2022年度（令和4年度）　修士論文

STR-MRRを用いた
機械翻訳の自動評価

鳥取大学大学院　持続性社会創生科学研究科
工学専攻　情報エレクトロニクスコース

自然言語処理研究室

M21J4043C　新田 玲輔

概要

本研究は機械翻訳の自動評価についての研究を行う．自動評価ではSTR(Sentence Translation Ratio)[8] といった手法が先行研究として取り組まれていた．これは正誤判定で翻訳評価をする手法である．しかし，STR は正解数が極端に少ない．さらに STR の評価では機械翻訳の出力文のうち 1-best しか翻訳評価をしない．そのため人手評価と差が生じている．そこで本研究では，先行研究の STR に加え，機械翻訳の出力文を N-best 評価する STR-MRR(Mean Reciprocal Rank)[9] という手法で人手評価との差を縮めることを目的とする．本研究では先行研究の STR と提案手法の STR-MRR を比較し，どちらがより人手評価に近い手法かを検証した．その結果，STR-MRR はSTR より人手評価に近い手法であると結論づけた．

目 次

図 目 次

表 目 次

第1章　はじめに

機械翻訳には翻訳評価が必要である．翻訳評価は人手で行うとコストがかかるため自動的に行う自動評価が存在する．自動評価については過去に石原 [8] が STR(Sentence translation ratio) の研究を行った．しかし，STR には問題点がある．STR は機械翻訳で出力される参照文と出力文との完全一致で正誤判定をする．完全一致の手法では正解がとても少なくなる．機械翻訳は一つの入力文に対し出力文を N-best 出力するが，STR は 1-best の出力文しか翻訳評価をしない．ゆえに自動評価値が極端に低くなり，人手評価と差が生じている．

それに対し N-best の出力文を自動評価すると，2-best 以降に正解が見つかり，人手評価との差を縮めることができると考える．そこで先行研究の STR に加え，N-best の出力順に着目した STR-MRR(Mean Reciprocal Rank)[9] という手法を提案する．本研究の目的は N-best の評価をすることで STR より人手評価との差を縮めることである．そして実験を通して先行研究の手法と提案手法を比較する．本研究ではそれぞれの自動評価と人手評価との相関係数をとり，より人手評価との相関が強い手法を優れたものとする．

本稿の構成は以下の通りである．第 2 章で従来の手法について説明する．第 3 章で先行研究の STR について説明し，第 4 章で提案手法の STR-MRR の説明をする．第 5 章で実験に用いる翻訳システム，人手評価の方法，実験結果の求め方を説明し，第 6 章で実験結果について述べる．第 7 章では考察，第 8 章で今後の課題，第 9 章で本研究のまとめについて述べる．

第2章　様々な自動評価

2.1　BLEU[1]

BLEU スコアは以下の式を用いて参照文と翻訳文との類似度を算出する．まず計算式を以下の式 (2.1) に示す.

$$BLEU = BP_{BLEU} \times \exp(\frac{1}{N}\sum_{n=1}^{N} \ln P_n) \tag{2.1}$$

ただし，

$$P_n = \frac{\sum i \text{ 翻訳文と参照文 } i \text{ で一致した } n-gram \text{ 数}}{\sum i \text{ 翻訳文 } i \text{ の全 } n-gram \text{ 数}} \tag{2.2}$$

ここで，BP_{BLEU} は参照文より文字列の長さが短い出力文が高い評価値にならないように補正するペナルティである．N とは N-gram のことである．BLEU の具体的な評価例を以下の表 2.1.1 とともに説明する.

表 2.1.1: BLEU の評価例

入力文	その窓はなかなか閉まらない。
参照文	The window won't shut.
出力文	The window won't close.

表 2.1.1 のような翻訳の場合，各 N-gram より，$P_1 = \frac{4}{5}$，$P_2 = \frac{2}{4}$，$P_3 = \frac{1}{3}$，$P_4 = 0$ つまり，BLEU スコアは以下のとおりとなる.

$$BLEU = 1 \times exp(\frac{1}{4}(\ln\frac{4}{5} + \ln\frac{2}{4} + \ln\frac{1}{3})) = 0.537 \tag{2.3}$$

2.2 NIST[2]

NIST は BLEU をベースとした手法であり，N-gram 適合率で評価を行う．NIST は 0 から∞のスコアを出力し，スコアが大きいほど大きい評価としている．計算式を次の式 (2.4) に示す．

$$NIST = \sum_{n=1}^{N} \left(\frac{\sum_i (\sum_{\text{出力文 } i \text{ と参照文 } i \text{ に共通する} \omega_1 ... \omega_n} Info(\omega_1 ... \omega_n)}{\sum_i \text{出力文 } i \text{ の中の全 } N-gram \text{ 数}} \right) \tag{2.4}$$

$$Info(\omega_1 ... \omega_n) = \log_2 \frac{\text{評価コーパス中の} \omega_1 ... \omega_{n-1} \text{数}}{\text{評価コーパス中の} \omega_1 ... \omega_n \text{数}} \tag{2.5}$$

NIST は BLEU と違い，各 $n-gram$ 長における適合率を計算する時，情報量で重み付けを行っている．それに加え，各 $n-gram$ 長における適合率を計算する時，単純和を用いていることと，短い文に対するペナルティ関数も BLEU と異なっている．

2.3 METEOR[3]

METEOR は適合率と再現率を考慮した自動評価である．単語属性が正しい場合に高いスコアを出す．METEOR は 0 から 1 までのスコアを出力し，スコアの大きい方が良い評価となる．スコアの算出方法を以下の式 (2.8) に示す．

$$F \text{ 値} = \frac{P \times R}{\alpha \times P + (1-\alpha) \times R} \tag{2.6}$$

$$Pen = \gamma \times (\frac{c}{m})^{\beta} \tag{2.7}$$

$$METEOR = F \times (1 - Pen) \tag{2.8}$$

METEOR は F 値と Pen を用いて評価値を算出する．F 値は適合率 P と再現率 R の調和平均で求められる．そして，Pen はペナルティ関数であるが，これは単語の非連続性に対するペナルティである．式 (2.7) において，m は参照文と出力文の単語の一致率であり，c は参照文と出力文を比較して，単語が一致したときの単語列を一つのまとまりとして，そのまとまりの数を示している．式 (2.6) と式 (2.7) において，α と β と γ はパラメータである．

α=0.8，β=2.5，γ=0.4 として METEOR の計算方法を表 2.3.1 を用いて説明する．

表 2.3.1: METEOR の評価例

入力文	その窓はなかなか閉まらない。
参照文	The window won't shut.
出力文	The window won't close.

表 2.3.1 より，A は出力文の単語数，B は参照文の単語数，C は参照文と出力文が一致した数として計算を行う．

$$\text{適合率 } P = \frac{C}{A} = \frac{4}{5} \tag{2.9}$$

$$\text{再現率 } R = \frac{C}{B} = \frac{4}{5} \tag{2.10}$$

$$F \text{ 値} = \frac{\frac{4}{5} \times \frac{4}{5}}{0.8 \times \frac{4}{5} + (1 - 0.8) \times \frac{4}{5}} = 0.8 \tag{2.11}$$

$$Pen = 0.4 \times (\frac{2}{4})^{2.5} = 0.070 \tag{2.12}$$

$$METEOR = 0.8 \times (1 - 0.070) = 0.744 \tag{2.13}$$

2.4 RIBES[4]

RIBES は参照文と出力文との間で，共通単語の出現順序を順位相関係数で評価する評価法である．計算式は以下のとおりである．

$$RIBES = NKT \times P^{\alpha} \tag{2.14}$$

$$NKT = \frac{\sum_{i=1}^{n-1} K_i - \sum_{i=1}^{n-1} L_i}{\frac{n(n-1)}{2}} \tag{2.15}$$

ここで，NKT はケンドールの順位相関係数である．P は共通単語が少ない場合のペナルティである．α はペナルティに対する重みとして使用され，値の範囲は $0 \leq \alpha \leq 1$ である．n は文の単語数，d_i は参照文の i 番目の単語と出力文の i 番目の単語の語順の差分である．

K_i は，以下の二つの単語列の共起数である．

- 出力文における，出力文 i 番目の単語以降の単語列
- 参照文における，出力文 i 番目の単語以降の単語列

L_i は，以下の二つの単語列の共起数である．

- 出力文における，出力文 i 番目の単語以前の単語列
- 参照文における，出力文 i 番目の単語以前の単語列

RIBES は，単語の出現順を順位相関係数を用いて評価することで，文全体の語順に着目することができる．RIBES は 0 〜 1 のスコアを出力し，スコアが大きい方が良い評価である．具体的な計算例を説明する．　表 2.4.1 より，各共起数は以下のとおりとなる．

表 2.4.1: $RIBES$ 算出のための例文

入力文	雨に濡れたため、彼は風邪を引いた。
参照文	He caught a cold because he got soaked in the rain.
出力文	He got soaked in the rain because he caught a cold.

$$K_1 = 5, K_2 = 4, K_3 = 3, K_4 = 2, K_5 = 1$$
$$K_6 = 0, K_7 = 0, K_8 = 3, K_9 = 2, K_10 = 1$$

$$L_1 = 5, L_2 = 5, L_3 = 5, L_4 = 5, L_5 = 5$$
$$L_6 = 5, L_7 = 4, L_8 = 0, L_9 = 0, K_10 = 0$$

$$NKT = \frac{21 - 34}{\frac{11 \times 10}{2}} = -0.23 \tag{2.16}$$

$$RIBES = \frac{-0.23 + 1}{2} = 0.3850 \tag{2.17}$$

2.5 TER[5]

TER とは，翻訳結果のエラー率を算出する自動評価である．具体的には，出力文が参照文に近づける時に行う修正 (置換，挿入，削除，シフト) の割合を算出する．つまり，TER は値が小さければ小さいほど良い評価なのである．TER が 30 %以下の数値になると，翻訳の品質が良いとされる．

第3章　先行研究：STR(Sentence Translation Ratio)[8]

3.1　従来の自動評価

まず，機械翻訳の過程について述べる．我々人間がテストの問題を解くために勉強することと同様に，コンピュータは翻訳をするために機械学習を行う．そして，入力文を入力した後，翻訳の模範解答となる参照文と出力文を出力する．基本的に自動評価は参照文と出力文を比較し，それぞれ違った計算方法でスコアリングを行い翻訳評価する．

大抵の自動評価は参照文と出力文の部分一致で翻訳評価する．しかしその場合，以下のようなことが起こる．

表 3.1.1: 1 単語の相違による影響

入力文	信号が青になった。
参照文	The signal turned green.
出力文	The signal turned red.

表 3.1.1 の場合，入力文の「青になった」の意味に対し，出力文は“ red ”と真逆の意味が出力されている．しかし部分一致の評価では,“ green ”と“ red ”以外の単語連語は一致しているので高い評価値がでてしまう．ここで人手評価と差が出てしまう．

3.2 STR の手法

石原 [8] は従来の自動評価では人手評価と差が出てしまうと考えた．そこで部分一致ではなく文全体を評価する完全一致が必要だと考え，STR(Sentence Translation Ratio) の研究を行った．

完全一致とは参照文と出力文を比較し，全単語が完全に一致か否かを正誤判定する．完全一致であれば正解，完全一致でなければ不正解とする．

表 3.2.1: STR の正誤判定

参照文	The signal turned green.	
出力文 1	The signal turned red.	不正解
出力文 2	The signal turned green.	正解

表 3.2.1 のより出力文 1 は参照文と比べて“ green ”と“ red ”で違っているので不正解，出力文 2 は完全一致しているので正解とする．このように STR は文を全体的に評価することを可能とする．

本研究では相関係数を求めるため STR を数値化する．数値化を行う時，STR が正解であれば $STR = 1$，不正解であれば，$STR = 0$ とする．

3.3 STRの問題点

STR は文全体の評価を可能としているが，問題点がある．STR は参照文と出力文との完全一致で翻訳評価している．つまり一単語でも違っていれば不正解になる．よってSTR の研究では正解数が極端に少なくなり，人手評価と差が出てしまう．

機械翻訳は入力文 1 個に対して N-best 出力文を出力する．しかし，STR は N-best の出力文のうち 1-best しか翻訳評価しない．すなわち以下の表 3.3.1 のような問題が起こる．

表 3.3.1: 機械翻訳の出力例

入力文	ブランコが揺れている。
参照文	The swing is swinging.
1-best	The stake is down.
2-best	The swing is on the ebb.
3-best	The swing is swung in balance.
4-best	The swing is swinging.

