タイトル	WMT20 評価タスクデータにおける自動評価法のメタ評価
著者	越前谷, 博; Echizen'ya, Hiroshi; 荒木, 健治; Araki, Kenji
引用	工学研究：北海学園大学大学院工学研究科紀要(21): 19-32
発行日	2021-12-24

研究論文

WMT20 評価タスクデータにおける自動評価法のメタ評価

越前谷　博* ・ 荒木健治**

Meta-evaluation of Automatic Evaluation Metrics by WMT20 Metrics Task Data

Hiroshi Echizen'ya* and Kenji Araki**

Abstract

本報告では機械翻訳分野の国際会議である WMT20 における Metrics Shared Task のデータを用いて行ったメタ評価について述べる．WMT20 の評価タスクでは様々な自動評価法が用いられており，自動評価法の現状を把握するために有効である．メタ評価はシステムレベルとセグメントレベルの 2 つの指標を用いて行った．また，我々が従来より提案している 2 つの自動評価法 IMPACT と WE_WPI も含めることで提案手法の他手法に対する位置付けについても検証した．その結果，WE_WPI は to-English（他言語から英語の翻訳）の言語ペアにおいては他手法に比べて比較的上位の評価精度を有していることを確認できた．

1　はじめに

機械翻訳のための自動評価法の研究は従来より盛んに行われており，機械翻訳の変遷と共に多様化している．特に，ここ 10 年においてはニューラル翻訳［1，2，3，4］の急速な普及により，文や単語の意味表現に基づく自動評価法の研究が盛んに行われている．自動評価法の研究は 1990 年代に登場した統計的機械翻訳［5，6，7］により，本格的に研究が始まった．統計的機械翻訳はソースコードが一般公開されており，膨大な対訳コーパスを共有することで容易に稼働できるため，多くの研究者が開発に取り組んだ．その結果，開発サイクルを早めるためにはシステムに対する評価速度が重要となり，自動でシステム訳を評価可能なシステムが不可欠となった．そのような背景のもと人手のように評価の揺れが生ぜずに，低コストかつ高速で評価が可能な自動評価法の研究が盛んに行われるようになった．

統計的機械翻訳が主流だった 1990 年代から 2000 年代にかけては，表層情報及び WordNet のような静的な語彙情報を用いた自動評価法が主流であった．自動評価法のディファクトスタンダートとなっている BLEU［8］は 2002 年に提案された．BLEU は参照訳と呼ばれる正解訳とシステム訳との間で n-gram 一致率を求めることでシステム訳に対する評価スコアを算出するシステムである．このように自動評価法では参照訳を用いた評価が一般的である．この BLEU がきっかけとなり多くの自動評価法が提案されるようになった．代表的なものとしては TER［9］，METEOR［10］，RIBES［11］などが挙げられる．我々も表層情報に基づく自動評価法として IMPACT（**I**ntuitive com**M**on **PA**rts **C**on**T**inum）［12，13，14，15，16，17］を提案している．IMPACT は最長共通部分列［18］を用いて，参照訳とシステム訳間に存在するチャンクを自動的に決定し，そのチャンクの位置と長さに基づいてチャンク列を再帰的に決定することで評価スコアを求める．

2010 年代に入ると統計的機械翻訳に取って代わり，ディープラーニングの技術発展に伴いニューラル翻訳の研究が急速に行われるように

* 北海学園大学大学院工学研究科電子情報生命工学専攻
Graduate School of Engineering (Electronics, Information and Life Science Eng.), Hokkai-Gakuen University
** 北海道大学大学院情報科学研究院
Faculty of Information Science and Technology, Hokkaido University

なった．ニューラル翻訳ではそれまでの表層情報や静的な言語資源に依存することなく，単語や文の意味をベクトルで表現することでより高度な言語処理が可能となった．このようなニューラル翻訳の進展に伴い，自動評価法もそれまでの表層情報や静的な言語知識を用いたものだけではなく，ディープラーニングより学習したモデルにより得られる単語分散表現や文ベクトルに基づく手法が提案されるようになった．我々もこのような状況下において単語分散表現に基づく自動評価法として WE_WPI（**W**ord **E**mbedding-based Automatic MT Evaluation using **W**ord **P**osition **I**nformation）[19] を提案している．WE_WPI は Earth Mover's Distance（EMD）[20, 21, 22] を言語処理に適応させることでシステム訳に対する評価スコアを得る．EMD を言語処理に適用するためには 3 つのパラメータを定義する必要がある．WE_WPI では特徴量，その特徴量に対する重み，そして，特徴量間における距離計算の 3 つのパラメータにそれぞれ単語分散表現，*tf・idf*，そして，コサイン距離を適用している．

本報告では，提案手法である IMPACT と WE_WPI を含めた様々な自動評価法を用いて行なったメタ評価について述べる．メタ評価には WMT20（The 2020 Conference on Machine Translation）[23, 24] の Shared Task の一つである Metrics のデータを使用する．国際会議 WMT は毎年開催され，機械翻訳に関する複数の Shared Task を扱っている．評価タスクは Shared Task としては開催初期から取り挙げられており，機械翻訳において非常に重要なタスクとして認識されている．この WMT20 の評価タスクデータによるメタ評価の結果，提案手法 WE_WPI は他言語から英語の翻訳を対象とする to-English においてシステムレベル，セグメントレベルの両方で他手法との比較において評価精度が上位に位置することを確認できた．

2　機械翻訳における自動評価法

WMT20 で使用された自動評価法と我々の 2 つの提案手法 IMPACT，WE_WPI のそれぞれの特徴をまとめたものを**表 1** に示す．表 1 の "learned" は推定モデルやベクトルを得るためのモデルを学習により構築する必要があるかどうかを意味し，"yes" はその必要があることを示している．モデルの学習を前提とした自動評価法は学習によりタスクに順応可能なため一般的には評価精度が高くなるが，学習のためのデータや学習時間が別途必要になることがデメリットとなる．また，表 1 の "types" における "src-based" はスコア計算において参照訳を必要としない自動評価法であることを示している．近年では参照訳を準備することのコストを回避するために原文のみを用いる自動評価法の研究も行われている．

2.1　WMT20 における自動評価法

2.1.1　ベースライン

表 1 の sentBLEU，BLEU，TER，chrF，そして，chrF＋＋は WMT20 ではベースラインとして位置付けられている．sentBLEU は BLEU をセグメントレベルでも評価スコアを算出できるように改良されたものである．TER は挿入，削除，移動，そして，置換の 4 つの編集操作の数に基づく自動評価法である．chrF は参照訳とシステム訳の比較において単語の n-grams ではなく文字の n-grams を用いた自動評価法である．また，chrF＋＋は単語の unigram と bigram に文字の n-grams を加えた自動評価法である．

2.1.2　BERT-base-L2, BERT-large-L2, mBERT-L2

BERT-base-L2，BERT-large-L2，そして，mBERT-L2 の 3 つの自動評価法はいずれも BERT［25］を fine-tuning することで得られる．その際には，WMT15 から WMT18 の評価タスクデータを用いている．BERT-base-L2 は 12 レイヤーの Transformer アーキテクチャに対して，英語データを用いて pre-trained を行ったものである．mBERT-L2 は 102 言語の Wikipedia データを用いて学習を行なっている．BERT-large-L2 は 24 レイヤーを用いており，WMT20 では to-English のみのメタ評価を行っている．

2.1.3　BLEURT, BLEURT-extended, Yisi-combi, bleurt-Yisi-combi

BLEURT は 2 つのデータを用いて BERT-based の回帰モデルのための学習を行なっている．一つはランダムに取得した 100 万ペアのデータであり，もう一つは WMT15 から WMT19 より取得したデータである．BLEURT-extended

