文レベルの機械翻訳評価尺度に関する調査

Graham Neubig[1,a)]

概要：本研究では，近年提案されている機械翻訳の自動評価尺度を翻訳文品質の判定能力の観点から調査する．具体的には，日英，英日，仏英翻訳の計 4 タスクにおいて，4 つのシステムの出力に対して人手評価を行い，その結果に基づいて，5 種類の自動評価尺度を分析する．最初の実験では，人手による誤り分析を行う前段階として，自動評価尺度を用いて誤り文を特定する可能性について調査する．次の実験では，システム統合などで必要となる複数のシステムによる翻訳候補の優劣判定能力について調査する．両方の調査の結果，すべての自動評価尺度は，別の評価者による人手評価から得られるアノテータ間一致を大幅に下回っており，文レベルの自動翻訳評価に大きな課題が残っていることが分かった．しかし，複数の参照文を用いることにより翻訳品質判定能力が文レベルでも向上する結果も見られた．

1. はじめに

機械翻訳の最大の課題の一つは人間の評価に沿うような自動評価尺度の開発である．近年様々な評価尺度が開発されており，ある程度大きなテストデータを用意すれば，システム間の優劣を正確に判定することができるという報告はある [3]．しかし，文レベルの翻訳品質を正確に判定する評価尺度にはどれぐらい近づいているのか？つまり，1 つの特定の翻訳候補に対して，品質の高い訳であれば高いスコア，品質の低い訳であれば低いスコアを与える翻訳尺度は存在するのか？

この文レベルの評価尺度の開発は非常に難しいながら，下記のような分野で非常に有用であると考えられる：

効率的な人手検証： 機械翻訳システムの開発を行う上で，誤った訳が生成された過程を人手で分析し，その改良作を検討することが有効である．この中で，特に誤りの可能性の高い訳をある程度自動的に特定できれば，これを翻訳システム開発者に提示し，詳細な分析を行うことが可能となる．

システム統合や識別学習： 機械翻訳の精度を向上させる方法として，複数のシステムの出力を組み合わせるシステム統合手法 [2] や，大規模な素性集合を最適化する識別学習手法 [5], [24] が提案されている．このような手法において「正解」として用いるオラクル訳が実際に良質な訳であれば，統合や学習によりこのオラクル訳に実際のシステム出力を近づかせることが可能である。しかし，オラクル訳の選択に自動評価尺度を利用することが一般的であり，不完全な評価尺度を用いてしまった場合，「正解」と信じ込みながら「不正解」に向かって最適化を行う恐れがある．

しかし，機械翻訳のメタ評価に関する研究が多く行われている [3], [16] 中で，そのほとんどは順位相関などの抽象的な評価値を用いており，このような明確なユースケースに着目した評価を行っていない．

本研究では，現在の機械翻訳評価尺度が文レベルの評価でどれぐらい有用であるかを様々な観点から検証する．特に，上記の最適化と人手検証の観点に基づいて以下のような調査を行う：

人間評価値の推定能力： 自動評価尺度の値に基づいて，人間評価の値を推定できるか？この推定が可能であれば，人手検証が必要な品質の低い訳，もしくはシステムが特に得意とする品質の高い訳を自動的に特定することが可能となる．

候補間の優劣判定能力： 同一の入力文に対する翻訳候補を 2 つ提示した場合，どれが人手評価で優れているかを推定できるか？この推定が可能であれば，最適化のオラクル訳選択などに使用可能である．

このような能力を調査する実験を 3 つの言語対（日英，英日，仏英），4 つのタスク，4 つのシステムに対して行い，その結果について議論する．両方の調査の結果，すべての自動評価尺度は，別の評価者による人手評価から得られるアノテータ間一致を大幅に下回っており，文レベルの自動翻訳評価に大きな課題が残っていることが分かった．

[1] 奈良先端科学技術大学院大学 情報科学研究科
[a)] neubig@is.naist.jp

表 1 5 段階人手評価の各段階の意味

5	意味は正確に理解でき，表現も母語話者なみに流暢
4	文法も正しく意味も正確に理解できるが，非母語話者の文のように不自然な箇所がある
3	文法は正しくないが，意味は理解できる
2	原文の情報が含まれているが，理解が難しい箇所や曖昧な箇所がある
1	原文の意味が一部読み取れない，もしくは肯定・否定が逆になっているような大きな誤りが含まれている

2. 評価尺度

本節では，本研究で対象とする評価尺度について述べる．

2.1 人手評価

本研究では，自動評価尺度が人間の評価にどの程度一致するかを調べるが，まず人手評価に用いた手順を説明する．上記の各タスクに対して構築された 4 つの異なる翻訳システムを用いて，まずテストデータに対して翻訳候補を生成する．この翻訳候補の中から，人手評価が行いやすいように，対象を比較的短い 1-30 単語からなる文に限定し，その中から無作為に 200 文を選択し，人手評価を行う．

人手評価の基準として，許容性 [9] を参考に，意味的妥当性と流暢性を同時に考慮した 5 段階評価を用いる．各段階の定義を**表 1** に示す．この評価基準に基づいて，評価者 2 名に 4 通りの評価タスクにおいて，200 文に対して，4 システムの出力を評価してもらった．具体的には，各タスクにおいて評価者は評価者 A と評価者 B と呼び，評価者 A は本論文の著者であり，評価者 B は企業に外注した翻訳評価結果である．以後，評価者 A による許容性評価を Human A，評価者 B による許容性を Human B と記述する．また，後述する実験において Human B を「正解」として扱い，Human A をアノテータ間一致を測るために用いる．

