機械翻訳向け原文編集の支援に向けた日英翻訳品質推定データセットの設計と構築

島田紗裕華[1] 山口大地[1] 宮田玲[2] 藤田篤[3] 梶原智之[4] 佐藤理史[1]
[1] 名古屋大学大学院工学研究科 [2] 東京大学大学院教育学研究科
[3] 情報通信研究機構 [4] 愛媛大学大学院理工学研究科
shimada.sayuka.y9@s.mail.nagoya-u.ac.jp

概要

機械翻訳向けの原文編集 (前編集) を支援する方法の 1 つとして、単語レベルの翻訳品質推定ラベル (OK/BAD) を原文側に表示する方法が考えられる。本研究では、そのような技術の実現に向けて品質ラベル付きデータセットを構築した。具体的には、前編集支援を目的とした品質ラベルの付与指針を定め、行政分野の日本語原文書 (18 文書) と日英機械翻訳出力 (2 種類) における計 2,090 文対に、人手で品質ラベルを付与した。また、構築したデータセットを用いて翻訳品質推定モデルを実装・評価することで、前編集支援における既存の自動ラベル付け手法の不十分さと現在の品質推定技術の課題を示した。

1 はじめに

機械翻訳 (MT) を活用する手法として、翻訳対象の原文を事前に編集する前編集 (pre-editing) がある。ニューラル MT に対する前編集の潜在的な有効性は示されているが [1]、実務において前編集を支援する技術の開発は十分進んでいない。これまでルールベース MT や統計的 MT においては、原文中の特定の言語表現特徴を規制する制限言語的アプローチが有効とされてきたが [2, 3, 4]、挙動が予測しづらいニューラル MT においては限界がある。特定の言語表現特徴を有することが常に翻訳品質の低下につながるとは限らず[1)]、事例ごとに翻訳品質に関する判断をする必要がある。

本研究では、とりわけ起点言語モノリンガルの前編集者に対して、編集を行うべきかの判断材料を提供しうるものとして、翻訳品質推定 (translation quality estimation; TQE) タスクにおける、原文側への品質ラベル付与タスクに注目する。TQE は人手の参照翻訳文を使わずに、機械翻訳の個々の出力の品質を予測するタスクで、文レベル・単語レベルの両方で取り組まれてきた [5, 6]。単語レベルの TQE は、MT 訳の単語列 (トークン列) に対して、翻訳品質に問題のないトークンに OK のラベル、問題のあるトークンに BAD のラベルを付与するタスクに加えて、それらに対応する原文側への OK/BAD のラベル付けを行うタスクもある [7]。

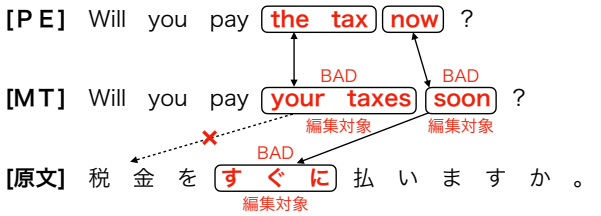

図 1 機械翻訳結果 (MT) に対する人手後編集結果 (PE) を利用した原文編集対象の同定; MT において PE との差分は編集対象 (BAD) であるが、原文においてそれに対応する箇所をそのまま編集対象と見なせるとは限らない

前編集の支援に有用な原文への品質ラベル付与の実現に向けて、TQE の訓練・評価データに関する課題が大きく分けて 2 つある。まず、根本的な課題として、既存のラベル付けの基準・方針が必ずしも前編集に適していないことが挙げられる。従来手法では、MT 訳における人手後編集（post-editing; PE）箇所や MQM 等の翻訳エラー体系に基づいて特定されたイシュー箇所に BAD のラベルを付与し、そのまま原文側にラベルを伝播させる [6, 8]。しかし、図 1 の日英 MT 訳の例に示すように、英語における単数形・複数形や冠詞の問題 (“your taxes” → “the tax”) は、直接対応する表現のない日本語側の前編集によって解消することは困難であると予想される。すなわち、言語依存性を考慮しながら、前編集の観点から MT 結果の改善につながりうる箇所にのみ BAD ラベルを付与する必要がある。さらに、人手で

1) 例えば、原文内の主語の不在という特徴は、日英 MT において常に主語の欠落や誤訳を引き起こすわけではなく、MT が文脈から正しく主語を補うことは十分ありえる。