例として表 3.3.1 のように機械翻訳が行われたとする．参照文に対して STR で正解となる文が 4-best に現れている．しかし 1-best のみを評価する STR では表 3.3.1 の場合でも正解が考慮されない．こうした要因がまた人手評価との差を引き起こしていると考えられる．

本研究では STR を用いて N-gram の出力文を翻訳評価し，先行研究の STR より人手評価との差を縮めることを目的とする．

第4章　提案手法：STR-MRR

4.1　STR-MRRの原理

本研究では先行研究のSTRに加えて，N-best評価するためにSTR-MRRを提案する．一つの入力文に対してN-best出力文がある場合，出力順にランク付けを行う．そして以下の式(4.1)で翻訳評価を行う．

$$STRMRR = \sum_{n=1}^{8} \frac{STR}{rank_n} \tag{4.1}$$

なお，STRは正解の場合1，不正解の場合0とする．$rank_n$はSTRが正解のときのランクである．具体的な計算例を表4.1.1とともに述べる．

表 4.1.1: 機械翻訳の出力例

入力文	彼は仕事で京都に行った。
参照文	He went to Kyoto on business.
$rank_1$	He went to Kyoto on business.
$rank_2$	He went to Kyoto on work.
$rank_3$	He went to Kyoto on business.
$rank_4$	He went over to Kyoto on business.

表4.1.1のように結果が出た時，STRが正解した出力文は下線部の出力文である．
したがって，STR-MRRの計算方法は以下の式(4.2)となる．

$$STRMRR = \frac{1}{1} + \frac{0}{2} + \frac{1}{3} + \frac{0}{4} \tag{4.2}$$

式(4.2)のようにして翻訳評価を行う．N-bestの評価をすることで人手評価との差を縮めることができると考える．

4.2 STR-MRRの実際の評価例

ここで実際に実験で行っているSTR-MRRの評価例を述べる.

表 4.2.1: STR-MRR の実例

入力文	彼は自身の権利を主張した。
参照文	He claims his own rights.
$rank_1$	He claims his rights.
$rank_2$	He supports his rights.
$rank_3$	He claims his own rights .
$rank_4$	He claims his rights in his claim .
$rank_5$	He claims his rights in his opinion .
$rank_6$	He claims his rights to his own rights .
$rank_7$	He claims his rights in his own claim .
$rank_8$	He claims his rights to his own .

$$STRMRR = \frac{0}{1} + \frac{0}{2} + \frac{1}{3} + \frac{0}{4} + \frac{0}{5} + \frac{0}{6} + \frac{0}{7} + \frac{0}{8} = \frac{1}{3} = 0.333 \tag{4.3}$$

表 4.2.2: STR-MRR の実例

入力文	彼は仕事で京都に行った。
参照文	He went to Kyoto on business .
$rank_1$	He went to Kyoto on business .
$rank_2$	He went to Kyoto by work .
$rank_3$	I went to Kyoto on business .
$rank_4$	He went to Kyoto on work .
$rank_5$	I went to Kyoto on his job .
$rank_6$	I went to Kyoto on his job . future .
$rank_7$	I went to Kyoto on his job . being .
$rank_8$	I went to Kyoto on his job . night .

$$STRMRR = \frac{1}{1} + \frac{0}{2} + \frac{0}{3} + \frac{0}{4} + \frac{0}{5} + \frac{0}{6} + \frac{0}{7} + \frac{0}{8} = \frac{1}{1} = 1 \tag{4.4}$$

表 4.2.3: STR-MRR の実例

入力文	ガラスは割れやすい。
参照文	Grass breaks eaily .
$rank_1$	Glass is liableto a break .
$rank_2$	Glass is liableto break .
$rank_3$	Glass is formed easily .
$rank_4$	Glass is liableto a leak .
$rank_5$	Glass is liableto cracked .
$rank_6$	Glass is liableto a break . being .
$rank_7$	Glass is liableto a leak . being .
$rank_8$	Glass is liableto a break . neck .

$$STRMRR = \frac{0}{1} + \frac{0}{2} + \frac{0}{3} + \frac{0}{4} + \frac{0}{5} + \frac{0}{6} + \frac{0}{7} + \frac{0}{8} = 0 \tag{4.5}$$

表 4.2.4: STR-MRR の実例

入力文	彼女の声は震えた。
参照文	Her voice wavered .
$rank_1$	Her voice wobbled .
$rank_2$	Her voice trembled .
$rank_3$	Her voice shook .
$rank_4$	Her voice had a quaver in it .
$rank_5$	Her voice wavered .
$rank_6$	Her voice had a quaver .
$rank_7$	Her voice had a quaver in me .
$rank_8$	Her voice had a quaver in him .

$$STRMRR = \frac{0}{1} + \frac{0}{2} + \frac{0}{3} + \frac{0}{4} + \frac{1}{5} + \frac{0}{6} + \frac{0}{7} + \frac{0}{8} = \frac{1}{5} = 0.2 \tag{4.6}$$

第5章　実験方法

5.1　実験手順

実際に行っている実験の手順は以下のとおりである．本研究の目的は先行研究の STR と提案手法の STR-MRR と比較することである．本研究ではそれぞれの自動評価と人手評価とで相関係数をとり，より人手評価との相関係数が高い手法を優れたものとする．なお，本研究では十分な信頼性を得るため実験を 3 回繰り返す．

手順 1

機械翻訳を行う．

手順 2

各テスト文に対する出力文 1-best を STR と人手で翻訳評価する．

手順 3

STR と人手評価との相関係数を求める．

手順 4

各テスト文に対する出力文 N-best を STR-MRR と人手で翻訳評価する．

手順 5

STR-MRR と人手評価との相関係数を求める．

手順 6

手順 3 と手順 5 によって得た相関係数を比較する．

5.2 翻訳システム

5.2.1 NMT について

本研究では翻訳システムとして NMT(Neural Macine Translation)[6] を用いる．NMT とは，人間の脳神経回路が情報伝達を行う仕組みを参考にしたものであり，人工的なニューラルネットワークが情報を収集し，自ら学習しながら単語の意味として正しい訳語を当てはめる翻訳システムである．NMT の手順を以下に示す．

手順 1

対訳学習文の日本語文と英語文をアテンションモデルに学習させる

手順 2

入力文をベクトル化する．

手順 3

ベクトル化された入力文とアテンションモデルをもとに出力文へと翻訳する．

NMT の概略図を以下の図 5.1 に示す．

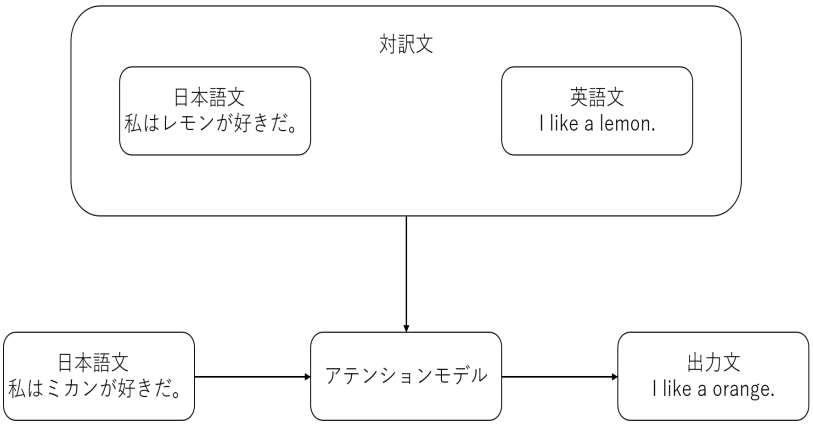

図 5.1: NMT

5.2.2 NMT のデータ

本研究に用いる NMT のデータを以下の表 5.2.2 に示す.

表 5.2.1: NMT データの内訳

対訳学習文数	160000
テスト文数	100
テスト文 1 文に対する N-best	8

本研究では実験を 3 回繰り返すが, 同じデータで行うわけではない. 本研究では実験に用いる NMT のデータを変更した. NMT の翻訳過程で変更した部分は NMT の学習ステップ数である. 学習ステップとは NMT が機械学習をする上で用いる乱数である. ステップ数を変更すれば翻訳精度も変わる. ステップ数が大きいほど翻訳精度は良くなる.

表 5.2.2: NMT データの内訳

	NMT ステップ数
実験 1 回目	25000
実験 2 回目	50000
実験 3 回目	100000

そして, 一つの指標として本実験に用いた NMT 各ステップ数の BLEU 値を以下の表 5.2.3 に示す.

表 5.2.3: NMT データの内訳

NMT ステップ数	25000	50000	10000
BLEU スコア	0.132	0.136	0.153

表 5.2.3 より, NMT のステップ数と翻訳精度が比例していることが分かる.

本実験では対訳学習文数 160000 対を機械翻訳に使用する．また，本実験で使用する対訳学習文対は日本語文と英語文の対である．使用する学習文対は電子辞書などの例文より抽出した単文データである．学習文対の例を表 5.2.4 に示す．

表 5.2.4: NMT データの内訳

学習文対例 (1)	
日本語文	ピアノの勉強にヨーロッパに行く。
英語文	Go to Europe to study the piano.

学習文対例 (2)	
日本語文	公園は川まで広がっている。
英語文	The park reaches to the river.

学習文対例 (3)	
日本語文	きょうは時折小雪のちらつく寒い一日だった。
英語文	It was a cold day today with occasional light snowfall.

5.3 人手評価

出力文の人手評価はすべて5段階評価で行う．数字が大きいほど評価が高いとする．人手評価の値は自動評価と相関係数を取るために人手評価値として計算に利用する．各人手評価値と評価の内容を以下の表5.3.1に示す．

表 5.3.1: 評価基準

人手評価値	評価判断
5	入力文が完璧に翻訳できている
4	入力文がほぼ翻訳できている
3	不備な部分もあるが入力文の意味を捉えることはできている
2	入力文をあまり翻訳できていない
1	入力文を全く翻訳できていない

本実験において，人手評価は相関係数の結果を出すために最重要な作業である．つまり明確な評価基準を持つことが重要である．しかし表5.3.1だけでは抽象度が高いため評価基準を明確に分けることができない．そこで各人手評価値となる事例を述べる．この事例は本実験で実際に行った人手評価とは別である．なお，本実験において人手評価をするとき，参照文との比較は行わない．

5.3.1 評価 5 の事例

評価 5 は完璧に翻訳ができていると言うことである．本実験の「完璧」の定義づけについて述べる．本研究では,

条件 1 入力文の重要な情報の 100 %を出力文で翻訳することができている

条件 2 文法的な間違いが一切ない

この二つ条件を両方満たしていることを「完璧」と定義する．ここで評価 5 となる具体的な事例を述べる．

表 5.3.2: 完璧な翻訳

入力文	彼は仕事で京都に行った。
出力文	He went to Kyoto on business.

