評価値・機械翻訳修正訳付き日英中韓対訳コーパスの構築

藤田 篤　隅田 英一郎
情報通信研究機構

1 はじめに

機械翻訳の性能向上は近年著しいが，正確かつ流暢な翻訳を常に生成する，ということは実現できていない．したがって，例えば我々の機関において開発・公開している音声翻訳アプリケーション VoiceTra[1] のように，機械翻訳の結果 (以下，MT 訳) を一般のユーザが直接即時に利用することを想定する場合は，非文法的な訳文によるユーザの失望や，誤訳によるミスリードを回避するための工夫が必要である．

この課題に対処するため，我々は，MT 訳の品質推定 (Quality Estimation; QE) 技術を実現し，メタ情報としての提示や人間による翻訳への切替等のワークフローの制御に利用することを検討している．本稿では，QE 技術の開発・評価のために構築した対訳コーパスについて述べる．このコーパスは，図 1 に示す要素で構成されるレコードからなる．原文としては，VoiceTra に対する日本語の音声入力 (概ね旅行会話に準ずる内容) の認識結果 8,789 発話，および病院における模擬被験者実験を通じて得た病院内会話の 1,676 発話を用いた．英語，中国語，韓国語を目標言語とし，VoiceTra と同じ翻訳モデルによって MT 訳を生成した．そして，

(1) 原文の内容を確認するために模範訳を作成
(2) MT 訳に対する評価値の付与
(3) MT 訳から必要最小限の修正 [5] によって正確かつ文法的な訳文 (修正訳) を作成

という手順で対訳コーパスを構築した．

2 機械翻訳結果の品質の推定

欧州言語の一部の言語間では，言語構造の類似性や多言語対訳コーパスの存在に支えられ，MT は早くからある程度高い精度を実現していた．またこれを受け，翻訳のワークフローにおいても，MT 訳を人間が後編集するアプローチ (Post-editing; PEMT) が試みられてきた．Blatz [1] は，PEMT 等の効率をさらに高める手段として，信頼度推定 (Confidence Estimation; CE) というタスクを提案した．信頼度という呼び方ではあるが，MT システム自身が学習に用いた対訳データ等に基づいて算出した尤度等ではなく，ユーザ目線での MT 訳の評価値のことを指す．また，参照訳を用いず原文を参照して MT 結果を評価するという点で，BLEU [4] 等の自動評価尺度とも異なる．その後，遅くとも文献 [6] 以降は品質推定 (Quality Estimation; QE) という呼称が定着し，2012 年からは評価型ワークショップ WMT のシェアードタスクとして取り上げられている．

[1] http://voicetra.nict.go.jp/

日本語原文	片道だけで買えますか。
目標言語 MT 訳	Can I buy just one way?
目標言語模範訳	May I get it for one way?
MT 訳の評価値	B
MT 訳の修正訳	Can I just buy a one way ticket?

図 1: 対訳コーパスのレコードの例.

これまで，MT 訳に対する QE は，次の 4 種類の粒度で取り組まれてきた [2].

語レベル: MT 訳中の個々の語の品質を推定して修正が必要な箇所を可視化することで，PEMT 等の効率を改善できると期待される．WMT では “OK” と “BAD” の 2 値分類の形式を取っている．

句レベル: 慣用表現や複合語，呼応表現等に見られるように，語は必ずしも原子的な意味の単位とは限らない．同様に，PEMT の処理単位とも限らない．そこで，固有表現抽出と同様に，MT 訳中の問題があるテキストスパンを同定するという形式のタスクが考えられている．

文レベル: 個々の文の品質については，少なくとも次の 2 種類の定義が考えられる．

a) PEMT を実施する場合に必要となる労力．例えば，編集の割合 (Human-Targeted Translation Edit Rate; HTER) [5] や編集に要する時間等．WMT のシェアードタスクでは，PEMT への応用を想定して HTER の値を予測する仕様となっている．
b) ユーザである人間にとっての正確さ，流暢さ等に基づく品質．MT 訳をそのままユーザに提示すべきか否かを判断するには，HTER よりもこちらの定義に沿った分類が適切である．