ラベル付けされたデータがないという課題もある。従来手法では、原文側へのラベルの伝播は、単語アライメントツールの結果を用いて自動で行われるが [8]、特に日英などの言語構造が大きく異なる言語対では、その精度が十分とはいえない。技術の発展のためには、定められた基準に沿って高い精度でラベル付けされたデータを整備することが重要である。

以上をふまえ、本稿では、日英 MT 向けの前編集支援技術の開発に向けた出発点として、原文への品質ラベル付けモデルの訓練・評価に使える人手ラベル付きデータセットを構築した試みについて報告する。まず、日英翻訳の特性を考慮した単語レベルのラベルの付与指針について説明する (2 節)。続いて、自治体分野の文書を対象に、2 種類の MT 訳と人手 PE 訳を用いて、MT 側・原文側ともに人手で品質ラベルを付与した過程と結果を説明する (3 節)。さらに、構築したデータセットを用いて TQE モデルを構築・評価することで、人手ラベル構築の必要性を示す (4 節)。構築したデータセットは、CC-BY ライセンスで公開する[2)]。

2 品質ラベル付与指針の作成

2.1 設計理念と基本指針

大きく以下の 3 つの設計理念のもとで、単語レベル品質ラベル付与の基本指針を定めた。

理念 1. 前編集対象箇所の同定に役立つ 1 節でも記したように、前編集を行うことで MT 訳品質の向上が見込める箇所に絞って原文の品質ラベルを付与する方針とする。そのためには、対象言語や翻訳の特性を考慮する必要がある。また、人間にとって理解・編集しやすいテキストスパンにラベルを付与することも重要である。

理念 2. 多様なモデルの訓練に使える TQE モデル実装時のトークナイザの種類によらず使えるように、原文側 (日本語) へのラベル付けは文字単位で行う。ただし、MT 側 (英語) へのラベル付けは、単語単位 (スペースおよび約物区切り) で行う。

理念 3. 将来的な自動データ構築を視野に入れる 将来的に PE 訳から自動で原文にラベル付けする技術の開発を想定し、PE 訳の編集スパンと乖離しないように、可能な限り最小のスパンで原文ラベルを付与する。ただし、人間にとっての理解・編集しやすさとのバランスをとる必要がある。

2) https://github.com/tntc-project/QEdatasetJaEn

表 1 手順 (2) 原文への品質ラベル付けの指針

2-a.	手順 (1) で BAD ラベルを付与した箇所が、単数・複数形、冠詞、所有格、人称などの言語学的特徴を持ち、その意味・機能が起点言語で明示的に現れない場合、手順 (2) では BAD ラベルを付与しない (ただし、明示的に対応する表現がある場合、BAD ラベルを付与する)
2-b.	手順 (1) で動詞や形容動詞のみに BAD ラベルを付与した場合、手順 (2) では対応する原文の動詞や形容動詞の語幹のみに BAD ラベルを付与し、付属語には BAD ラベルを付与しない
2-c.	原文中の体言や用言の語幹に接続され、文意に大きく影響しない助動詞 (「です」「ます」等) には BAD ラベルを付与しない
2-d.	手順 (1) で句・節に渡るスパンで BAD ラベルを付与した場合、手順 (2) では明確に対応する前置詞等の箇所がなくても適宜助詞・助動詞も含めて BAD ラベルを付与する
2-e.	手順 (1) で BAD ラベルを付与した箇所と対応する原文箇所がない場合、文の自然さを保つ範囲内で、妥当な挿入位置に GAP タグを挿入する

2.2 具体的指針

品質ラベルの付与は、(1) MT 訳と PE 訳の差分への BAD ラベルの付与 (MT 側)、(2) 原文への BAD ラベルの伝播 (原文側)、の 2 段階で行う。BAD が付かなかったトークンには全て OK を付与する。手順 (2) では、原文には明示的に含まれないが MT 側で BAD が付いた情報に対応する原文位置 (文字と文字の間) に GAP タグを挿入する。

2.1 節の基本指針を前提に、第 1 著者が実際の事例に対してラベル付けを行いながら、各手順の指針の具体化を行った。その結果を一部の共著者が確認し、最終的な指針としてまとめた。原文へのラベル付けに関わる手順 (2) の指針を表 1 に示す[3)]。

