Vol.2010-FI-97 No.3
Vol.2010-NL-195 No.3
2010/1/28

単語情報及びフレーズによる大局的情報を用いた機械翻訳自動評価手法

小山田崇[†]　越前谷博[††]　荒木健治[†]

本稿では，MT 訳と参照訳間において単語レベルの局所的な情報だけでなく，大局的な情報としてフレーズの情報も利用した機械翻訳自動評価のための新たな手法を提案する．性能評価実験の結果，フレーズ情報を利用した提案手法が人手評価との間の文単位の相関において，他手法に比べ最も高い相関を示した．この結果は提案手法の有効性を示すものである．

Automatic Evaluation of Machine Translation Using both Words Information and Comprehensive Phrases Information

Takashi Oyamada[†]　Hiroshi Echizen-ya[††]　and Kenji Araki[†]

In this paper, we propose a new method for automatic evaluation of a machine translation using both words information and comprehensive phrases information. As the result of the evaluation experiments, the proposed method using the information of phrases obtained the highest correlation values in sentence-level correlation, comparing with other methods. These results show the effectiveness of the proposed method.

1. はじめに

近年，機械翻訳分野の研究の進展に伴い，開発された機械翻訳システムをより高い精度で自動評価することが求められている．そのため，BLEU[1]を始め MT 訳と人手により作成された参照訳との間の類似度に基づく評価手法の研究が盛んに行われるようになった．その結果，複数の文から構成されるドキュメント単位の自動評価においては人手評価との間で高い相関が得られるようになった．しかし，文単位の自動評価においては十分な相関が得られるに至っておらず，そのことが問題点として指摘されている[2].

そこで，本稿では文単位での自動評価の精度向上を目的とした新たな機械翻訳自動評価のための手法を提案する．これまでに提案されている多くの手法は N グラムマッチ率のように単語レベルの部分的な一致に基づいている．しかし，そのような局所的な情報のみの利用では文全体の大局的な観点からの評価が不十分であると考えられる．そこで，構文情報による大局的な情報を用いた手法が提案されている．しかし，それらの手法は構文解析ツールに依存しているためその精度の影響を強く受けることになる[3][4].

そこで，我々は MT 訳と参照訳間における単語レベルの部分的な一致とフレーズレベルの大局的な一致に基づく新たな機械翻訳手法を提案する．提案手法では単語レベルの部分的な一致を反映したスコアとフレーズレベルの大局的な一致を反映したスコアを組み合わせることで，文単位においてより高い精度を得ることのできる自動評価を行う. 性能評価実験の結果, 人手評価との間の相関において, 提案手法は mBLEU[1], ROUGE-L[5]，mPER[6]，mWER[7]，METEOR[8]，IMPACT[9]を用いた場合よりも高い相関を導き出した．この結果により，提案手法の有効性が確認された．

2. 提案手法

提案手法は単語レベルの局所的な一致に基づくスコアとフレーズレベルの大局的な一致に基づくスコアの組み合わせにより最終的なスコアを求める．

2.1 単語情報に基づくスコア付け

本節では単語レベルの局所的な一致に着目した単語スコアの計算について述べる．単語スコアの計算法には IMPACT[9]を用いる．IMPACT は特許翻訳文を用いたメタ評価[10]において最も高い相関が得られることから，単語レベルのスコアを求めるうえで最適な計算法と考えられる．始めに，MT 訳と参照訳間の LCS(Longest Common Subsequence : 最長一致部分列)を一意に定める．そして，定められた LCS を除き，LCS を再帰的に決定する．複数の LCS 経路が存在する場合，式(1)，式(2)を用いて最もスコアが高い経路を選択する．

$$pos_w = \left(1.0 - \left|\frac{posM(c)}{m} - \frac{posR(c)}{n}\right|\right)^{\alpha} \quad (1)$$

† 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University
†† 北海学園大学工学部
Faculty of Engineering, Hokkai-Gakuen University

