2-3-4　機械翻訳結果の品質推定
2-3-4　Quality Estimation of Machine Translation Outputs

ルビノ ラファエル　藤田 篤
RUBINO Raphael and FUJITA Atsushi

機械翻訳（MT）システムの研究開発においてはMTシステムの出力（MT訳）の品質を人間が作成した参照訳と比較して評価することが一般的であるが、MTシステムを実際に利用する場面でMT訳を採用するか、人間が修正するか、破棄するかという判断を行うには、参照訳なしにMT訳の品質を評価する必要がある。本稿では、参照訳なしにMT訳の品質を推定するQuality Estimation（QE）と呼ばれる技術について述べ、NICTで開発した手法を紹介する。この手法は、2021年に開催された国際ワークショップEval4NLPのExplainable QEシェアードタスクにおいて良好な成績を収めた。その結果についても報告する。

In the research and development of machine translation (MT) systems, the quality of MT outputs is evaluated in general by comparing them with human-produced reference translations. However, in the real use cases of MT systems, their users should judge the quality of the MT outputs without reference translations and determine whether the MT outputs can be used, they need revision, or they should be disposed. This paper describes the methods for MT Quality Estimation (QE), the task of automatically estimating the quality of MT outputs without reference translations, and presents the method developed at the National Institute of Information and Communications Technology. The Explainable QE shared task held at Eval4NLP 2021 and our results are also presented.

1　はじめに

ある言語（起点言語）のテキストに対して、それと同じ意味内容を表す別の言語（目標言語）のテキストを生成する技術・アルゴリズムを機械翻訳（Machine Translation: MT）あるいは自動翻訳という。多言語コミュニケーションの円滑化を目指して、NICTでは長年にわたりMTの研究開発を推進してきた。研究成果を音声翻訳アプリVoiceTra® [1]、みんなの自動翻訳@TexTra® [2]などのサービスとして実現するとともに、技術移転を通じた社会実装を行ってきた。

技術の研究開発においては、その評価が不可欠である。大量の対訳（起点言語のテキストと目標言語のテキストの対）のデータを用いてMTシステムを学習する過程では、MTシステムが生成する翻訳（以下、MT訳）の品質を頻繁かつ高速に計測する必要がある。その際、学習用の対訳データとは別の評価用対訳データを用いて自動評価を行うことが一般的である。具体的には、評価用対訳データ中の起点言語のテキスト（以下、原文）に対するMT訳を対訳データ中の目標言語のテキスト（人間が作成した参照訳）と比較し、それらの類似度で翻訳の品質を近似する手法が用いられている。例えば、BLEUスコア[3]がもっともよく用いられている[4]。ただし、これは品質の近似に過ぎない。MTのサービスを提供したり更新したりする際には、誤訳によって生じるリスクやサービスの劣化を低減するために、MT訳の品質をより正確に評価することが不可欠である。そのためには例えば、MT訳が原文の意味や内容を過不足なく誤りなく伝えているか、目標言語の表現として流暢（りゅうちょう）であるか、固有表現や専門用語を正確に訳しているか、所定の文体や記法を遵守しているか、などを人間が評価することが考えられる。

一方で、MTシステムを実際に利用する場面で、ユーザは、原文をシステムに入力してMT訳を得た後に、それをそのまま採用するか、修正して用いるか、破棄するかという判断を行う。その際、MT訳の品質について知る必要があるが、当然参照訳はないし、都度人間が評価することも金銭的・時間的に現実的ではない。このような背景で、参照訳や人手での評価なしに、所与の原文に対するMT訳の品質を自動的に推定する技術（MT Quality Estimation: MTQE; 以下単にQEと記す）[5]の研究開発が行われてきた。特にこれ

原文	Game 1 of the World Series will start Saturday evening.
MT訳	ワールドシリーズの第1戦は土曜日の夜に始まります。
修正訳	ワールドシリーズの第1戦は土曜日の夜に始まります。
	文単位のスコア: 人手評価品質 1.00, chrF 1.000, TER 0.000, BLEU 1.000
原文	John Paul's six-day tour was hugely popular.
MT訳	ジョン・ポールの6日間のツアーは非常に人気がありました。
修正訳	法王ヨハネ・パウロの６日間の訪問は非常に人気がありました。
	文単位のスコア: 人手評価品質 0.30, chrF 0.500, TER 0.222, BLEU 0.605

図 1　機械翻訳の品質推定の例：MT 訳の黄色は誤訳、緑色は適訳という語単位の品質を表す

までは、開発と性能評価の基盤となるデータや評価尺度を共有して参加者の技術を競うシェアードタスク [6][7] が研究開発を活性化してきた。

MT 訳の品質を測る単位としては、文単位、語単位の 2 種類がよく研究されている。QE の入出力の例を図 1 に示す。文単位と語単位のいずれも、入力は < 原文 , MT 訳 > の対である。文単位の品質としては、訳文全体の適否や MT 訳を修正するコストの多寡が考えられる。例えば、音声翻訳アプリ VoiceTra® [1] を用いる場合に、訳文の適否を 0 ～ 100 点に定量化して示すことができれば、MT 訳の品質が十分でない場合にそのまま使用することによるリスクを回避できるだろう。あるいは、近年多くの翻訳会社が採用している、MT 訳を下訳とみなし、それを人手で修正したものを翻訳成果物とする翻訳制作工程 [8] について考えてみよう。例えば、原文中のどの部分が翻訳誤りを生じているか、MT 訳中のどの部分に修正が必要であるか（語単位の品質）、文全体でどれくらい修正が必要であるか（文単位の品質）が推定できれば、翻訳制作工程の効率化に役立つ。実際に、Memsource [9] などの翻訳支援ツールにも QE の機能が実装されつつある。