表 1 ：WMT20 における自動評価法と提案手法のそれぞれの特徴

metric	features	learned	types
sentBLEU	n-grams		Reference-based
BLEU［8］	n-grams		Reference-based
TER［9］	edit-distance		Reference-based
chrF［26］	character n-grams		Reference-based
chrF＋＋［27］	character n-grams		Reference-based
parbleu［28］	paraphrases		Reference-based
parachrf＋＋［28］	paraphrases		Reference-based
paresim［28］	paraphrases	yes	Reference-based
prism［29］	paraphrases		Reference-based
CharacTER［30］	character edit distance		Reference-based
EED［31］	character edit distance		Reference-based
SWSS＋METEOR［32］	semantic similarity		Reference-based
MEE［33］	word embeddings		Reference-based
YiSi［35］［36］	contextual word embeddings		Reference-based
BERT-base-L2［25］	contextual word embeddings	yes	Reference-based
BERT-large-L2［25］	contextual word embeddings	yes	Reference-based
mBERT-L2［25］	contextual word embeddings	yes	Reference-based
BLEURT［25］	contextual word embeddings	yes	Reference-based
BLEURT-extended［37］	contextual word embeddings	yes	Reference-based
Yisi-combi［25］	contextual word embeddings	yes	Reference-based
bleurt-combi［25］	contextual word embeddings	yes	Reference-based
COMET［38］	predictor-estimator model	yes	Reference-based
COMET-Rank［38］	predictor-estimator model	yes	Reference-based
COMET-HTER［38］	predictor-estimator model	yes	Reference-based
COMET-2R［38］	predictor-estimator model	yes	Reference-based
COMET-MQM［38］	predictor-estimator model	yes	Reference-based
BAQ，EQ	?	?	Reference-based
COMET-QE［38］	predictor-estimator model	yes	src-based
OpenKiwi-Bert［39］	predictor-estimator model	yes	src-based
OpenKiwi-XLMR［39］	predictor-estimator model	yes	src-based
YiSi-2［40］	contextual word embeddings		src-based
IMPACT	longest common subsequence		Reference-based
WE_WPI	word embeddings		Reference-based

は WMT15 から WMT19 の人手評価に基づき BERT-based の回帰モデルを学習したものである．Yisi-combi は WMT データで fine-tuned された mBERT モデルに対して YiSi-1 を用いたものである．また，bleurt-combi は複数参照訳を用いて算出されたスコアに対して，YiSi-1，YiSi-2，そして，BLEURT のスコアを組み合わせたものである．

2.1.4　CharacTER

CharacTER は TER を文字レベルで適用した自動評価法である．編集操作は参照訳と一致するまで行われ，システム訳の長さで正規化される．ただし，移動の操作においては単語レベルで行われる．したがって，参照訳と移動の操作後のシステム訳との間のレーベンシュタイン距離は文字レベルで求められることになる．さらに CharacTER は英語―ロシア語の言語ペアを除き，トークン化されていない参照訳とシステム訳に対して適用される．

2.1.5　COMET

COMET は推定モデルおよび翻訳ランキングモデルを用いて構築される．ニューラルモデルは参照訳とシステム訳をエンコードするために XLM-RoBERTa を用いる．さらに WMT17 から WMT19 における Direct Assessments（DA）に基づいて学習される推定モデルとなっている．COMET-2R は複数参照訳を処理するために学習された COMET の改良版である．COMET-Rank はシステム訳と原文との間の距離及びシステム訳と参照訳との間の距離を最適化するための翻訳ランキングアーキテクチャを用いている．そして，COMET-HTER と COMET-MQM は同じアーキ

テクチャを用いているが，COMET-HTER は Human-mediated Translation Edit Rate（HTER），COMET-MQM は多次元品質推定に基づいている．

2.1.6 EED

EED は “jump” の操作を含めた拡張版の編集距離を character-based で求めることで評価スコアを算出する．“jump” の操作はスペース文字が見つかったタイミングで行われる．

2.1.7 MEE

MEE のスコアはシステム訳と参照訳の単語分散表現の間の類似度を求めることで得られる．単語分散表現による類似度計算に用いられる単語分散表現は Facebook より提供されている fasttext［34］を用いることで取得する．さらに MEE ではスコア計算には表層的な一致，原形の一致（root match），そして，同義語の一致の 3 つのモジュールを用いている．MEE は既存研究の中で学習済みの fasttext に基づく単語分散表現を用いていることから提案手法 WE_WPI に最も近い自動評価法と言える．しかし，提案手法では文全体の類似度を EMD に基づいて求めている点や EMD に語順情報を導入している点が大きく異なっている．

2.1.8 esim

esim はニューラルモデルに基づく自動評価法である．システム訳と参照訳をベクトルで表現するためにセンテンスアテンションとセンテンスマッチングについてのヒューリスティックスを用いている．

2.1.9 OpenKiwi-Bert，OpenKiwi-XLMR

OpenKiwi-Bert と OpenKiwi-XLMR は品質推定モデルを WMT 評価タスクデータで学習することで実現している．OpenKiwi は WMT20 QE shared task において最先端のモデルとして位置付けられている．

2.1.10 parbleu，parchrf＋＋，paresim

parbleu，parchrf＋＋，そして，paresim は与えれた人手評価と自動生成された言い換えのセットに対して BLEU，charf＋＋，esim の自動評価法をそれぞれ適用することで評価スコアを算出している．

2.1.11 prism

prism は 39 言語の言語ペアからなるデータを学習して構築された多言語のニューラル翻訳に基づいている．セグメントレベルの評価スコアはシステム訳と参照訳の間のデコーディングにより得られる．また，システムレベルの評価スコアはセグメントレベルのスコアの平均値を用いて得られる．

2.1.12 SWSS＋METEOR

SWSS＋METEOR は文中の意味的に重要な単語を抽出することで評価スコアを得る．そこでは意味的に重要な単語を定義するために意味表現フレームワークの UCCA を用いる．そのうえで重み付けされた SWSS と METEOR の組み合わせにより評価スコアを算出する．

2.1.13 YiSi-0，YiSi-1，YiSi-2

YiSi は統一的な意味レベルの機械翻訳の品質評価及び利用可能な様々なレベルの言語資源を有する言語に対する推定尺度に基づいた自動評価法である．YiSi-1 は事前学習されたモデルから単語分散表現を抽出し，語彙の意味レベルでの類似度を求める．そのうえでシステム訳と参照訳の間の意味的な類似度を求める．YiSi-2 は参照訳を必要としない自動評価法となっている．事前学習済みの言語モデルから抽出される単語分散表現を多言語へ写像することで意味的な類似度を求め，評価スコアを得ている．YiSi-0 は迅速に評価スコアを算出できるように YiSi-1 を簡素化したものである．そこでは表層情報を用いており，語彙的な類似度を最長共通部分列を文字レベルで適用することで評価スコアを求める．

2.2 提案手法

本節では我々が提案している自動評価法である IMPACT と WE_WPI について述べる．

2.2.1 IMPACT

IMPACT は単語を単位としたチャンクを再帰的に求めることで語順に着目した自動評価法となっている．システム訳と参照訳の間の最長共通部分列を一意に決定するために共通部分の長さと相対的な位置関係を用いる．一意に決定された共通部分をチャンクと見なし，チャンクの長さに依

存した重みを用いることでチャンクが長く，かつチャンクの出現位置がシステム訳と参照訳間で類似しているほど評価スコアが高くなるようにパラメータにより制御している．また，このチャンクの決定処理を再帰的に行うことで参照訳に対して語順が異なるシステム訳の評価スコアが小さくなるようにパラメータを用いて制御している．

この IMPACT は表層情報のみに基づいているため容易に評価スコアを算出することが可能である．その一方で，表層情報のみに依存するため同義語や活用形による表層上の違いを考慮することができないという問題点もある．

2.2.2　WE_WPI

WE_WPI は単語の意味表現である単語分散表現を Earth Mover's Distance（EMD）に適用した自動評価法である．さらにシステム訳と参照訳の間の語順の違いを EMD の距離行列に反映させることで語順についても考慮した評価が可能となっている．具体的には単語分散表現で表された単語間の距離をコサイン距離より求めるが，その距離に単語アライメントにより得られた単語間の位置の相対的なずれを負の重みとして付与することで対応関係にある単語間であっても出現位置が大きく異なる場合には，距離が大きくなる．