2.2 自動評価

本研究では, BLEU+1, WER, TER, RIBES, METEOR という 5 通りの評価尺度を調査する．本節ではこれらについて簡単に述べる．

2.2.1 BLEU+1

BLEU は機械翻訳で最も広く用いられている自動評価尺度である [21]．システム出力と参照文を比較し，n-gram 適合率に基づいて翻訳の精度を評価する．

定式化するために，ある K 文からなる参照訳 $\mathcal{E}^* = \{\boldsymbol{e}_1^*, \ldots, \boldsymbol{e}_K^*\}$ とシステム出力 $\hat{\mathcal{E}} = \{\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_K\}$ が与えられた場合を考える．この場合，各文対 $\langle \boldsymbol{e}_k^*, \hat{\boldsymbol{e}}_k \rangle$ に対して，n-gram 数 $\mathrm{c}_n(\hat{\boldsymbol{e}}_k)$ と n-gram 一致数 $\mathrm{h}_n(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}}_k)$ を計算する関数を定義する．n-gram 数は単純に $\hat{\boldsymbol{e}}_k$ の中で長さ n の単語列の数となっており，

$$\mathrm{c}_n(\hat{\boldsymbol{e}}_k) = |\hat{\boldsymbol{e}}_k| - n + 1 \tag{1}$$

として定義される．n-gram 一致数は $\hat{\boldsymbol{e}}_k$ の中に含まれている長さ n-gram x の内，いくつが $\boldsymbol{e}_k^*$ の中に含まれているかを表す数字である．ある n-gram x が $\hat{\boldsymbol{e}}_k$ に現れた回数を表す関数 $\mathrm{c}(\hat{\boldsymbol{e}}_k, x)$ を定義すれば，n-gram 一致数を以下のように定義する．

$$\mathrm{m}_n(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}}_k) = \sum_x \min(\mathrm{c}(\hat{\boldsymbol{e}}_k, x), \mathrm{a}(\boldsymbol{e}_k^*, x)) \tag{2}$$

この関数を用いて n-gram 適合率 $\mathrm{a}_n(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}}_k)$ をコーパスごとに計算する．

$$\mathrm{a}_n(\mathcal{E}^*, \hat{\mathcal{E}}) = \frac{\sum_{k=1}^K \mathrm{m}_n(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}}_k)}{\sum_{k=1}^K \mathrm{c}_n(\boldsymbol{e}_k^*)} \tag{3}$$

また，n-gram 適合率を最適化しようと思えば，正解であると確信している n-gram のみを含む非常に短い文を出力する戦略も考えられる．このように短い文が不当に高い評価値とならないように，BLEU では参照文より短いシステム出力に対する簡潔ペナルティ (BP) も設ける．

$$\mathrm{BP}(\mathcal{E}^*, \hat{\mathcal{E}}) = \begin{cases} 1 & \text{if } |\hat{\mathcal{E}}| > |\mathcal{E}^*| \\ e^{1-|\mathcal{E}^*|/|\hat{\mathcal{E}}|} & \text{otherwise} \end{cases} \tag{4}$$

BP と $n = 1$ から $n = 4$ の確率を組み合わせることで BLEU が計算される．

$$\mathrm{BLEU}(\mathcal{E}^*, \hat{\mathcal{E}}) = \mathrm{BP}(\mathcal{E}^*, \hat{\mathcal{E}}) \prod_{n=1}^4 \mathrm{a}_n(\mathcal{E}^*, \hat{\mathcal{E}})^{1/4} \tag{5}$$

このように BLEU をコーパス全体に対して計算するが，本研究で扱うような文ごとの評価には向かない．その理由として，多くの文では，$n = 4$ のような高次の n-gram が 1 つも一致せず，式 2 の n-gram 一致数がゼロとなり，その影響で式 3 の n-gram 適合率と式 5 の BLEU スコアが 0 となってしまう．この問題を解決するために，[16] では，$n = 2$ 以上の n-gram に対して分子，分母ともに 1 を足すことで，高次元の n-gram が一致しなくても 0 とならない BLEU+1 を提案している．これで，1 文に対して以下のように n-gram 適合率を計算し，

$$\mathrm{a}{+}1_n(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}_k}) = \begin{cases} \frac{\sum_{k=1}^K \mathrm{m}_n(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}}_k)}{\sum_{k=1}^K \mathrm{c}_n(\boldsymbol{e}_k^*)} & \text{if } n = 1 \\ \frac{\sum_{k=1}^K \mathrm{m}_n(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}}_k)+1}{\sum_{k=1}^K \mathrm{c}_n(\boldsymbol{e}_k^*)+1} & \text{otherwise} \end{cases} \tag{6}$$

これを用いて BLEU+1 を計算する：

$$\mathrm{BLEU}{+}1(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}_k}) = \mathrm{BP}(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}_k}) \prod_{n=1}^4 \mathrm{a}{+}1_n(\boldsymbol{e}_k^*, \hat{\boldsymbol{e}_k})^{1/4} \tag{7}$$

BLEU の特徴として，局所的に流暢な文，参照文に表現法やスタイルが一致する文などに高い評価値を与えることが挙げられる．しかし，意味的妥当性との相関が低いなど，

様々な問題点が指摘されており，これを解決するために次節以降説明する評価尺度を含めて，多くの代替案が提案されている．

2.2.2 単語誤り率

音声認識の評価などで広く用いられる尺度として単語誤り率 (WER) がある．単語誤り率は，まず「挿入 (I)」「削除 (D)」「置換 (S)」という 3 種類の編集操作を定義し，システム出力を参照文へと変更するのに必要な編集操作を参照文の長さ R で割ったものとして求められる．

$$WER = \frac{I + D + S}{R} \tag{8}$$

例えば，出力文「the taro visit friend」と参照文「taro visited his friend」が与えられた時，「the」を削除し，「visit → visited」と置換し，「his」を挿入することで出力文を参照文へと変更できるため，編集操作数が 3 である．これを参照文の長さ 4 で割れば，0.75 という WER 値が求まる．出力文を参照文へ変更する最小の編集操作数を動的計画法により効率的に計算可能である [15].