Vol.2010-FI-97 No.3
Vol.2010-NL-195 No.3
2010/1/28

$$RS = \left(\sum_{c \in LCS} \left(length(c)^{\beta} \times pos_w \right) \right)^{\frac{1}{\beta}} \quad (2)$$

α と βはパラメータで, それぞれ特許翻訳での精度が高かった[9]0.1 と 1.1 をそのまま用いる．m は参照訳の全単語数であり，n は MT 訳の全単語数である．式(1)は LCS 経路を構成する共通部分の相対位置のずれを示している．そして，式(1)の相対位置のずれを負の重みとして用い，共通部分の構成単語数に基づくスコア RS を式(2)により求める．式(2)により共通部分の相対位置のずれが小さく，単語数の多い LCS 経路を選択することができる．以上の処理で決定した LCS を除き，残った部分の LCS を決定するという処理を LCS が存在しなくなるまで繰り返す．

このようにして決定した LCS に基づき式(3)の再現率 $R_{IP\text{-}w}$ と式(4)の適合率 $P_{IP\text{-}w}$ を求める．そして式(5)，式(6)より $R_{IP\text{-}w}$ と $P_{IP\text{-}w}$ の F 値を求めることで単語レベルのスコアを求める．

$$R_{IP-w} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^{\beta} \right)}{m^{\beta}} \right)^{\frac{1}{\beta}} \quad (3)$$

$$P_{IP-w} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^{\beta} \right)}{n^{\beta}} \right)^{\frac{1}{\beta}} \quad (4)$$

$$score_{IP-w} = \frac{\left(1+\gamma^2\right) \cdot R_{IP-w} \cdot P_{IP-w}}{R_{IP-w} + \gamma^2 P_{IP-w}} \quad (5)$$

$$\gamma = \frac{P_{IP-w}}{R_{IP-w}} \quad (6)$$

式(3)，(4)における CC は一意に決定された LCS を構成する共通部分の数を表す．i は再帰処理の回数をカウントする変数であり，$length(c)$は共通部分の単語数を示している．スコアは 0.0～1.0 で表され，値が大きいほど良質な MT 訳であることを意味する．また，複数参照訳を用いる場合，MT 訳と全ての参照訳との間で適合率，再現率を求め，それぞれの最大値を用いて求めた F 値をスコアとする．

単語情報に基づくスコア付けの具体例を図 1 に示す．図 1 ではまず，MT 訳と参照訳の間の LCS を求める．このとき複数の LCS 経路が存在すれば，式(1)と式(2)を用いて共通部分の相対位置のずれが小さく，単語数の多い LCS 経路である{は 、}，{な 回復}，{に ある}，{。}を選択する．次に先ほど決定した LCS 経路を除き，残った部分の LCS 経路である{消費}を LCS 経路として決定する．そして，決定した LCS 経路と式(3)を用いて(再現率)=0.2953，式(4)を用いて(適合率)=0.4895 を求め，最後に式(5)，式(6)を用いて(F 値):$score_{IP\text{-}w}$=0.3302 を求める．

(1) LCSを求める
MT訳:彼 {は 、} 個人 消費 が 一般 に ゆるやか {な 回復} 基調 {に ある} と 言い ました {。}
参照訳:私的 消費 {は 、} おおむね 緩やか {な 回復} 傾向 {に ある} {。}
lcs=7
MT訳:彼 個人 {消費} が 一般 に ゆるやか 基調 と 言い ました
参照訳:私的 {消費} おおむね 緩やか 傾向
lcs=1
(2) 適合率と再現率からF値を求める

$$(\text{再現率}) = \left(\frac{0.1^0 \cdot 2^{1.1} + 0.1^0 \cdot 2^{1.1} + 0.1^0 \cdot 2^{1.1} + 0.1^0 \cdot 1^{1.1} + 0.1^1 \cdot 1^{1.1}}{19^{1.1}} \right)^{\frac{1}{1.1}} = \left(\frac{7.5306}{25.5052} \right)^{\frac{1}{1.1}} = 0.3299$$