指針 2-a は、MT 訳の品質向上のための前編集という目的に対応したものである (図 1 も参照)。指針 2-b, 2-c, 2-d は、ラベルを付与するスパンの認定に関するものである。なるべく必要最小限のスパンにラベルを付けること (指針 2-b, 2-c) が原則であるが、人間の理解しやすさを考慮し編集のまとまりを分断しすぎないこと (指針 2-d) も重要である。指針 2-e は原文で明示的に現れていない要素を、可能な範囲で追加することを促すためのものであり、指針 2-a にも関わる。ただし、例えば、MT 訳で単数形の名詞が PE により複数形に修正されていても、その情報を原文で明示するために「〜たち」のような表現を無理に GAP として挿入する必要はない。

3) 手順 (1) を含めた指針の一覧と適用事例は付録 A を参照。

表 2　日英翻訳品質推定データセットの基本統計

ラベル付与対象	使用 MT	文書数	総文数	BAD, GAP あり文数	総トークン数	BAD トークン数	GAP タグ数
原文側 (日本語)	TexTra	18	1,045	419 (40.1%)	24,270	1,661 (6.8%)	84
	Google	18	1,045	483 (46.2%)	24,262	2,278 (9.4%)	284
MT 側 (英語)	TexTra	18	1,045	372 (35.6%)	13,871	945 (6.8%)	–
	Google	18	1,045	522 (50.0%)	13,242	1,785 (13.5%)	–

3　データセットの構築

3.1　構築方針と手順

翻訳方向は日英を対象とした。TQE の研究分野では、[9] を除き、日英方向の TQE データセットの構築・公開はほとんどなされていない。また、対象文書は行政文書とした。近年の研究で使われるデータセット [6, 8] は、Wikipedia や産業分野のテキストから抽出したものが多いが、行政分野における MT や前編集のニーズ [4, 10] を考慮し、自治体が住民向けに発信する行政文書を対象とした。

また、従来の TQE データセットと異なり、サンプリングした文の集合ではなく、文書の単位を保持する。例えば、図 1 における「すぐに」を “soon” と訳すか、“now” と訳すかは、文書全体における前後の文脈を確認しなければ本来判断できない。近年の文脈を考慮した MT の進展 [11] もふまえ、文書単位でデータセットを構築する方針とした。

以上の方針のもと、下記の手順で構築を行った。

1. 対象文書の選定　第 3 著者が愛知県名古屋市から提供を受けた対訳文書群[4]から、一定の内容的なまとまりがあり、かつ、人手の英訳版が存在する日本語文書に限定した上で、全体の分野多様性が確保できるように選定した。従来の単語レベル TQE の人手ラベルデータセットは概ね 1K〜数 K 文であることをふまえ、原文が 1,000 文を超えるまで文書を集め、最終的に 18 文書 (1,045 文) を選定した。

2. MT 訳の取得　日本の自治体でも利用されている TexTra[5]および Google 翻訳[6]をデフォルトの設定で用いて、原文書を英語に翻訳した。

3. PE 訳の作成　プロの日英翻訳者が MT 訳に対する後編集を行った。作業者は、各原文書の全体を参照しながら、次の要件をすべて満たす訳文を作成するために、必要最低限の修正を施した。

- 原文書の情報を過不足なく、誤りなく伝える
- 文法的な誤りを含まない
- 同一文書内での用語の一貫性を保つ

4. 品質ラベルの付与　2 節で定義した指針に従い、第 1 著者が手作業で単語レベルの品質ラベル (原文側・MT 側) を付与した。作業時の疑問点は、共著者らと随時解消した。

表 3　原文側の人手ラベルと自動ラベルの一致度

(a) TexTra

		自動 BAD	自動 OK
人手	BAD	525	1,136
	OK	321	22,288

(b) Google

		自動 BAD	自動 OK
人手	BAD	1,236	1,044
	OK	1,294	20,696

3.2　構築結果

構築したデータセットの基本統計を表 2 に示す。原文側の統計に注目すると、BAD や GAP が 1 つ以上付与された文は TexTra で 419 文 (40.1%)、Google 翻訳で 483 文 (46.2%) と、全 1,045 文の半数未満であった。また、BAD のラベル数は TexTra、Google 翻訳ともに 10%以下であり、OK ラベルに比べて大幅に少ない。TexTra と比べて、Google 翻訳の方が、原文への BAD ラベル数、GAP タグ数ともに多い。