WER は BLEU が開発される前から，音声認識などの評価で広く使用されていた．しかし，WER では参照文と出力文の語順の違いに非常に厳しい評価尺度となっており，例えば「brown dog」と「dog brown」のような比較的人間に理解しやすい小さな語順の誤りを完全に誤った訳と判定してしまう．このため，翻訳評価のために WER の代わりに BLEU が利用されていたが，日英・英日翻訳などで BLEU より WER の方がシステムレベルで人間の評価と相関が高い報告もある [6].

2.2.3 翻訳編集率

翻訳編集率 (Translation Edit Rate; TER) は人間が機械翻訳結果の後編集を行った際のコストに着目した評価尺度である [22]．WER と基本的に同じ考え方であるが，WER の並べ替えに対する厳しい罰則を緩める．具体的には，通常の WER で対象となる挿入，削除，置換以外に，「並べ替え」操作も加える．これにより，「brown dog」を「dog brown」に変更するために，2 回の置換ではなく，「brown を dog の後へ並べ替える」という 1 回の操作だけで済む．

2.2.4 RIBES

BLEU の弱点の 1 つとして，語順の誤りに対してそれほど敏感ではないことが取り上げられる．例えば，「taro visited hanako」という文に対して，システム 1 が「taro visits hanako」，システム 2 が「hanako visited taro」を出力した場合，BLEU が文の並びより単語の表層的な一致度に引っ張られ，システム 2 に高い評価を与えてしまう．しかし，日英・英日翻訳など，並び替えが多く発生する言語対において，文の正しい並びを実現することが文の意味を正確に伝えるのが重要となる．この並べ替えの情報を重視する評価尺度として，RIBES が提案されている [10]．この並べ替えを自動評価可能な形に落とし込むために，出力文と参照文に対して，ケンダルの τ 順位相関係数 [11] を用いる．

その計算例として，正解の単語列が (a, b, c)，システム出力が (b, c, a) であった場合を考えよう．ケンダルの τ を計算する上で，「出力の単語列の中で，全ての単語対を比較した際に，正解の単語列と同じ順番となっている単語対の割合」をまず計算する．上記の計算例で単語対を列挙すると「a-b」「b-c」「a-c」が存在する．その中で，「b-c」はシステム出力で正解の単語列と同じ順番となっているが，「a-b」と「a-c」は正解と逆順となっている．このため，順位の正解率は 1/3 である．

RIBES はこの順位正解率をもとに作られた評価尺度であるが，順位正解率は同一の単語数，同一の単語の場合のみに適用可能である．この制約を緩めるため，RIBES は参照文とシステム出力の間で一致する単語のみに対して順位相関を計算し，語彙選択精度や文の長さを 2.2.1 節で説明した 1-gram 適合率と簡潔ペナルティを用いて評価する．この 3 つの評価使用を組み合わせたものが以下の式の通り，RIBES の評価値となる．

$$\mathrm{RIBES}(\boldsymbol{e}^*_k, \hat{\boldsymbol{e}_k}) = \mathrm{KT}(\boldsymbol{e}^*_k, \hat{\boldsymbol{e}_k}) * \mathrm{a}_1(\boldsymbol{e}^*_k, \hat{\boldsymbol{e}_k})^{\alpha} * \mathrm{BP}(\boldsymbol{e}^*_k, \hat{\boldsymbol{e}_k})^{\beta} \tag{9}$$

ここで KT は順位正解率であり，α と β は 1-gram 適合率と簡潔ペナルティの影響をコントロールするパラメータである．通常 $\alpha = 0.25$ と $\beta = 0.1$ が用いられる．

2.2.5 METEOR

上記の評価尺度は全て言語の完全一致に基づくものであり，微妙な表現や活用の違いに敏感である．例えば，「taro visited hanako」の例で，比較的近い「taro visits hanako」でも完全に異なる「taro entertained hanako」でも同等の評価値となる．この問題を克服する方法として，上記の評価尺度で複数の参照訳を用意し，評価の際に全ての参照訳を参考しながら評価値を計算することは可能である [21]．しかし，多くの場合複数の参照訳を用意することは現実的ではない．

複数の参照訳を用意しなくても表現の微妙な違いを吸収する評価尺度として METEOR が提案されている [1]．様々な言語で類義語集を用意したり，語幹だけのマッチを許したり，厳密に単語が一致しなくても，単語のマッチと判定する仕組みである．これにより，より正確な評価が可能となる一方，評価する言語に対して類義語集を用意する必要がある．[*1]