$$(\text{適合率}) = \left(\frac{0.1^0 \cdot 2^{1.1} + 0.1^0 \cdot 2^{1.1} + 0.1^0 \cdot 2^{1.1} + 0.1^0 \cdot 1^{1.1} + 0.1^1 \cdot 1^{1.1}}{12^{1.1}} \right)^{\frac{1}{1.1}} = \left(\frac{7.5306}{15.3851} \right)^{\frac{1}{1.1}} = 0.5223$$

$$\gamma = \frac{0.5223}{0.3299} = 1.5832$$

$$score_{IP-w} = \frac{\left(1+1.5832^2\right) \times 0.3299 \times 0.5223}{0.3299 + 1.5832^2 \times 0.5223} = \frac{0.6042}{1.6391} = 0.3686$$

図 1 単語情報に基づくスコア付けの具体例

2.2 フレーズ情報に基づくスコア付け

フレーズレベルの大局的な一致に基づくスコア付けは MT 訳と参照訳中の名詞句に着目することで行う．名詞句は文中に最も数多く出現し，かつ，その決定が比較的容易と考えられるためである．

(1) 名詞句の決定

本節では名詞句の決定は係り受け解析器である CaboCha[11]を用いて行った．以下に名詞句決定の処理の詳細を述べる．

Vol.2010-FI-97 No.3
Vol.2010-NL-195 No.3
2010/1/28

① 係り受け関係の存在しない後置詞句の抽出

MT 訳及び参照訳を CaboCha により解析し，係り受けが連続していない部分を抽出する．その中で名詞を含み，動詞，助動詞を含まない部分のみを抽出する．

② 係り受け関係の存在する後置詞句の抽出

係り受けが連続している部分を抽出する．そして抜き出した部分を末尾の係り受けより遡って調べ，最初の句から一番後ろにある名詞を含み，動詞・助動詞を含まない句までを抽出する．

③ 助詞，記号の削除

最後に，上記①と②で抽出した部分の末尾の助詞，記号を削除したものを名詞句とする．

解析器による解析結果から名詞句を決定する具体例を図 2 に示す．図 2 ではまず CaboCha[11]を用いて“私的 消費 は 、 おおむね 緩やか な 回復 傾向 に ある 。”を解析を行う．そして，その解析結果から，係り受け関係の存在しない単独の後置詞句である“私的 消費 は 、”を抽出する．続いて，係り受け関係が存在する部分“おおむね 緩やか な 回復 傾向 に ある 。”を抽出し，末尾から遡って調べたとき，最後尾の後置詞句が“回復 傾向 に”であるため，係り受けの最初の部分である“おおむね”から“回復 傾向 に”までを抽出する．このようにして抽出した部分の末尾にある助詞や記号を削除し，“私的 消費”と“おおむね 緩やか な 回復 傾向”を名詞句として決定する．

(2) 名詞句の対応付け

MT 訳と参照訳に対して名詞句を決定した後，MT 訳と参照訳間において名詞句の対応付けを行う．本手法では全ての名詞句間に対して PER を計算し，その値が最も高いものを対応する名詞句として決定する．以下にその処理手順の詳細を述べる．

① 決定された名詞句ごとに MT 訳と参照訳間で PER を求める．

② PER の値が最大となる名詞句のペアを対応する名詞句と位置付ける．ただし，PER が最大である組み合わせが複数存在する場合には対応する名詞句を一意に決定できないため未対応とする．

図 3 に名詞句の対応付けの具体例を示す．図 3 では MT 訳から抽出された[彼]，[個人 消費]，[一般]，[ゆるやか な 回復 基調]の 4 つの名詞句と，参照訳から抽出された[私的 消費]，[おおむね 緩やか な 回復 傾向]の 2 つの名詞句の間で PER を求める．次に，それぞれの名詞句と PER が最大となる組み合わせを選択する．その結果得られた組み合わせである[個人 消費]⇔[私的 消費]，[ゆるやか な 回復 基調]⇔[おおむね 緩やか な 回復 傾向]は一意であるので対応する名詞句とする．