また，EMD は輸送問題における最適解を求めるための手法であり，荷物を輸送するための最適な輸送経路を探索する．そのため EMD では荷物を重みとして定義する必要がある．WE_WPI ではその重みについては文レベルの *tf*・*idf* を計算することで得ている．また，個々の単語の分散表現は事前学習された fasttext を用いて取得する．そのため，単語分散表現を得るための学習済みのモデルさえ取得できれば容易に評価スコアを得ることが可能である．

3　性能評価実験

3.1　実験データ及び実験方法

本報告では自動評価法のメタ評価を行うために WMT20 の評価タスクデータを用いた．WMT は機械翻訳のみを対象とした，最先端の研究発表が行われる国際会議として広く認知されている．さらに WMT では機械翻訳に関する様々なタスクごとにコンペティションを実施することで機械翻訳分野の進展を図っている．評価タスクは 2008 年より WMT で取り挙げられており，非常に重要なタスクとして位置付けられている．

WMT20 の評価タスクでは to-English（他言語から英語への翻訳）については 10 言語ペア，out-of-English（英語から他言語への翻訳）については 8 言語ペアのニュースに関するシステム訳と参照訳のデータが提供されている．to-English においてはチェコ語から英語への翻訳（cs-en），ドイツ語から英語への翻訳（de-en），日本語から英語への翻訳（ja-en），ポーランド語から英語への翻訳（pl-en），ロシア語から英語への翻訳（ru-en），タミル語から英語への翻訳（ta-en），中国語から英語への翻訳（zh-en），イヌクティトゥット語から英語への翻訳（iu-en），クメール語から英語への翻訳（km-en），そして，パシュトー語から英語への翻訳（ps-en）により得られたシステム訳を用いている．out-of-English については英語からチェコ語への翻訳（en-cs），英語からドイツ語への翻訳（en-de），英語から日本語への翻訳（en-ja），英語からポーランド語への翻訳（en-pl），英語からロシア語への翻訳（en-ru），英語からタミル語への翻訳（en-ta），英語から中国語への翻訳（en-zh），そして，英語からイヌクティトゥット語への翻訳（en-iu）により得られたシステム訳を用いている．WMT20 ではこれらの言語ペアのシステム訳に対する参照訳も提供されているため自動評価法を用いて評価スコアを算出することができる．

また，メタ評価ではシステムレベルとセグメントレベルの 2 つの観点からのメタ評価を実施した．具体的には自動評価法が算出した評価スコアと人手評価との間で相関係数を求めた．システムレベルのメタ評価ではピアソンの相関係数を用いた．セグメントレベルのメタ評価ではケンドール τ を用いた．システムレベルのメタ評価においてもケンドール τ を求めているが，ケンドール τ はその特性から少ないデータにおいては複数の自動評価法で同等な値が得られる傾向があり，差別化を図る際には不十分となる．したがって，WMT20 においてはシステムレベルのケンドール τ は参考程度として求めているため，本報告でも同様の扱いとする．

3.2 実験結果

表 2 と**表 3** にシステムレベルのメタ評価の結果を示す．表 2 は to-English におけるシステムレベルのピアソンの相関係数，表 3 は out-of-English におけるシステムレベルのピアソンの相関係数である．また，**表 4** と**表 5** にセグメントレベルのメタ評価の結果を示す．表 4 は to-English におけるセグメントレベルのケンドール τ，表 5 は out-of-English におけるセグメントレベルのケンドール τ である．また，**表 6** と**表 7** にはシステムレベルのケンドール τ を示す．表 6 は to-English，表 7 は out-of-English となっている．

表 2 から表 5 は各言語ペアにおいて “all” と “-out” の 2 種類の相関係数が示されている．“all” は全てのシステム訳を用いた相関係数である．それに対して “-out” は “outlier（外れ値）” を意味し，任意の機械翻訳システムのスコアが他の機械翻訳システムのスコアから大きく離れている場合には過度に高い相関係数を導く可能性があるため “all” から取り除いており，その場合の相関係数となっている．また，メタ評価では統計的有意性に基づく相関係数の有意差を示すために Williams の有意検定［41］を使用している．表中の太字の相関係数は同じ言語ペアにおいて他の自動評価法に対して統計的有意差がなかったことを示している．

3.3 考察

我々が提案している IMPACT と WE_WPI のメタ評価における位置付けについて述べる．IMPACT は表層情報にのみ基づく自動評価法である．そのため，同様のアプローチの自動評価法は表 1 の features が n-grams，edit distance，character n-grams，そして，character edit distance となっている BLEU，TER，chrF，chrF＋＋，CharacTER，そして，EED である．また，WE_WPI は事前学習された単語分散表現モデルの fasttext を用いている．したがって，同様のアプローチとしては表 1 の EED が該当する．そこで，IMPACT と WE_WPI と類似している自動評価法と比較するために上述した自動評価法について全言語ペアの相関係数の平均を求めた．**表 8** に相関係数の平均を示す．

表 8 よりシステムレベルのメタ評価においては表層情報を用いた自動評価法の中では EED が最も高い評価精度を示している．IMPACT はその EED に比べると低い値であるが上位に位置している．特に out-of-English の “-out” は IMPACT が最も高い値を示した．また，単語分散表現を用いる自動評価法においては WE_WPI は MEE よりも高い評価精度を示した．なお，out-of-English の平均値が未記入となっているが，これは相関係数が得られなかった言語ペアが EED，WE_WPI それぞれに存在していたためである．WE_WPI においてはイヌクティトゥット語の fasttext のモデルが取得できなかったため en-iu の評価値が得られず，それに伴い相関係数が得られなかった．

表 8 のセグメントレベルのメタ評価においては表層情報を用いた自動評価法では EED が to-English で最も値が高く，out-of-English では chrF が最も高い値を示した．IMPACT は下位に位置し，十分な評価精度を示すことができなかった．この結果から IMPACT は特にセグメントレベルの評価精度が不十分であることが明らかとなった．一方，単語分散表現を用いた自動評価法においては我々が提案する WE_WPI がセグメントレベルにおいても MEE よりも高い評価精度を示した．WE_WPI は表層情報に基づく自動評価法を含めた場合の比較においても to-English の中で最も高い値を有しており，セグメントレベルの評価においては有効であることが確認できた．

4 まとめ

本報告では，WMT20 の評価タスクデータを用いたメタ評価の結果について述べた．メタ評価の結果，提案手法の WE_WPI は to-English の言語ペアにおいて WMT20 における様々な先行研究に対して比較的高い評価精度が得られることを確認できた．しかし，out-of-English においては先行研究と比べて上位に位置しているとは言えず不十分であった．今後はその原因について精査し，WE_WPI の評価精度の向上のための改良を進める予定である．

References

［1］Ilya Sutskever, Oriol Vinyals and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks.