NICT では、VoiceTra® 及びみんなの自動翻訳 @ TexTra® の一般公開後、2015 ～ 2019 年度に実施した総務省委託研究「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」[10] において QE 技術の研究開発に着手し、これまで研究開発を進めてきた [11]–[13]。本稿では、QE 技術について述べ、NICT で開発した最新の手法を紹介する。この手法は、2021 年に開催された国際ワークショップ Eval4NLP の Explainable QE シェアードタスク [7] において良好な成績を収めた [14]。その結果についても報告する。

2　ニューラルネットワークを用いた近年の QE 技術

2.1　機械学習に基づく QE 技術の定式化

改めて、QE とは、所与の < 原文 , MT 訳 > の対に対する品質スコアを推定する技術である。QE のモデルは、他の自然言語処理タスク向けのモデルと同様に、正解事例（以下、QE データ）からの機械学習によって実現される。QE データ中の個々の事例は次のような形式のものである。

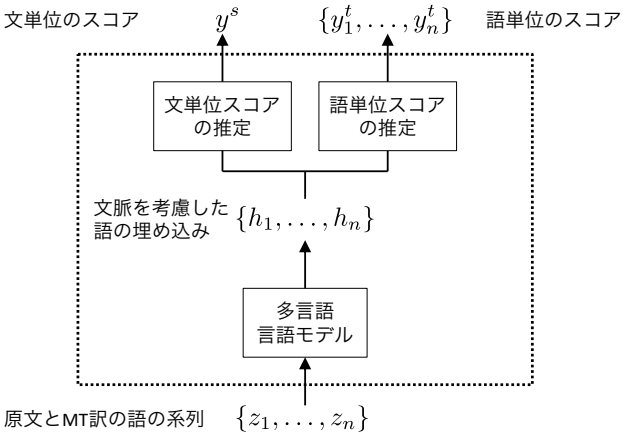

図 2　基本的な QE モデルの構成

- 文単位の QE データ：
  < 原文 , MT 訳 , 文単位の品質スコア >
- 語単位の QE データ：
  < 原文 , MT 訳 , 語単位の品質スコア >

例えば、図 1 の 2 つ目の例に対する品質スコアは次のように表せる。

- 文単位の品質スコア：
  人手評価品質 = 0.30, BLEU = 0.605, …
- 語単位の品質スコア：
  “ジョン・ポール” = Bad, “の” = Good, “6 日間” = Good, “の” = Good, “ツアー” = Bad, …

語単位の QE データの作成コストは文単位の QE データの作成コストに比べてはるかに高い。そこで近年、文単位の QE データのみを用いて語単位の品質スコアを推定する Explainable QE というタスク [7] が関心を集めている。

機械学習の基盤としては近年では、ニューラルネットワークを用いることが主流である。さらに、QE の対象として入力される < 原文 , MT 訳 > の対をベクトルに変換する（エンコードする）際に多言語言語モデルを用いること、人手で作成された QE データに加えて自動生成した大規模な疑似 QE データを用いることも一般的である [11][15][16]。ニューラルネットワークを用いた自然言語処理向けのモデルでは、大語彙を頑健に扱うために、個々の語をサブワードと呼ばれる短い文字列に分割して扱う [17][18] が、本稿では簡単のため語とサブワードを区別せず「語」と記す。

ニューラルネットワークに基づく近年の QE モデルは、図 2 に示すように、多言語言語モデルを用いたエ

ンコーダ、文単位の品質スコアの推定器、語単位の品質スコアの推定器で構成される。エンコーダは、入力された原文と MT 訳の合計 n 語の各語（$z_1, \ldots, z_n$）を、文内の文脈を考慮して d 次元ベクトル（$h_1, \ldots, h_n$）に変換する。原文と MT 訳は異なる言語で記述されているため、両者をエンコードするには多言語言語モデルが不可欠である。文単位の品質スコアの推定器は、原文と MT 訳の各語に対する合計 n 個のベクトルをプーリングした後にスコア y^s に変換する。一方、語単位の品質スコアの推定器は、原文と MT 訳の各語に対するベクトルを品質スコア $\{y_1^t, \ldots, y_n^t\}$ に変換する。

文単位と語単位のいずれについても、最後の変換処理の実装は品質スコアの定義によって異なる。例えば文全体または個々の語に対するスコアを 0 ～ 1 の実数値で得る場合は、d 次元のベクトルを 1 つの実数値に変換する回帰モデルとして実装する。文全体または個々の語を {Good, Fair, Bad, Critical} のような複数のクラスに分類する場合は、d 次元のベクトルをクラスの個数の実数値に変換する分類モデルとして実装する。

2.2　学習に使われる言語資源

QE モデルの学習には 2 種類の QE データが用いられる。1 つ目は、人手で作成された QE データである。このデータは、MT 訳の文全体の品質や各語の品質を人間が評価したり、MT 訳を人手で修正することを通じて修正が必要な部分とそうでない部分を同定したりすることで作成できる。ニューラルネットワークに基づくモデルの性能を向上させるには、大規模なデータを学習に用いることが望ましいが、QE データの作成は金銭的・時間的コストが高いため、これまでの研究では特定の翻訳方向及び特定の文書分野のテキストに対して 7,000 ～ 40,000 文ほどしか作られていない [6][7][12]。そこで 2 つ目のデータとして、人手で作成されたデータほど正確ではないものの大規模な QE データを自動的に生成して、学習に活用することが一般的である。そのような擬似 QE データは、例えば対訳データを用いて生成できる [11][15][16]。