例：私的 消費 は 、おおむね 緩やか な 回復 傾向 に ある 。

CaboChaによる解析結果

```
* 0 4D 1/2 3.50009742
私的    シテキ  私的    名詞-形容動詞語幹              O
消費    ショウヒ 消費    名詞-サ変接続                  O
は      ハ      は      助詞-係助詞                    O
、      、      、      記号-読点              O
* 1 2D 0/0 0.50509486
おおむね オオムネ おおむね 副詞-一般          O
* 2 3D 0/1 1.54496132
緩やか  ユルヤカ 緩やか  名詞-形容動詞語幹              O
な      ナ      だ      助動詞  特殊・ダ  体言接続 O
* 3 4D 1/2 0.00000000
回復    カイフク 回復    名詞-サ変接続                  O
傾向    ケイコウ 傾向    名詞-一般              O
に      ニ      に      助詞-格助詞-一般               O
* 4 -1O 0/0 0.00000000
ある    アル    ある    動詞-自立 五段・ラ行 基本形 O
。      。      。      記号-句点              O
EOS
```

(1) 係り受け関係の存在しない後置詞句の抽出

→”私的 消費 は 、” を獲得

(2) 係り受け関係の存在する後置詞句の抽出

→”おおむね 緩やか な 回復 傾向 に ある 。” を抽出

末尾から調べ，最後尾の後置詞句までを抽出

→“おおむね 緩やか な 回復 傾向 に”を獲得

(3) 助詞，記号の削除

“私的 消費 は 、”→“私的 消費”

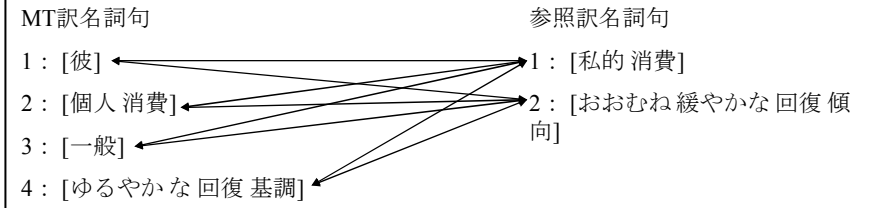
“おおむね 緩やか な 回復 傾向 に”→“おおむね 緩やか な 回復 傾向 に”

図 2 名詞句決定の具体例

Vol.2010-FI-97 No.3
Vol.2010-NL-195 No.3
2010/1/28

MT訳:[彼] は 、[個人 消費] が [一般] に [ゆるやか な 回復 基調] に あると 言い ました 。
参照訳:[私的 消費] は 、[おおむね 緩やか な 回復 傾向] に ある 。

(1) 全名詞句についてPERスコアを求める



(2) 各フレーズから見たときのPERスコアが最大となる組み合わせを求める
(括弧内はPERスコアの値)

MT訳名詞句
1： [彼] ⇒なし
2： [個人 消費] ⇒[私的 消費](0.5)
3： [一般] ⇒ なし
4： [ゆるやか な 回復 基調] ⇒[おおむね 緩やか な 回復 傾向](0.4)
参照訳名詞句
1： [私的 消費] ⇒[個人 消費](0.5)
2： [おおむね 緩やか な 回復 傾向] ⇒[緩やか な 回復 基調](0.4)

(3) MT訳側から見たときと参照訳側から見たときのPERの値が最大である名詞句が一致した場合，対応関係とみなす

対応名詞句1： [個人 消費] ⇔ [私的 消費]
対応名詞句2： [ゆるやか な 回復 基調] ⇔ [おおむね 緩やか な 回復 傾向]

図 3 名詞句対応の決定の具体例

(3) フレーズ情報に基づくスコア付け

MT 訳と参照訳中の名詞句のみを抽出し，フレーズレベルでのスコア付けを行う．その際，名詞句を一般化し，対応する名詞句を共通単語とみなす．次に，MT 訳と参照訳間においてスコア付けを行う．式(7)，(8)にその計算式を示す．

$$R_{IP-p} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^\beta \right)}{\left\{ cph\_m \times \left(\log_2 uph\_m + 1.0 \right) \right\}^\beta} \right)^{\frac{1}{\beta}} \quad (7)$$

$$P_{IP-p} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^\beta \right)}{\left\{ cph\_n \times \left(\log_2 uph\_n + 1.0 \right) \right\}^\beta} \right)^{\frac{1}{\beta}} \quad (8)$$