文書レベル: 文書全体に対する QE は，例えば，機械翻訳によって得られた文書が，要旨を掴む目的 (gisting) で使えるか否かの判断に資する．

3 コーパスの構築手順

我々は，Snover らの研究 [5] や WMT における試みをふまえ，日本語を起点言語，英語，中国語，韓国語を目標言語とし，図 1 の 5 つの要素をレコードとする対訳コーパスを構築した．以下本節では，作業手順 (図 1

表 1: 原文と MT 訳の記述統計.

		旅行会話 (8,789 発話)		病院内会話 (1,676 発話)	
データ	単位	合計	平均	合計	平均
日本語原文	文字	105,642	12.0	33,979	20.3
英語 MT 訳	ワード	44,614	5.1	14,844	8.9
中国語 MT 訳	文字	65,734	7.5	21,974	13.1
韓国語 MT 訳	文字	94,597	10.8	30,283	18.1

も参照されたい) に沿って，日本語原文の獲得，MT 訳の生成，模範訳の作成，MT 訳の評価値の付与，MT 訳の修正，整合性の確認について述べる．

3.1 日本語原文の獲得

日本語の原文としては，次の 2 種類を使用した．

旅行会話: VoiceTra に実際に入力された音声の自動認識結果を無作為に抽出した．音声入力時に特段の制限は設けていないが，概ね旅行会話に準ずる内容であり，話し言葉である．

病院内会話: 我々の機関と東大病院による音声翻訳の模擬被験者実験において，実際に発話された会話文．上記の旅行会話の発話に比べるとフォーマルではあるが，主語の省略の程度やくだけた文末表現は同様に見られる．

意味が分からない発話や公序良俗に反する発話等を除外し，旅行会話の 8,789 発話と病院内会話の 1,676 発話を使用した．なお，個々の発話が複数の文を含む場合もあるが，以下では簡単のため「原文」と記す．

3.2 MT 訳の生成

前節で得た原文の各々に対して，VoiceTra と同じ翻訳モデルを用いて MT 訳を付与した．日英，日中，日韓の翻訳モデルはいずれもフレーズベースの統計的機械翻訳 [3] に基づくものであるが，日英のモデルは 2013 年 10 月時点のもの，日中，日韓のモデルは 2016 年 9 月時点のものを使用した．表 1 に原文および MT 訳の分量を示す．

3.3 模範訳の作成

これ以降の作業には，目標言語を母語とし，かつ日本語原文を正しく解釈できる者が従事した．

まず，日本語原文のみを参照し，それに対する目標言語の模範訳を作成した．個々の文の前後の文脈は存在しないため，作業者に対しては，必要に応じて発話の状況を想定して訳文を作成して構わないこと，ただし原文およびそれが発話されると想定される状況から明らかでない情報を付加してはならないと指示した．省略されている代名詞等を補う場合は，可能な複数の候補を列挙するのではなく，1 つのみを示すこととした．

内容について
- 原文の内容を過不足なく翻訳しているか?
- 原文の文意を考慮して，文中のスペースや読点，および文末の句点，疑問符，感嘆符等を適切に指定しているか?
- 綴り誤りはないか?

形式およびスタイルについて
- 訳文にエスケープシーケンス等が含まれていないか?
- 訳文の先頭や末尾に不要なスペースや改行コードが含まれていないか?
- 訳文が引用記号で終わっている場合に，文末記号を引用符よりも前に置いているか?
- 引用記号としてダブルクオートとシングルクオートを統一的に使用しているか?
- 全角記号と半角記号を適切に使用しているか?
- (目標言語が中国語の場合) 複数の文を含む場合に文間に余分なスペース等を含めていないか?
- (目標言語が韓国語の場合) 文中のスペースは 1 つに統一されているか?
- (目標言語が韓国語の場合) 複数の文を含む場合に文間はスペース 1 つで統一されているか?