表 5.3.2 が評価 5 となる理由を解説する．まず，入力文を解剖すると，入力文において重要な情報は

情報 1 主語を表す「彼」

情報 2 目的を表す「仕事」

情報 3 地名を表す「京都」

情報 4 行動，そして時制を表す「行った」

この 4 つである．表 5.3.2 の場合，入力文におけるこれらの情報を間違いがなく翻訳ができているため完璧の条件 1 を満たしている．そして，文法的な間違いもないため条件 2 も満たしている．よって評価 5 とする．

また，本実験において人手評価は翻訳の実用性にも着目する．そのため入力文の情報のうち，数字の情報と地名や人名といった固有名詞は最重要である．これらを間違えた場合，他の要素を翻訳できていたとしても実用性においては致命的であるので注意が必要である．

5.3.2　評価 4 の事例

評価 5 と評価 1 については評価をはっきり分けることができるが，評価 2〜4 は評価をする境界線が曖昧になる．つまり具体的な事例を増やす必要がある．
表 4 は「ほぼ完璧」という評価であるが，この定義づけも行う．本実験において

定義 1 入力文の情報に対し微量の翻訳漏れ，あるいは誤りがある

定義 2 文法的な間違いが一部存在する

定義 3 意味に大きな影響はないが余計な単語が加わっている

以上の 3 つの定義が一つでも当てはまれば評価 4 となる．評価 4 となる具体的な事例をいくつか述べる．

表 5.3.3: 時制の誤り

入力文	私は車で海に行った。
出力文	I go to sea by car.

表 5.3.3 において入力文の必要な情報は，

情報 1 主語を表す「私」

情報 2 手段を表す「車」

情報 3 場所を表す「海」

情報 4 時制を表す「行った」

である．表 5.3.3 において，誤っている情報は情報 4 のみである．加えて，手段や場所と比べると時制の情報は優先度が低いと考える．よって評価 4 となる．

表 5.3.4: 文法の小さな誤り

入力文	私はりんごが好きです。
出力文	I like apple.

基礎的な文法として，冠詞のつかない可算詞は複数形にする決まりがある．しかし表 5.3.4 の場合,“ apple ”は前置詞がついていないのにもかかわらず，複数形になっていない．つまり文法上の誤りである．しかし入力文のりんごが“ apple ”と翻訳できているため，このような文法上の誤りはほんの一部分に過ぎないと考えるため評価 4 とする．

表 5.3.5: ピリオド後の余計な単語

入力文	彼はアメリカに住んでいる。
出力文	He lives in America. in.

表 5.3.5 において入力文の必要な情報は,

情報 1 主語を表す「彼」

情報 2 場所を表す「アメリカ」

情報 3 行動，時制を表す「住んでいる」

である．表 5.3.5 ではすべての情報が完璧に翻訳できている．しかし，出力文のピリオドの後ろをみると余計な単語“ in ”が入っている．実際，NMT で翻訳するときにこのようなことはよく起こる．しかし，ピリオドで一度区切っているのであれば“ in ”という単語は出力文の意味自体にはさほど影響はしないと判断するためこの場合は評価 4 とする．

5.3.3　評価 3 の事例

評価 3 は幅が広い．とても長い文章が出力されるとき，評価 3 が多い傾向がある．本実験において

条件 1 入力文の情報に対しいくつかの翻訳漏れ，あるいは誤りが存在する

条件 2 文法的な間違いがやや多く存在する

条件 3 文全体の意味に影響が出る誤った単語や句の出現

以上の 3 つの条件が主に評価 3 となる要因である．評価 3 となる具体的な事例をいくつか述べる．

表 5.3.6: 情報の部分的欠落

入力文	富樫は背が低いが、スピードがありとても優れたバスケ選手だ。
出力文	Togashi is nice, but he is a nice basketball player.

表 5.3.6 において入力文の必要な情報は,

情報 1 主語を表す「富樫」

情報 2 特徴を表す「背が低い」「スピード」

情報 3 属性を表す「バスケ」

である．機械翻訳では表 5.3.6 のように長い文を完璧に翻訳するのは難しい．この事例では重要度の高い主語と属性は翻訳ができている．しかし，情報 2 の重要度も高いが，こちらは翻訳できていない．この事例をまとめると，入力文で翻訳できている情報は半分程度であると考える．本実験ではこういった出力文を評価 3 としている．

これまでは入力文の情報に着目して厳密な減点方式での事例を挙げてきた．しかし，表5.3.6 より難しく複雑な文を機械翻訳する時，ほとんど翻訳ができてない支離滅裂な結果が出力されることも少なくない．そこで，複雑な入力文の翻訳には加点方式で人手評価を行うこともある．

表 5.3.7: 単語の足し算

入力文	広陵高校が甲子園に歩を進めた。
出力文	Koryo approach walk Koshien.

入力文の中には「歩を進めた」のような慣用句が含まれていることもある．学習の足りない機械翻訳では慣用句を翻訳することができず表5.3.7 のように直訳してしまうことが頻繁に起こる．しかし，表5.3.7 の出力文は文法が良くないが，重要な情報の“ Koryo ”と“ Koshien ”が翻訳できている．そして“ approach ”が加わることで単語のみで文全体の意味を推測することが可能であると判断できる．この場合，評価は3 としている．

表 5.3.8: 誤解を招きかねない余計な句

入力文	これはペンである。
出力文	This is a pen of the pen.

機械翻訳の中には表5.3.8 のような出力も見られる．この出力文には“ of the pen ”と余計な連語がピリオドで区切られずに含まれている．表5.3.8 のような誤解を招く可能性があると判断した場合，評価は3 とする．

5.3.4 評価 2 の事例

評価 2 は評価 3 との境界線が曖昧である．そこで以下の点で評価 2 と評価 3 との差別化を行う．入力文の意味を部分的に含んでいる出力文のうち，

選別 1 文中の重要度が高い情報を翻訳できているか

選別 2 文中の単語は入力文の全体的な意味をつかめるものか

以上の 2 点より条件を満たしていれば評価 3，満たしていなければ評価 2 とする．
評価 2 の事例を以下の表 5.3.9 に示す．

表 5.3.9: 情報の少ない出力文

入力文	愛人を失った彼女は泣いていた。
出力文	She cried.

表 5.3.9 では「彼女が泣いた」は翻訳ができているが，原因を表す「愛人を失った」の部分が翻訳できていない．原因は文において重要な部分であると考える．なぜならば「彼女が泣いた」だけでは「なぜ？」という疑問点が出てくる．出力文の 2 単語だけでは入力文の全体的な意味を掴むことができない．つまり上の選別 2 の条件を満たせていないため評価は 2 となる．

5.3.5　評価 1 の事例

評価 1 は入力文の意味が全く翻訳できていない出力文である．そこで，全く翻訳できていないについて本実験では次のように定義づけする．

定義 1 入力文と出力文が反対の意味をもつこと

定義 2 入力文の必要不可欠な情報が欠落している

定義 3 出力文自体が理解不能な翻訳になっていること

ここで評価 1 の事例を以下に示す．

表 5.3.10: 反対の意味

入力文	私は右利きです。
出力文	I am left handed.

表 5.3.10 の場合，出力文は「私は左利きです.」と入力文と反対の意味になっている．この場合，本実験では評価 1 とする．続いて評価 1 の事例を紹介する．

表 5.3.11: 最重要な情報の誤り

入力文	今日は東京で雨が降るでしょう.
出力文	It will be rain at Tottori.

表 5.3.11 の場合，地名を表す「東京」は必要不可欠な情報である．特に固有名詞は他の変えがきかない重要な単語であるので固有名詞を間違えると，まるきり別の文になってしまう．よって表 5.3.11 の時，評価は 1 である．

5.4 STR と人手評価との相関係数の求め方

まず NMT によって翻訳した出力文の 1-best のみを STR と人手評価で評価する．そして，以下の表 5.4.1 のようなデータを生成する．

表 5.4.1: STR の実験データ

テスト文		
参照文		
1-best	STR	人手評価値

ここで表 5.4.1 の評価例を表 5.4.2 と表 5.4.3 に示す．STR は正解なら 1，不正解なら 0 を出力する．

表 5.4.2: 先行研究の実験データの例 1

入力文: 腹がゴロゴロ鳴った。		
参照文:His stomach rumbled .		
	STR	人手評価値
My stomach growled softly .	0	3

表 5.4.3: 先行研究の実験データの例 2

入力文: 指輪はどこにも見つからなかった。		
参照文:The ring was nowhere to be found .		
	STR	人手評価値
The ring was nowhere to be found .	1	5

そして，表 5.4.1 のデータを 100 個集め，次式のように相関係数を算出する．

$$r_{STR} = \frac{\frac{1}{n}\sum_{i=1}^{n}(STR_i - \overline{STR})(\text{人手評価値}_i - \overline{\text{人手評価値}})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(STR_i - \overline{STR})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{人手評価値}_i - \overline{\text{人手評価値}})^2}} \tag{5.1}$$

5.5 STR-MRR と人手評価との相関係数の求め方

5.5.1 人手 MRR の求め方

本研究では STR-MRR は 8-best の評価を行う．そして 8-best を評価し STR-MRR と相関を取る．人手評価は STR-MRR に揃えるため，以下の式 (5.4) によって人手 MRR として数値化する．

$$\text{人手 } MRR = \sum_{n=1}^{8} \frac{\text{人手評価値}}{rank_n} \tag{5.2}$$

ここで人手 MRR の求め方を，例を用いて述べる．

表 5.5.1: 人手 MRR の例

	人手評価値
1-best	5
2-best	5
3-best	5
4-best	4
5-best	2
6-best	3
7-best	3
8-best	2

表 5.5.1 のように人手評価を行ったとすると，人手 MRR の値は次式のようになる．

$$\begin{aligned} \text{人手 } MRR &= 5 \times \frac{1}{1} + 5 \times \frac{1}{2} + 5 \times \frac{1}{3} + 4 \times \frac{1}{4} + 2 \times \frac{1}{5} + 3 \times \frac{1}{6} + 3 \times \frac{1}{7} + 2 \times \frac{1}{8} \\ &= 5 + 2.5 + 1.67 + 1 + 0.4 + 0.5 + 0.429 + 0.25 \\ &= 11.7 \end{aligned} \tag{5.3}$$

5.5.2 STR-MRR と人手評価との相関

まず，それぞれのテスト文において以下の表 5.5.2 をもとに STR-MRR と人手 MRR を算出する．

表 5.5.2: 提案手法の実験データ

テスト文		
参照文		
1-best	$STR/1$	人手評価値/1
2-best	$STR/2$	人手評価値/2
3-best	$STR/3$	人手評価値/3
4-best	$STR/4$	人手評価値/4
5-best	$STR/5$	人手評価値/5
6-best	$STR/6$	人手評価値/6
7-best	$STR/7$	人手評価値/7
8-best	$STR/8$	人手評価値/8
	$STRMRR$	人手 MRR

ここで表 5.5.2 の評価例を表 5.5.3 に示す．

表 5.5.3: 提案手法の実験データ

入力文: 値段は大きさで違う。		
参照文:The prices differ according to size .		
Prices differ according to size .	0/1	5/1
Prices vary according to size .	0/2	5/2
Prices differ by size .	0/3	5/3
The prices differ according to size .	1/4	5/4
Prices differ according to the size .	0/5	5/5
Prices differ according to the size of the size .	0/6	4/6
Prices differ according to the size of size .	0/7	4/7
rices differ according to the size of a size .	0/8	4/8
	0.25	13.1548

そして，表 5.5.2 のデータを 100 個集め，次式のように相関係数を算出する．

$$r_{STRMRR} = \frac{\frac{1}{n}\sum_{i=1}^{n}(STRMRR_i - \overline{STRMRR})(\text{人手 } MRR_i - \overline{\text{人手 } MRR})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(STRMRR_i - \overline{STRMRR})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{人手 } MRR_i - \overline{\text{人手 } MRR})^2}} \tag{5.4}$$

第6章　実験結果

6.1　人手評価

5.3 節に示した例は，実際に人手評価を行う前に参考にするための作例である．本節は実際に行った人手評価の結果を解説する．人手評価は本稿の著者のみであるが，合計 2400 文の出力文の評価を 2ヶ月の間隔を開け 2 回繰り返した．その結果，k 係数は 0.89 と安定した人手評価となっている．

6.1.1　評価 5

評価が 5 となった例文をいくつか解説する．

表 6.1.1: 評価 5 の実例 1

入力文	彼は額の汗を拭った。
出力文	He wiped the sweat on his forehead.

