翻訳時に参照すべき情報が欠けることで生じる問題: ニュース記事の英日機械翻訳・ポストエディットを例題に

藤田 篤
情報通信研究機構

1 はじめに

翻訳は一般に文書およびその内容を扱う．ある文書を翻訳する際は，その文書に含まれるテキストだけでなく，翻訳の目的や用途 (スコポス)，文書が所属する言語使用域 (レジスタ) や目標言語における規範，文書の構造，テキスト要素とそれ以外の構成要素の関係などを考慮する必要がある．

一方機械翻訳 (MT) は，文書の外側の情報を参照せず，文書中のテキストのみを対象とし，それを文単位で目標言語のテキストに変換する，という問題の単純化のもとで発展してきた[1]．性能の改善に向けて誤り分析が行われることがあるが，MT の出力そのものの分析や別途 (同様に文単位で) 作成された参照訳との比較だけでは，真の翻訳との差異や，人間の訳者の好みと MT の出力における表現の偏りの間の乖離と真の誤りを区別できない．さらに，翻訳の際に参照すべき情報源や遵守すべき規範との関係について論じることも困難である．

MT の実用化への期待が増す一方で翻訳過誤による問題も生じている．翻訳の道具として MT を活用するには，翻訳時に参照すべき情報や遵守すべき規範が参照されないことで生じる問題，ひいては社会的リスクについて整理・共有する必要がある．本稿では，それに向けた一つの例として，特に，文単位で解決すべき課題と文書単位ではじめて解決できるようになる課題を切り分けた MT 訳の誤りの分析手法と，ニュース記事の英日翻訳を題材とした実施例について述べる．

2 分析対象データの作成

誤り分析の際，人間の選好と MT の偏りの差異と，翻訳としての真の誤りを区別する必要がある．そこで，MT 訳を起点とし，人間による必要最低限のポストエディット (PE) によって最終訳を得ることにした．より具体的には，次の工程で分析用のデータを作成した．

(1) 文単位の MT 訳の生成
(2) 文単位の PE
(3) 文書単位の PE

[1]近年，対訳データや単言語データからの事前学習，テキスト以外のモーダル情報や前方文脈の参照などの検討はなされている．

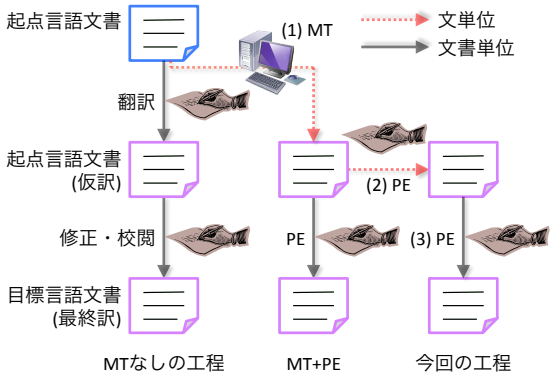

図 1: 翻訳工程の対比．図中の人間の作業者は実際には対象文書以外の情報も参照することに注意されたい．

既存の翻訳工程と今回のものを図 1 に示す．今回の工程で作成したデータは次の 2 種類の分析を可能にする．

- 作業工程 (2) の文単位の PE の結果から，文を処理単位とする MT の課題を知る．
- 文単位では受理可能な訳文に対して実施された作業工程 (3) の文書単位の PE の結果から，文書を扱う実際の翻訳としての課題を知る．

2.1 分析対象文書

著者自身が分析を行えるよう，英語を起点言語，日本語を目標言語とした．また，必要以上の PE を避けるために，MT 訳とは独立に作成された参照訳を利用する．これらの条件を満たす題材として，今回は，Asian Language Treebank プロジェクト [5][2] で作成されたニュース記事の翻訳 (以下，ALT) を用いた．英語原文，日本語訳を内製のトーカナイザを用いて分かち書きした後の ALT の諸元を表 1 に示す．

2.2 文単位の MT 訳の生成 (図 1 の (1))