ここで未対応の名詞句が 1 つも存在しない場合, $\log_2 0$ となり分母の計算を行えないため，分母の式は cph_m^β，cph_n^β とする．($\log_2 uph\_m$+1.0)，($\log_2 uph\_n$+1.0)は未対応の名詞句に基づく負の重みであり，その数が大きいほどスコアはより小さくなる．式(7)は再現率, 式(8)は適合率を示す. ここで *cph_m*, *cph_n* は対応する名詞句の数, *uph_m*, *uph_n* は未対応の名詞句数である．式(5)と式(6)と同様の計算式を用いて F 値を求めることでフレーズレベルでのスコアを求める．複数の参照訳を用いる場合には，個々の参照訳とのスコアの相加平均を用いる．図 4 にフレーズ情報に基づくスコア付けの具体例を示す.

図 4 では MT 訳，参照訳でそれぞれの名詞句のみを抽出し，MT 訳を[未対応名詞句 1] [対応名詞句 1] [未対応名詞句 2] [対応名詞句 2]，参照訳を[対応名詞句 1] [対応名詞句 2]とする．このように名詞句を一般化した後，式(7)から(再現率)=0.4665，式(8)から(適合率)=0.4353 を求め，単語レベルのスコアと同様に式(5)，式(6)から(F 値):$score_{IP\text{-}p}$ =0.4493 を求め，その値をフレーズレベルのスコアとする．

2.3 提案手法に基づくスコア付け

2.1 で求めたスコアと **2.2** で求めたスコアを以下の式(9)を用いることでスコアを算出する．

$$score = \frac{\left(score_{IP-w} \right) + w_p \cdot \left(score_{IP-p} \right)}{1.0 + w_p} \quad (9)$$

$score_{IP\text{-}w}$，$score_{IP\text{-}p}$ はそれぞれ **2.1** で求めた単語レベルのスコア，**2.2** で求めたフレーズレベルのスコアを示し，w_p はフレーズに対する重みのパラメータであり，性能評価実験に用いた英日対訳コーパスと 1 つの MT システムを用いた予備実験において精度が高かった 0.3 を用いた．

Vol.2010-FI-97 No.3
Vol.2010-NL-195 No.3
2010/1/28

[]内は抽出された名詞句
MT訳:[彼] は 、[個人 消費] が [一般] に [ゆるやか な 回復 基調] に あると 言い ました 。
参照訳:[私的 消費] は 、[おおむね 緩やか な 回復 傾向] に ある 。
⬇フレーズのみを抽出し，一般化
MT 訳 : [未対応名詞句] [対応名詞句1] [未対応名詞句] [対応名詞句2]
参照訳 : [対応名詞句1] [対応名詞句2]

$$(\text{再現率}) = \left(\frac{0.1^0 \times 1^{1.1} + 0.1^0 \times 1^{1.1}}{2^{1.1}} \right)^{\frac{1}{1.1}} = \left(\frac{2}{2.1435} \right)^{\frac{1}{1.1}} = 0.9389$$

$$(\text{適合率}) = \left(\frac{0.1^0 \times 1^{1.1} + 0.1^0 \times 1^{1.1}}{\{2 \times (\log_2 2 + 1.0)\}^{1.1}} \right)^{\frac{1}{1.1}} = \left(\frac{2}{4.5948} \right)^{\frac{1}{1.1}} = 0.4695$$

$$score_{IP-p} = \frac{(1.0 + 0.5000^2) \times 0.9389 \times 0.4695}{0.9389 + 0.5000^2 \times 0.4695} = \frac{0.6612}{1.0559} = 0.6262 \qquad \gamma = \frac{0.4695}{0.9383}$$