入力文において重要な情報は

情報 1 主語を表す「彼」

情報 2 目的語を表す「額の汗」

情報 3 行動，そして時制を表す「拭った」

である.「彼」に対して“ He ”,「額の汗」に対して“ the sweat on his forehead ”,「拭った」に対して“ wiped ”と完璧に翻訳できている．そして文法的な誤りも一切ないため評価 5 である．

表 6.1.2: 評価 5 の実例 2

入力文	どうぞお楽にしてください。
出力文	Make yourself at home.

この文では単語ではなく,“ make yourself at home ”この組み合わせで「くつろいでください」という意味なので，入力文の意味を完璧に訳しているため，評価 5 とした．翻訳評価をするときは，表 6.1.2 のように熟語や飛訳にも注意して評価をする必要がある．

表 6.1.3: 評価 5 の実例 3

入力文	我々は特別番組を放送した。
出力文	We broadcast a special program.

入力文において重要な情報は

情報 1 主語を表す「我々」

情報 2 目的語を表す「特別番組」

情報 3 行動，そして時制を表す「放送した」

である.「我々」に対して“ We ”,「特別番組」に対して“ a special program ”,「放送した」に対して“ broadcast ”と入力文の情報を完璧に翻訳することができている．ここで注目すべきなのが“ broadcast ”である．こちらは現在形と同じなのだが例外的に過去形にも変化できる．つまり“ broadcast ”は過去形としても使える単語である．このように中には，過去形で綴りが変化しない単語もあるため注意が必要である．文法的な間違いもないため評価は 5 である．

6.1.2 評価 4

評価が 4 となった例文をいくつか解説する.

表 6.1.4: 評価 4 の実例 1

入力文	当社商品はよく売れる。
出力文	Our product is selling well . being .

入力文において重要な情報は

情報 1 主語を表す「当社商品」

情報 2 状態を表す「よく売れる」

である.「当社商品」に対して,“ Our product ”,「よく売れる」に対して“ is sellling well ”と翻訳ができている．そして文法に間違いはない．ここまでは完璧な翻訳であったが，ピリオドの後に“ being ”という余計な単語が入ってしまったため減点である．よって，評価は 4 である.

表 6.1.5: 評価 4 の実例 2

入力文	私は酒で頭がぼんやりしてしまった。
出力文	My head was distracted with drink .

入力文において重要な情報は

情報 1 主語を表す「私」

情報 2 原因を表す「酒で」

情報 3 状態を表す「頭がぼんやり」

情報 4 時制を表す「してしまった」

である.「酒で」に対して“ drink ”,「頭がぼんやり」に対して“ distracted ”を過去形で翻訳できているため情報の翻訳はできている．しかし,“ drink ”の前の“ with ”単語は文法上正しくない．しかし，文の意味への影響は小さいため評価は 4 となる.

表 6.1.6: 評価 4 の実例 3

入力文	車輪が上着を掠った。
出力文	The wheels trailed their jacket .

入力文において重要な情報は

情報 1 主語を表す「車輪」

情報 2 目的語を表す「上着」

情報 3 出来事，時制を表す「掠った」

である.「車輪」に対して“ wheel ”,「上着」に対して“ jacket ”とこの二つは翻訳することができている．しかし,「掠った」が翻訳できていない.“ trailed ”は「かじる」という意味である．ただ,「かじる」は「掠った」の比喩表現の可能性も考えられる．つまり評価は 4 となった.

表 6.1.7: 評価 4 の実例 4

入力文	彼女の声は震えた。
出力文	Her voice had a quaver in me .

入力文において重要な情報は

情報 1 主語を表す「彼女の声」

情報 2 状態を表す「震えた」

である.「彼女の声」に対して“ Her voice ”,「震えた」に対して“ had a quaver ”と入力文の情報は翻訳できている．しかし表 6.1.6 の例文ではピリオドの前，つまり文中に“ in me ”といった余計な単語列が入っている．しかし“ in me ”自体に意味を持つ事例はないので影響力の小さい余分な単語列として評価は 4 とした.

6.1.3 評価 3

評価が 3 となった例文をいくつか解説する.

表 6.1.8: 評価 3 の実例 1

入力文	雨で試合が流れた。
出力文	The game was rained away .

入力文において重要な情報は

情報 1 主語を表す「試合」

情報 2 状態を表す「雨天」

情報 3 結果を表す「流れた」

である.「試合」に対して“ The game ”,「雨天」に対して“ rained ”が翻訳できている. しかし，重要な結果の情報である「流れた」が翻訳できていない. しかし,“ away ”という単語があることから，なくなったと遠回しの表現をしている可能性もあると考えた. よって，評価は 3 である.

表 6.1.9: 評価 3 の実例 2

入力文	彼は戦闘中にそこで殺された。
出力文	He was killed there by the battle.

入力文において重要な情報は

情報 1 主語を表す「彼」

情報 2 状態を表す「戦闘中」

情報 3 結果を表す「殺された」

である.「彼」に対して“ He ”,「殺された」に対して“ was kille ”は翻訳できている．しかし「戦闘中」は翻訳できていないと考える．入力文の意味から考えると，彼を殺したのは戦争そのものではなく戦争の中の誰かであると考えるからである．しかし，比喩表現としては合っている可能性もある．しかし今回は評価 3 とした.

表 6.1.10: 評価 3 の実例 3

入力文	お先に失礼します。
出力文	Excuse me, but I am leaving now.

入力文において重要な情報は

情報 1 お先に失礼します (先に出ていきます)

である．直訳すると「すみません、今私は出ていってます。」となり文章の成り立ちとしては好ましくない．しかし，出力文のうち,“ Excuse me ”,“ I ”,“ leaving ”といった単語から，これから出ていくという入力文の意図を推測することが可能であるので評価は 3 である．

表 6.1.11: 評価 3 の実例 4

入力文	田園生活を知った。
出力文	I found of life in rural life ago.

入力文において重要な情報は

情報 1 目的語を表す「田園生活」

情報 2 行動，時制を表す「知った」

である.「田園生活」に対して“ rural life ”,「知った」に対して“ I found ”と情報はしっかりと翻訳できている．しかし，6.1.11 の出力文のうち,“ found ”に続く語に“ of life in ”や“ rural life ”の後に続く“ ago ”と不要な単語が加わっている．このため，全体的に入力文の意図が伝わりにくい文章になってしまった．故に評価は 3 とした．

表 6.1.12: 評価 3 の実例 5

入力文	その新人女優は体当たりの演技で新人賞を獲得した。
出力文	The new actress won the prize performance by a performance performance .

入力文において重要な情報は

情報 1 主語を表す「新人女優」

情報 2 手段を表す「体当たりの演技」

情報 3 目的語を表す「新人賞」

情報 4 行動，時制を表す「獲得した」

である.「新人女優」に対して,“ new actress ”,「獲得した」に対して“ won ”と情報 1 と情報 4 は完璧に翻訳できている．しかし,「体当たりの演技」に対しては“ performance ”,「新人賞」に対しては“ the prize ”と入力文の細かい内容に対して出力文が抽象的になっている．とくに「新人賞」などといった目的語は重要度が高いと考える．つまり表 6.1.12 の出力文は入力文の意味を十分に翻訳できているとは言えない．しかし，出力文のうち,“ new actress ”,“ won ”,“ the prize ”,“ by a performance ”の単語が存在することによって，出力文から大まかな内容を掴むことは可能である．よって部分的な欠落として評価を 3 とする．

6.1.4 評価 2

評価が 2 となった例文をいくつか解説する．

表 6.1.13: 評価 2 の実例 1

入力文	彼女は思いがけない質問にまごついたようだった。
出力文	She seemed to be embarrassed.

入力文において重要な情報は

情報 1 主語を表す「彼女」

情報 2 原因を表す「思いがけない質問」

情報 3 状態，時制を表す「まごついた」

である．主語と状態，時制は翻訳できているが，表 6.1.13 において入力文の肝となる要素は，原因である．5.3.4 節の条件より，表 6.1.13 の出力文は，重要度の高い単語の翻訳がまだ不十分である．また出力文の語だけでは入力文の全体的な意味を掴むことができないと判断したため，評価 2 とする．

表 6.1.14: 評価 2 の実例 2

入力文	転職を考えています。
出力文	I think about my job .

入力文において重要な情報は

情報 1 目的語を表す「転職」

情報 2 行動，時制を表す「考えています」

である．出力文は「仕事について考えている」という意味にはなっている．確かに，転職も仕事に関連する話である．間違いではないのかもしれない．しかし，出力文の「仕事について考えている」だけでは，転職の話をしていると推測するのは困難であると考える．よって評価は 2 とした．

表 6.1.15: 評価 2 の実例 3

入力文	富士山は昔からたくさんの絵描きに描かれた山だ。
出力文	Mount Fuji has been a lot of admires who were a young man.

入力文において重要な情報は

情報 1 主語を表す「富士山」

情報 2 時を表す「昔から」

情報 3 対象を表す「たくさんの絵描き」

である．このうち，翻訳ができている情報は，主語だけである．出力文を直訳すれば入力文の意味とは全く違った文である．しかし,「描かれた山」は「絵師たちの憧れ」からだと仮定すれば，入力文と出力文は全くかけ離れた文とは言い難いと考え評価 2 とした．

表 6.1.16: 評価 2 の実例 4

入力文	子供たちの旅行の話は食べ物に集中した。
出力文	The children's jorney on the chileren centered on food.

入力文において重要な情報は

情報 1 主語を表す「子供たちの旅行の話」

情報 2 目的語を表す「食べ物」

である.「子供たちの旅行」,「食べ物に集中」といった熟語は出力文中には存在するが,語順や単語がバラバラであり文としての完成度は低い．よって評価は 2 となる．

6.1.5　評価 1

評価が 1 となった例文をいくつか解説する．

表 6.1.17: 評価 1 の実例 1

入力文	その窓はなかなか閉まらない。
出力文	He is quite close to the window.

表 6.1.17 の場合，入力文に対して出力文の意味は「彼はその窓にとても近い。」となっており，意味が全くかけ離れているため評価は 1 である．

表 6.1.18: 評価 1 の実例 2

入力文	音波で鼓膜が振動する。
出力文	The eardrum vibrated with moonlight.

表 6.1.18 の場合，入力文において最も重要となるキーワードが「音波」である．しかし，出力文では入力文の「音波」が翻訳できておらず,“ moonlight ”，つまり「月光」と翻訳されている.「本日の朝ごはんはパンです。」が“ Today’s breakfast is rice. ”と翻訳されるように実用性の観点から見れば最重要な部分が翻訳できていない出力文は致命的である．よって評価は 1 とする．

表 6.1.19: 評価 1 の実例 2

入力文	その窓はなかなか閉まらない。
出力文	The window will close.