< 原文 , MT 訳 > の対を精度良くエンコードするには、大規模な多言語テキストデータを用いて長時間かけて学習した多言語言語モデルを用いることが望ましい。近年では、そのような学習を経て得られた様々な種類の多言語言語モデルが一般公開されている。それらをそのまま活用することにより、コストを抑えつつ高い性能を実現することができる。実際に自然言語処理に関する多くの研究においてそのようなことが行われている。

2.3　QE モデルの学習手順

QE モデルの学習手順を図 3 に示す。以下では 4 つのステップの各々について述べる。

ステップ 1　エンコーダの初期化 : 事前学習済多言語言語モデルのパラメタをコピーして QE モデルのエンコーダのパラメタを初期化する。

ステップ 2　疑似 QE データの自動生成 : < 原文 , MT 訳 , 品質スコア > の組の形式のデータを機械的かつ大規模に生成する。まず、事前に用意した対訳データの一部を用いて MT モデルを学習する。既存の MT モデルが利用できる場合は、それを用いても良い。次に、対訳データのうち MT モデルの学習に使用していない部分の原文を翻訳モデルで翻訳して MT 訳を得る。最後に、MT 訳を対訳データにおける目標言語の訳文（MT 訳とは独立に作成された人間訳）と比較して品質スコアを得る。文単位のスコアの尺度としては、例えば、MT の自動評価に用いられる BLEU スコア [3]、chrF スコア [19]、TER スコア [20] などを用いることができる。また、人間訳を MT 訳の人手修正訳とみなして MT 訳と比較し、例えば MT 訳中の修正された語を Bad, 修正されていない語を Good というように分類したものを語単位の品質スコアとすることもできる [11][15][16]。

ステップ 3　QE 向け事前学習 : ステップ 2 で生成した疑似 QE データを用いて、文単位及び語単位の品質スコアの推定器のパラメタを学習するとともに、多言語言語モデルのパラメタを更新する。具体的にはまず、学習用の QE データ中の個々の < 原文 , MT 訳 , 品質スコア > というデータのうち、< 原文 , MT 訳 > の対を QE モデルに入力し、得られた予測スコアと正解の品質スコアの誤差を計算する。誤差としては、回帰モデルの場合は二乗平均誤差、分類モデルの場合は交差エントロピーが用いられる。そして、誤差逆伝播法によって、この誤差が小さくなるようにパラメタを更新する。図 2 のように単一のモデルで文単位の QE と語単位の QE の両方を行う場合は、マルチタスク学習が用いられる。例えば、各スコアの誤差に基づいてパラメタを更新する操作を交互に行うことや、両方の誤差を統合してパラメタを更新することが考えられる。