TQE 研究において一般に、原文ラベルは、MT 訳・PE 訳間の差分を自動で原文に伝播させる手法で付与される [8]。このような自動手法で付与したラベル (自動ラベル) と今回人手で付与したラベル (人手ラベル) の比較結果を表 3 に示す[7]。BAD ラベルに関して、人手ラベルを正解としたときの自動ラベルの F1 は TexTra で 0.419、Google 翻訳で 0.514 と低く、本研究で提案したラベル付け指針と現在の自動ラベル付け手法との乖離が明らかになった。

4)　この文書群は「名古屋市翻訳資源」として、CC-BY ライセンスで公開予定である。https://github.com/tr4lg/nagoya-dataset/

5)　https://mt-auto-minhon-mlt.ucri.jgn-x.jp/　(2023/1/31 訳文取得; ver. GPMT-3.9_220930_nmt)

6)　https://translate.google.com/ (2023/03/05 訳文取得)

7)　自動ラベルの付与に当たって、MT 訳・PE 訳間の単語アライメントは tercom [12] (https://www.cs.umd.edu/~snover/tercom/) で、PE 訳・原文間の単語アライメントは awesome-align [13] (https://github.com/neulab/awesome-align) で行った。アライメント結果を用いた品質ラベルの付与は、qe-corpus-builder (https://github.com/deep-spin/qe-corpus-builder) で行った。

表 4　各種ラベルデータを用いた単語レベル翻訳品質推定 (原文側) の評価結果 (F1-Mult=F1-BAD×F1-OK)

タスク	訓練データ設定	MCC		F1-BAD		F1-OK		F1-Mult	
		人手	自動	人手	自動	人手	自動	人手	自動
GAP あり	(1) 人手	0.200	0.139	0.190	0.148	0.982	0.969	0.187	0.144
GAP なし	(2) 人手 (w/o GAP)	0.189	0.182	0.189	0.185	0.966	0.970	0.183	0.180
	(3) 自動	0.087	0.362	0.035	0.235	**0.969**	0.991	0.034	0.233
	(4) 自動+人手 (w/o GAP)	0.136	0.185	0.097	0.189	**0.969**	0.971	0.094	0.183
	(5) 疑似	0.000	0.000	0.000	0.000	**0.969**	0.990	0.000	0.000
	(6) 疑似+人手 (w/o GAP)	**0.206**	0.239	**0.222**	0.226	0.964	0.968	**0.214**	0.218
	(7) 疑似+自動	-0.008	**0.444**	0.000	**0.333**	**0.969**	**0.992**	0.000	**0.331**

4　翻訳品質推定モデルの実装と評価

4.1　実験の枠組み

3 節で提案した人手ラベルデータを用いて、原文ラベル付与がどれだけできるか検証する。また、自動ラベルデータ (3.2 節参照) や疑似的に生成したラベルを用いる従来手法と比較する。

疑似ラベルデータ生成には、対訳コーパス中の参照訳を疑似的に PE 訳と見なす手法を用いた [14, 15, 16, 17]。原文・MT 訳・疑似 PE 訳の組に対するラベル付与の方法は、3.2 節で示した自動ラベル付与の方法と同様である。本研究では WMT'22 で配布された日英対訳コーパスをもとに、1.3M 組からなる疑似ラベルデータを作成した (詳細は付録 B)。

単語レベル TQE タスクは、原文と MT 訳のトークン系列を連結して入力し、各トークンの品質ラベル (OK/BAD) を予測する系列ラベリングタスクと捉えられる。先行研究 [6, 18] を参考に、事前学習済み言語モデル XLM-RoBERTa [19] を微調整して単語レベル TQE モデルを実装した (詳細は付録 C)。データ構築のタイミングの都合上、人手・自動ラベルには TexTra のデータ (訓練 835 件、開発 105 件、評価 105 件) のみを使用した。訓練データの組み合わせに関する設定は表 4 に示す計 7 通りである。2 種類のデータセットによる 2 段階訓練の設定も含まれる。例えば、(6) の「疑似+人手」は、疑似ラベルで訓練をした後に、人手ラベルで追加訓練する方法を指す。なお、自動ラベル、疑似ラベルとの統合・比較の都合上、設定 (2)〜(7) については GAP タグを使わないタスクとして実装したため、訓練および評価の際に人手ラベルデータから GAP タグは除外した。