図 4 フレーズ情報に基づくスコア付けの具体例

3. 性能評価実験

3.1 実験方法

実験は始めにロイターの新聞記事による英日の対訳コーパス[12]中の対訳文 150 文の英文を原文として，3 つのルールベースの MT システムにより日本語訳文を得た．本稿ではこれら 3 つの MT システムをそれぞれ MT1,MT2,MT3 と記す．また，150 文の MT 訳に対してはそれぞれ 4 つの参照訳を用意した．本実験で用いた自動評価システムは mBLEU[1]，ROUGE-L[5]，mPER[6]，mWER[7]，METEOR[8]，IMPACT[9]，そして提案手法の 7 つである．これら 7 つの自動評価システムを用いて，3 つの MT システムそれぞれが出力した日本語訳文に対するスコア付けを行った．

更に，自動評価システムにより得たスコアと人手評価との間の文単位での相関を求めた．人手評価は 3 つの MT システムが出力した全ての日本語訳文に対して，3 人のバイリンガルが Adequacy と Fluency の観点から 5 段階で評価し，その結果において，MT 訳ごとに得られた評価値のメジアン値を用いた．また，相関は Pearson の相関係数と Spearman の順位相関係数を求めることで得た．

3.2 実験結果

実験結果を表 1 から表 4 に示す．表 1 から表 3 は MT1，MT2，MT3 のそれぞれの相関を示している．表 4 は表 1 から表 3 に示す相関の平均である．提案手法は表 1 において，Adequacy の相関が ROUGE-L を下回る結果となったが，それ以外では他手法を上回る結果が得られた．また，提案手法と最も類似している自動評価手法 IMPACT に対しては表 4 より，Pearson の Adequacy，Fluency ではそれぞれ 0.013，0.017 高い値を示し, Spearman の Adequacy, Fluency ではそれぞれ 0.026, 0.032 高い値となった. Pearson, Spearman 共に Adequacy に比べ Fluency の方がより改善されたことがわかる. これは，提案手法がフレーズ情報を利用していることが原因と考えられる．

3.3 考察

実験の結果，6 つの他手法との比較において，提案手法が最も高い相関を得た．より詳細に提案手法の有効性を確認するために，人手評価を 0.0～1.0 に正規化し，提案手法によるスコアとの差が±0.2 以内となる, 比較的高い相関の文がどの程度存在していたかを調査した．

その結果，MT1 では Adequacy が 45 文，Fluency が 93 文，MT2 では Adequacy が 47 文，Fluency が 83 文，そして，MT3 では Adequacy が 109 文，Fluency が 128 文であった．IMPACT と比べると，MT1 では Adequacy が 1 文，Fluency が 5 文増加していた．MT2 では Adequacy が 1 文，Fluency が 6 文増加していた．そして MT3 では Adequacy に変化はなく，Fluency が 2 文減少していた．したがって，提案手法により Fluency の相関がより改善されたことが明らかとなった．

一方, 0.0～1.0 に正規化した人手評価と提案手法によるスコアの差が±0.4 以上となる，比較的低い相関の文の数についても調査を行った．その結果，MT1 では Adequacy が 27 文，Fluency が 7 文，MT2 では Adequacy が 40 文，Fluency が 9 文，そして MT3 では Adequacy で 7 文, Fluency で 1 文であった. また, IMPACT との比較においては，MT1 では Adequacy が 2 文，Fluency が 1 文減少していた．MT2 では Adequacy が 4 文減少していたが，Fluency が 1 文増加していた．MT3 では Adequacy が 1 文，Fluency が 2 文減少していた．したがって，提案手法により IMPACT では低い相関であった文の数は Adequacy の方が Fluency に比べより多く減少したことになる．