ステップ 4　パラメタの洗練 : 人手で作成された QE データを用いて、ステップ 3 と同様にして QE モデルのパラメタを洗練する。

2.4　課題

既存の QE モデルにおいて、文単位のスコアの推定器と語単位のスコアの推定器は、< 原文 , MT 訳 > の各語のベクトルを共有するものの、互いにはインタラ

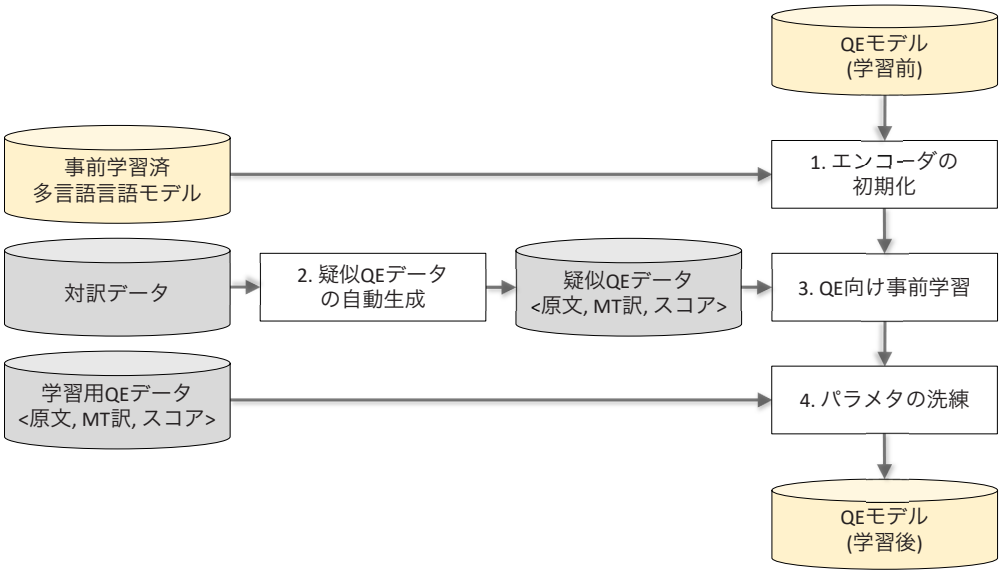

図 3　QE モデルの学習手順

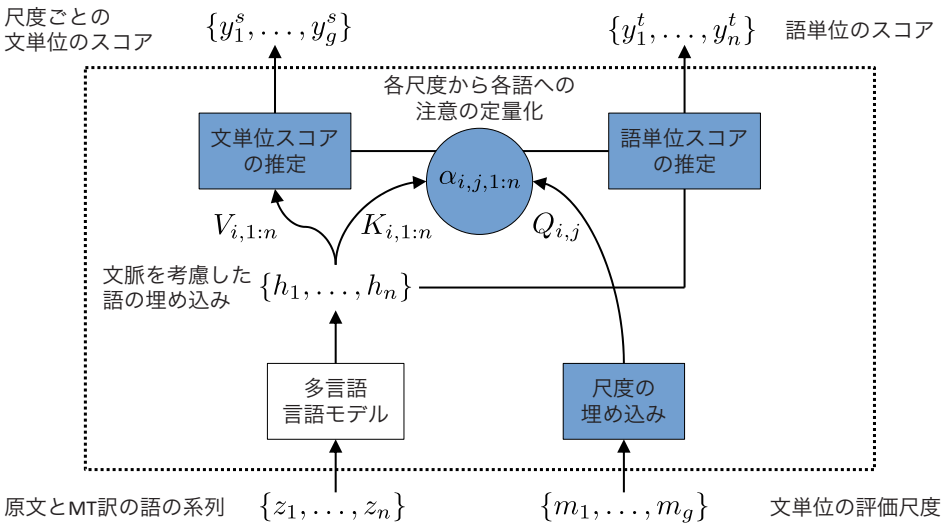

図 4　NICT で開発した QE モデルの構成

クションを行わずに学習が行われる。そのため、文単位のスコアの推定結果と語単位のスコアの推定結果が矛盾してしまうこともある。例えば文単位のスコアの推定器は、原文と MT 訳の合計 n 個のベクトルをプーリングして参照するのみであるため、MT にとって翻訳が困難な固有表現などの低頻度語や、学習用の QE データにおける言語表現やスコアの偏りに過度に影響を受けてしまう。

3　NICT で開発した QE モデル

2.4 で述べた課題に対処するために、我々は図 4 に示す新たな QE モデルを提案した。このモデルは、文単位の品質スコアと語単位の品質スコアの推定根拠として、複数の文単位の評価尺度の各々が < 原文 , MT 訳 > 中のどの語にどの程度注意しているかを捉えて利用する。これにより、従来手法と比べて両スコアの一貫性を担保することをねらっている。また、語単位の QE データがない場合でも語単位の品質スコアの予測が可能になる。

提案手法では従来手法と同様に、多言語言語モデルを用いて < 原文 , MT 訳 > の対を d 次元ベクトルの列 $(h_1, \ldots, h_n)$ に変換する。この点以外（図中の青塗りの部分）は提案手法で新たに導入した要素である。以下、各々について述べる。

尺度の埋め込み（metric embedding）: 正規の QE データで用いられる主観評価のスコアや、疑似 QE データで用いられる自動評価尺度のスコアなど、文単位の複数の評価尺度を学習に利用する。まず、g 個の尺度の各々を表すワンホットベクトル $\{m_1, \ldots, m_g\}$

を定める。そしてそれらを d 次元ベクトルに変換する。

$$e_j = m_j \cdot W^m \tag{1}$$

尺度からの語への注意の定量化（metric attention）：各尺度が、そのスコアの根拠として＜原文，MT 訳＞中のどの語をどの程度注意しているかを定量化する。ここで尺度ごとに注意機構を持たせ、各々において u 個の注意ヘッドに d 次元のうち異なる d/u 次元ずつを参照させる。

- 尺度埋め込みを変換して、個々の尺度（$1 \leq j \leq g$）、個々の注意ヘッド（$1 \leq i \leq u$）のクエリを生成する。

$$Q_{i,j} = e_j \cdot W_i^Q \tag{2}$$

- 注意対象である個々の語のベクトルを変換して注意機構のキー、バリュー対を得る。

$$ff_{1:n} = ReLU(h_{1:n} \cdot W^{s,D_1}) \cdot W^{s,D_2} \tag{3}$$

$$K_{i,1:n} = ff_{1:n} \cdot W_i^K \tag{4}$$

$$V_{i,1:n} = ff_{1:n} \cdot W_i^V \tag{5}$$

ここで $ReLU(\cdot)$ は活性化関数の一種である。

- 各尺度の各注意ヘッドから各語への注意の度合いを定量化する。

$$\alpha_{i,j,1:n} = \sigma\left(\frac{Q_{i,j}K_{i,1:n}^T}{\sqrt{d/u}}\right) \tag{6}$$

ここで $\sigma(\cdot)$ はシグモイド関数である。

文単位の品質スコアの予測：尺度ごとの予測スコアを次の手順で計算する。

- 注意の度合いに応じて語の埋め込みのバリューを混合する。

$$attn_{i,j} = \sum \alpha_{i,j,1:n} \cdot V_{i,1:n} \tag{7}$$

- u 個の注意ヘッドの情報を統合し、全尺度に共通の変換及び各尺度に固有の変換を施す。

$$y_j^s = (attn_{1,j} \oplus \cdots \oplus attn_{u,j}) \cdot W^O \cdot W_j^s \tag{8}$$

2 つの変換行列に分割することによって、可能な限り多くのパラメタを個々の尺度に依存しないようにしている。また、文単位のスコアの予測に注意機構を用いることにより、プーリングとは異なり、すべての語に対する情報を集約して使用できる。

語単位の品質スコアの予測：各語に対する予測スコアを次の手順で計算する。

- 各尺度について u 個の注意ヘッドの情報を統合して注意量を計算する。

$$y_{j,1:n}^{t'} = (\alpha_{1,j,1:n} \oplus \cdots \oplus \alpha_{u,j,1:n}) \cdot W^{t,H} \tag{9}$$

