機械翻訳を使った実務翻訳への挑戦：
日英翻訳におけるプリエディットとポストエディットの研究

立見みどり（Dublin City University博士課後期程修了）
山田優（立教大学博士課後期程修了）

2012年9月9日
日本通訳翻訳学会第13回年次大会
＠京都橘大学

概要

- 機械翻訳とは
- 機械翻訳の実務への適用可能性
- プリエディットとポストエディット
- ポストエディット
  - 英日
  - 日英

機械翻訳とは

◆ RBMT（ルールベース機械翻訳）

◆ SMT（統計的機械翻訳）

- 無料（オンライン）
  - Google Translate
  - bing Translator
  - エキサイト翻訳
  - Yahoo!翻訳
  - Moses
- 有料（スタンドアロン）
  - Systran
  - ATLAS
  - The 翻訳
  - PC-Transer
  - LogoVista

機械翻訳の実務への適用可能性

機械翻訳結果の品質は一般に良いと思われていないが、ジャンルによって期待できる品質は異なる

【品質】

- Google 翻訳結果（英日）例：

  We are all in the gutter, but some of us are looking at the stars.
  我々は、樋内のすべてのですが、私たちの一部は、星を見ている。

  To save the file with a new name, click File menu and select Save As....
  新しい名前でファイルを保存するには、[ファイル]メニューをクリックし、[名前を付けて保存]を選択....

【適性】

- 定型文の多い文書や技術的な内容の文書
- 繰り返しの多い文書
  - 文学作品や口語体の文章と比較して、機械翻訳による品質がある程度期待できる
  - また、用語を統一する必要のある文書などでは、機械翻訳を使用することによって、人手によるチェックを省けるなど、人手翻訳を上回るメリットもある

プリエディットとポストエディット

プリエディット
機械翻訳にかける前に原文を編集する

Google 翻訳結果（英日）例：

- プリエディット無し

The status will be changed to Done and the ticket will be expired.

ステータスが完了するとチケットが期限切れになるか変更されます。

- プリエディット有り

The status will be changed to "Done", and the ticket will be expired.

ステータスが“完了”に変更され、チケットの有効期限が切れます。

※ 引用符とコンマを追加することで、正しい翻訳が実現する

ポストエディット
機械翻訳の結果を後編集する

Google 翻訳結果（英日）例：

（原文）

Search for items in the folder.

（機械翻訳）

フォルダ内**の**アイテムを検索します。

↓ ポストエディット

フォルダ内でアイテムを検索します。

※「in」の訳によって意味が若干変わってしまっているところが修正されている

英→日機械翻訳におけるポストエディット研究 (Tatsumi, 2010)

■ 背景

- 米国グローバル企業（ソフトウェア開発会社）
- ユーザーマニュアルを英語から多国語化 → 機械翻訳の採用
- 品質を維持するため、ポストエディティングを実施

■ 目的

- ポストエディティング速度を遅くしている要因を調べる

■ 研究方法

- 企業の通常のワークフローを実験で再現
- プロの翻訳者にポストエディティングを依頼
- 作業の様子をスクリーン動画キャプチャソフト（BB Flashback）で記録
- センテンスごとの作業時間を記録（1分あたりの編集語数で速度を計算）
- 編集前後のテキストを比較
- 定量的・定性的に分析

英→日機械翻訳におけるポストエディット研究 (Tatsumi, 2010)

- 修正量とポストエディティング速度の相関性は0.6程度（相関係数）

- 原文の特徴がポストエディティング速度に与える影響
  - センテンスの構造
    - 単文 → 複文・重文 → 不完全文 の順に速度が遅くなる
    - （編集量は単文 → 不完全文 → 複文・重文 の順に多くなる）
    - センテンスの長さよりも影響が大きい
    - ➢ なるべく単文を使用することによってポストエディティングが高速化？
  - センテンスの機能
    - 手順説明文は、セクションタイトル、箇条書き項目、機能説明文より速い
    - ➢ 選択的に機械翻訳を使用することで利用効果が向上？
  - 製品画面用語の有無
    - 修正したかどうかに関係なく、画面用語が存在する文は遅くなる
    - ➢ 画面用語の翻訳を完全自動化することでポストエディティング労力を軽減？

- 作業の観察からの定性分析
  - ポーズ分析：ポストエディティング を極端に低速化させていたのは、修正量の多さよりも、熟考や調査のため作業が停止している「ポース」時間の長さ
  - 再編集：特に画面用語を修正するために、同じセンテンスに何度も戻って編集を重ねるケース

日→英機械翻訳におけるポストエディット研究 (Yamada, forthcoming)

■ 背景

- グローバル化を背景に、日英翻訳需要は依然として高い
- 英語ノンネイティブ翻訳者によるポストエディットの必要性
- MT+PEの導入は効果的か？
- 英日のポストエディットとの違い