全ての設定について、評価は同じ原文・MT 訳 (105 件) に対する人手ラベルおよび自動ラベルを用いた。評価指標には、MCC、BAD ラベルの F1 (F1-BAD)、OK ラベルの F1 (F1-OK)、F1-BAD と F1-OK の積 (F1-Mult) を用いた [20]。

4.2　評価結果

表 4 に評価結果 (原文側の結果のみ) を示す。GAP なしのタスクでは、「疑似＋人手」手法が、人手ラベル評価データの MCC, F1-BAD, F1-Mult の指標で最も良い性能を示した。「疑似＋自動」手法は、自動ラベル評価データでは全指標で最高スコアを示したが、人手ラベルの評価データではスコアが低かった。3.2 節で示したように、人手ラベルと自動ラベルに乖離があるため、自動ラベルで訓練したモデルが人手ラベルで良い結果を得られないのは当然である。本来解くべきタスクに応じてラベルを設計し、高品質の教師データを構築することが重要である。

なお、今回の実験において、MCC の最大値は 0.206、F1-Mult の最大値は 0.214 と低い。単語レベル TQE は、MT 側でも実用水準には達しておらず、後編集の支援には F1-Mult で 0.8 以上が必要であると指摘されている [21]。より難易度が高いと想定される原文側のタスクは、研究の余地が大きい。

5　おわりに

本研究では、MT 向け前編集に役立つ原文品質ラベルの付与方針を策定し、それに従い人手でラベル付けしたデータセットを構築した。TQE モデルの構築実験により提案データセットの有効性を確認し、TQE 研究で広く使われるラベル付け手法が前編集の用途には必ずしも適合しないことを示した。

今後の課題として、提案指針に基づく原文へのラベル付けが実際の前編集作業に役立つかの検証がある。現時点では TQE モデルの性能には限界があるため、[21] を参考に、まずは人手ラベルを活用し求められる性能水準を見極める。並行して、データの拡張を行い、TQE モデルの性能改善を目指す。

謝辞

本研究は JSPS 科研費（課題番号：19H05660, 23H03689）の支援を受けた。

参考文献

[1] Rei Miyata and Atsushi Fujita. Understanding pre-editing for black-box neural machine translation. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, pp. 1539–1550, Online, 2021.

[2] Eric Nyberg, Teruko Mitamura, and Willem-Olaf Huijsen. Controlled language for authoring and translation. In Harold Somers, editor, **Computers and Translation: A Translator's Guide**, pp. 245–281. John Benjamins, Amsterdam, 2003.

[3] Takako Aikawa, Lee Schwartz, Ronit King, Monica Corston-Oliver, and Carmen Lozano. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In **Proceedings of the Machine Translation Summit XI**, pp. 1–7, Copenhagen, Denmark, 2007.

[4] Rei Miyata. **Controlled Document Authoring in a Machine Translation Age**. Routledge, London, 2020.

[5] Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. **Quality Estimation for Machine Translation**. Springer, Cham, 2018.

[6] Frederic Blain, Chrysoula Zerva, Ricardo Ribeiro, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. Findings of the WMT 2023 Shared Task on Quality Estimation. In **Proceedings of the 8th Conference on Machine Translation (WMT)**, pp. 629–653, Singapore, 2023.

[7] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. Findings of the WMT 2020 Shared Task on Quality Estimation. In **Proceedings of the 5th Conference on Machine Translation (WMT)**, pp. 743–764, Online, 2020.

[8] Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. MLQE-PE: A multilingual quality estimation and post-editing dataset. In **Proceedings of the 13th Language Resources and Evaluation Conference (LREC)**, pp. 4963–4974, Marseille, France, 2022.

[9] Atsushi Fujita and Eiichiro Sumita. Japanese to English/Chinese/Korean datasets for translation quality estimation and automatic post-editing. In **Proceedings of the 4th Workshop on Asian Translation (WAT)**, pp. 79–88, Taipei, Taiwan, 2017.

[10] Anthony Pym, Nune Ayvazyan, and Jonathan Prioleau. Should raw machine translation be used for public-health information? Suggestions for a multilingual communication policy in Catalonia. **Just. Journal of Language Rights & Minorities, Revista de Drets Lingüístics i Minories**, Vol. 1, No. 1-2, pp. 71–99, 2022.