表 6.1.19 の場合，入力文と出力文が反対の意味になっているので評価 1 とした．

6.2　自動評価と人手評価の分散図

STR と STR-MRR それぞれの自動評価と人手評価との散布図を報告する．なお，STR の分散図において，プロット付近の数字は重複した数を示している．

6.2.1　実験 1 回目

実験 1 回目の各自動評価と人手評価との散布図を以下に示す．

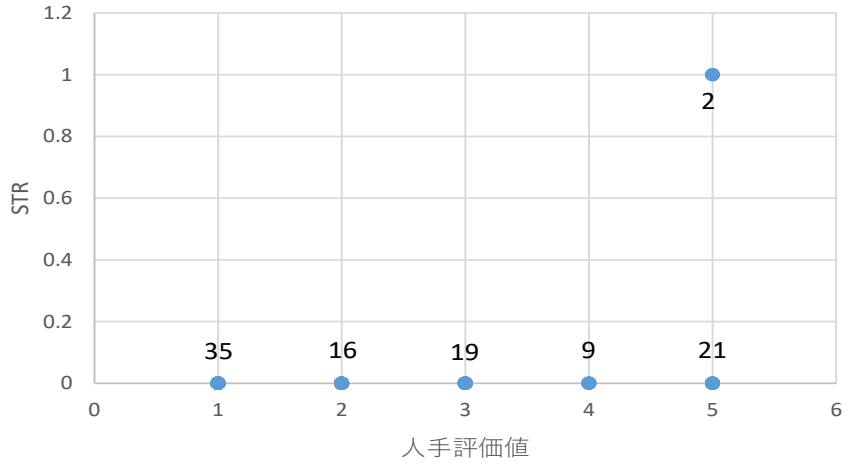

図 6.1: STR と人手評価との散布図 (1 回目)

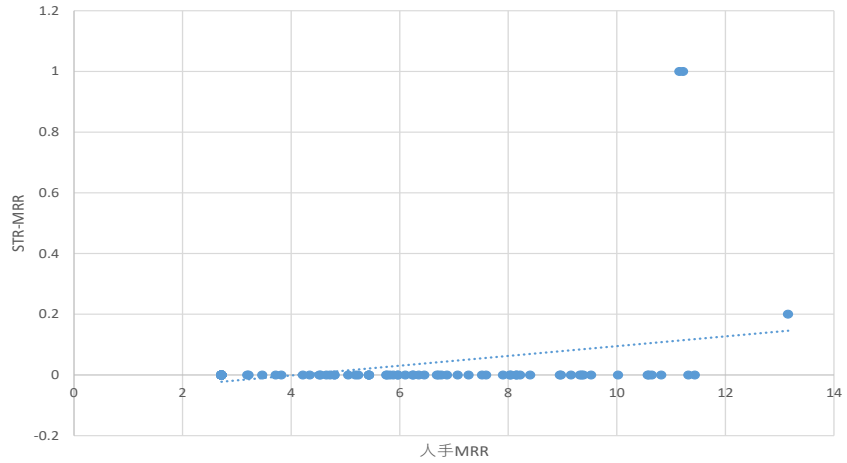

図 6.2: STR-MRR と人手評価との散布図 (1 回目)

6.2.2　実験 2 回目

実験 2 回目の各自動評価と人手評価との散布図を以下に示す.

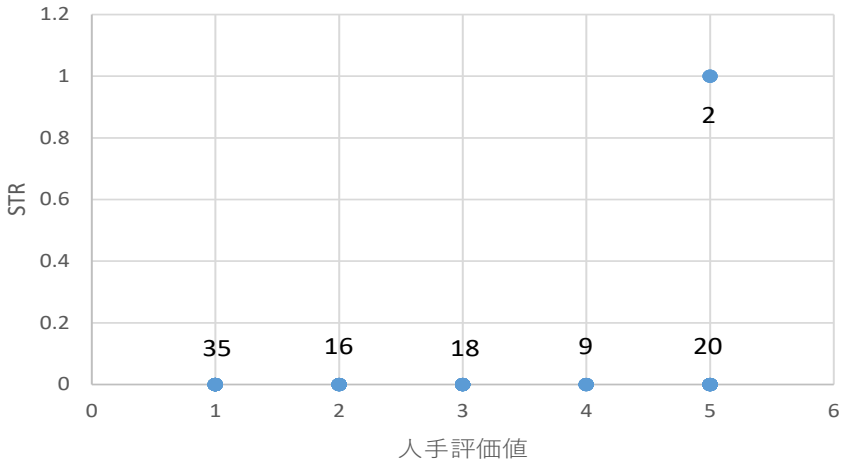

図 6.3: STR と人手評価との散布図 (2 回目)

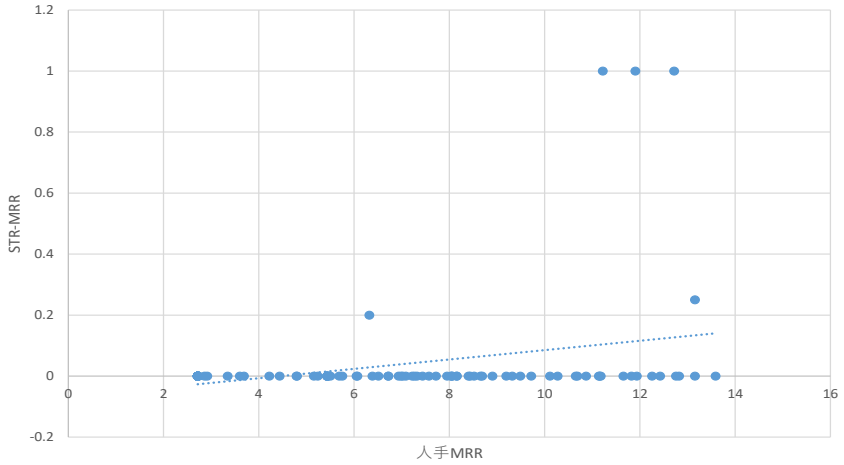

図 6.4: STR-MRR と人手評価との散布図 (2 回目)

6.2.3　実験 3 回目

実験 3 回目の各自動評価と人手評価との散布図を以下に示す.

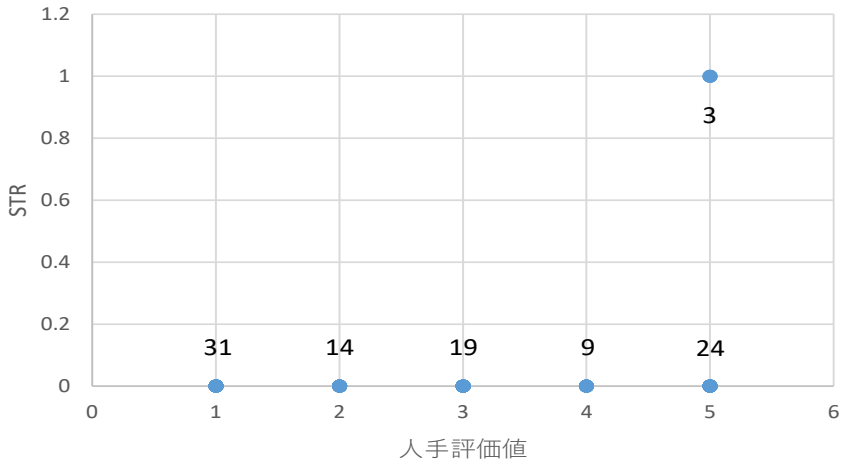

図 6.5: STR と人手評価との散布図 (3 回目)

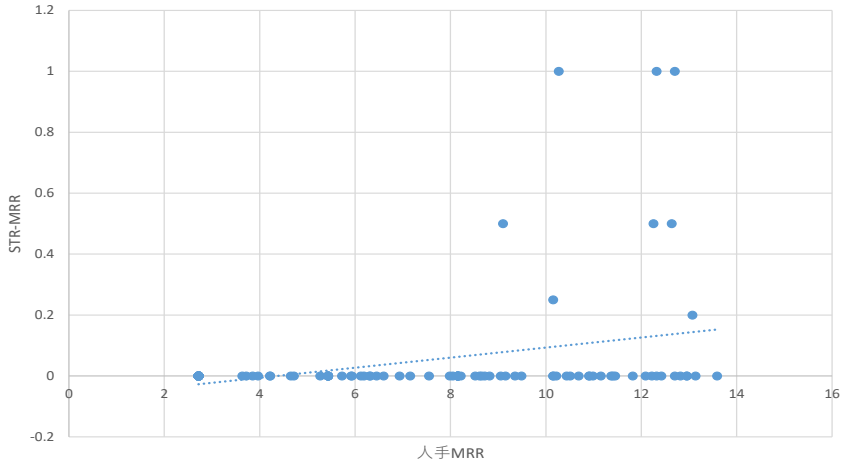

図 6.6: STR-MRR と人手評価との散布図 (3 回目)

6.3 自動評価と人手評価との相関係数

高い信頼性を得るために，本研究では実験を 3 回繰り返した．提案手法と先行研究それぞれの人手評価との相関係数は以下の表 6.3.1 に示す．

表 6.3.1: 各自動評価と人手評価との相関係数

	実験 1 回目	実験 2 回目	実験 3 回目
提案手法 STR-MRR	0.324	0.298	0.310
先行研究 STR	0.301	0.263	0.235

表 6.3.1 よりいずれの実験でも提案手法の STR-MRR は先行研究の STR より相関係数が上回る結果となった．

第7章　考察

7.1　N-best 評価

STR は 1-best のみの完全一致で評価をしていた．では N-best ではどのくらい完全一致が存在するのか調査をした．表 7.1.1 に各実験において STR が正解した $rank_n$ を示す．

表 7.1.1: 各自動評価と人手評価との相関係数

	実験 1 回目	実験 2 回目	実験 3 回目
$rank_1$	2	3	3
$rank_2$	0	0	3
$rank_3$	0	0	0
$rank_4$	0	1	1
$rank_5$	1	1	1
$rank_6$	0	0	0
$rank_7$	0	0	0
$rank_8$	0	0	0

1-best のみ評価する STR は正解数が極端に小さくなってしまう．それが人手評価との差が生まれる原因である考えられている．表 7.1.1 より $rank_2$ 以降にも STR が正解となる出力文が存在することがわかった．つまり，STR においては N-best 翻訳を評価することで人手評価との差を縮めることができたと考えられる．しかし，それだけでは N-best が優れている理由としては不十分である．なぜならば本研究において N-best 評価が人手評価との差を縮められたのは，先行研究の自動評価値が極端に低いからである．つまり極端に低い自動評価値を出さない手法には N-best 評価は効果がないということとなってしまう．

本実験で人手評価をしたうち，$rank_1$ よりも $rank_2$ 以降の方が優れていると判断した数を以下の表 7.1.2 に示す．

表 7.1.2: 第 1 位より第 2 位以降が人手評価が高い割合

	実験 1 回目	実験 2 回目	実験 3 回目
$rank_2$ 以降○	24/100	27/100	28/100

表 7.1.2 のとおり，全体の約 25 %，いわば 4 つのうち 1 つは $rank_1$ よりも $rank_2$ 以降の方が優れているのである．さらに 3 回の実験ともおよそ 25 %の確率を保っている．要するに翻訳を行えば確かに一定数は $rank_1$ よりも $rank_2$ 以降の方が優れている事例が存在するのである．つまり N-best の翻訳は STR に限らずに必要であると考える．

7.2 提案手法と 1-best 人手評価との相関

本実験では提案手法と相関を取る人手 MRR は 8-best の評価であったが，8-best では人手評価の手間がかかる．そこで，本実験の提案手法と 1-best のみの人手評価値との分散と相関を調査した．以下に散布図と相関係数を示す．なお，散布図では重複したプロットがあるが，その場合，プロット上に数字を示している．

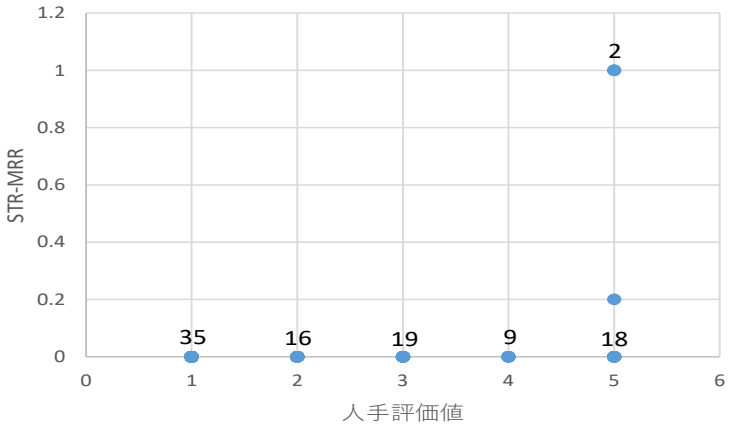

図 7.1: STR-MRR と 1-best 人手評価との散布図 (実験 1 回目)

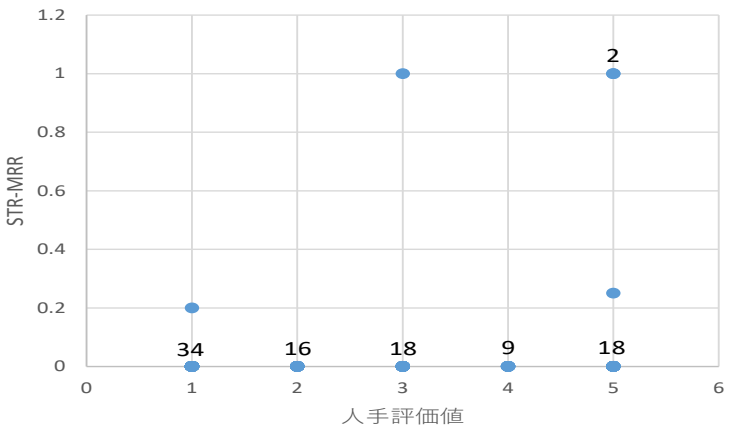

図 7.2: STR-MRR と 1-best 人手評価との散布図 (実験 2 回目)