表２：to-English におけるシステムレベルのピアソンの相関係数

	cs-en		de-en		ja-en		pl-en		ru-en		ta-en		zh-en		iu-en		km-en	ps-en
	All	-out	All	-out	All	-out	All	-out	All	-out	All	-out	All	-out	All	-out	All	All
	12	10	12	9	10	7	14	13	11	10	14	12	16	15	11	9	7	6
BAQ_dyn		−		−		−		−		−		−	0.956	0.928		−	−	−
BAQ_static		−		−		−		−		−		−	**0.960**	**0.933**		−	−	−
BERT-base-L2	0.775	0.693	0.997	0.791	0.971	0.789	**0.552**	**0.328**	0.919	**0.836**	0.909	0.746	**0.967**	0.929	0.704	0.145	0.967	**0.945**
BERT-large-L2	**0.784**	0.695	0.990	0.800	**0.974**	**0.784**	0.520	0.282	**0.925**	**0.843**	0.901	0.760	0.962	0.928	0.744	0.211	0.959	**0.950**
BLEU	**0.851**	**0.800**	0.985	0.778	**0.969**	0.826	**0.549**	**0.355**	0.884	0.761	0.916	**0.807**	0.956	**0.957**	0.569	0.348	0.969	0.888
BLEURT	**0.792**	0.725	0.996	0.770	**0.978**	**0.820**	**0.591**	**0.371**	**0.924**	**0.844**	0.906	0.768	**0.966**	**0.931**	0.771	0.320	0.984	**0.955**
BLEURT-extended	0.771	0.668	0.985	**0.818**	0.961	**0.772**	**0.551**	0.298	0.900	0.797	0.897	0.743	0.945	0.931	0.789	0.359	0.985	**0.942**
CharacTER	0.844	**0.812**	**0.998**	0.687	**0.970**	**0.895**	**0.522**	**0.325**	**0.927**	**0.869**	**0.965**	**0.880**	0.964	**0.950**	0.763	0.410	0.977	0.841
chrF	**0.872**	**0.806**	**0.997**	0.687	0.968	**0.861**	**0.528**	**0.312**	0.890	0.831	**0.951**	**0.828**	**0.976**	**0.954**	0.729	0.337	0.978	0.898
chrF＋＋	**0.867**	**0.804**	0.997	0.699	**0.974**	**0.871**	**0.538**	**0.328**	0.894	**0.833**	**0.953**	**0.830**	**0.975**	**0.955**	0.726	0.392	0.983	0.900
COMET	0.783	0.694	**0.998**	0.773	0.964	**0.828**	**0.591**	**0.345**	**0.923**	**0.836**	0.880	**0.764**	0.952	**0.931**	0.852	0.605	0.971	0.941
COMET-2R	0.777	0.697	**0.998**	0.772	0.964	**0.818**	**0.584**	**0.332**	**0.924**	**0.843**	0.881	**0.770**	0.949	**0.928**	0.872	0.644	0.970	**0.949**
COMET-HTER	0.738	0.661	0.995	**0.767**	0.912	**0.702**	0.446	0.231	0.867	0.741	0.726	0.595	0.809	0.873	0.770	0.464	0.901	0.862
COMET-MQM	0.728	0.612	0.991	**0.684**	0.906	0.707	0.424	0.222	0.858	0.746	0.767	0.617	0.784	0.862	0.841	0.631	0.914	0.880
COMET-QE	0.755	0.622	0.939	**0.805**	0.892	**0.585**	0.447	0.218	0.883	0.773	0.795	0.672	0.847	0.887	0.685	0.661	0.896	0.832
COMET-Rank	0.705	0.534	0.964	**0.757**	0.923	**0.793**	0.483	**0.284**	0.868	0.732	0.787	0.664	0.877	**0.909**	0.158	0.214	0.911	0.855
EED	**0.884**	**0.838**	**0.997**	**0.752**	**0.974**	**0.904**	**0.538**	**0.299**	**0.926**	**0.872**	**0.958**	**0.862**	0.956	0.932	0.821	0.587	0.990	**0.930**
esim	**0.790**	0.716	**0.998**	0.808	**0.983**	**0.822**	**0.591**	**0.358**	**0.928**	**0.834**	0.885	**0.801**	**0.963**	0.910	0.807	0.514	0.929	0.929
IMPACT	**0.848**	**0.801**	0.996	**0.799**	**0.973**	**0.879**	**0.536**	**0.289**	**0.934**	**0.870**	0.911	**0.840**	0.952	0.916	0.714	0.587	0.981	**0.910**
mBERT-L2	**0.798**	0.715	0.995	**0.824**	**0.969**	**0.811**	**0.555**	**0.302**	**0.908**	0.805	0.887	0.740	0.959	**0.935**	0.837	0.530	0.980	**0.938**
MEE	**0.861**	**0.822**	0.995	0.712	**0.982**	**0.900**	**0.464**	**0.295**	**0.927**	**0.878**	**0.950**	**0.835**	0.952	**0.948**	0.771	0.562	0.970	0.878
OpenKiwi-Bert	0.726	**0.698**	0.989	**0.741**	0.735	0.546	0.355	0.187	**0.862**	0.695	0.645	0.469	0.625	0.774	−0.126	−0.671	0.751	0.753
OpenKiwi-XLMR	0.760	**0.680**	0.995	0.701	0.931	0.714	0.442	0.171	0.859	0.697	0.792	0.659	0.905	0.899	0.271	−0.577	0.880	0.865
parbleu	**0.834**	**0.774**	0.986	**0.838**	**0.970**	**0.833**	**0.562**	**0.342**	0.877	0.744	0.908	**0.801**	0.958	**0.953**	0.624	0.398	0.971	**0.939**
parchrf＋＋	**0.865**	**0.810**	**0.998**	0.708	**0.974**	**0.877**	**0.551**	**0.347**	0.885	0.823	**0.942**	**0.825**	**0.976**	**0.956**	0.720	0.296	0.985	0.899
paresim-1	**0.788**	0.712	**0.998**	**0.835**	**0.983**	**0.819**	**0.591**	**0.363**	**0.926**	**0.828**	0.885	**0.797**	**0.963**	0.910	0.800	0.495	0.929	0.929
prism	0.818	0.720	**0.998**	**0.775**	**0.974**	**0.869**	**0.502**	**0.269**	**0.908**	**0.839**	0.898	**0.788**	0.957	**0.945**	0.833	0.616	0.950	**0.966**
sentBLEU	0.844	**0.800**	0.978	**0.786**	**0.974**	**0.851**	**0.502**	**0.284**	**0.916**	**0.833**	0.925	**0.829**	0.948	**0.950**	0.649	0.469	0.969	0.888
SWSS＋METEOR		−		−	**0.978**	**0.919**	**0.472**	**0.212**	**0.925**	**0.876**	**0.967**	**0.862**	0.959	0.926	0.766	0.545	**0.990**	**0.946**
TER	**0.845**	**0.783**	0.993	**0.766**	**0.974**	0.752	**0.586**	**0.346**	0.904	**0.829**	0.805	**0.795**	0.956	0.911	0.733	0.616	0.973	**0.935**
WE_WPI	**0.838**	0.757	**0.998**	**0.743**	**0.973**	**0.890**	**0.573**	**0.335**	**0.939**	**0.861**	0.933	**0.878**	**0.965**	0.924	0.776	0.654	**0.993**	**0.922**
YiSi-0	**0.876**	**0.825**	**0.998**	**0.786**	**0.972**	**0.867**	0.453	**0.207**	**0.938**	**0.874**	**0.968**	**0.861**	0.956	0.918	0.831	0.563	0.986	**0.932**
YiSi-1	**0.832**	**0.746**	**0.998**	**0.783**	**0.982**	**0.868**	**0.543**	**0.316**	**0.915**	**0.833**	0.925	**0.797**	0.961	**0.942**	0.834	0.590	0.977	**0.953**
YiSi-2	0.764	0.640	0.988	0.404	**0.971**	**0.776**	**0.437**	**0.230**	**0.825**	**0.814**	0.849	**0.761**	**0.964**	0.933	0.676	0.371	0.790	**0.942**