図 2: 訳文に関する検査項目の例.

模範訳を作成した後，翻訳の作成者とは別の者が図 2 の項目に照らして検査した．

3.4 MT 訳の評価値の付与

日本語原文および模範訳を参照しながら，MT 訳の品質を表 2 に示す 6 段階で評価した．ただし，自動評価における参照訳 の (本来の) 位置付けと同様に，模範訳のみが正しい訳文であるという強い制約ではない．作業者には，あくまで MT 訳の品質を評価するように指示した．

3.5 MT 訳の修正

日本語原文および模範訳を参照して内容を確認した後，MT 訳に対する修正訳を作成した．その際，文献 [5] にならい，次に示す 4 種類の操作を編集の単位とし，必要最低限の修正で正確かつ文法的な訳文 (表 2 の S または A に相当) を作成するよう，作業者に指示した．

語の削除: 不要な語を削除する: e.g., “the an” → “the”

語の挿入: 必要な語を必要な位置に挿入する: e.g., “We will stay at hotel.” → “We will stay at the hotel.”

語句の置換: 語句を別の表現に置き換える (語形の修正も含む): e.g., “Can you teach me the way to the station?” → “Can you tell me the way to the station?”

語句の移動: 語句の順序を変更する: e.g., “I’ll send a card my friend.” → “I’ll send my friend a card.”[2]

HTER の定義に従って修正の程度を「必要最低限」としたいが，真に最低限の修正を追求することは現実的ではない．そこで，MT 訳と模範訳の比較に基づいて

[2] この例を “I’ll send a card to my friend.” と修正してもよいが，その場合は,「語の挿入」の操作をしたことになる．

表 2: 評価値の種類と判断基準.

評価値	呼称	説明
S	ネイティブ並み	原文の情報が漏れなく翻訳されており，訳文に文法的な間違いがない．使われている語彙もネイティブから見て自然である．
A	申し分ない	使われている語彙はネイティブから見て不自然であるが，原文の情報が漏れなく翻訳されており，訳文に文法的な間違いがない．句読点 (あるいはピリオド，コンマ) の抜けやアルファベットの大文字と小文字の違い等の軽微な誤りも，意味が通じるならば許容する．
B	まずまず	原文のあまり重要でない情報が一部漏れていたり，訳文に綴り誤りや，文法的な誤り，句読点の誤りがあったりするが，容易に理解できる．
C	許容範囲	原文の重要な情報が漏れていたり，訳文に文法的な間違いが多数あったりしてかなり崩れた訳文ではあるが，原文の意味をなんとか理解することはできる．
D	意味不明	重要な情報が誤訳されており，原文の意味を正しく理解することは不可能である．
E	訳文なし	翻訳結果が与えられていない．

編集回数の上限を定めた．具体的には，MT 訳に対して k 回の編集で模範訳が得られるとき，修正訳の作成の際の編集回数の上限を k とした [3]．MT 訳から模範訳または修正訳を得るための編集回数は，TER のパッケージ [4] を用い，“-r” オプションで MT 訳を，“-h” オプションで模範訳または修正訳を指定して算出した [5]．その際，対象テキストは下記のツールを用いて事前にトーカナイズした．

- 英語: Moses のトーカナイザ [6] による処理結果．
- 中国語: 文字．
- 韓国語: MeCab[7] と mecab-ko-dic[8] を用いた分かち書き結果．ただし，平文において明示されているスペースは特殊なトークンに置き換えて処理した．

なお，この作業においては，3.4 節で付与した評価値は参照しないこととし，修正の有無をもって評価値のダブルチェックを行うこととした (詳細は 3.6 節)．また，模範訳と同様に，修正訳の作成者とは別の者が，図 2 の項目に照らして検査した．