[11] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level neural machine translation: Methods and evaluation. **ACM Computing Surveys**, Vol. 54, No. 2, pp. 1–36, 2021.

[12] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In **Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers (AMTA)**, pp. 223–231, Cambridge, Massachusetts, USA, 2006.

[13] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, pp. 2112–2128, Online, 2021.

[14] Lemao Liu, Atsushi Fujita, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Translation quality estimation using only bilingual corpora. **IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)**, Vol. 25, No. 9, pp. 1762–1772, 2017.

[15] Yi-Lin Tuan, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Francisco Guzmán, and Lucia Specia. Quality estimation without human-labeled data. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, pp. 619–625, Online, 2021.

[16] Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, Gyeongmin Kim, Jungseob Lee, and Heuiseok Lim. QUAK: A synthetic quality estimation dataset for Korean-English neural machine translation. In **Proceedings of the 29th International Conference on Computational Linguistics (COLING)**, pp. 5181–5190, Gyeongju, Republic of Korea, 2022.

[17] Zhen Yang, Fandong Meng, Yuanmeng Yan, and Jie Zhou. Rethinking the word-level quality estimation for machine translation from human judgement. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 2012–2025, Toronto, Canada, 2023.

[18] Dongjun Lee. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In **Proceedings of the 5th Conference on Machine Translation (WMT)**, pp. 1024–1028, Online, 2020.

[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 8440–8451, Online, 2020.

[20] Varvara Logacheva, Michal Lukasik, and Lucia Specia. Metrics for evaluation of word-level machine translation quality estimation. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 585–590, Berlin, Germany, 2016.

[21] Raksha Shenoy, Nico Herbig, Antonio Krüger, and Josef van Genabith. Investigating the helpfulness of word-level quality estimation for post-editing machine translation output. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 10173–10185, Online and Punta Cana, Dominican Republic, 2021.

[22] Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook AI's WMT21 News Translation Task submission. In **Proceedings of the 6th Conference on Machine Translation (WMT)**, pp. 205–215, Online, 2021.

A　ラベル付け指針の全体像と事例

手順 (1) MT 側への品質ラベルの付与に関する指針と適用事例は以下の通りである (赤字は BAD ラベル)。

1-a. 文頭か否かによる大文字・小文字の違いには BAD ラベルを付与しない
例) **[MT]** Ashtrays always contain water. **[PE]** (...) ashtrays always contain water. [原文] 灰皿はいつも水をいれます。

1-b. 記号の全角・半角の違いには BAD ラベルを付与しない
例) **[MT]** FAX: **[PE]** FAX：[原文] FAX：

1-c. 数や序数が意味的には同じでも表記が異なる場合は BAD ラベルを付与する
例) **[MT]** **Two** people **[PE]** **2** people [原文] 2 人

1-d. 冠詞と冠詞が付属する語はそれぞれ独立に考えて BAD ラベルを付与する
例) **[MT]** **an** document **[PE]** **a** document [原文] 書類

手順 (2) 原文側への品質ラベルの付与に関する指針と適用事例は以下の通りである (赤字は BAD ラベル)。

2-a. 手順 (1) で BAD ラベルを付与した箇所が、単数・複数形、冠詞、所有格、人称などの言語学的特徴を持ち、その意味・機能が起点言語で明示的に現れない場合、手順 (2) では BAD ラベルを付与しない (ただし、明示的に対応する表現がある場合、BAD ラベルを付与する)
例) **[MT]** **place** **[PE]** **places** [原文] 場所

2-b. 手順 (1) で動詞や形容動詞のみに BAD ラベルを付与した場合、手順 (2) では対応する原文の動詞や形容動詞の語幹のみに BAD ラベルを付与し、付属語には BAD ラベルを付与しない
例) **[MT]** the family members you **are supporting** **[PE]** the family members you **have as dependents** [原文] あなたが**扶養し**ている家族

2-c. 原文中の体言や用言の語幹に接続され、文意に大きく影響しない助動詞 (「です」「ます」等) には BAD ラベルを付与しない
例) **[MT]** I **need** proof of income. **[PE]** I **want** proof of income. [原文] 所得証明が**ほしい**です

2-d. 手順 (1) で句・節に渡るスパンで BAD ラベルを付与した場合、手順 (2) では明確に対応する前置詞等の箇所がなくても適宜助詞・助動詞も含めて BAD ラベルを付与する
例) **[MT]** **repayment of debt** **[PE]** **debt repayment** [原文] **借金の返済**