表 3 ：out-of-English におけるシステムレベルのピアソンの相関係数

	en-cs		en-de		en-ja	en-pl		en-ru	en-ta		en-zh	en-iu_full		en-iu_news	
	All	-out	All	-out	All	All	-out	All	All	-out	All	All	-out	All	-out
	12	10	14	11	11	14	11	9	15	12	12	11	8	11	8
BAQ_dyn	−		−		−	−		−	−		0.904	−		−	
BAQ_static	−		−		−	−		−	−		0.958	−		−	
BLEU	0.825	0.390	0.928	0.825	0.945	0.943	**0.743**	**0.980**	0.880	0.829	0.928	0.163	0.131	0.074	**0.111**
BLEURT-extended	0.989	0.960	**0.969**	**0.870**	0.944	**0.982**	**0.828**	**0.980**	0.940	0.814	0.928	0.823	**0.122**	0.762	**0.155**
bleurt-Yisi-combi	−		0.971	0.868	−	−		−	−		−	−		−	
CharacTER	0.807	0.269	0.961	**0.868**	0.951	0.935	**0.726**	0.961	0.957	0.851	0.905	0.503	0.008	0.515	**0.121**
chrF	0.826	0.313	**0.962**	**0.862**	0.951	0.957	**0.793**	**0.982**	0.937	**0.890**	0.923	0.350	**0.122**	0.336	**0.091**
chrF + +	0.833	0.349	0.958	**0.850**	0.952	0.956	**0.783**	**0.983**	0.929	**0.880**	0.878	0.328	0.128	0.315	**0.098**
COMET	0.978	0.926	**0.972**	**0.863**	0.974	**0.981**	**0.800**	0.925	0.944	0.798	0.007	0.860	0.028	**0.858**	0.152
COMET-2R	0.983	0.942	**0.972**	**0.869**	0.986	**0.982**	**0.803**	0.872	0.959	0.852	−0.066	0.848	−0.008	**0.867**	**0.177**
COMET-HTER	0.976	0.917	0.951	**0.852**	0.989	**0.974**	**0.763**	0.803	0.925	0.681	−0.073	**0.900**	0.142	**0.888**	0.092
COMET-MQM	0.974	0.910	0.881	0.840	0.974	0.967	**0.766**	0.788	0.910	0.641	0.084	0.870	0.129	**0.867**	0.172
COMET-QE	0.989	0.974	0.903	0.831	0.953	0.969	**0.804**	0.807	0.887	0.622	0.375	**0.905**	**0.578**	**0.928**	**0.651**
COMET-Rank	0.959	0.868	0.877	**0.860**	0.931	0.957	**0.760**	0.676	0.876	0.511	0.540	0.283	0.099	0.392	**0.252**
EED	0.817	0.271	**0.965**	**0.869**	0.955	**0.962**	**0.789**	**0.980**	0.959	**0.913**	0.928	0.519	0.043	0.483	**0.122**
EQ_dyn	−		−		−	−		−	−		0.948	−		−	
EQ_static	−		−		−	−		−	−		**0.976**	−		−	
esim	0.908	0.575	**0.979**	**0.894**	0.993	**0.969**	0.698	**0.967**	0.937	0.833	0.972	0.814	**0.365**	0.760	0.418
IMPACT	0.861	0.482	0.932	0.797	0.932	0.939	**0.679**	0.961	0.954	**0.896**	0.913	0.405	0.122	0.430	**0.200**
mBERT-L2	0.946	0.782	**0.970**	**0.861**	0.977	**0.976**	**0.775**	0.946	0.944	0.834	0.934	−		−	
MEE	0.875	0.495	0.954	0.820	−	0.952	**0.733**	0.724	0.906	0.861	−	0.287	0.094	0.242	**0.113**
OpenKiwi-Bert	0.920	0.830	0.852	**0.829**	0.363	0.903	0.450	0.834	0.846	0.370	0.551	0.573	−0.602	0.808	**0.194**
OpenKiwi-XLMR	0.972	0.911	**0.968**	0.814	0.992	0.957	0.638	0.875	0.910	0.676	−0.010	0.513	−0.668	0.680	−0.358
parbleu	0.870	0.543	0.910	0.774	0.869	0.948	**0.760**	**0.959**	0.871	**0.849**	0.962	0.194	**0.464**	0.126	**0.306**
parchrf + +	0.860	0.438	0.957	**0.845**	0.955	0.953	**0.818**	**0.975**		−	0.948	−		−	
paresim-1	0.919	0.635	**0.974**	**0.886**	0.989	**0.968**	0.705	**0.964**	0.937	0.833	**0.983**	0.814	**0.365**	0.760	0.418
prism	0.949	0.805	0.958	**0.851**	0.932	0.958	**0.742**	0.724	0.863	0.452	0.221	**0.957**	**−0.043**	**0.945**	**0.088**
sentBLEU	0.840	0.436	0.934	0.823	0.946	0.950	**0.772**	**0.981**	0.881	0.852	0.927	0.129	0.047	0.075	**0.172**
TER	0.814	0.339	0.941	**0.848**	0.297	0.893	0.553	0.064	0.870	**0.883**	−0.213	0.384	**0.133**	0.357	**0.083**
WE_WPI	0.879	0.566	0.941	**0.808**	0.964	0.894	0.526	0.945	0.936	**0.875**	0.911	−		−	
YiSi-0	0.797	0.270	0.953	**0.889**	0.967	0.953	**0.783**	**0.971**	0.929	**0.897**	0.362	0.525	0.015	0.505	**0.095**
YiSi-1	0.922	0.664	**0.971**	**0.887**	0.969	**0.964**	**0.714**	0.926	0.973	**0.909**	0.959	0.554	−0.217	0.523	−0.014
YiSi-2	0.714	0.353	0.899	0.552	0.854	0.470	−0.107	0.584	0.922	**0.923**	−0.215	0.802	**−0.257**	0.830	**0.065**
Yisi-combi	−		0.971	0.868	−	−		−	−		−	−		−	