3. 実験設定

まず，本節実験に用いたデータや評価方法について述べる．

[*1] 2013 年 6 月現在，METEOR にはまだ日本語の類義語が含まれてないため，日本語に対する評価は METEOR の「言語非依存」設定で評価する．

表 3 各タスク (Task) における翻訳方式 (System)，チューニングに用いた評価尺度 (Tune)，コーパスごとの自動評価 (BLEU,BLEU+1,WER,TER,RIBES,METEOR)，2 名の評価者による許容性評価 (Human A,Human B)．太字は最も精度の高いシステムを示す

Task	System	Tune	BLEU	BLEU+1	WER	TER	RIBES	METEOR	Human A	Human B
IWSLT fr-en	PBMT	BLEU	**27.7**	32.4	54.4	52.3	82.2	32.2	**3.07**	2.55
	Hiero	BLEU	**27.7**	32.5	54.3	52.3	81.9	**32.4**	3.00	2.51
	PBMT	TER	27.3	**33.3**	51.3	49.5	**83.5**	31.9	3.03	**2.57**
	Hiero	TER	26.8	32.7	**51.1**	**49.3**	83.2	31.8	2.85	2.48
KFTT en-ja	PBMT	BLEU	25.4	30.3	78.2	69.3	68.2	43.5	2.24	1.98
	Hiero	BLEU	26.7	31.7	73.8	66.7	71.9	43.4	2.32	2.13
	F2S	B+R	**27.7**	**34.0**	68.6	63.2	75.2	**46.1**	2.70	2.42
	F2S	RIBES	25.3	32.3	**65.1**	**60.7**	**75.9**	42.3	**2.78**	**2.45**
MED en-ja	PBMT	BLEU	21.3	24.3	84.2	78.1	64.0	37.4	2.15	2.17
	Hiero	BLEU	**24.0**	26.8	78.5	73.2	69.0	**39.8**	**2.65**	**2.57**
	F2S	B+R	22.9	26.5	76.5	73.0	68.6	37.8	2.60	2.54
	F2S	RIBES	21.4	**27.8**	**69.1**	**66.8**	**69.2**	36.1	2.52	2.39
MED ja-en	PBMT	BLEU	17.6	26.1	77.6	71.6	63.3	23.9	2.22	**2.03**
	Hiero	BLEU	**20.1**	**26.9**	**74.0**	**68.9**	**64.2**	**24.4**	**2.33**	2.02
	T2S	B+R	14.0	23.6	85.9	79.7	55.1	21.6	1.98	1.84
	T2S	RIBES	13.8	23.4	82.5	76.9	57.6	21.2	2.01	1.82

表 2 各実験設定で用いた翻訳モデル学習データ (TM)，言語モデル学習データ (LM)，チューニングデータ (tune)，テストデータ (test) の単語数

	IWSLT	KFTT	MED
TM (ja/fr)	65.4M	9.41M	36.9M
TM (en)	58.8M	9.12M	25.4M
LM (ja/fr)	—	9.41M	38.2M
LM (en)	1.10G	—	716M
tune (ja/fr)	20.6k	26.8k	12.5k
tune (en)	20.2k	24.3k	10.0k
test (ja/fr)	3.64k	3.47k	3.11k
test (en)	3.52k	3.18k	2.20k

3.1 データと翻訳システム

評価の題材として 3 通りのデータ，4 通りの翻訳タスクを用いた：

IWSLT: IWSLT2012 ワークショップ [7] として配布されたデータを仏英翻訳システムの構築と評価に利用する．対象として TED[*2]の講演．おおよそのデータは [19] に説明されているとおりであるが，GIGA コーパスを利用しない．

KFTT: 情報通信研究機構により構築された日英京都関係 Wikipedia 記事を京都フリー翻訳タスク [17] で指定された学習・開発・テストセットを利用する．

MED: 医療に関する文書の日英・英日翻訳タスク．学習データに我々が収集した医療関係の文章に加えて，英辞郎辞書と例文[*3]，や上記の KFTT 学習データ，BTEC[23] などを利用する．

各コーパスの諸元を**表 2** に示す．

このデータに対して，翻訳システムを構築し，翻訳仮説を生成する．翻訳システムは Moses[12] を用いたフレーズベース [13] か階層的フレーズベース [4] システムや，Travatar[18] を用いた tree-to-string と forest-to-string システムを利用する．トークン化には，英語やフランス語で Moses に含まれるスクリプト，日本語では KyTea[20] を用いた．構文解析を用いるシステムでは，英語の構文解析器として Egret[*4]，日本語の構文解析器として Eda[8] と Travatar に含まれる，日本語の係り受け解析器を句構造機へと変換するルールを用いた．システムではデフォルトの設定を利用するが，最適化の際に BLEU や RIBES，BLEU と RIBES の線形補間から得られた評価関数 (B+R と記述)，TER などで最適化した．KFTT の英日，IWSLT の仏英，MED の日英・英日タスクに対して，4 つずつシステムを構築し，その詳細を**表 3** に示す．

なお，人手評価に関しては，KFTT と MED の翻訳タスクでは，評価者は両言語に精通しており，評価者への指示では原言語文と目的言語の参照訳の意味に隔たりがある場合，原言語文を優先するように指示した．IWSLT の場合，評価者は目的言語の英語のみに精通していたため，原言語文を参考にせず，目的言語参照文のみに基づいて評価を行うように支持した．これにより，IWSLT は KFTT と MED と若干異なる傾向がみられる可能性はあるが，両言語が理解できる評価者と目的言語文のみが理解できる評価者による評価に相関があることも先行研究により，ある程度確認されている [21]．

[*2] http://www.ted.org
[*3] http://www.eijiro.jp
[*4] http://code.google.com/p/egret-parser/