3.6 整合性の確認

MT 訳に対する評価値と修正訳が得られた時点で次の 3 点を確認し，一貫性が担保されていない場合は 3.4 節と 3.5 節の作業をやり直した．

- 評価値が S または A であるが，修正訳として MT 訳がそのまま採用されていない場合，評価値が誤っているか，不要な修正が施されている．
- 評価値が S でも A でもないが，修正訳として MT 訳がそのまま採用されている場合，評価値が誤っているか，必要な修正が施されていない．
- 修正訳を作成する際に要した編集回数が，模範訳に基づいて定めた編集回数の上限 k を超えている場合，より少ない編集回数で評価値が A 以上の訳文を作成できる [9]．

4 構築した対訳コーパスの諸元

構築したコーパスにおける評価値ごとの文の数を表 3 に示す．日英翻訳においては，旅行会話の方が病院内会話に比べて全体的な訳質が高かったが，日中・日韓翻訳においては，病院内会話の方が訳質が高かった．

MT 訳の評価値と修正訳作成時の修正の有無は，3.6 節に示した観点では整合している．ただし，2 節で述べたように，評価値と HTER は異なる概念である．実際に，図 3 に例示するように，大まかな相関は見られるものの，評価値が B の MT 訳の方が評価値が D の MT 訳よりも HTER が低い，ということは必ずしも成立していない．次の 2 例は，病院内会話の日英翻訳において，わずかな編集で修正訳が得られたにもかかわらず評価値が D であったもの (図 3 左下の外れ値) である．

(1) 原文: 多額の現金は持ってこないでください。
MT 訳: Please bring a lot of cash.
修正訳: Please don’t bring a lot of cash.

(2) 原文: 首が痛くありませんか。
MT 訳: Do you have pain in my neck?
修正訳: Do you have pain in your neck?

例 (1) の MT 訳は，原文における否定の意味を適切に表現できておらず，まったく逆の意味を持ってしまっている．例 (2) は，病院内会話ということを考慮する

[3] ただしこの制約は，模範訳をコピーして修正訳とすることで容易に満たせてしまうという問題がある．一方で，編集回数の上限を $(k-1)$ 回とすると制約としては強すぎる．

[4] http://www.cs.umd.edu/~snover/tercom/, version 0.7.25

[5] WMT のシェアードタスクにおけるデータは，修正訳に対する MT 訳の逸脱の度合いを測るため，“-r” オプションと “-h” オプションを我々とは逆に (正しく) 指定している．また，語句の移動を編集の基本操作とはみなさず，“-d 0” オプションを指定している．

[6] http://statmt.org/moses/, RELEASE-2.1.1

[7] http://taku910.github.io/mecab/, version 0.996

[8] https://bitbucket.org/eunjeon/mecab-ko-dic/, version 2.0.1-20150920

[9] 編集回数の把握が困難になるため，作成済の修正訳を MT 訳に近づけるように編集することは禁止し，再度 MT 訳から始めて修正訳を作成するよう指示した．

表 3: 構築した対訳コーパスの記述統計.

	旅行会話 (8,789 発話)						病院内会話 (1,676 発話)					
	日英		日中		日韓		日英		日中		日韓	
評価値	文数	割合	文数	割合	文数	割合	文数	割合	文数	割合	文数	割合
S	1,961	22.3%	2,829	32.2%	3,466	39.4%	95	5.7%	708	42.2%	903	53.9%
A	1,462	16.6%	1,875	21.3%	2,326	26.5%	107	6.4%	514	30.7%	482	28.8%
B	1,269	14.4%	1,275	14.5%	1,361	15.5%	181	10.8%	172	10.3%	166	9.9%
C	1,067	12.1%	899	10.2%	724	8.2%	333	19.9%	107	6.4%	97	5.8%
D	3,026	34.4%	1,909	21.7%	908	10.3%	960	57.3%	175	10.4%	28	1.7%
E	4	0.0%	2	0.0%	4	0.0%	0	0.0%	0	0.0%	0	0.0%