表 4 ：to-English におけるセグメントレベルのケンドール τ

	cs-en		de-en		iu-en		ja-en		km-en	pl-en		ps-en	ru-en		ta-en		zh-en	
	all	all-out	all	all-out	all	all-out	all	all-out	all	all	all-out	all	all	all-out	all	all-out	all	all-out
	14018	9461	16584	6185	8162	5381	15193	6286	3706	21121	17979	3507	14024	11020	12789	8749	62586	53610
BAQ_dyn	−		−		−		−		−	−		−	−		−		0.119 / 0.089	
BAQ_static	−		−		−		−		−	−		−	−		−		0.119 / 0.087	
BERT-base-L2	0.103 / 0.087		0.454 / 0.026		0.238 / 0.229		0.263 / 0.129		0.295	0.032 / 0.013		0.159	0.087 / 0.037		0.223 / 0.135		0.141 / 0.113	
BERT-large-L2	0.102 / 0.087		0.456 / 0.025		0.251 / 0.249		0.262 / 0.114		0.314	0.044 / 0.027		0.151	0.094 / 0.047		0.245 / 0.157		0.133 / 0.102	
BLEURT	0.126 / 0.118		0.456 / 0.015		0.258 / 0.256		0.265 / 0.123		0.327	0.057 / 0.040		0.207	0.093 / 0.046		0.230 / 0.145		0.137 / 0.107	
BLEURT-extended	0.127 / 0.113		0.448 / 0.004		0.259 / 0.259		0.271 / 0.124		0.330	0.044 / 0.019		0.161	0.101 / 0.057		0.246 / 0.165		0.137 / 0.107	
CharacTER	0.090 / 0.087		0.440 / 0.011		0.214 / 0.220		0.221 / 0.106		0.248	0.023 / −0.002		0.172	0.057 / 0.028		0.138 / 0.078		0.123 / 0.093	
chrF	0.086 / 0.072		0.438 / 0.018		0.254 / 0.260		0.242 / 0.109		0.267	0.028 / 0.003		0.144	0.049 / 0.012		0.186 / 0.096		0.132 / 0.098	
chrF + +	0.090 / 0.075		0.435 / 0.013		0.246 / 0.251		0.245 / 0.115		0.275	0.034 / 0.009		0.145	0.054 / 0.018		0.186 / 0.098		0.130 / 0.096	
COMET	0.129 / 0.112		0.485 / 0.090		0.281 / 0.271		0.274 / 0.127		0.298	0.099 / 0.085		0.158	0.156 / 0.117		0.241 / 0.163		0.171 / 0.142	
COMET-2R	0.120 / 0.107		0.479 / 0.101		0.257 / 0.251		0.268 / 0.120		0.308	0.098 / 0.085		0.144	0.148 / 0.110		0.253 / 0.177		0.163 / 0.136	
COMET-HTER	0.103 / 0.087		0.481 / 0.088		0.198 / 0.199		0.241 / 0.095		0.269	0.080 / 0.067		0.116	0.131 / 0.098		0.227 / 0.151		0.135 / 0.113	
COMET-MQM	0.108 / 0.097		0.483 / 0.100		0.215 / 0.209		0.259 / 0.112		0.282	0.080 / 0.066		0.141	0.137 / 0.102		0.227 / 0.158		0.141 / 0.117	
COMET-QE	0.091 / 0.072		0.410 / 0.042		0.031 / 0.020		0.153 / 0.048		0.148	0.039 / 0.029		0.092	0.084 / 0.049		0.163 / 0.099		0.088 / 0.070	
COMET-Rank	0.099 / 0.096		0.470 / 0.061		0.188 / 0.181		0.235 / 0.086		0.228	0.073 / 0.057		0.107	0.118 / 0.082		0.199 / 0.112		0.142 / 0.117	
EED	0.091 / 0.078		0.440 / 0.013		0.256 / 0.258		0.235 / 0.116		0.271	0.045 / 0.022		0.149	0.053 / 0.018		0.198 / 0.103		0.129 / 0.093	
esim	0.110 / 0.103		0.454 / 0.031		0.241 / 0.233		0.239 / 0.119		0.300	0.58 / 0.045		0.147	0.084 / 0.044		0.208 / 0.117		0.138 / 0.108	
IMPACT	0.070 / 0.057		0.427 / −0.012		0.188 / 0.177		0.194 / 0.069		0.243	−0.007 / −0.031		0.097	0.009 / −0.032		0.191 / 0.095		0.103 / 0.068	
mBERT-L2	0.119 / 0.111		0.442 / 0.001		0.244 / 0.235		0.251 / 0.120		0.312	0.047 / 0.029		0.151	0.083 / 0.036		0.227 / 0.139		0.133 / 0.104	
MEE	0.063 / 0.045		0.402 / −0.04		0.134 / 0.126		0.187 / 0.064		0.206	−0.084 / −0.105		0.078	−0.041 / −0.084		0.114 / 0.032		0.083 / 0.050	
OpenKiwi-Bert	0.036 / 0.029		0.379 / 0.013		−0.005 / −0.009		0.110 / 0.000		0.168	−0.033 / −0.043		0.076	−0.033 / −0.067		0.118 / 0.052		0.029 / 0.020	
OpenKiwi-XLMR	0.093 / 0.079		0.463 / 0.074		0.056 / 0.031		0.220 / 0.086		0.244	0.059 / 0.051		0.106	0.092 / 0.065		0.188 / 0.109		0.115 / 0.089	
parbleu	0.058 / 0.038		0.415 / −0.039		0.167 / 0.161		0.198 / 0.074		0.203	−0.025 / −0.049		0.100	−0.011 / −0.052		0.159 / 0.064		0.095 / 0.059	
parchrf + +	0.096 / 0.082		0.436 / 0.009		0.232 / 0.235		0.247 / 0.117		0.267	0.027 / 0.002		0.147	0.044 / 0.007		0.184 / 0.095		0.132 / 0.099	
paresim-1	0.105 / 0.098		0.464 / 0.051		0.249 / 0.241		0.242 / 0.121		0.292	0.066 / 0.055		0.149	0.089 / 0.049		0.213 / 0.123		0.139 / 0.111	
prism	0.143 / 0.135		0.475 / 0.057		0.255 / 0.254		0.272 / 0.146		0.304	0.109 / 0.093		0.165	0.145 / 0.111		0.237 / 0.151		0.167 / 0.138	
sentBLEU	0.068 / 0.057		0.413 / −0.025		0.182 / 0.170		0.188 / 0.061		0.226	−0.024 / −0.046		0.096	−0.005 / −0.038		0.162 / 0.069		0.093 / 0.060	
SWSS+METEOR	−		−		0.226 / 0.218		0.228 / 0.086		0.264	0.011 / −0.016		0.130	0.048 / 0.010		0.205 / 0.120		0.133 / 0.099	
TER	−0.04 / −0.06		0.355 / −0.137		0.021 / 0.012		0.044 / −0.077		0.125	−0.172 / −0.196		−0.036	−0.117 / −0.154		0.046 / −0.063		−0.01 / −0.047	
WE_WPI	0.102 / 0.087		0.474 / 0.077		0.218 / 0.224		0.238 / 0.133		0.239	0.080 / 0.062		0.134	0.133 / 0.096		0.222 / 0.139		0.151 / 0.126	
YiSi-0	0.072 / 0.065		0.441 / 0.024		0.261 / 0.263		0.241 / 0.121		0.268	0.035 / 0.013		0.140	0.065 / 0.030		0.183 / 0.089		0.127 / 0.093	
YiSi-1	0.117 / 0.103		0.468 / 0.051		0.253 / 0.260		0.277 / 0.128		0.316	0.042 / 0.023		0.147	0.091 / 0.049		0.248 / 0.162		0.146 / 0.115	
YiSi-2	0.068 / 0.054		0.413 / 0.006		0.039 / 0.028		0.204 / 0.074		0.214	0.048 / 0.042		0.073	0.070 / 0.056		0.199 / 0.113		0.116 / 0.084	

表 5 ：out-of-English におけるセグメントレベルのケンドール τ

	en-cs		en-de		en-iu		en-ja	en-pl		en-ru	en-ta		en-zh
	all	all-out	all	all-out	all	all-out	all	all	all-out	all	all	all-out	all
	21121	10283	9339	4637	13159	5490	12830	17689	9316	8330	9087	3695	12652
BAQ_dyn	−		−		−		−	−		−	−		0.351
BAQ_static	−		−		−		−	−		−	−		0.344
bleurt-combi	−		0.399 / 0.224		−		−	−		−	−		−
BLEURT-extended	0.689 / 0.517		0.447 / 0.278		0.359 / 0.112		0.533	0.430 / 0.271		0.305	0.643 / 0.419		0.460
CharacTER	0.413 / 0.195		0.311 / 0.179		0.309 / 0.108		0.471	0.198 / 0.107		0.143	0.525 / 0.270		0.339
chrF	0.472 / 0.229		0.379 / 0.224		0.344 / 0.095		0.506	0.250 / 0.150		0.153	0.589 / 0.359		0.400
chrF + +	0.478 / 0.228		0.367 / 0.215		0.338 / 0.075		0.506	0.255 / 0.154		0.156	0.579 / 0.349		0.388
COMET	0.668 / 0.487		0.468 / 0.324		0.322 / 0.078		0.624	0.462 / 0.316		0.344	0.671 / 0.457		0.432
COMET-2R	0.669 / 0.512		0.463 / 0.321		0.326 / 0.078		0.630	0.445 / 0.294		0.343	0.676 / 0.463		0.434
COMET-HTER	0.665 / 0.500		0.440 / 0.303		0.331 / 0.088		0.601	0.427 / 0.274		0.292	0.640 / 0.411		0.411
COMET-MQM	0.666 / 0.490		0.423 / 0.275		0.313 / 0.078		0.588	0.424 / 0.271		0.281	0.635 / 0.413		0.388
COMET-QE	0.614 / 0.470		0.347 / 0.233		−0.04 / −0.051		0.470	0.360 / 0.211		0.264	0.514 / 0.320		0.346
COMET-Rank	0.629 / 0.408		0.379 / 0.217		0.297 / 0.097		0.569	0.388 / 0.207		0.229	0.588 / 0.342		0.380
EED	0.458 / 0.210		0.363 / 0.203		0.361 / 0.109		0.515	0.248 / 0.151		0.155	0.587 / 0.342		0.393
EQ_dyn	−		−		−		−	−		−	−		0.356
EQ_static	−		−		−		−	−		−	−		0.409
esim	0.469 / 0.253		0.347 / 0.195		0.122 / −0.018		0.522	0.312 / 0.203		0.224	0.599 / 0.363		0.391
IMPACT	0.457 / 0.209		0.306 / 0.152		0.209 / -0.11		0.486	0.181 / 0.076		0.068	0.525 / 0.293		0.391
mBERT-L2	0.567 / 0.359		0.361 / 0.202		−		0.541	0.350 / 0.212		0.246	0.587 / 0.334		0.432
MEE	0.411 / 0.157		0.289 / 0.128		−0.074 / −0.272		−	0.125 / 0.025		0.027	0.373 / 0.168		−
OpenKiwi-Bert	0.262 / 0.142		0.168 / 0.058		−0.115 / -0.233		−0.529	0.153 / 0.035		0.164	0.169 / 0.022		0.077
OpenKiwi-XLMR	0.607 / 0.417		0.369 / 0.224		0.060 / 0.009		0.553	0.347 / 0.189		0.279	0.604 / 0.354		0.377
parbleu	0.460 / 0.226		0.299 / 0.136		0.212 / −0.051		0.052	0.183 / 0.088		0.062	0.340 / 0.178		0.356
parchrf + +	0.492 / 0.253		0.355 / 0.192		−		0.527	0.272 / 0.167		0.176	−		0.398
paresim-1	0.475 / 0.257		0.343 / 0.197		0.122 / −0.018		0.510	0.324 / 0.209		0.230	0.599 / 0.363		0.396
prism	0.619 / 0.414		0.447 / 0.280		0.452 / 0.195		0.579	0.414 / 0.274		0.283	0.448 / 0.211		0.397
sentBLEU	0.432 / 0.194		0.303 / 0.155		0.206 / −0.084		0.479	0.153 / 0.067		0.051	0.398 / 0.206		0.396
TER	0.317 / 0.067		0.182 / 0.044		−0.071 / −0.337		−0.591	0.003 / −0.094		−0.121	0.203 / 0.019		−0.36
WE_WPI	0.477 / 0.267		0.331 / 0.197		−		0.502	0.276 / 0.182		0.192	0.376 / 0.200		0.379
YiSi-0	0.432 / 0.191		0.349 / 0.212		0.362 / 0.101		0.484	0.233 / 0.132		0.151	0.547 / 0.336		0.319
YiSi-1	0.550 / 0.320		0.427 / 0.263		0.251 / 0.082		0.568	0.349 / 0.209		0.256	0.669 / 0.440		0.463
YiSi-2	0.187 / 0.104		0.296 / 0.171		0.146 / 0.073		0.383	0.115 / 0.052		0.146	0.545 / 0.332		0.152
Yisi-combi	−		0.399 / 0.224		−		−	−		−	−		−