4. 人間評価の特定精度

4.1 許容性の推定精度

本節では，各自動評価尺度が誤った翻訳結果や良質な翻訳結果を文レベルで特定できるかどうかについて調査する．具体的には，許容性の 1 から 5 の各段階において，各自動評価尺度の中央値と 0.25 と 0.75 の信頼区間を調べる．この結果において，自動評価の中央値が人間評価と同じ昇順に並べば，自動評価は人間の評価と同様な情報を捉えており，ある程度翻訳文の優劣の判別に利用可能であることが分かる．さらに，信頼区間が小さければ小さいほど，ある人間評価の値に対して自動評価のばらつきが少なく，自動評価尺度の評価を信頼できると言えよう．

このような中央値と信頼区間を前節で述べた実験設定においてまとめた結果を**図 1** に示す．各行は 3.1 で説明したタスクを表し，左の 5 列は各自動評価尺度，右の 1 列は別の評価者に許容性を評価した値を用いている．グラフの 5 つの棒は正解として用いた評価者 B が 1 から 5 の評価値を付与した文に対する自動評価の中央値を指している．例えば，最も左上のグラフの最も左の棒が 0.22 となっているということは，IWSLT の仏英翻訳タスクにおいて，評価者が許容性 1 と評価した文の中で，BLEU+1 の中央値が 0.22 であったという意味となる．

この結果から様々な結論が読み取れる．まず，全体的な傾向として，全ての自動評価尺度における評価値は人間の評価と平均的に相関があることが分かる．中央値のおおよその傾向を見ると，BLEU，METEOR，RIBES は人間評価とともに上昇し，WER，TER は人間の評価が上昇するとともに低下する．[*5]

また，自動評価尺度の中央値は 1，2-4，5 という 3 つのグループに分かれる傾向も見られる．この結果から，自動評価尺度のほとんどは入力文の意味をなさない訳 (1)，完璧ではないがある程度意味が伝わる訳 (2-4)，完璧な訳 (5) という 3 つのグループをある程度判別できるが，2-4 の間の微妙な違いを判別することが困難であることが分かる．その中で，比較的 2-4 の間の中央値に差が見られるのは IWSLT と KFTT における RIBES と METEOR，MED en-ja における RIBES と，MED ja-en における BLEU と METEOR である．最後に別の評価者による許容性の評価を見ると，自動評価尺度より文の優劣を十分に評価できていることが分かる．MED en-ja を除いて，全ての段階において評価値の中央値がアノテータ間で一致していることが分かる．

次は，各許容性の段階における自動評価値の信頼区間に着目すると，全ての自動評価尺度において，信頼区間が重なっていることが多いことが分かる．従って，本研究で対象にした自動評価尺度がある評価値となったからと言って，人間の評価が悪い，もしくは良いと言い切ることができるほど判別能力が高いというわけではない．

[*5] MED ja-en において，5 と評価された文はこの傾向と逆方向に動いているが，MED ja-en は他のタスクと比べて全体的に精度が低かったため，5 と評価される文が少なく，中央値としての信頼度が低いためであると考えられる．

表 4 各評価尺度の誤り文特定効率

	IWSLT fr-en	KFTT en-ja	MED en-ja	MED ja-en
BLEU+1	**56.7%**	**51.5%**	59.2%	63.3%
WER	57.8%	52.2%	60.5%	63.3%
TER	58.2%	52.2%	61.2%	62.5%
RIBES	61.2%	53.6%	**58.1%**	**61.6%**
METEOR	60.1%	54.2%	58.5%	62.7%
Human A	46.5%	41.2%	39.2%	50.1%
Human B	32.6%	30.7%	32.1%	47.7%

4.2 誤り文の特定効率

この分析を更に深めるために，誤り分析のために誤訳を特定する評価尺度としての利用可能性の観点から見た統計を示す．**表 4** にまとめた統計は，自動評価尺度に基づいて，「評価の悪い順にシステム出力を見ていった際，許容性 1 の誤訳を 75%特定するまで，全文の何%を見る必要があるか」を表した数字である．完璧な評価尺度が存在した場合，この値は 1 と評価された文のちょうど 75%となる．つまり，無駄なく，閲覧した文がすべて誤訳であり，誤り分析を効率良く行うことが出来ると言える．

各自動評価尺度，別の評価者による許容性判定，オラクルの誤り文特定効率を示す．この結果から分かる通り，IWSLT と KFTT において誤り文を特定する効率は BLEU が最も良く，MED の ja-en と en-ja において RIBES が最も良かったが，最も効率の良い尺度と効率の悪い尺度の間にわずかな差しか見られなかった．これに比べて，別の評価者による人間評価はオラクルに近づき，他の評価尺度を大幅に上回っている．

5. システム間の文選択性能

前節の分析は，誤り分析などで用いられる翻訳システムにおける誤訳や良質な訳の特定に着目した．評価尺度のもう 1 つの役割として，複数の翻訳システムが同一の入力に対して出力した文を比較し，その優劣を判定するタスクがある．このような訳の判定はオンライン学習によるチューニング [24] や，システム統合 [2] などで重要となる．

本節では，各評価尺度が，同一の入力文に対する複数の翻訳候補の優劣を判定できるかどうかを調査する．調査方法として，まず各翻訳タスクにおいて，3.1 節で述べた 4 つの翻訳システムを用いて翻訳候補を生成する．次は，ある評価尺度の評価を行うために，各入力文に対して，4 つの候補の中からその評価尺度が最も良いと判定した候補を選択する．最後に，評価尺度によって選択された文を人間の

