知による省力化について研究を行う。

6 参考文献

“ホームページの多言語化に向けた機械翻訳とコミュニティによる後編集の活用”、相川孝子、井佐原均、言語処理学会第１７回年次大会発表論文集.pp.615-618