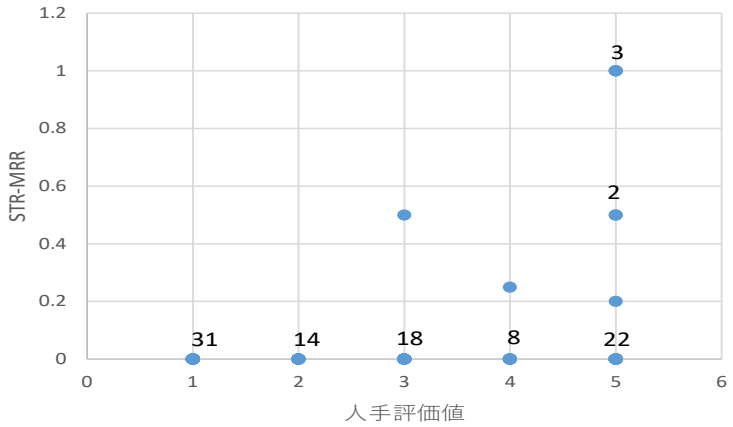

図 7.3: STR-MRR と 1-best 人手評価との散布図 (実験 3 回目)

そして，STR-MRR と 1-best のみの人手評価値との相関係数を STR と 1-best のみの人手評価値との相関係数と比較した．以下の表 7.2.1 に結果を示す．

表 7.2.1: STR-MRR と 1-best 人手評価との相関係数

	実験 1 回目	実験 2 回目	実験 3 回目
$STRMRR$-1best 人手	<u>0.329</u>	0.199	<u>0.307</u>
STR	0.301	<u>0.263</u>	0.235

表 7.2.1 より提案手法と相関をとる人手評価が 1-best のみであっても，実験 3 回分の合計では先行研究 STR の手法より相関が強いことが分かる．しかし，2 回目の実験では STR-MRR と 1-best 人手評価との相関係数が STR と人手評価との相関係数を下回っている．さらに，$STRMRR$ と 1-best 人手評価との相関係数と，$STRMRR$ と 人手 MRR との相関係数を比較した結果を表表 7.2.2 に示す．

表 7.2.2: STR-MRR と 1-best 人手評価との相関係数

	実験 1 回目	実験 2 回目	実験 3 回目
$STRMRR$-1best 人手	<u>0.329</u>	0.199	0.307
$STRMRR$-人手 MRR	0.324	<u>0.298</u>	<u>0.310</u>

表 7.2.2 の結果より，$STRMRR$ と 1-best 人手評価との相関係数は $STRMRR$ と 人手 MRR との相関係数より相関が低い．よって N-best 評価と相関を取る人手評価は N-best の方が好ましいと考えられる．しかし，3 回のうち 1 回は $STRMRR$ と 1-best 人手評価との相関係数が上回っている．翻訳精度によって変動がある可能性もあるため，今後は NMT のステップ数を変更して実験量を更に増やす必要があると考える．

7.3　提案手法と従来自動評価との比較

本実験では先行研究の STR と提案手法の STR-MRR を比較した．本実験を通して提案手法が先行研究より優れた手法であると結論づけることはできた．ここで，従来の自動評価と人手評価との相関を求めたので以下に示す．それぞれテスト文は 100 文で行った．また，本節では y 座標のプロットが密集しているため．箱ヒゲ図によって分布を示す．

7.3.1　実験 1 回目 (分布図)

実験 1 回目におけるそれぞれの自動評価と人手評価との分布図を示す．

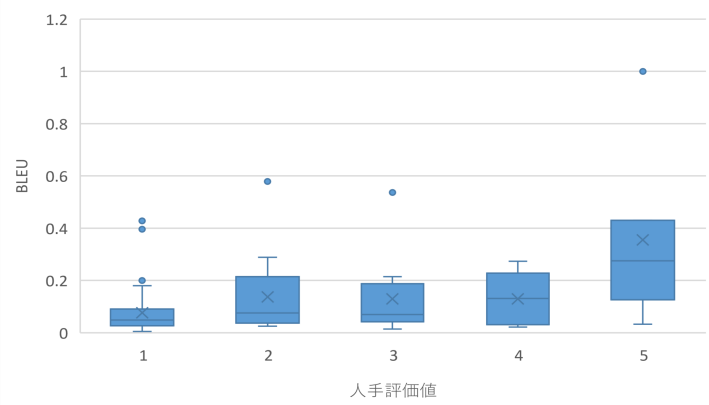

図 7.4: BLEU と人手評価との分布図 (1 回目)

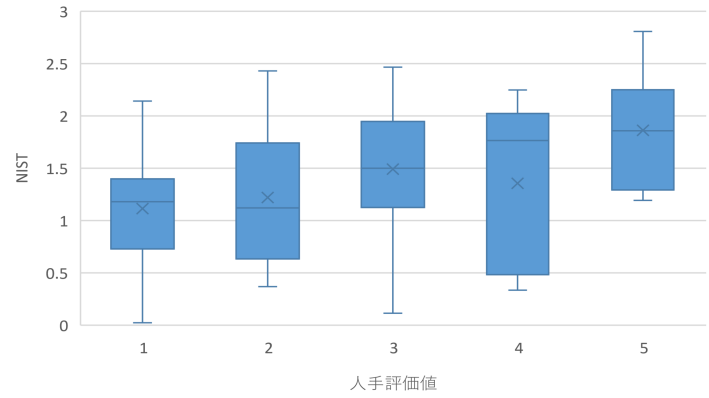

図 7.5: NIST と人手評価との分布図 (1 回目)

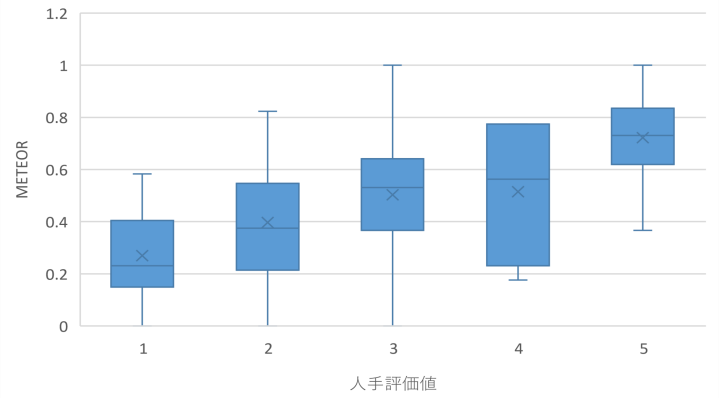

図 7.6: METEOR と人手評価との分布図 (1 回目)

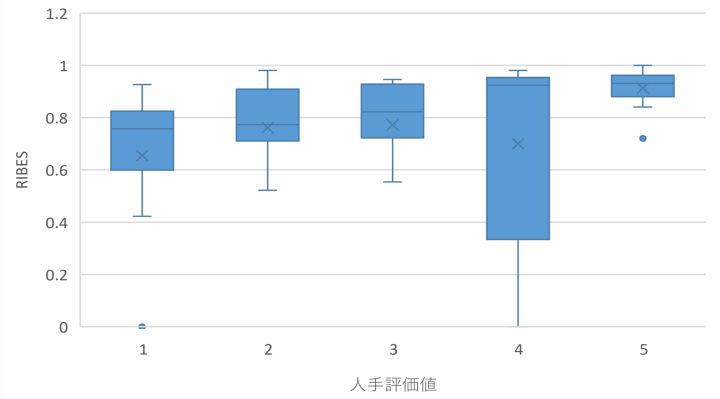

図 7.7: RIBES と人手評価と分布図 (1 回目)

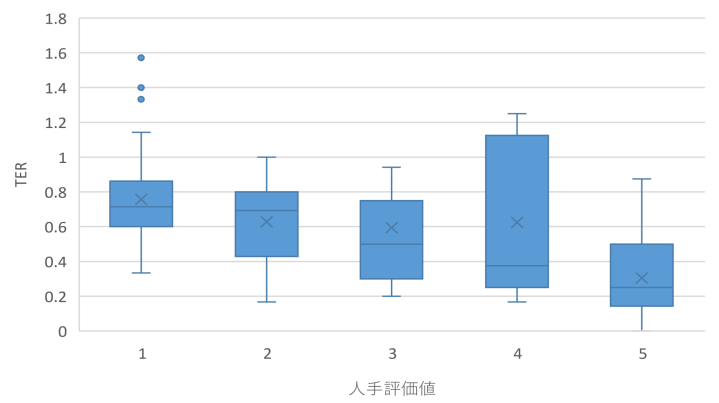

図 7.8: TER と人手評価との分布図 (1 回目)

7.3.2 実験 2 回目 (分布図)

実験 2 回目におけるそれぞれの自動評価と人手評価との分布図を示す.

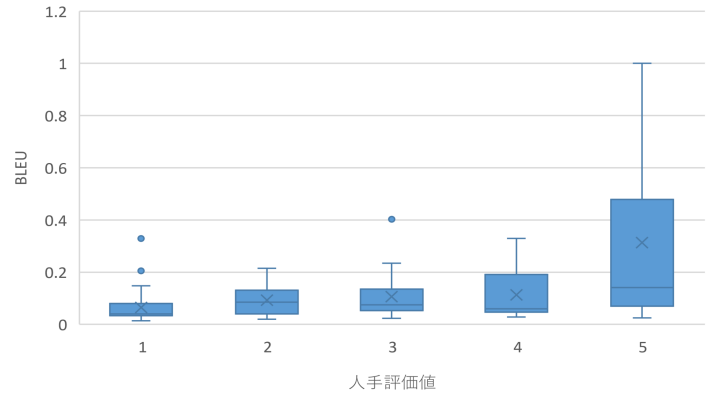

図 7.9: BLEU と人手評価との分布図 (2 回目)

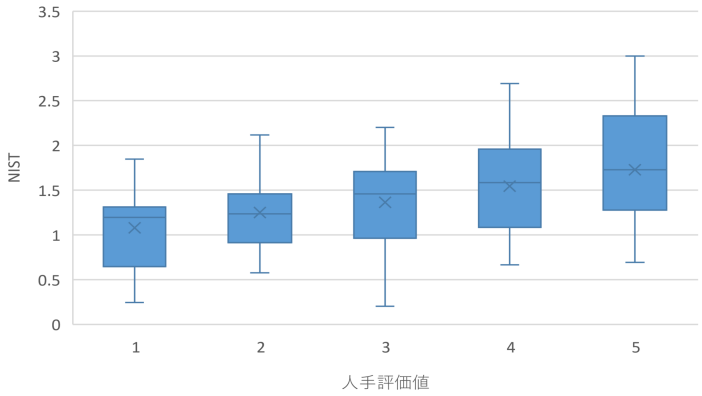

図 7.10: NIST と人手評価との分布図 (2 回目)

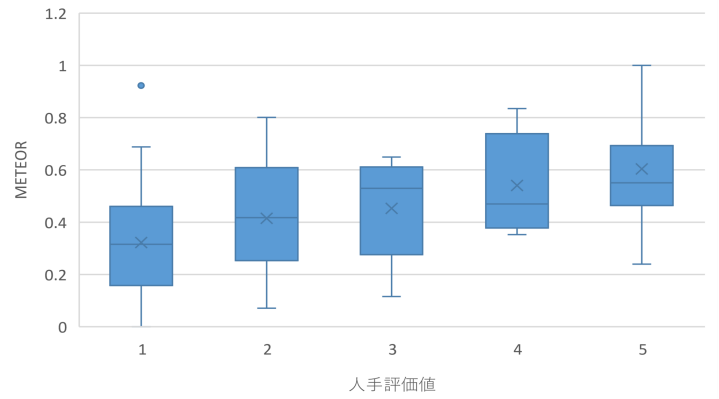

図 7.11: METEOR と人手評価との分布図 (2 回目)