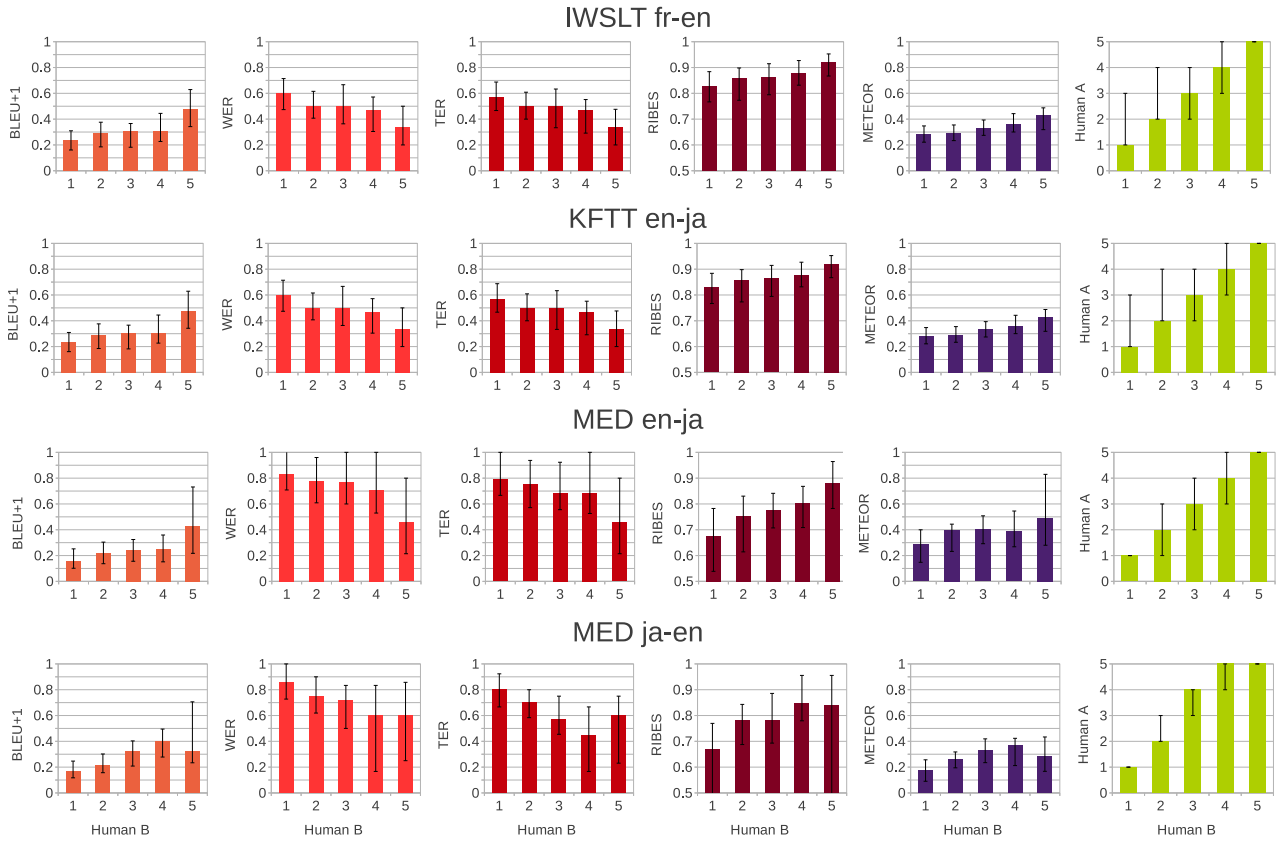

図 1 各タスクと人間評価値における自動評価尺度の中央値と 0.25，0.75 の信頼区間

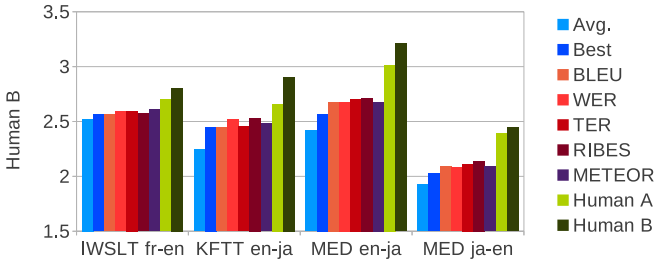

図 2 各評価尺度によって選択された文の平均許容性

許容性評価と照らし合わせて，選択された文の許容性の平均値を計算する．この平均値が高ければ高いほど，自動評価尺度は複数の候補の中から許容性の高い文を選択していると言える．

実験の結果を**図 2** に示す．左側の青で書かれている値は，4 システム全ての許容性の平均と最も許容性の高かったシステムの訳を既に選択した際の許容性の平均である．次の赤で書かれた 5 つの値はそれぞれの自動評価尺度に基づいて文の選択を行った際の結果であり，最後の 2 つの緑で書かれた値は別の評価者の人間評価値に基づいて文を選択した場合と，オラクルを示す．

まず，各自動評価尺度により選択された訳の平均的な許容性を，文選択を行わずに計算した全てのシステムの許容性の平均と比較すると，全てのタスクと評価尺度において，平均値を上回っていることが分かる．この結果から，全ての自動評価尺度は文レベルでもある程度システム間で許容性の高い訳と低い訳を弁別する能力があることが分かる．