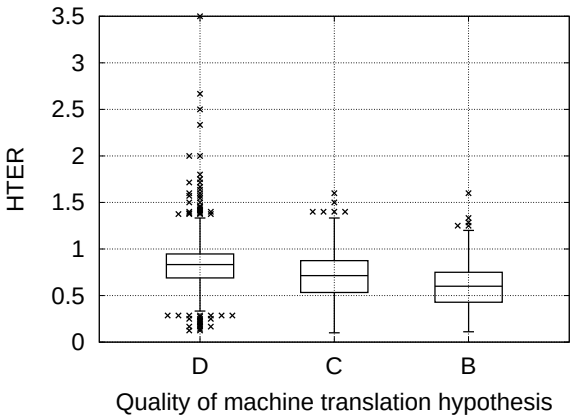

図 3: 病院内会話の日英翻訳において評価値が B，C，または D となった MT 訳に対する HTER の分布.

と，医療従事者から患者に対する問いかけと解釈できる．したがって，この文脈の「首」は，発話者ではなく患者の「首」を指す．いずれの例においても MT 訳の誤訳は致命的だが，わずかな編集で修正訳を作成できるため，HTER の観点では「品質が高い」ということになる．

逆に，評価値が B であるにもかかわらず多くの編集がなされた例 (図 3 右側中央の外れ値) を見てみよう．

(3) 原文: 素晴らしい景色だね
MT 訳: It's beautiful scenery.
修正訳: The scenery's beautiful, isn't it ?

例 (3) の MT 訳からは，末尾のピリオドを "isn't it?" に置換するだけ (HTER は 0.8) で修正訳を得られる．しかし，作業者は主節の構文も修正したため，HTER は 1.4 となった．3.5 節の作業要領では，このような編集過多を完全には回避できない．一方で，MT 訳から修正訳を得る際の編集回数がどうしても減らせない場合は，評価値の方を疑うべきかもしれない．例えば，例 (4) の MT 訳は正しくは D と評価されるべきであった．

(4) 原文: 筆談でお願いします
MT 訳: Brush please.
修正訳: Could you write it down, please?

5 おわりに

本稿では，MT 訳の品質を自動的に推定する技術 (QE 技術) の開発・評価のために構築した，評価値および修正訳付き対訳コーパスについて述べた．

我々は現在，本コーパスを用いて語レベルおよび文レベルの QE 技術について研究している．また，フレーズベースの統計的機械翻訳による MT 訳を用いて構築した本コーパスが，ニューラル翻訳による MT 訳に対する品質推定でも有効であるかについて調査している．今後は，コンパラブルコーパスや単言語コーパスから自動抽出できる対訳文対の候補を QE 技術を活用してクリーニングする等の応用を予定している．

謝辞: 本研究は総務省の情報通信技術の研究開発「グローバルコミュニケーション計画の推進–多言語音声翻訳技術の研究開発及び社会実証– I. 多言語音声翻訳技術の研究開発」の一環として行われた．

参 考 文 献

[1] J. Blatz, E. Fitzgerald, G. Foster, S. Grandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING*), pp. 315–321, 2004.

[2] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the 1st Conference on Machine Translation* (*WMT*), pp. 131–198, 2016.

[3] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 311–318, 2002.

[5] M. Snover, N. Madnani, B. J. Dorr, and R. Schwartz. Fluency, adequacy, or HTER? exploring different human judgments with a tunable MT metric. In *Proceedings of the 4th Workshop on Statistical Machine Translation* (*WMT*), pp. 259–268, 2009.

[6] L. Specia, D. Raj, and M. Turchi. Machine translation evaluation versus quality estimation. *Machine Translation*, Vol. 24, No. 1, pp. 39–50, 2010.