表 6 ：to-English におけるシステムレベルのケンドール τ

	cs-en 12	de-en 12	ja-en 10	pl-en 14	ru-en 11	ta-en 14	zh-en 16	iu-en 11	km-en 7	ps-en 6
BAQ_dyn	−	−	−	−	−	−	0.817	−	−	−
BAQ_static	−	−	−	−	−	−	0.867	−	−	−
BERT-base-L2	0.758	0.848	0.822	0.407	0.491	0.604	0.633	0.564	1.000	0.867
BERT-large-L2	0.758	0.848	0.867	0.341	0.564	0.626	0.700	0.527	1.000	0.867
BLEU	0.848	0.697	0.778	0.407	0.455	0.692	0.833	0.309	0.714	0.600
BLEURT	0.758	0.788	0.822	0.407	0.600	0.604	0.650	0.527	1.000	0.867
BLEURT-extended	0.727	0.848	0.778	0.341	0.455	0.582	0.617	0.527	0.905	0.867
CharacTER	0.758	0.758	0.822	0.341	0.745	0.692	0.800	0.527	0.810	0.733
chrF	0.818	0.727	0.822	0.363	0.709	0.714	0.833	0.418	0.619	0.733
chrF + +	0.818	0.697	0.778	0.407	0.673	0.714	0.850	0.418	0.619	0.733
COMET	0.727	0.758	0.778	0.407	0.564	0.626	0.733	0.636	1.000	0.867
COMET-2R	0.727	0.788	0.778	0.451	0.527	0.582	0.717	0.600	1.000	0.867
COMET-HTER	0.667	0.788	0.822	0.275	0.491	0.604	0.533	0.564	1.000	0.867
COMET-MQM	0.667	0.727	0.822	0.275	0.455	0.582	0.517	0.636	1.000	1.000
COMET-QE	0.697	0.788	0.778	0.297	0.455	0.516	0.550	0.491	0.905	0.733
COMET-Rank	0.576	0.727	0.822	0.341	0.455	0.626	0.650	0.309	0.810	1.000
EED	0.788	0.727	0.733	0.297	0.782	0.758	0.833	0.636	0.714	0.733
esim	0.727	0.848	0.822	0.451	0.491	0.670	0.717	0.636	1.000	0.867
IMPACT	0.727	0.818	0.689	0.319	0.673	0.802	0.833	0.587	0.714	0.733
mBERT-L2	0.758	0.818	0.822	0.429	0.564	0.604	0.750	0.673	1.000	0.867
MEE	0.758	0.697	0.867	0.363	0.709	0.692	0.783	0.636	0.714	0.733
OpenKiwi-Bert	0.697	0.667	0.733	0.187	0.455	0.429	0.450	−0.055	0.714	0.467
OpenKiwi-XLMR	0.727	0.636	0.822	0.275	0.418	0.560	0.567	0.018	1.000	0.867
parbleu	0.809	0.779	0.778	0.420	0.491	0.685	0.807	0.404	0.714	0.867
parchrf + +	0.818	0.727	0.822	0.407	0.709	0.714	0.817	0.491	0.619	0.733
paresim-1	0.727	0.879	0.822	0.451	0.491	0.670	0.700	0.636	1.000	0.867
prism	0.758	0.727	0.867	0.341	0.564	0.648	0.800	0.673	0.714	0.867
sentBLEU	0.788	0.758	0.733	0.297	0.564	0.692	0.850	0.455	0.619	0.600
SWSS + METEOR	−	−	0.822	0.341	0.818	0.736	0.817	0.491	0.714	0.733
TER	0.758	0.788	0.689	0.287	0.600	0.780	0.800	0.514	0.878	0.867
WE_WPI	0.788	0.667	0.778	0.451	0.673	0.736	0.850	0.673	0.714	0.733
YiSi-0	0.758	0.758	0.689	0.231	0.782	0.802	0.833	0.600	0.714	0.733
YiSi-1	0.758	0.758	0.778	0.451	0.564	0.692	0.817	0.673	1.000	0.867
YiSi-2	0.576	0.515	0.778	0.319	0.527	0.582	0.750	0.491	0.810	0.867

Neural Information Processing Systems.

［2］Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals and Wojciech Zaremba. 2015. Addressing the Rare Word Problem in Neural Machine Translation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. pp.11–19.

［3］Minh-Thang Luong, Hieu Pham and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp.1412–1421.

［4］Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. Attention Is All You Need. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). pp.6000–6010.

［5］Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics, Vol.16, No.2*. pp.79–85.

［6］Peter F. Brown, Vincent J. Della Pietr, Stephen A. Della Pietra and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation *Computational Linguistics, Vol. 19, No.2*. pp.263–311.

［7］Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-Based Statistical Machine Translation. LNAI 2479, pp.18–32. *Springer-Verlag Berlin Heidelberg*