既存の MT+PE の翻訳工程 (図 1 の中央の工程) と同様に文単位の MT を利用するが，起点の MT 訳としてはできる限り品質が高いものを生成しておきたい．そこで，ALT の訓練・開発データに加えて，NICT が有する大規模な英日対訳データ[3](以下，TexTra データ) も活用して，次の手順で英日 NMT モデルを学習した．

[2]http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/
[3]NICT が運営する翻訳サービス『みんなの自動翻訳@TexTra』(http://textra.nict.go.jp/) の『汎用 NMT』モデルの訓練に用いられている対訳データ．ただし，ALT の対訳をすべて除外して使用した．

表 1: ALT における英日対訳データの記述統計.

用途	文書数	文数	トークン数	
			英語	日本語
訓練データ	1,698	18,088	2,572k	3,743k
開発データ	98	1,000	139k	202k
評価データ	97	1,018	143k	208k

フェーズ 1. 事前学習: まず，TexTra データのみを用いて NMT モデルを学習した．起点言語および目標言語の語彙もこのデータから決定した．ALT と同様に分かち書きした後，バイト対符号化 [6] によって各言語の 32,000 語のサブワードを定めた．

フェーズ 2. 混合パラメタ調整: 次に，文献 [1] の手法を援用し，ALT の訓練データを k 倍したものと，TexTra データから無作為に抽出した同数の文対を混合したものを用いて，さらに学習を行った．

フェーズ 3. 非混合パラメタ調整: 最後に，ALT の訓練データのみを用いて，さらにパラメタを調整した．

NMT モデルの学習には Marian NMT[4] を用いた．モデルの構造および学習に関するハイパーパラメタは，Transformer base モデル [9] のものを採用した．各フェーズの学習は，一定ステップ t ごとに ALT の開発データを用いてモデルのパープレキシティを評価し，評価値が 5 回連続で改善しない場合に停止させた．t の値は，フェーズ 1 では 5,000，フェーズ 2 と 3 では 10 とした．フェーズ 2 における TexTra データのサンプル規模 k の値としては {1,2,4,8,16,32,64} を試し，フェーズ 3 を終えた時点で開発データに対する BLEU スコア [3] が最も高くなった 32 を選択した．

得られた英日 NMT モデルを用いて，ALT の評価データ中の各文をデコードした．その際，ビーム幅は 10，長さに関する正規化重みは ALT の開発データに対する BLEU スコアが最大となる値を用いた．

2.3 文単位のポストエディット (図 1 の (2))

次に，MT 訳を最大限に踏襲した文単位の修正訳を作成した．すべての文書のすべての文を無作為に並び替えて前後の文脈を参照できなくした上で，MT 訳に対し，原文の内容を過不足なく伝える文法的な訳文にするために必要最低限の PE を施した．その際，PE の分量を必要最低限に抑制するために次の制約を課した．

$$dist(\text{MT 訳}, \text{PE 訳}) \leq dist(\text{MT 訳}, \text{参照訳})$$

ここで $dist(\cdot,\cdot)$ は，2 つの文の表層的な距離を表す関数である．今回は HTER [7] を使用した．ただし，修正の分量について筆者と作業者の間で齟齬が生じることを避けるため，分かち書きには，内製のトーカナイザではなく，OSS である MeCab[5] を使用した．なお，ALT の参照訳は，PE の作業者には開示していない．

この工程では，970 文 (95%) に対して何らかの修正が行われた．tercom[6] を用いて同定した元の MT 訳における修正対象は 9,244 形態素，HTER は 26.8 であった．

2.4 文書単位のポストエディット (図 1 の (3))

上述の作業を経て文単位では受理可能となった訳文を，文書全体を参照しながら，さらに修正した．ここでは，元々の MT 訳の代わりに文単位の PE の結果を MT 訳とみなすことにより，文書単位の翻訳における文単位の MT 訳の PE，という LSP において近年一般的に行われている工程 (図 1 の中央の工程) と等価な設定を実現した．作業者には，文単位の PE の際の要件に加えて，結束性が高く日本語のニュース記事として適切な文章を作成するよう指示した．