しかし，これはそもそも IMPACT において低い相関だった文の数が Fluency において 10 文未満と非常に少なく，改善の余地がそれほど大きくなかったと考えられる．これらの調査結果より提案手法は IMPACT に対し，高い相関へと移行した文が増加し，低い相関であった文が減少したことが明らかとなり，より良い自動評価に向けての改善が見られた．フレーズを用いることで全体の中の 1 割ほどの文で評価の改善が見られた．以下の文は IMPACT では高い相関が得られず，提案手法で MT1 と MT2 の adequacy と fluency の両方，MT3 の adequacy において高い相関が得られた例である．([]内はフレーズ抽出された部分)

MT1 訳 : [2 0 0 人 の 工事 幹部 社員 の 調査] は 毎月 編集 さ れ ます 。
MT2 訳 : [２００ 人 の 建設 経営 者 の 調査] は 毎月 編集 さ れる 。
MT3 訳 : [2 0 0 人 の 構造 エグゼクティブ の 調査] は 毎月 編集 さ れる 。
参照訳 : [2 0 0 人 の 建設 経営 者 たち の 調査] は 毎月 まとめ られる 。

“工事幹部社員”と“建設経営者”のように単語レベルの一致度のみで評価をしている

Vol.2010-FI-97 No.3
Vol.2010-NL-195 No.3
2010/1/28

表 1 MT1 における実験結果

MT1	Pearson		Spearman	
	Adequacy	Fluency	Adequacy	fluency
提案手法	0.5057	**0.5355**	0.5206	**0.5215**
mBLEU(1-gram)	0.3059	0.2937	0.3031	0.2647
mBLEU(2-gram)	0.4132	0.3614	0.4079	0.3388
mBLEU(3-gram)	0.3727	0.3108	0.3701	0.2935
mBLEU(4-gram)	0.3415	0.2741	0.3374	0.2074
mBLEU(相乗平均)	0.3479	0.2320	0.3664	0.2494
mPER	0.3644	0.2719	0.3486	0.2077
mWER	0.4834	0.5011	0.4619	0.4618
ROUGE-L	**0.5535**	0.5354	**0.5242**	0.4683
METEOR	0.2598	0.2185	0.2700	0.2647
IMPACT	0.5334	0.5306	0.5194	0.4741

表 2 MT2 における実験結果

MT2	Pearson		Spearman	
	Adequacy	Fluency	Adequacy	fluency
提案手法	**0.4918**	**0.5871**	**0.5175**	**0.5975**
mBLEU(1-gram)	0.2785	0.3918	0.2703	0.3832
mBLEU(2-gram)	0.3590	0.4820	0.3416	0.4578
mBLEU(3-gram)	0.3654	0.5001	0.3721	0.4783
mBLEU(4-gram)	0.2965	0.4487	0.2992	0.4127
mBLEU(相乗平均)	0.3655	0.4902	0.3506	0.4678
mPER	0.2748	0.2699	0.2184	0.1990
mWER	0.3902	0.4885	0.3876	0.4713
ROUGE-L	0.4611	0.5378	0.4521	0.5160
METEOR	0.3912	0.4512	0.3606	0.4390
IMPACT	0.4558	0.5507	0.4707	0.5523

表 3 MT3 における実験結果

MT3	Pearson		Spearman	
	Adequacy	Fluency	Adequacy	Fluency
提案手法	**0.5351**	**0.5645**	**0.5195**	**0.5553**
mBLEU(1-gram)	-0.0038	0.0103	0.0627	0.0358
mBLEU(2-gram)	0.1607	0.1416	0.1561	0.1500
BLEU(3-gram)	0.1664	0.1375	0.1315	0.1173
mBLEU(4-gram)	0.1814	0.1476	0.1139	0.0996
mBLEU(相乗平均)	0.1089	0.0855	0.0981	0.0760
mPER	0.2744	0.2905	0.2608	0.3312
mWER	0.4177	0.4919	0.4059	0.4986
ROUGE-L	0.4769	0.5326	0.4662	0.5345
METEOR	0.2095	0.2491	0.2100	0.2308
IMPACT	0.5059	0.5545	0.4896	0.5523