- 語のベクトルと g 個の尺度の各々の注意量を統合した上で、スコアに変換する。

$$y_{1:n}^t = (h_{1:n} \oplus y_{1,1:n}^{t'} \oplus \cdots \oplus y_{g,1:n}^{t'}) \cdot W^{t,O} \tag{10}$$

このQEモデルも、**2.3** で述べた手順で学習できる。学習ステップ 3 及び 4 で新たに学習するパラメタは、式（1）～（5）及び式（8）～（10）におけるすべての変換行列 W である。これらのうち、各尺度に固有のパラメタは式（1）の W^m と式（8）の W_j^s のみである。したがって、疑似 QE データを用いる学習ステップ 3 では人間の主観評価のスコアは利用できないが、人間の主観評価のスコアが使える学習ステップ 4 では、この新たな尺度に専用のパラメタのみを追加学習すれば良い。また、語単位のスコアの予測に固有のパラメタは式（9）の $W^{t,H}$ と式（10）の $W^{t,O}$ のみである。語単位の QE データは作成コストが高いため、人手で作成されたデータを用いて QE モデルを学習できるとは限らないが、これらのパラメタは学習ステップ 3 において疑似 QE データからも学習できる。

4 Explainable QE シェアードタスクにおける評価実験

2021 年、国際ワークショップ Eval4NLP（Evaluation & Comparison of NLP Systems）において Explainable QE というシェアードタスクが開催された [7]。我々は、**3** で述べたシステムを用いてこのシェアードタスクに参加し、良好な成績を収めた。

4.1　シェアードタスクの仕様

Explainable QE は、所与の＜原文，MT 訳＞の対に対して、文単位及び語単位の品質スコアを予測するタスクである。今回のシェアードタスクではモデルの学習のためのデータとして＜原文，MT 訳，文単位の品質スコア＞という形式の QE データのみが配布された。すなわち、文単位の QE モデルを教師あり学習によって得ながらも、その推定の根拠として語単位の品質スコアを精度良く推定することが求められた。

タスクの主催者から配布された QE データの記述統計を表 1 に示す。エストニア語 → 英語とルーマニア語 → 英語の 2 つの翻訳方向については、評価用の QE データに加えて、モデルの学習用及び検証用の QE データが配布された。一方、ロシア語 → ドイツ語とドイツ語 → 中国語の 2 つの翻訳方向については、評価用の QE データのみが配布された。これらの翻訳方向についてはゼロショット学習が想定されている。

表 1　公式の QE データの記述統計：“k” は 1,000 を表す。中国語は文字を単位として算出。

翻訳方向	用途	文数	のべ語数	語の異なり数
エストニア語 → 英語（ET→EN）	学習	7 k	98.1 k / 136.6 k	28.9 k / 14.6 k
	検証	1 k	14.4 k / 20.1 k	6.9 k / 4.7 k
	評価	1 k	14.0 k / 19.6 k	6.9 k / 4.7 k
ルーマニア語 → 英語（RO→EN）	学習	7 k	120.2 k / 123.3 k	23.5 k / 15.2 k
	検証	1 k	17.3 k / 17.7 k	6.4 k / 4.8 k
	評価	1 k	17.4 k / 17.8 k	6.3 k / 4.8 k
ロシア語 → ドイツ語（RU→DE）	評価	1 k	25.4 k / 28.8 k	10.2 k / 7.5 k
ドイツ語 → 中国語（DE→ZH）	評価	1 k	24.9 k / 52.8 k	8.4 k / 2.2 k

表 2　疑似 QE データの記述統計：“k” は 1000、“M” は 100 万を表す。中国語は文字を単位として算出。

翻訳方向	文数	のべ語数	語の異なり数
エストニア語 → 英語（ET→EN）	24.9 M	322.5 M / 411.0 M	4.8 M / 2.8 M
ルーマニア語 → 英語（RO→EN）	42.1 M	600.5 M / 601.2 M	4.0 M / 3.6 M
ロシア語 → ドイツ語（RU→DE）	19.5 M	256.9 M / 262.7 M	4.4 M / 4.4 M
ドイツ語 → 中国語（DE→ZH）	19.8 M	422.8 M / 708.1 M	4.5 M / 3.3 k

4.2　使用した言語資源

3 で述べたモデルを学習するために、2 種類の事前学習済モデル及び対訳データを使用した。1 種類目の事前学習済モデルは多言語言語モデルである。HuggingFace Transformers ライブラリ [21] として公開されている XLM-RoBERTa [22] のモデル *xlm-roberta-large* を使用した。2 種類目の事前学習済モデルは、対訳データを翻訳して疑似 QE データを生成するために使用した学習済の MT モデルである。具体的には、エストニア語 → 英語及びルーマニア語 → 英語の 2 つの翻訳方向については、国際ワークショップ WMT 2020 の QE シェアードタスク [6] において主催者から提供されたニューラル MT モデルを使用した。ロシア語 → ドイツ語については、mBART50 [23][24] を使用した。ドイツ語 → 中国語については、mBART50 による翻訳の品質が低かったため、mBART50 のパラメタを OPUS [25] における NewsCommentary 及び MultiUN の 2 種類の対訳データを用いて洗練してから使用した。