この工程では，86 文書 (89%) 中の 320 文 (31%) に対して何らかの修正が行われた．この工程における修正対象は 1,168 形態素で，PE の結果を参照訳としたときの元の MT 訳および文単位の PE の結果の HTER はそれぞれ 28.7，3.4 であった．

3 誤り分析

3.1 分析の焦点と手順

2 節で述べた 2 種類の分析のうち，今回は後者を実施した．すなわち，文書を扱う実際の翻訳としての課題を明らかにするために，文単位の PE の結果と文書単位の PE の結果を比較し，作業工程 (3) で行われた修正内容を次の手順で分析した．

1. 事例の列挙: 各文対を比較し，修正前の表現と修正後の表現の対を抽出した．その際，文献 [2] にならい，互いに依存していないと考えられる修正事例を複数の事例に分解した．この結果，530 件の修正事例を得た．

2. 事例の分類: 個々の修正事例を，(人間による) 英日翻訳の校閲のために作成された校閲カテゴリ体系およびカテゴリ分類のための決定木 [8] に従って分類した．

3. 文単位での修正可否の判定: 個々の修正事例について，文書中の他の文を参照せずに修正できた内容か否かを判定した．これは，作業工程 (2) の文単位の PE において目標とした必要最低限の訳と文単位での最適訳との差を明らかにするものである．

[4] https://github.com/marian-nmt/marian

[5] https://taku910.github.io/mecab/

[6] https://github.com/jhclark/tercom

表 2: 各校閲カテゴリの事例数
「可」: 文書中の他の文を参照せずに修正できた事例
「不可」: 他の文を参照してはじめて修正が可能になる事例

カテゴリ名	事例数 可	不可
Lv 1 **未完成**		
【X4a 未翻訳】	16	0
【X6 曖昧さ未解消】	0	2
Lv 2 **起点言語テキストの要素に対する過不足や誤解**		
【X7 用語の訳出誤り】	48	30
【X1 原文内容の欠落】	16	0
【X2 原文にない要素の付加】	3	0
【X3 原文内容の歪曲】	57	41
Lv 3 **目標言語の文法的・統語的な問題**		
【X8 コロケーションの誤り】	5	0
【X10 前置詞や助詞の誤り】	2	0
【X11 活用の誤りや数・性などの不一致】	0	0
【X12 綴り誤り・誤変換】	0	0
【X13 句読法に関する誤り】	7	0
【X9 その他の文法的・統語的な誤り】	1	0
Lv 4 **目標言語テキストの質的な問題**		
【X16 結束性違反】	34	75
【X4b 直訳調】	57	0
【X15 表現のぎこちなさ】	37	3
Lv 5 **納品・公表に際しての問題**		
【X14 レジスタ違反】	83	13
合計	366	164

3.2 修正事例の分類結果と考察

修正事例の分類結果を表 2 に示す．文単位で可能だった修正事例は合計 366 件 (69%) あった．このうち 85 件は，作業工程 (2) の目標である「原文の内容を過不足なく伝える文法的な訳文」が未達成であった．目標言語の文法的・統語的な問題 (Lv 3) はいずれもこれに該当する．その他の事例について，PE の際に必要だったと思われる情報と事例を示す．

対象分野の専門知識: 今回扱ったニュース記事の中には，政治，災害，宗教，スポーツなどの多様な内容が含まれていた．PE の際に分野の内容や表現に関する知識が欠けていることで問題が生じていた．