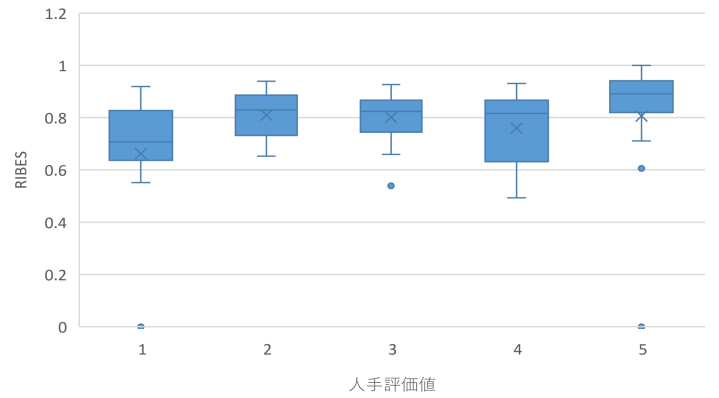

図 7.12: RIBES と人手評価との分布図 (2 回目)

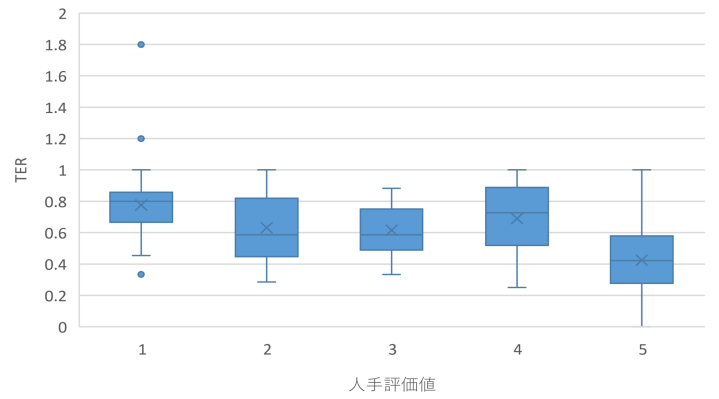

図 7.13: TER と人手評価との分布図 (2 回目)

7.3.3　実験 3 回目 (分布図)

実験 3 回目におけるそれぞれの自動評価と人手評価との分布図を示す.

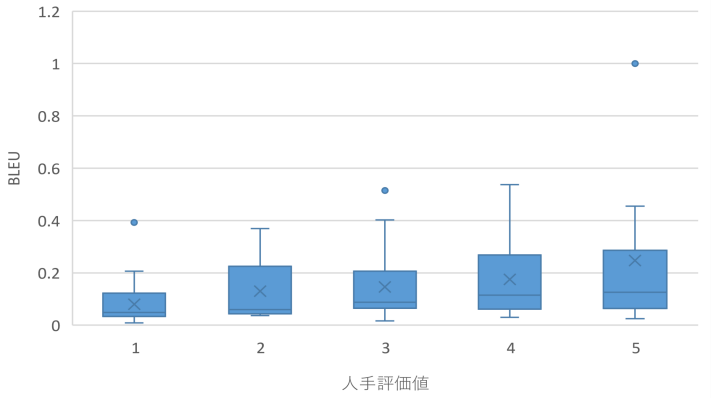

図 7.14: BLEU と人手評価との分布図 (3 回目)

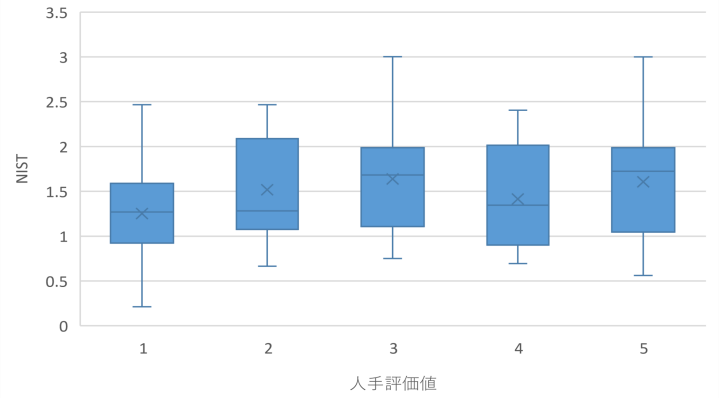

図 7.15: NIST と人手評価との分布図 (3 回目)

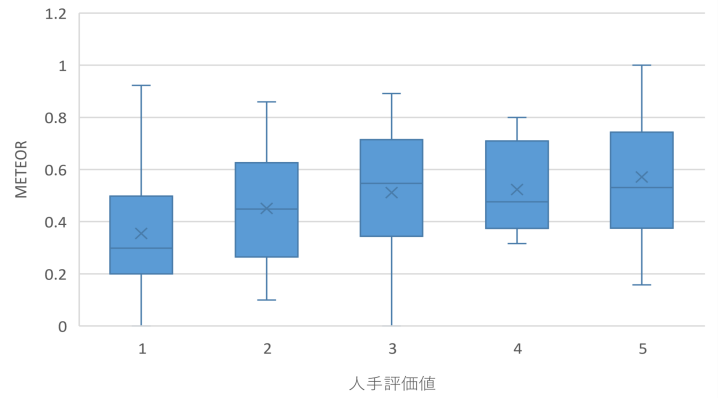

図 7.16: METEOR と人手評価との分布図 (3 回目)

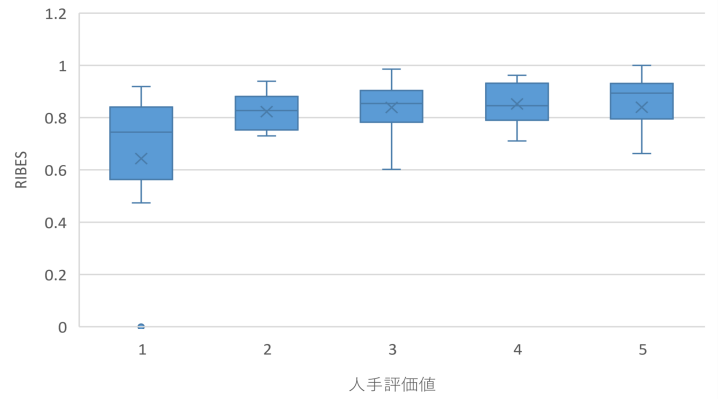

図 7.17: RIBES と人手評価との分布図 (3 回目)

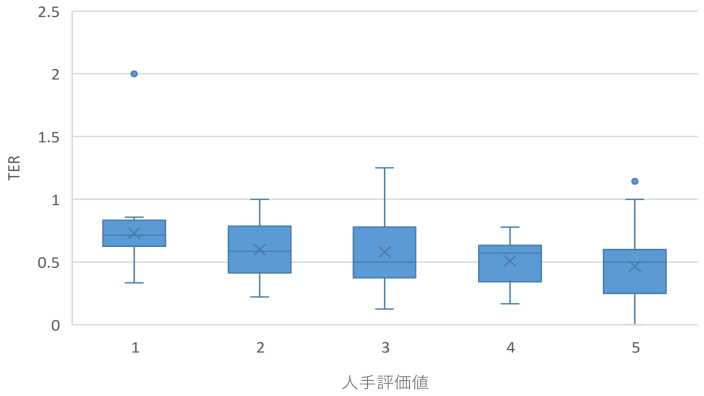

図 7.18: TER と人手評価との分布図 (3 回目)

7.3.4 提案手法との比較

提案手法と各自動評価とで人手評価との相関係数を比較する.

表 7.3.1: 各自動評価と人手評価との相関係数

	実験 1 回目	実験 2 回目	実験 3 回目
$STRMRR$-人手 MRR	0.324	0.298	0.310
$STRMRR$-1best 人手	0.329	0.199	0.307
STR	0.301	0.263	0.235
BLEU	0.423	0.448	0.334
NIST	0.374	0.436	0.214
METEOR	0.605	0.487	0.356
RIBES	0.290	0.229	0.352
TER	-0.419	-0.445	-0.364

本研究の提案手法 STR-MRR は先行研究 STR より人手評価に近づけることができた. しかし, 表 7.3.1 より, 提案手法は汎用性の高い従来手法よりも人手評価に近づけることができなかった. これまで分布図より, 部分評価の従来手法の値は人手評価の値に比例している. しかし, 提案手法は先行研究と同様, 完全一致を用いて計算をしている. その結果, 従来手法と比べて, 自動評価値が 0 になる数がとても多く, 人手評価と比例関係を作ることができていない. つまり自動評価は完全一致より部分一致の方が人手評価に近い評価方法であると結論に至った.

加えて, 本研究においては人手評価との相関係数は BLEU よりも METEOR のほうが優れていることが分かった.

7.3.5 BLEU と METEOR との差

BLEU と METEOR との決定的な違いは，連語に着目するか単語に着目するかである．BLEU は 4-gram までの一致率で翻訳評価をする．つまり，単語のみが一致していても連語が一致していなければ高い数値が出ない．一方 METEOR は適合率と再現率で翻訳評価を行う．つまり連語に囚われず参照文と出力文とで共通して出現した単語の数が多ければ高い値が出る．こうした各手法の評価方法の違いにより，評価にばらつきが出た．その例を以下に示す．

表 7.3.2: BLEU と METEOR との差

入力文	沖の方に船の明かりが見える。
参照文	You can see ships' lights toward the offing.
出力文	He can see the lights of the ship.

表 7.3.2 について解説する．まず人手評価について解説する．
入力文において重要な情報は

情報 1 主語を表す「船の明かり」

情報 2 場所を表す「沖の方」

出力文において情報 1 は翻訳できているが，情報 2 は翻訳できていない．しかし，情報 1 は出力文において重要度が高いと考える．そのため部分的な欠落があるとして評価を 3 とした．

続いて，BLEU について解説する．BLEU の評価では参照文と出力文を比較して，一致した $N-gram$ は $1-gram$ の“ can ”と“ see ”と“ lights ”と“ . ”，そして $2-gram$ の“ can see ”である．$1-gram$ つまり単語自体は一致数が多い．しかし $2-gram$ 以降は“ can see ”の一つしか一致していない．故に BLEU スコアは 0.057 と低い評価値が出た．このように，BLEU は単語の一致が多くても単語列が一致していなければ数値が低くなる．

続いて，METEOR について解説する．METEOR は主に単語の一致率で評価する．つまり BLEU のように単語列は考慮しないうえに語順にも囚われない．表 7.3.2 では出力文の単語数が 9 のうち，参照文と一致した単語数は 4 と適合率の値が高く，METEOR の値は 0.560 とまずまず高い評価値を出した．単語列とは“ of the ship ”と句としてだけではなく,“ see ships ”のような句としてではなく単純に連なっただけのものも含まれる．人手評価を行うときは，単語列よりも単語そのものや句に着目して評価をするため，METEOR の方が人手評価に近かったと考えられる．ただし，表 7.3.2 の場合のように本研究では連語が一致している数が少なかったため結果として METEOR の方が人手評価に近かったに過ぎない可能性もある．こちらも翻訳精度次第で変動する可能性も考えられる．

7.3.6 BLEU と人手評価との差 (人手評価が高い場合)

本研究の散布図より BLEU と人手評価にも差が生まれた例文も存在している．本節ではその例について解説する．

表 7.3.3: BLEU と人手評価との差 1

入力文	政府はこの事業への具体的な内容を示した。
参照文	The government presented some of the features of the project that are concrete .
出力文	The government demonstrated a specific policy for this enterprise .