疑似 QE データを生成するための対訳データとして、WMT2020 のニュース翻訳タスク [26] において主催者から提供された対訳データ及び OPUS [25] の対訳データを使用した。ただし、ドイツ語と中国語の言語対に関しては、次の手順で新たに疑似対訳データを生成して用いた。まず、CommonCrawl 及び NewsCrawl の 2018 ～ 2020 年の中国語テキストデータを収集した。次に、ニューラル MT の学習フレームワーク Marian [27] 及び WMT 2020 の QE シェアードタスク [6] において主催者から提供された中国語・英語の対訳データを用いて、中国語 → 英語のニューラル MT モデルを学習し、それを用いて上記の中国語テキストデータを英語に翻訳した。続けて、WMT 2020 の QE シェアードタスク [6] において主催者から提供された英語 → ドイツ語のニューラル MT モデルを用いて上で得た英語の MT 訳をドイツ語に翻訳し、元の中国語の各文と対応付けたものを対訳データとした。

4.3　QE モデルの学習とアンサンブル

2.3 で述べた手順で複数の QE モデルを学習した。

ステップ 1　*xlm-roberta-large* でエンコーダを初期化した。

ステップ 2　疑似 QE データの自動生成：**4.2** で述べた対訳データの起点言語の文を、同じく **4.2** で述べた MT モデルに入力して MT 訳を生成し、対訳データの目標言語の人間訳と比較して文単位・語単位の品質スコアを計算し、QE データとした。文単位の品質スコアの尺度としては、BLEU スコア [3]、chrF スコア [19]、TER スコア [20] を用いた。また、語単位の品質スコアとしては、原文と MT 訳、原文と人間訳、MT 訳と人間訳の各々の語の対応付けに基づいて原文及び MT 訳の各語を {Bad, Good} のいずれかに分類して用いた。語の対応付けには TER 及び fast_align [28] を用いた。疑似 QE データの記述統計を表 2 に示す。表 1 の学習用データと比べて、文数で約 3,000 ～ 6,000 倍、のべ語数で約 3,000 ～ 5,000 倍の規模である。

ステップ 3　疑似 QE データ（表 2）を用いて合計 12 個

表 3　アンサンブルして使用したモデルの一覧

語単位の QE（原文と MT 訳の各々）	文単位の QE
翻訳方向専用（シード 1）の AUC 最大	翻訳方向専用（シード 1）の ρ 最大
翻訳方向専用（シード 2）の AUC 最大	翻訳方向専用（シード 2）の ρ 最大
多言語（シード 1）の AUC 最大	多言語（シード 1）の ρ 最大
多言語（シード 2）の AUC 最大	多言語（シード 2）の ρ 最大
多言語（シード 3）の AUC 最大	多言語（シード 3）の ρ 最大
多言語（シード 4）の AUC 最大	多言語（シード 4）の ρ 最大
翻訳方向専用（シード 1）の AP 最大	翻訳方向専用（シード 1）の RMSE 最小
翻訳方向専用（シード 2）の AP 最大	翻訳方向専用（シード 2）の RMSE 最小

の QE モデルを学習した。まず、4 つの翻訳方向の各々について、2 種類のランダムシードと学習率の組み合わせを用いて合計 8 個の QE モデルを得た。これらの学習は、2 エポックで停止した。次に、4 つの翻訳方向のすべての疑似 QE データを用いて、4 種類のランダムシードと学習率で合計 4 個の多言語 QE モデルを得た。これらのモデルの学習は 1 エポックで停止した。

ステップ 4　学習用及び検証用の QE データ（表 1）が配布されたエストニア語 → 英語及びルーマニア語 → 英語の 2 つの翻訳方向についてのみ、ステップ 3 で得た各翻訳方向に固有の 2 つの QE モデル及び 4 つの多言語 QE モデルを洗練した。文単位の品質スコアの尺度として、正規の QE データにおいて人間が与えた 0 ～ 1 の実数の評価値（Direct Assessment; DA）に加えて、ステップ 3 で用いた BLEU スコア、chrF スコア、TER スコアも用いた。一方、**4.1** で述べたタスク仕様に従い語単位の品質スコアの正解は参照しなかった。各 QE モデルについてパラメタの洗練を 20 エポック行った。

評価用データに対して QE を実施する際は、上述の手順で得た複数のモデルをアンサンブルして使用した。アンサンブルの構成要素を表 3 に示す。各モデルは、検証用データに対する各評価指標に基づいて選択した。

- 語単位の評価指標（4 つ）: 原文と MT 訳の各々に対する再現率 – 精度曲線の下側面積（AUC）及び平均精度（AP）
- 文単位の評価指標（2 つ）: Pearson の積率相関係数（ρ）及び二乗平均平方根誤差（RMSE）

ロシア語 → ドイツ語とドイツ語 → 中国語の 2 つの翻訳方向については、ステップ 2 で生成した疑似 QE データから 2,000 文をサンプルして用い、ステップ 3 で得た QE モデルからモデルを選択した。エストニア語 → 英語及びルーマニア語 → 英語の 2 つの翻訳方向については、ステップ 4 で得た QE モデルからモデルを選択した。またこれらの 2 つの翻訳方向については、表 3 のアンサンブルに加えて、表 3 の先頭の 2 つの QE モデルのみをアンサンブルしたものをベースラインとして評価した。

4.4　評価結果

シェアードタスクにおける公式の性能評価にならい評価用データに対する語単位の QE の性能を AUC 及び AP で、文単位の QE の性能を Pearson の積率相関係数（ρ）で評価した結果を表 4 に示す。シェアードタスクの主催者が用意した 3 種類の公式ベースラインの評価結果も合わせて示す。いずれの尺度についても、値が大きいほど性能が良いことを表す（最大値は 1）。語単位の QE については、我々のシステムは 4 つの翻訳方向すべてにおいて、3 種類のベースラインよりも優れた性能を達成した。また、文単位の QE においても、ドイツ語 → 中国語以外のタスクにおいてベースラインよりも優れた性能を達成した。