しかし，自動評価尺度が選択した訳と，最も許容性の高いシステムの訳を比較すると必ずしも許容性の向上が見られるわけではない．具体的には，MED では評価尺度により選択された訳は最も良いシステムの訳を上回っているが，IWSLT と KFTT において顕著な差が見られなかった．このような状況において，システム統合などで自動評価尺度が最大となるようにシステムの出力を選択したとしても，自動評価尺度の値が増えても実際の人間評価に差が見られない恐れがある．

更に，各評価尺度のを比較すると，顕著な差が見られないが，おおよその傾向として仏英で METEOR により選択された文が最も高い許容性の平均となり，日英と英日翻訳において，RIBES が最も高い文選択能力を示した．しかし，別の評価者による許容性評価の選択能力に比べて，全ての自動評価尺度の選択性能が大幅に下回っていることが分かる．この結果から，全ての自動評価尺度は文レベルでは人間の評価に及ばないことが分かる．

6. 参照文数の影響

翻訳の自動評価の初期から，複数の参照文を用いて自動評価の信頼性を上げることが提案されている [21]．しかし，

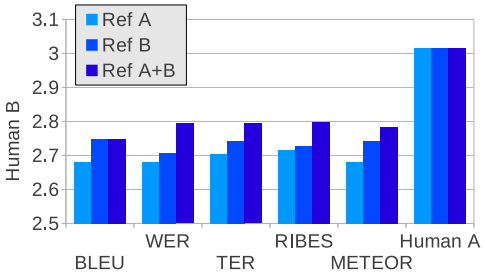

図 4 1 つもしくは複数の参照文を用いて選択された文の平均許容性

多くの場合，既存の翻訳データを用いて翻訳の学習と評価を行い，更に両言語に精通した翻訳者を雇って更なる参照文を作成するコストが高い．このため，ほとんどの翻訳タスクは 1 つの参照文しか使用しない．

本節で，更なる参照文の追加が上記のような誤り文特定能力やシステム間の翻訳結果優劣判定能力に与える影響を調べる．この調査では，最も自動評価と人間評価の差が大きかった MED en-ja に焦点を絞る．MED en-ja のデータに対して，もともとの参照文を参考にせずに，新たな参照文を作成した．下記の分析で，もともとの参照文に対して計算した自動評価値を Ref A，我々が作成した新たな参照文に対して計算した自動評価値を Ref B と呼ぶ．また，Ref A と Ref B 両方の最大値を取った複数の参照文を用いた評価値を Ref A+B と記述する．

まず，Ref A と Ref B，Ref A+B を参照文として用いた際の許容性特定能力に関する結果を**図 3** に示す．Ref A と Ref B を比較すると，Ref B の方が若干自動評価と人手評価の相関が高いことが分かる．その理由として考えられるのは我々が参照文を作成した際，Ref A に比べて忠実な（直訳に近い）訳し方を行っていることが考えられる．このような忠実な訳は機械翻訳システムが生成しうる訳に近く，より中の単語がマッチする確率が高いため，訳文の質の判断に有効であると考えられる．Ref A+B に関しては，中央値の順番と差や分散が Ref B とほぼ同等である．つまり，Ref B のみを用いた場合に比べて，文の評価特定精度が大幅に上回っているわけではないが，複数の参照文の中で比較的精度の良いものに沿った評価が行えると考えられる．

次に，システム間の優劣判定の結果を**図 4** に示す．この結果から，全ての評価尺度において，Ref B で計算された値が Ref A で計算された値より正確に許容性の高い文を選択している．更に，許容性の特定タスクと異なって，Ref A+B は BLEU 以外の評価尺度で 1 文しか用いない Ref A と Ref B を上回ったことも分かる．ここで特に注意すべき点として，評価尺度の選択と関係なく，2 つの参照文を用いた方が高い判定精度が実現可能である．つまり，本研究の対象となった評価尺度の差より，もう 1 つの参照文を用意することによる差の方がはるかに大きいことが分かる．とはいえ，2 つの参照文を用いた自動評価は人間の評価にまだ及ばず，2 つの参照文を用意するだけでは人間の評価者と同様の判定精度を実現できているわけではない．

7. おわりに

本研究では，機械翻訳の自動評価尺度を文レベルの評価に適用し，誤り文の特定性能とシステム間の文選択性能という 2 つの観点でその能力を検証した．その結果，調査対象とした BLEU，WER，TER，RIBES，METEOR の内，全ての評価尺度は人間の評価を大幅に下回っていることが分かり，文レベルの機械翻訳評価に大きな課題が残っていることが分かった．

今後の課題として，文レベルでも正確に翻訳結果の質を評価できる自動評価尺度の提案が残る．この問題を解決するために，新たな評価尺度の提案のみならず，タスクに特化した評価尺度の学習 [14] や既存の評価尺度の組み合わせなどを視野に入れていきたい．