まず，人手評価について解説する．入力文において重要な情報は

情報 1 主語を表す「政府」

情報 2 目的語を表す「この事業への具体的な内容」

情報 3 時制と行動を示す「示した」

である.「政府」に対して“ The government ”,「この事業への具体的な内容」に対して“ a specific policy for this enterprise ”,「示した」に対して“ demonstrated ”と重要な情報を完璧に翻訳することができている．そして，文法的な間違いも無いため評価は 5 とした．

続いて，BLEU について解説する．参照文と出力文を比較して，一致した $N-gram$ は $1-gram$ の“ the ”と“ government ”と“ . ”，$2-gram$ の“ The government ”のみである．このため，BLEU スコアは 0.033 と低い値が出た．

表 7.3.3 では，参照文の“ features ”に対して“ policy ”，参照文の“ the project ”に対して“ this enterprise ”と出力されている．

表 7.3.4: BLEU と人手評価との差 2

入力文	その部屋はむんむんしていた。
参照文	It was stuffy in the room .
出力文	The room was stuffy .

まず，人手評価について解説する．入力文において重要な情報は

情報 1 主語を表す「その部屋」

情報 2 状態を表す「むんむん」

情報 3 時制を示す「していた」

である.「その部屋」に対して“ The room ”,「むんむん」に対して“ stuffy ”，時制に対して“ was ”と重要な情報を完璧に翻訳することができている．そして，文法的な間違いも無いため評価は 5 とした.

続いて，BLEU について解説する．参照文と出力文を比較して，一致した $N-gram$ は $1-gram$ の“ was ”と“ stuffy ”と“ room ”,“ . ”である．$2-gram$ は“ the room ”と“ was stuffy ”が一致した．$1-best$ と $2-best$ は一致度が高いが，一致した $3-best$ と $4-best$ は一つもない．そのため BLEU スコアは 0.09 と高くない値が出た.

このように違う単語でも同じ意味，あるいは言い換えになっていることもある．その場合，人手評価では高い評価をしても，BLEU スコアは低い値が出てしまい，差が生まれてしまう結果となった.

7.3.7　BLEU と人手評価との差 (人手評価が低い場合 1)

前節では人手評価値が BLEU スコアに比べて高い例であったが，本節では逆に人手評価が低く，BLEU スコアが高かった例について紹介する．

表 7.3.5: BLEU と人手評価との差 3

入力文	そのけんかは何年も糸を引いた。
参照文	The quarrel lasted for years .
出力文	The quarrel went out for years .

まず，人手評価について解説する．入力文において重要な情報は

情報 1 主語を表す「そのけんか」

情報 2 状態を表す「何年も糸を引いた」

である．表 7.3.5 の出力文は主語は翻訳することはできている．しかし，肝心の「何年も糸を引いた」に対しては,“ went out ”と翻訳されている.“ went out ”は日本語で「なくなった」という意味である．つまり，入力文の「けんかが続いた」に対して「けんかがなくなった」という真逆の意味と鳴っているため，評価は 1 とした．

続いて，BLEU について解説する．$1-gram$ は“ The”と“ quarrel”,“ for”,“ years”,“ .”が一致している．$2-gram$ は“ The quarrel”と“ for years”,“ years .”が一致している．そして $3-gram$ は“ for years .”が一致している．$4-gram$ は一致していない．しかし，$3-gram$ の一致が存在している上に，$1-gram$ と $2-gram$ において一致率が高い．そのため，BLEU スコアは 0.206 と低くない値が出ている．

このように入力文と出力文が反対の意味を持っている時に BLEU と人手評価との差が生じる．

7.3.8 BLEU と人手評価との差 (人手評価が低い場合 2)

表 7.3.6: BLEU と人手評価との差 4

入力文	その険しい山腹には木が 1 本も生えていなかった。
参照文	No trees grew on the steep mountainside .
出力文	Not a lamp was growing on the steep mountainside .

まず，人手評価について解説する．入力文において重要な情報は

情報 1 主語を表す「木」

情報 2 場所を表す「その険しい山腹」

情報 1 状態を表す「1 本も生えていなかった」

である.「その険しい山腹」に対して“ the steep mountainside ”,「1 本も生えていなかった」に対して“ Not a was growing ”と翻訳できている．しかし，入力文において最重要である主語の「木」が“ lamp ”となっている.“ lamp ”とは「灯り」という意味である．つまり出力文を日本語に翻訳すると,「その険しい山腹には灯りが 1 本も生えていなかった。」と入力文とかけ離れた意味になっている．よって評価は 1 とした．

BLEU について解説する．$1-gram$ は,“ No ”と“ on ”,“ the ”,“ steep ”,“ mountainside ”,“ . ”が一致している．$2-gram$ は,“ on the ”と“ the steep ”,“ steep mountainside ”,“ mountainside . ”が一致している．$3-gram$ は,“ on the steep ”と“ the steep mountainside ”,“ steep mountainside . ”が一致している．$4-gram$ は“ on the steep mountainside ”と“ the steep mountainside . ”が一致している．表 7.3.6 の出力文は $3-gram$ と $4-gram$ の一致数が多いため，BLEU スコアは 0.393 と表 7.3.5 の出力文よりも高い値が出ている．

BLEU のような部分一致は完全一致と比べて人手評価との相関が強い．しかし実際，表 7.3.5 や表 7.3.6 のように，一つの単語あるいは句によって文全体の意味がまるきり変わる場合には部分一致が裏目に出るという事実も否めない．

7.4 BLEU と人手 MRR との相関

7.3.4 節の結果より BLEU と 1-best 人手評価との相関は提案手法より強く優れた自動評価であることが分かった．では，人手 MRR の信頼性を確かめるために BLEU との分散と相関を調査した．以下に散布図と相関係数を示す．

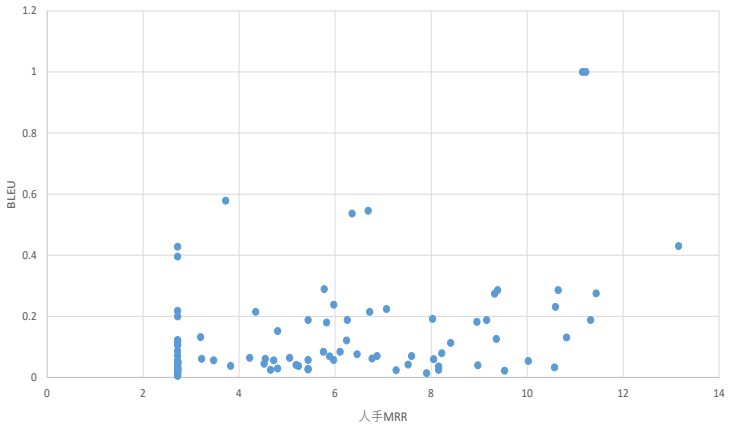

図 7.19: BLEU 人手 MRR との散布図 (実験 1 回目)

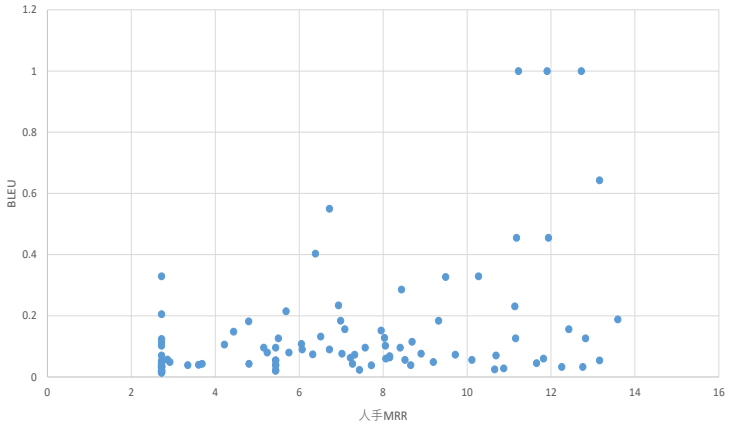

図 7.20: BLEU 人手 MRR との散布図 (実験 2 回目)

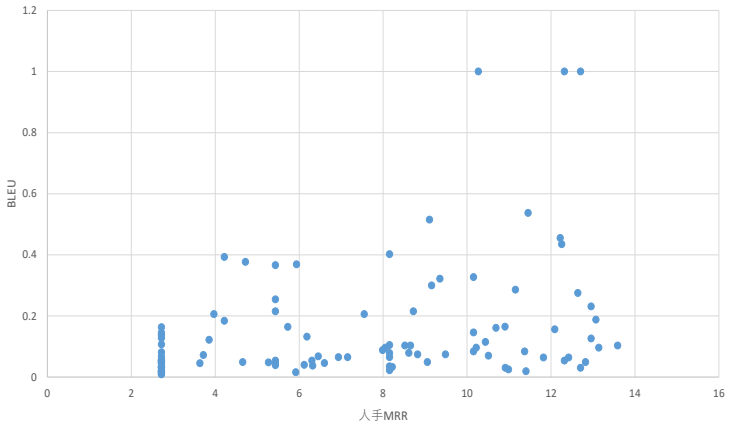

図 7.21: BLEU 人手 MRR との散布図 (実験 3 回目)

BLEU と人手 MRR との相関係数を以下の表 7.4.1 に示す．参考として BLEU と 1-best の人手評価値との相関係数，提案手法と人手 MRR との相関係数と比較する．

表 7.4.1: 各自動評価と人手評価との相関係数

	実験 1 回目	実験 2 回目	実験 3 回目
BLEU-人手 MRR	0.404	0.419	0.330
BLEU-1best 人手	0.423	0.448	0.334
$STRMRR$-人手 MRR	0.324	0.298	0.310

表 7.4.1 より BLEU と相関をとる人手評価が人手 MRR であったとしても，提案手法より安定して強い相関を得ることができる．しかし，BLEU は人手 MRR より，1-best のみの人手評価の方が強い相関を得ることが分かった．

第8章　今後の課題

本研究では，実験の結果より N-best の翻訳評価をすることの必要性が分かった．また，翻訳評価は完全一致よりも部分一致の評価手法の方がより人手評価との差が小さいことも実験を通して分かった．つまりこの二つの観点から，部分評価の N-best 評価が優れた評価であると予想する．ところが部分評価の N-best 評価についての研究はすでに行われている．そこで今後は N-best 評価の N の変動による人手評価との差の変動を調査することが課題になっていくと考える．また，本実験では 5 段階の人手評価について，場合分けを行ったが，未だに曖昧性が残っている．今後は人手評価をするときには，チェック項目を作るなど曖昧性を減らすことも課題であると考える．

第9章　おわりに

本研究では機械翻訳の自動評価についての研究を行った．そこで先行研究の手法よりも人手評価との差を縮めることを目的として実験を行った．実験より，本研究の提案手法は先行研究の手法よりも人手評価との差が小さく本研究の目的を達成することができた．ここで N-best の翻訳評価は人手評価との差を縮めることができると結論づける．しかし従来手法と比べると本研究の提案手法は人手評価との差が出てしまったため未だに課題は残されている．本研究を通じて従来の自動評価の信頼性を改めて思い知らされた．

謝辞

最後に, 三年間に渡り, 本研究のご指導をいただきました鳥取大学工学部知能情報工学科自然言語処理研究室の村上仁一准教授, 村田真樹教授に深く感謝すると共に, 厚く御礼申し上げます. そして, 日常の議論を通じて多くの知識や示唆を頂いた同研究室の皆様に深謝いたします.

参考文献

[1] “a Method for Automatic Evaluation of Machine Translation”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp.311-318, 2002.

[2] NIST, “Automatic Evaluation of Machine Translation Quality Using n-gram Co-Occurence Statistics” Proceedings of the Human Language Technology Conference (HLT), pp.128-132, 2002.

[3] Lavie Alon, and Denkowski Michael “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, 2007.

[4] Hideki Isozaki, “Automatic Evaluation of Translation Quality for Distant Language Pairs”, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp.944-952, 2010.

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

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014.

[7] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-source toolkit for neural machine translation. ArXiv e-prints, 2017.

[8] 石原 雅文: “文一致数を用いた機械翻訳の自動評価”，卒業論文 鳥取大学, 2012

[9] MRR:https://softwaredoug.com/blog/2021/04/21/compute-mrr-using-pandas.html