■ 実践データの分析

- 大規模プロジェクトに機械翻訳（Google翻訳）の訳文を参照訳として使用
- 共有型翻訳メモリ（memsource）でデータ収集

日→英機械翻訳におけるポストエディット研究 (Yamada, forthcoming)

事例検証：実務プロジェクトでの機械翻訳参照割合

日英（ITマニュアル）
TM有り、翻訳者4名　原文文字数：90,507

	割合（%） MT	セグメント文字数	50%以上平均
	19		
	–		
100%	2	8	10
95%–99%	2	18	
85%–94%	1	25	
75%–84%	1	18	
50%–74%	5	33	
0%–49%	9	38	

日英（IT系の仕様書）
TM無し、翻訳者8名　原文文字数：181,285

	割合（%） MT	セグメント文字数	50%以上平均
	37		
	–		
100%	14	16	30
95%–99%	8	31	
85%–94%	1	29	
75%–84%	2	25	
50%–74%	5	28	
0%–49%	6	40	

日→英機械翻訳におけるポストエディット研究 (Yamada, forthcoming)

事例検証：例文

原文	最終訳	Google 翻訳	GTM	TM	原文文字数
製品名の確認における注意点	Notes on confirmation of the product name	Notes on confirmation of the product name	1.000	100%	13
製品名取得用API	API for retrieving product name	API for retrieving product name	1.000	100%	9
最大保持期間	Maximum retention period	Maximum retention period	1.000	100%	6
[リセット]ボタン押下	Press the [Reset] button	Press the Reset button	0.750	85%	11
プリンターグループファイルの取得	Retrieves printer group file	Getting the printer group file	0.667	75%	16
生成されるハッシュコード	Hash code to be generated	Hash code that is generated	0.600	70%	12
同じ製品名のグループが登録されています。	A group with the same product name is already registered.	Group of the same product name has been registered.	0.526	55%	20
以下のデータ構造にて、同一ユーザー／グループに属する格納先を追加します。	Add a storage location belonging to the same user/group in the following data structure.	At the following data structure, add a storage location belonging to the user / group the same.	0.581	60%	36
印刷指示キューの先頭の印刷データから印刷指示要求を送信する。	Send print instruction request from print data at the top of print instruction queue.	Print instruction to send a request from the head of the queue print data print instruction.	0.533	50%	30
同時接続数オーバー時は、24[セッション]目以降は、接続自体が確立されていないため、{j}クライアントPC側で以下の図に示すようなタイムアウトエラーが発生し、再接続処理が実施されます。	In case the number of possible concurrent connections is exceeded, {j}the timeout error as indicated in the figure below occurs on the client PC side and reconnecting process is performed as connection itself is not established for the 24th and subsequent [sessions].	When over the number of concurrent connections, after the first 24 [session], because the connection itself has not been established, time-out error, as shown in the figure below on the {j} client PC has occurred and is performed reconnection process You.	0.506	50%	92

日→英機械翻訳におけるポストエディット研究 (Yamada, forthcoming)

暫定考察

- MT訳が参照される割合は、全体の19～37%
- 文字数が少ない文（単文だけでなく、不完全文、タイトル等）では、修正量は少ない
- サブセグメントのレベルでは、（S）MT訳を参照にしている可能性がある

今後の分析

- 修正量、時間、認知負荷の関係性
- プロセス分析
- サブセグメント化した原文での実験検証

参考文献

- Hartley, T., Tatsumi, M., Isahara, H., Kageura, K., & Miyata, R. (2012). Readability and translatability judgments for 'Controlled Japanese'. EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy, May 28-30 2012, ed. Mauro Cettolo, Marcello Federico, Lucia Specia, Andy Way; pp.237-244 [PDF, 465KB]
  Retrieved Aug 31, 2012 from http://www.mt-archive.info/EAMT-2012-Hartley.pdf

- O'Brien, S. (2006). Controlled language and post-editing. MultiLingual, October/November, 17-19. Retrieved June 10, 2010 from https://216.18.156.115/multiligual/downloads/ screenSupp83.pdf

- O'Brien, S. & Roturier, J. (2007) . How portable are controlled language rules? A comparison of two empirical MT studies. MT Summit XI, 10-14 September 2007, Copenhagen, Denmark. Proceedings; pp.345-352
  Retrieved Aug 31, 2012 from From http://www.mt-archive.info/MTS-2007-OBrien.pdf

- Tatsumi, Midori. (2010). Post-Editing Machine Translated Text in a Commercial Setting: Observation and Statistical Analysis. Unpublished doctoral thesis. Dublin City University: Dublin

- Yamada, Masaru (2011). Revising text:An empirical investigation of revision and the effects of integrating a TM and MT system into the translation process. Unpublished doctoral thesis. Rikkyo University: Tokyo.