シェアードタスクの主催者による公式の性能評価結果 [7] によると、我々が提出したシステム（表 3 のアンサンブル）は、4 つの翻訳方向のうちエストニア語 → 英語を除く 3 つにおいて、提出された他の 7 つのシステムよりも顕著に良い性能を達成した。エストニア語 → 英語に関しては、我々が提出したシステムの成績は 3 位であった。

4.5　QE 結果の説明性

Explainable QE シェアードタスクが開催された背景に、文単位の品質スコアの予測結果に対する説明性という課題があった。そこで我々は、各尺度から各語への注意量（式 (9)）を、文単位の品質スコアの根拠として使用することを提案した。図 5 に、Eval4NLP のシェアードタスクの検証用 QE データから抽出したエストニア語 → 英語の 2 例（上段）とルーマニア語 → 英語の 2 例（下段）に対する各尺度の注意量を示す。横軸には < 原文 , MT 訳 > の各語を、縦軸には我々が QE モデルの訓練に使用した各尺度（DA，TER，chrF，

表 4　評価用データに対する QE の性能評価結果：太字は最も良い数値

翻訳方向	モデル	原文の語単位		MT 訳の語単位		文単位
		AUC	AP	AUC	AP	ρ
ET→EN	ランダムベースライン	0.488	0.338	0.496	0.358	-0.029
	公式ベースライン 1	0.545	0.440	0.624	0.536	0.772
	公式ベースライン 2	0.535	0.370	0.616	0.441	0.494
	我々のベースライン	0.926	0.848	0.887	0.808	0.793
	我々の提出システム	**0.932**	**0.852**	**0.896**	**0.824**	**0.845**
RO→EN	ランダムベースライン	0.501	0.281	0.515	0.312	0.017
	公式ベースライン 1	0.478	0.351	0.635	0.523	0.899
	公式ベースライン 2	0.535	0.293	0.667	0.536	0.695
	我々のベースライン	0.937	0.826	0.942	0.860	0.855
	我々の提出システム	**0.947**	**0.851**	**0.946**	**0.869**	**0.918**
RU→DE	ランダムベースライン	0.506	0.340	0.494	0.309	-0.017
	公式ベースライン 1	0.535	0.427	0.403	0.263	0.498
	公式ベースライン 2	0.522	0.356	0.523	0.329	0.252
	我々の提出システム	**0.922**	**0.804**	**0.927**	**0.829**	**0.679**
DE→ZH	ランダムベースライン	0.499	0.300	0.495	0.293	0.000
	公式ベースライン 1	0.486	0.317	0.461	0.271	**0.335**
	公式ベースライン 2	0.474	0.288	0.545	0.333	0.176
	我々の提出システム	**0.847**	**0.645**	**0.849**	**0.679**	0.286

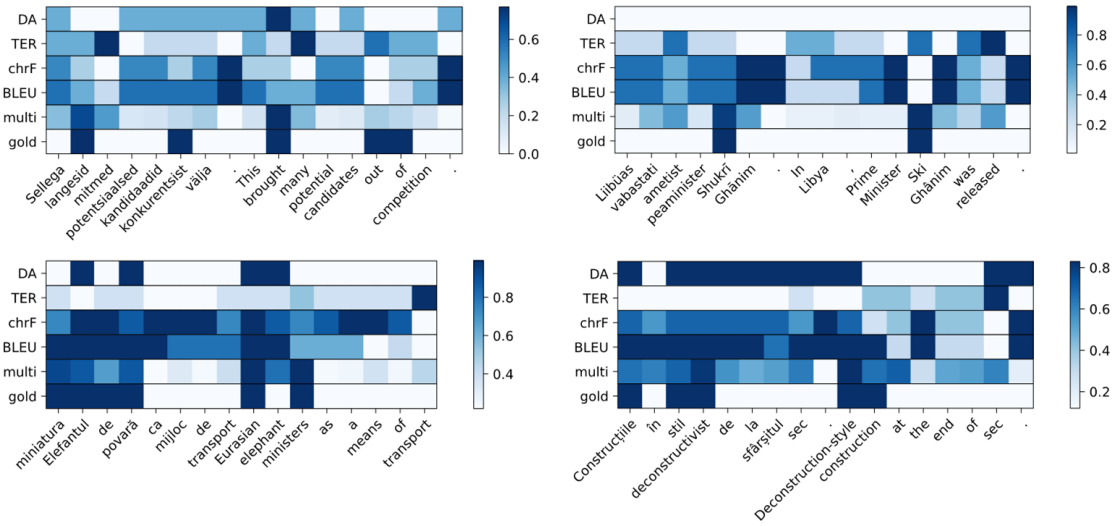

図 5　各尺度から各語への注意量

BLEU; 式 (9))、QE モデルの予測値 (multi; 式 (10)) 及び正解 (gold) を表示している。正解については 0 または 1 の 2 値であり、それ以外は 0 ～ 1 の実数値である。QE モデルの学習に使用した 4 つの尺度の注意の分布が必ずしも類似していないことは、尺度ごとに異なる特徴を捉えていることを示唆する。またこれらの尺度は、正解とは大きく異なる。しかしながら、我々の QE モデルはそれらを統合することにより、正解に比較的近い品質スコア (multi) を予測できている。

5　おわりに

NICT では現在、MT 技術の研究開発と並行して、MT システムを実際に利用する際のサポート技術としての QE 技術の研究開発に取り組んでいる。本稿では、QE 技術の概要及び近年の一般的なアプローチについて述べ、我々が開発した最新の手法を紹介した。国際ワークショップ Eval4NLP の Explainable QE シェアードタスクにおける成果をふまえ、今後は NICT におけ

るサービスを視野に、日本語と英語の間の翻訳を対象としたQE技術の実現に取り組む予定である。

【参考文献】

1 松田 繁樹, 林輝 昭, 葦苅 豊, 志賀 芳則, 柏岡 秀紀, 安田 圭志, 大熊 英男, 内山 将夫, 隅田 英一郎, 河井 恒, 中村 哲, “多言語音声翻訳システムVoiceTraの構築と実運用による大規模実証実験,” 電子情報通信学会論文誌D, vol.J96-D, no.10, pp.2549–2561, 2013.