表 7 ：out-of-English におけるシステムレベルのケンドール τ

	en-cs 12	en-de 14	en-ja 11	en-pl 14	en-ru 9	en-ta 15	en-zh 12	en-iu_full 11	en-iu_news 11
BAQ_dyn	−	−	−	−	−	−	0.697	−	−
BAQ_static	−	−	−	−	−	−	0.788	−	−
BLEU	0.515	0.802	0.818	0.582	0.889	0.829	0.727	0.236	0.236
BLEURT-extended	0.879	0.802	0.782	0.780	0.833	0.771	0.848	0.382	0.345
bleurt-Yisi-combi	−	0.824	−	−	−	−	−	−	−
CharacTER	0.515	0.890	0.782	0.560	0.944	0.771	0.697	0.236	0.345
chrF	0.485	0.868	0.818	0.604	0.889	0.810	0.727	0.345	0.309
chrF + +	0.485	0.868	0.782	0.604	0.889	0.829	0.727	0.309	0.309
COMET	0.909	0.846	0.745	0.736	0.722	0.771	0.606	0.382	0.382
COMET-2R	0.909	0.890	0.891	0.714	0.611	0.790	0.606	0.309	0.418
COMET-HTER	0.909	0.802	0.818	0.736	0.667	0.619	0.576	0.491	0.491
COMET-MQM	0.909	0.802	0.818	0.736	0.667	0.619	0.545	0.527	0.455
COMET-QE	0.848	0.802	0.709	0.802	0.667	0.543	0.576	0.600	0.673
COMET-Rank	0.848	0.780	0.782	0.692	0.556	0.524	0.515	0.127	0.345
EED	0.545	0.868	0.782	0.604	0.833	0.867	0.727	0.273	0.273
EQ_dyn	−	−	−	−	−	−	0.727	−	−
EQ_static	−	−	−	−	−	−	0.818	−	−
esim	0.606	0.912	0.855	0.692	0.833	0.752	0.788	0.382	0.455
IMPACT	0.515	0.736	0.818	0.560	0.722	0.867	0.697	0.257	0.236
mBERT-L2	0.788	0.846	0.782	0.736	0.778	0.752	0.909	−	−
MEE	0.576	0.802	−	0.582	0.667	0.829	−	0.273	0.382
OpenKiwi-Bert	0.758	0.780	0.236	0.538	0.722	0.314	0.606	−0.273	0.200
OpenKiwi-XLMR	0.909	0.780	0.818	0.692	0.667	0.657	0.545	0.018	0.200
parbleu	0.504	0.736	0.611	0.633	0.761	0.842	0.718	0.404	0.345
parchrf + +	0.515	0.846	0.818	0.670	0.889	−	0.727	−	−
paresim-1	0.667	0.890	0.818	0.692	0.833	0.752	0.818	0.382	0.455
prism	0.818	0.868	0.818	0.670	0.611	0.562	0.576	0.418	0.600
sentBLEU	0.515	0.802	0.855	0.604	0.944	0.867	0.727	0.236	0.273
TER	0.515	0.824	0.018	0.641	0.556	0.752	0.242	0.309	0.309
WE_WPI	0.515	0.780	0.818	0.495	0.833	0.771	0.667	−	−
YiSi-0	0.545	0.846	0.818	0.604	0.944	0.790	0.515	0.236	0.345
YiSi-1	0.606	0.868	0.782	0.626	0.833	0.810	0.758	0.091	0.273
YiSi-2	0.485	0.582	0.527	0.077	0.444	0.886	0.121	0.309	0.455
Yisi-combi	−	0.824	−	−	−	−	−	−	−

表 8 ：相関係数の平均

	システムレベル				セグメントレベル			
	to-English		out-of-English		to-English		out-of-English	
	All	-out	All	-out	All	-out	All	-out
BLEU	0.765	0.652	0.741	0.505	0.139	0.034	0.302	0.108
CharacTER	0.793	0.667	0.833	0.474	0.173	0.078	0.339	0.172
chrF	0.789	0.651	0.803	0.512	0.183	0.084	0.387	0.211
chrF + +	0.791	0.661	0.792	0.515	0.184	0.084	0.383	0.204
EED	0.804	0.698	0.841	0.501	0.187	0.088	0.385	0.203
IMPACT	0.785	0.689	0.814	0.529	0.152	0.049	0.328	0.124
TER	0.777	0.673	0.490	0.473	0.024	−0.090	−0.055	−0.060
MEE	0.787	0.683	−	−	0.114	0.034	−	−
WE_WPI	0.799	0.696	−	−	0.199	0.118	−	−

［8］K. Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp.311–318.

［9］Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. Proceedings of the Association for Machine Transaltion in the Americas. pp.223–231.

［10］A. Lavie and A. Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. Proceedings of the Second Workshop on Statistical Machine Translation. pp.228–231.

［11］Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp.944–952.

［12］Hiroshi Echizen-ya and Kenji Araki. 2007. Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum. Proceedings of the Eleventh Machine Translation Summit. pp.151–158.

［13］Hiroshi Echizen-ya, Terumasa Ehara, Sayori Shimohata, Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro and Noriko Kando. 2009. Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR-7. Proceedings of the 3rd Workshop on Patent Translation pp.9–16.

［14］Hiroshi Echizen-ya and Kenji Araki. 2010. Automatic Evaluation Method for Machine Translation using Noun-Phrase Chunking. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp.108–117.

［15］Hiroshi Echizen'ya, Kenji Araki, and Eduard Hovy. Optimization for Efficient Determination of Chunk in Automatic Evaluation for Machine Translation. Proceedings of the 1th International Workshop on Optimization Techniques for Human Language Technology. pp.17–30.

［16］Hiroshi Echizen'ya, Kenji Araki, and Eduard Hovy. 2013. Automatic Evaluation Metric for Machine Translation that is Independent of Sentence Length. Proceedings of the 9th Recent Advances in Natural Language Processing. pp.230–236.

［17］Hiroshi Echizen'ya, Kenji Araki, and Eduard Hovy. 2014. Application of Prize based on Sentence Length in Chunk-based Automatic Evaluation of Machine Translation. Proceedings of the Ninth Workshop on Statistical Machine Translation. pp.381–386.

［18］A. Apostolico and C. Guerra. 1987. The Longest Common Subsequence Problem Revisited. *Algorithmica, Volume 2, issue 1-4*. pp.315–336, Springer.

［19］Hiroshi Echizen'ya, Kenji Araki, and Eduard Hovy. 2019. Word Embedding-Based Automatic MT Evaluation Metric using Word Position Information. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp.1874–1883.

［20］Yossi Rubner, Carlo Tomashi and Leonidas J. Guibas. 1998. A Metric for Distributions with Applications to Image Database. Proceedings of the 1998 IEEE International Conference on Computer Vision. pp.59–66.

［21］Yossi Rubner, Carlo Tomashi and Leonidas J. Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision 40 (2), pp.99-121* Kluwer Academic Publishers.

［22］柳本豪一，大松繁．Earth Mover's Distance を用いたテキスト分類．2007．The 21st Annual Conference of the Japanese Society for Artificial Intelligence．1G3–4.

［23］Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). Proceedings of the 5th Conference on Machine Translation (WMT). pp.1–55.

［24］Nitika Mathur, Johnny Tian-Zheng Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 Metrics Shared Task. Proceedings of the 5th Conference on Machine Translation (WMT). pp.688–725.

［25］Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp.4171–4186.

［26］Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. Proceedings of the Tenth Workshop on Statistical Machine Translation. pp.392–395.

[27] Maja Popović. 2017. chrF++: words helping character n-grams. Proceedings of the Second Conference on Machine Translation. pp.612–618.

[28] Rachel Bawden, Biao Zhang, Andre Tättar, and Matt Post. 2020. ParBLEU: Augmenting metrics with automatic paraphrases for the WMT'20 metrics shared task. Proceedings of the Fifth Conference on Machine Translation (Volume 2: Shared Task Papers). pp.887–894.

[29] Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp.90–121.

[30] Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. pp.505–510.

[31] Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. Eed: Extended edit distance measure for machine translation. Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers). pp.514–520.

[32] Jin Xu, Yinuo Guo, and Junfeng Hu. 2020. Incorporate semantic structures into machine translation evaluation via ucca. Proceedings of the 5th Conference on Machine Translation (WMT). pp.934–939.

[33] Manish Shrivastava Ananya Mukherjee, Hema Ala, and Dipti Misra Sharma. 2019. Mee: An automatic metric for evaluation using embeddings for machine translation. Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics. pp.292–299.

[34] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics, Volume 5*. pp.135–146.

[35] Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers). pp.507–513.

[36] Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers. pp.895–902.

[37] Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020. Learning to evaluate translation beyond english: Bleurt submissions to the wmt metrics 2020 shared task. Proceedings of the Fifth Conference on Machine Translation. pp.921–927.

[38] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the wmt20 metrics shared task. Proceedings of the Fifth Conference on Machine Translation. pp.911–920.

[39] Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp.117–122.

[40] Chi-kiu Lo and Samuel Larkin. 2020. Machine translation reference-less evaluation using YiSi-2 with bilingual mappings of massive multilingual language model. Proceedings of the Fifth Conference on Machine Translation. pp.903–910.

[41] Evan James Williams. 1959. Regression analysis. wiley.