2-e. 手順 (1) で BAD ラベルを付与した箇所と対応する原文箇所がない場合、文の自然さを保つ範囲内で、妥当な挿入位置に GAP タグを挿入する
例) **[MT]** How many **tickets** do you need? **[PE]** How many **copies** do you need? [原文] **[GAP]** **何枚必要ですか**。

B　疑似ラベルデータの構築手順

本研究では、対訳コーパス中の参照訳を MT 訳に対する疑似 PE 訳とみなす。以下の手順で日本語原文、英語参照訳 (疑似 PE 訳)、英語 MT 訳の組 (計 1.3M) を収集した。

1. WMT'22 のウェブサイト[8]から英日対訳文データを取得 (計 33.9M)
2. WMT'21 の英日ペアの翻訳タスクで高成績を収めた MT モデル [22][9]を用いて、日本語文を英訳
3. 組になる日本語文、英語参照訳、英語 MT 訳のいずれかにおいて、トークン数[10]が 120 を超えた場合、その組を除外 (計 33.7M)
4. PE らしい事例、すなわち、英語参照訳と英語 MT 訳がある程度似ている事例を得るため、両者の間の文レベル TER が 0.05 以下のものを選定（計 1.3M）

8) https://statmt.org/wmt22/mtdata/index.html
9) https://github.com/facebookresearch/fairseq/tree/main/examples/wmt21
サンプル文に対する評価に基づき、よりよい性能が期待できる "wmtdata otherdomain" というタグを特定し、データのデコード時に利用した。

表 5　各データセットのラベルの分布 (*: w/o GAP)

データセット	訓練 BAD	訓練 OK	検証 BAD	検証 OK	評価 BAD	評価 OK
人手	2,184	36,996	114	4,172	199	5,592
人手*	2,123	24,098	107	2,771	191	3,688
自動	1,914	38,709	228	4,535	165	4,574
疑似	188,991	46,661,987	118	36,284	131	34,891

表 6　各設定のハイパーパラメータ (*: w/o GAP)

設定	学習率	エポック数	α
人手	3.0×10^{-5}	15	16.9
人手*	3.0×10^{-5}	6	11.4
自動	3.0×10^{-5}	5	20.2
自動+人手*	9.0×10^{-5}	10	1
疑似	3.0×10^{-8}	21	1
疑似+人手*	1.0×10^{-5}	18	1
疑似+自動	3.0×10^{-5}	12	1

C　翻訳品質推定モデルの実装詳細

XLM-RoBERTa [19] に全結合層を 1 層追加し、微調整することで TQE モデルを実装した。疑似ラベルデータでの訓練には large モデル[11]を使用し、それ以外には base モデル[12]を使用した。モデル訓練時には、各データセットを訓練・開発・評価用に分割した。人手ラベルデータと自動ラベルデータはそれぞれ、訓練 835 件、開発 105 件、評価 105 件とした。疑似ラベルデータは、1.3M 件の内、開発 1K 件、評価 1K 件とし、残りを訓練とした[13]。各データセットのラベル分布を表 5 に示す。

全ての訓練設定でバッチサイズは 8、最適化手法は AdamW とした。損失関数には BAD の損失を α で重み付けした、以下の交差エントロピー損失を用いた。

$$loss = -\alpha y \log \hat{y} - (1-y)\log(1-\hat{y})$$

その他のハイパーパラメータを表 6 に示す。表中の α が 1 以外の場合には、訓練データ内の原文と MT 訳のトークンに付与された BAD と OK の合計の比を α とした。なお、ここで示したハイパーパラメータ以外は、Huggingface Transformers のデフォルト値を使用した。訓練はそれぞれのデータセットの開発データにおいて、原文と MT 訳のトークンに付与されたラベルから算出した MCC が 5 エポック連続して改善しなかったときに終了し、開発データでの MCC が最も高いモデルを使用した。

10) 日本語文は形態素数 (MeCab+IPAdic に基づく) を利用。
11) https://huggingface.co/xlm-roberta-large
12) https://huggingface.co/xlm-roberta-base
13) ただし、本稿では、人手・自動ラベルデータによる評価のみを報告しており、疑似ラベルデータの評価 1K 件は用いていない。