(1) 原文: Clemens (3-0, 1.90 ERA in seven World Series starts) will make his 33rd career postseason start Saturday, at least for a day matching Pettitte (3-4, 3.90 in 10 World Series starts) for the most ever.
文単位の PE: クレメンス（ワールドシリーズ７回出場で３対０、防御率１．９０）は、少なくとも１日[ ]ペティット（ワールドシリーズ１０回出場で３対４、３．９０）と[ ]組んで、土曜日に[ ]３３回目のポストシーズンのスタートを切る。
文書単位の PE: クレメンス（ワールドシリーズ７回出場で３勝０敗(X3)、防御率１．９０）は、少なくとも１日は(X15)ペティット（ワールドシリーズ１０回出場で３勝４敗(X3)、３．９０）と史上最多で(X1)並び(X3)、土曜日に生涯で(X1)ポストシーズン 33 回目(X3)の先発登板を行う(X3)。

残りの 281 件は，今回の作業工程 (2) において目標とした「原文の内容を過不足なく伝える文法的な訳文」と「文単位の翻訳の範囲で十分に高品質な訳文」の品質の差異を示すものと言える．例えば，【X4a 未翻訳】の事例はすべて，アルファベット表記のまま残っている固有名をカタカナで訳出したものであった[7]．作業工程 (2) では必要最低限の修正のみを指示したため，MT 訳中の内容が伝わる表現はそのまま残されていた．【X4b 直訳調】の事例や【X7 用語の訳出誤り】の一部の事例も同様である．これらから，翻訳/PE の際には次の 2 種類の情報が不可欠であると言える．

詳細な翻訳の仕様: アルファベットの表記の固有名をカタカナ訳するか否か (X6)，使用可能な漢字に関する指針，作品の題名などに対す括弧の使い方，社名の表し方，数字につく「か」(いずれも X14)，など．スタイルと呼ばれるものも含む．

用語集: 用語の認定基準や訳出の一貫性を保証するには用語集が不可欠である．ただし，姓のみから個人を特定できない場合や地名と人名の曖昧性など，正しい訳出が困難な場合もある．

文書内の他の文を参照してはじめて修正が可能になる事例 (164 件，全体の 31%) は，16 カテゴリ中 6 カテゴリのみで観察された．全体で最も多かった【X16 結束性違反】の事例 (109 件，全体の 21%) が，文書内の他の文を参照する必要がある修正事例の中でも最も多かった (75 件，46%)．下位分類の一部を示す．

明示化 (explicitation, 要文脈 40 件/文内 12 件): 指示詞や人称代名詞,「後者」などの相対表現の参照先の名詞句などの明示．

省略 (要文脈 13 件/文内 13 件): 動詞の格要素となっている指示詞や人称代名詞の中で自明なものの削除．

主題化/非主題化 (要文脈 9 件/文内 3 件): 前後の文脈に応じた「が」と「は」の適切な使い分け．

訳出の統一 (要文脈 8 件): 同一文書中の同じ原文表現の訳出を統一する修正．

明示化の中には，原文の文内には現れていない情報を補う次のような事例もあった．

(2) 原文: Rother says that price indexing has to be linked to Bush's private investment accounts.
文単位の PE: ローザーは、物価指数をブッシュ[ ]の個人投資口座と関連させる必要があると述べている。
文書単位の PE: ローザーは、物価指数をブッシュ提案(X16)の個人投資口座と関連させる必要があると述べている。

[7] 【X7 用語の訳出の誤り】や【X14 レジスタ違反】ともみなせるが，決定木において優先度が高い X4a に分類される．

上述の統一の事例に類似の，同一の実体の訳出に関する修正は他のカテゴリにも見られた．

【X7 用語の訳出誤り】： 訳出の修正/統一 (要文脈 22 件/文内 48 件).

【X3 原文内容の歪曲】： 参照先が曖昧な場合の誤訳の修正 (要文脈 21 件). e.g., “track work” に対する「保線工事」と「コースでの馬の調教」，“relief work” に対する「救済活動」と「臨時便」.

【X6 曖昧さ未解消】： 2 件とも文書中の参照先を特定しなければ正しい訳出が困難なものであった．“brother” に対する「兄弟」と「の兄」，“limbs” に対する「手足」と「腕」.