2 内山 将夫, “みんなの自動翻訳@TexTra®,” 情報通信研究機構研究報告, 本特集号, 2-3-1, 2022.

3 K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp.311–318, 2002.

4 B. Marie, A. Fujita, and R. Rubino, “Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 756 Papers,” Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp.7297–7306, 2021.

5 J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing, “Confidence Estimation for Machine Translation,” Proceedings of the 20th International Conference on Computational Linguistics, pp.315–321, 2004.

6 L. Specia, F. Blain, M. Fomicheva, E. Fonseca, V. Chaudhary, F. Guzmán, and A. F. T. Martins, “Findings of the WMT 2020 Shared Task on Quality Estimation,” Proceedings of the 5th Conference on Machine Translation, pp.743–764, 2020.

7 M. Fomicheva, P. Lertvittayakumjorn, W. Zhao, S. Eger, and Y. Gao, “The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results,” Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems, pp.165–178, 2021.

8 International Organization for Standardization, “ISO 18587. Translation Services – Post-editing of Machine Translation Output – Requirements,” International Organization for Standardization, 2017.

9 Memsource, “機械翻訳品質評価(MTQE),” 参照日: July 21, 2022. https://www.memsource.com/ja/features/translation-quality-estimation/

10 総務省, “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,” 参照日: May 1, 2015. https://www.soumu.go.jp/main_content/000356284.pdf

11 L. Liu, A. Fujita, M. Utiyama, A. Finch, and E. Sumita, “Translation Quality Estimation Using Only Bilingual Corpora,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.25, no.9, pp.1762–1772, 2017.

12 A. Fujita and E. Sumita, “Japanese to English/Chienese/Korean Datasets for Translation Quality Estimation and Automatic Post-Editing,” Proceedings of the 4th Workshop on Asian Translation, pp.79–88, 2017.

13 R. Rubino and E. Sumita, “Intermediate Self-supervised Learning for Machine Translation Quality Estimation,” Proceedings of the 28th International Conference on Computational Linguistics, pp.4355–4360, 2020.

14 R. Rubino, A. Fujita, and B. Marie, “Error Identification for Machine Translation with Metric Embedding and Attention,” Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems, pp.146–156, 2021.

15 D. Lee, “Two-Phase Cross-Lingual Language Model Fine-Tuning for Machine Translation Quality Estimation,” Proceedings of the 5th Conference on Machine Translation, pp.1024–1028, 2020.

16 Y.-L. Tuan, A. El-Kishky, A. Renduchintala, V. Chaudhary, F. Guzmán, and L. Specia, “Quality Estimation without Human-labeled Data,” Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp.619–625, 2021.

17 R. Sennrich, B. Haddow, and A. Birch, “Neural Machine Translation of Rare Words with Subword Units,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.1715–1725, 2016.

18 Taku Kudo and John Richardson, “SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp.66–71, 2018.

19 M. Popović, “chrF: Character n-gram F-score for Automatic MT Evaluation,” Proceedings of the 1st Conference on Machine Translation, pp.392–395, 2015.

20 M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A Study of Translation Edit Rate with Targeted Human Annotation,” Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pp.223–231, 2006.

21 T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp.38–45, 2020.

22 A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised Cross-lingual Representation Learning at Scale,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.8440–8451, 2020.

23 Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, “Multilingual Denoising Pre-training for Neural Machine Translation,” Transactions of the Association for Computational Linguistics, vol.8, pp.726–742, 2020.

24 Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, “Multilingual Translation with Extensible Multilingual Pretraining and Finetuning,” arXiv preprint arXiv:2008.00401, 2020.

25 J. Tiedemann, “OPUS: Parallel Corpora for Everyone,” Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products, p.384, 2016.

26 L. Barrault et al., “Findings of the 2020 Conference on Machine Translation (WMT20),” Proceedings of the 5th Conference on Machine Translation, pp.1–55, 2020.

27 M. Junczys-Dowmunt et al., “Marian: Fast Neural Machine Translation in C++,” Proceedings of ACL 2018, System Demonstrations, pp.116–121, 2018.

28 C. Dyer, V. Chahuneau, and N. A. Smith, “A Simple, Fast, and Effective Reparameterization of IBM Model 2,” Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.644–648, 2013.

ルビノ ラファエル

ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的翻訳技術研究室
主任研究員
Ph.D.
自然言語処理

【受賞歴】
2021年　Best Overall Approach, Eval4NLP 2021 Explainable Quality Estimation Shared Task
2021年　Outstanding Paper Award, ACL-IJCNLP 2021

藤田 篤（ふじた あつし）

ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的翻訳技術研究室
主任研究員
博士(工学)
計算言語学、自然言語処理

【受賞歴】
2021年　Best Overall Approach, Eval4NLP 2021 Explainable Quality Estimation Shared Task
2021年　Outstanding Paper Award, ACL-IJCNLP 2021