表 4 MT1,MT2,MT3 の平均

Avg.	Pearson		Spearman	
	Adequacy	Fluency	Adequacy	Fluency
提案手法	**0.5109**	**0.5624**	**0.5192**	**0.5581**
mBLEU(1-gram)	0.1935	0.2319	0.2120	0.2279
mBLEU(2-gram)	0.3110	0.3283	0.3019	0.3155
mBLEU(3-gram)	0.3015	0.3161	0.2912	0.2964
mBLEU(4-gram)	0.2731	0.2901	0.2502	0.2399
mBLEU(相乗平均)	0.2741	0.2692	0.2717	0.2644
mPER	0.3045	0.2774	0.2759	0.2460
mWER	0.4304	0.4938	0.4185	0.4772
ROUGE-L	0.4972	0.5353	0.4808	0.5063
METEOR	0.2868	0.3063	0.2802	0.3115
IMPACT	0.4984	0.5453	0.4932	0.5262

Vol.2010-FI-97 No.3
Vol.2010-NL-195 No.3
2010/1/28

システムでは正当な評価ができない部分をフレーズレベルでの対応関係を一般化して用いることで，言い換えに相当する部分の評価を正当に行うことができている．これは，提案手法が単語レベルの一致だけでなく，フレーズレベルの一致も考慮した自動評価を行っているためと考えられる．

4. まとめ

本稿では，MT 訳と参照訳間において単語レベルの局所的な一致に基づくスコアだけでなく，フレーズレベルの大局的な一致に基づくスコアも考慮した新たな自動評価手法を提案した．性能評価実験の結果，他手法に比べ，文単位での Pearson の相関係数と Spearman の順位相関係数の両方において，3 つの MT の平均では他手法の中でも高い精度である ROUGE-L や IMPACT より 0.2 前後の高い相関が得られた．今後は更なる精度向上のため，名詞句の自動抽出の精度向上，シソーラスを用いる等の改良を行う予定である．

参考文献

1 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “BLEU: a method for automatic evaluation of machine translation.” Annual Meeting of the ACL, pp. 311–318, Philadelphia, Pennsylvania 2002.
2 Andrew Mutton, Mark Dras, Stephen Wan and Robert Dale. “GLEU: Automatic Evaluation of Sentence-Level Fluency”, the 45th Annual Meeting of the ACL, pp.344–351, Prague, Czech Republic, June 2007.
3 Dennis N. Mehay and Chris Brew. BLEUÂTRE: Flattening Syntactic Dependencies for MT Evaluation. In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI). pp.122-131. 2007.
4 Michael Pozar and Eugene Charniak. : Bllip : An Improved Evaluation Metric for Machine Translation, Brown University Master Theses, 2006.
5 Chin-Yew Lin, Franz Josef Och, “Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics”, ACL-2004, pp.606-613, 2004.
6 Keh-Yih Su, Ming-Wen Wu, Jing-Shin Chang, “A New Quantitative Quality Measure for Machine Translation Systems”, COLING’92, pp.433-439, 1997.
7 Gregor Leusch, Nicola Ueffing and Hermann Ney : A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation, Proc . of MT Summit IX, pp.240-247, 2003.
8 Banerjee Satanjeev and Lavie Alon, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”, ACL-2005, pp. 65-72, 2005.
9 Hiroshi Echizen-ya and Araki Kenji, “Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum”, Proceedings of the Eleventh Machine Translation Summit (MT SUMMIT XI), pp.151-158, Copenhagen, Denmark, 2007.
10 Hiroshi Echizen-ya, Terumasa Ehara, Sayori Shimohata, Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro and Noriko Kando. “Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NICIR-7”, Proceedings of the 3rd Workshop on Patent Translation, pp.9-16, Ottawa, Canada, 2009.8
11工藤拓，松本裕治，チャンキングの段階適用による係り受け解析，情報処理学会論文誌，Vol 43 No.6, pp.1834-1842,June 2002.
12 Masao Utiyama and Hitoshi Isahara. (2003) “Reliable Measures for Aligning Japanese-English News Articles and Sentences.” ACL-2003, pp. 72-79, 2003