同一の実体を参照する表現 (共参照表現) を扱うこれらの事象は，文書内の他の文中の情報を要する修正事例のうち 70% (115/164) を占めていた．

4 現状認識と提言

本稿で述べた分析作業では，2 節で述べたように，文単位とはいえできる限り品質が高い MT 訳を起点とした．表 3 に示すように，2.2 節で述べた手続きによって比較的高い BLEU スコアの MT 訳が得られていた．ALT の参照訳に対する 36.0 という BLEU スコアは，PE の最終訳の 36.8 と比べても遜色ない．しかしながら，作業工程 (2) における高い編集率 (95%) を鑑みると，実際には，文単位のファクトレベルの訳出にもまだ多くの課題が残されていると言わざるを得ない．

翻訳の際に参照すべき情報 (例えば 3.2 節で挙げたもの) を MT に組み込むことが今後の課題である．用語の訳出や文体などの統一，すなわち首尾一貫性 (coherence) の観点では，指定された特定の表現を強制的に訳出する仕組みが必要である．NMT の拡張であれば，制約付きデコーディング [4] が参考になる．既存の CAT ツールのように，用語集を用いたプレースホルダ化や一貫性担保に特化した後編集も有用であろう．主題の展開，接続表現や指示詞の選択など，結束性 (cohesiveness) の観点では，前方文脈の文を参照する単純な手法では限界がある．後方文脈との接続の良さを評価すべき場合もある．修辞構造理論や共参照解析，自然言語生成における参照表現生成に関する取り組みが参考になる．

一方，MT の利用時には，依然としてファクトレベルで誤りを生じる可能性を念頭に置く必要がある．例えば，対訳データに頻出する専門用語の語彙選択やコロケーションは，既存の NMT でも訳出できる可能性が高いが，低頻度の (あるいは初見の) 用語・固有名を適切に訳出できる可能性は低い．CAT ツールや Web ブラウザから容易に利用できる MT エンジンは種々あるが，使用する MT エンジンが参照している情報と自らが扱う文書，翻訳戦略などに照らして，用途を吟味する必要がある．

表 3: ALT の参照訳および文書単位の PE の結果 (PE 訳) に対する各訳出物の BLEU スコア.

訳出物	BLEU 計算時の参照訳	
	ALT の参照訳	PE 訳
ALT のみを用いた場合の MT 訳	14.6	15.9
フェーズ 1 終了時の MT 訳	29.0	40.3
フェーズ 2 終了時の MT 訳	35.8	57.6
フェーズ 3 終了時の MT 訳	36.0	58.6
文単位の PE の結果	36.8	95.0
文書単位の PE の結果	36.8	100.0

5 おわりに

本稿では，MT 訳に対する文単位，文書単位の 2 段階の PE を通じて収集した誤りの修正事例を分析し，翻訳時に参照すべき情報を整理した．今後は，文単位の PE の内容も分析し，MT の性能向上に役立てたい．

謝辞: 本研究の一部は科研費基盤研究 (S)「翻訳規範とコンピテンスの可操作化を通した翻訳プロセス・モデルと統合環境の構築」（課題番号：19H05660，代表：影浦峡）の支援を受けた．

参 考 文 献

[1] C. Chu, R. Dabre, and S. Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In *Proc. of ACL*, pp. 385–391, 2017.

[2] 宮田玲, 藤田篤. 機械翻訳向けプリエディットの有効性と多様性の調査. 通訳翻訳研究への招待, Vol. 18, pp. 53–72, 2017.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proc. of ACL*, pp. 311–318, 2002.

[4] M. Post and D. Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proc. of NAACL*, pp. 1314–1324, 2018.

[5] H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, R. Sun, S. Sam, S. Seng, K. M. Soe, K. T. Nwet, M. Utiyama, and C. Ding. Introduction of the Asian Language Treebank. In *Proc. of O-COCOSDA*, pp. 1–6, 2016.

[6] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proc. of ACL*, pp. 1715–1725, 2016.

[7] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*, pp. 223–231, 2006.

[8] 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.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. of NIPS*, pp. 5998–6008, 2017.