謝辞：本研究の一部は，JSPS 科研費 25730136 の助成を受け実施したものである．

参考文献

[1] Banerjee, S. and Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, *Proc. ACL Workshop* (2005).
[2] Bangalore, S., Bordel, G. and Riccardi, G.: Computing consensus translation from multiple machine translation systems, pp. 351–354 (2001).
[3] Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R. and Specia, L.: Findings of the 2012 Workshop on Statistical Machine Translation (2012).
[4] Chiang, D.: Hierarchical phrase-based translation, *Computational Linguistics*, Vol. 33, No. 2 (2007).
[5] Chiang, D., Marton, Y. and Resnik, P.: Online large-margin training of syntactic and structural translation features, *Proc. EMNLP*, pp. 224–233 (2008).
[6] Echizen-ya, H. and Araki, K.: Automatic evaluation of machine translation based on recursive acquisition of an intuitive common parts continuum, *Proc. MT Summit*, pp. 151–158 (2007).
[7] Federico, M., Cettolo, M., Bentivogli, L., Paul, M. and Stüker, S.: Overview of the IWSLT 2012 Evaluation Campaign, *Proc. IWSLT*, Hong Kong, HK (2012).
[8] Flannery, D., Miyao, Y., Neubig, G. and Mori, S.: Training Dependency Parsers from Partially Annotated Corpora, *Proc. IJCNLP*, Chiang Mai, Thailand, pp. 776–784 (2011).
[9] Goto, I., Lu, B., Chow, K. P., Sumita, E. and Tsou, B. K.: Overview of the patent machine translation task at the ntcir-9 workshop, *Proceedings of NTCIR*, Vol. 9, pp. 559–578 (2011).
[10] Isozaki, H., Hirao, T., Duh, K., Sudoh, K. and Tsukada, H.: Automatic Evaluation of Translation Quality for Distant Language Pairs, *Proc. EMNLP*, pp. 944–952 (2010).
[11] Kendall, M. G.: A new measure of rank correlation, *Biometrika*, Vol. 30, No. 1/2, pp. 81–93 (1938).
[12] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and

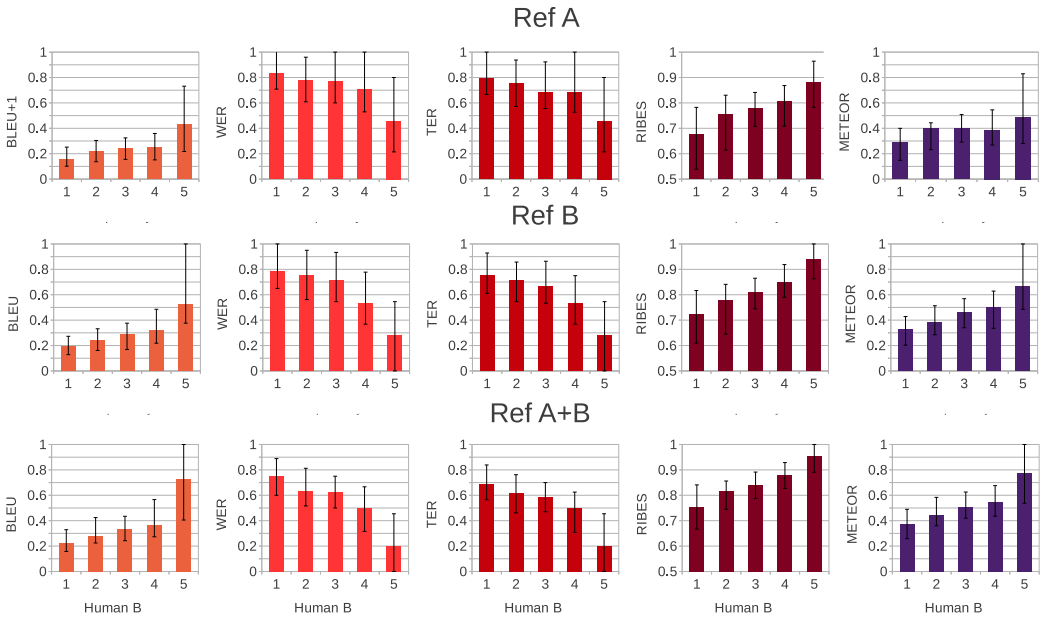

図 3 各参照文と人間評価値における自動評価尺度の中央値と 0.25，0.75 の信頼区間

Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. ACL*, Prague, Czech Republic, pp. 177–180 (2007).

[13] Koehn, P., Och, F. J. and Marcu, D.: Statistical phrase-based translation, *Proc. HLT*, Edmonton, Canada, pp. 48–54 (2003).

[14] Kulesza, A. and Shieber, S. M.: A learning approach to improving sentence-level MT evaluation, *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 75–84 (2004).

[15] Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions and reversals., *Soviet Physics Doklady.*, Vol. 10, No. 8, pp. 707–710 (1966).

[16] Lin, C.-Y. and Och, F. J.: Orange: a method for evaluating automatic evaluation metrics for machine translation, *Proc. COLING*, pp. 501–507 (2004).

[17] Neubig, G.: The Kyoto Free Translation Task, `http://www.phontron.com/kftt` (2011).

[18] Neubig, G.: Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers, *Proc. ACL Demo Track*, Sofia, Bulgaria (2013).

[19] Neubig, G., Duh, K., Ogushi, M., Kano, T., Kiso, T., Sakti, S., Toda, T. and Nakamura, S.: The NAIST Machine Translation System for IWSLT 2012, *Proc. IWSLT* (2012).

[20] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proc. ACL*, Portland, USA, pp. 529–533 (2011).

[21] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, *Proc. ACL*, Philadelphia, USA, pp. 311–318 (2002).

[22] Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J.: A study of translation edit rate with targeted human annotation, pp. 223–231 (2006).

[23] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S.: Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world, *Proc. LREC*, pp. 147–152 (2002).

[24] Watanabe, T., Suzuki, J., Tsukada, H. and Isozaki, H.: Online Large-Margin Training for Statistical Machine Translation, *Proc. EMNLP*, pp. 764–773 (2007).