HOKUGA 北海学園学術情報リポジトリ
学校法人北海学園
北海学園大学
北海商科大学

タイトル	WMT2012データとWMT2013データにおける機械翻訳のための自動評価法の性能について
著者	越前谷, 博; 荒木, 健治; Echizen'ya, Hiroshi; Araki, Kenji
引用	工学研究 : 北海学園大学大学院工学研究科紀要(14): 13-22
発行日	2014-09-30

Hokkai-Gakuen Organization of Knowledge Ubiquitous through Gaining Archives.

研究論文

WMT2012 データと WMT2013 データにおける機械翻訳のための自動評価法の性能について

越前谷　博* ・ 荒 木 健 治**

Performance of Automatic Evaluation Metrics for Machine Translation in WMT2012 data and WMT2013 data

Hiroshi Echizen'ya* and Kenji Araki**

概要

近年，機械翻訳システムの急速な進展に伴い，その評価を自動的に行うための自動評価法の研究が盛んに行われている．その結果，多くの自動評価法が提案され，利用可能となっている．このような状況において，自動評価法の評価精度における現状把握とより高い評価精度を持つ自動評価法の開発を目的としたコンテスト型ワークショップ WMT が毎年開催されている．そして，そこで使用されるデータは一般公開されており，常時使用可能である．本報告では，2012 年と 2013 年に開催された WMT において使用されたデータに基づき，我々が従来より提案している自動評価法の性能及び自動評価タスクにおける提案手法の位置付けについて述べる．

1　はじめに

機械翻訳分野では近年，統計翻訳の研究[1] が盛んに行われている．統計翻訳は基本的に原言語文とその訳文である目的言語文のペアのセットである対訳コーパスから言語モデルと翻訳モデルを統計手法に基づき学習し，未知の原言語文を翻訳するものである．その際，必要となるのは対訳コーパスのみであるため，様々な言語間の対訳コーパスを構築することで翻訳対象となる言語を制限することなくシステムの構築が可能である．更に，一般的には対訳コーパスのデータサイズが大きくなればなるほど翻訳精度は向上する．この統計翻訳の問題点の 1 つは対訳コーパスのデータサイズが大きくなるのに伴い，モデルの学習に時間がかかることであった．しかし，近年の計算機におけるハードウェアの性能向上により，この問題も大幅に改善されてきた．そのことが，統計翻訳研究の急速な進展の大きな要因となっている．

統計翻訳は機械翻訳分野において最も活発に行われている研究テーマであるが，統計翻訳研究の進展に伴い，大きな問題となっているのが評価方法である．通常，基準となる統計翻訳システム，即ち，ベースラインシステム（改良対象となるシステムはベースラインシステムと呼ぶ.）に対して改良を行った場合，ベースラインシステムと提案手法の翻訳文を評価し，その評価結果を比較することで提案手法の有効性を明らかにする.その際，翻訳文に対する最も正確な評価方法は人手による評価だと考えられる．しかし，人手評価はコストと時間を要することが大きな問題であり，そのことが改良，実験，評価といった開発サイクルの速度向上の大きな妨げとなる．そこで，短時間かつ低コストで評価が可能な新たな評価法へのニーズが高まっている．そして，このニーズに応えるために開発されたのが自動評価法である．自動評価法は機械翻訳システムが出力する翻訳文を自動的に評価することを目的とした評価システムであ

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

る．具体的には，自動評価法は翻訳文と人手により作成される正翻訳（自動評価法の分野では，「参照訳」と呼ばれることが多い．）を比較することで，翻訳文が正翻訳に対してどれだけ近いかを自動的に数値化し，それを評価値とする．また，近年は参照訳を必要としない評価システムの研究も「品質推定タスク」として行われている．

このような自動評価法を含む統計翻訳分野の急速な発展を背景として，コンテスト型ワークショップが開催されるようになった．そして，そのようなワークショップの中でも精力的な活動を行っているものの１つとして「**W**orkshop on Statistical **M**achine **T**ranslation（略して，以降，WMT と呼ぶ．）」が挙げられる．本ワークショップでは機械翻訳に関する研究がタスクごとに分類されており，それぞれのタスク毎に，参加者が提供するシステムの結果を評価し，システムの順位付けが行われる．そして，対象タスクにおけるシステムの現状把握及び今後の研究のための課題を明らかとする．開催初年度の 2006 年は「翻訳タスク」のみであったが，2008 年より「自動評価タスク」が加わり，2013 年には更に「品質推定タスク」が加わった．また，言語については 2008 年はドイツ語と英語間，フランス語と英語間，スペイン語と英語間，スペイン語とドイツ語間，チェコ語と英語間，そして，ハンガリー語と英語間の双方向での翻訳データが作成された．2013 年には，フランス語と英語間，スペイン語と英語間，ドイツ語と英語間，チェコ語と英語間，そして，ロシア語と英語間の双方向での翻訳データが作成された．このように WMT では例年，数多くの言語に対する翻訳データが利用可能である．

機械翻訳分野の研究者は本ワークショップに参加することにより，自らが開発したシステムが対象タスクの中でどの程度のレベルにあるのかを明確にすることができる．また，本ワークショップで利用されるデータはオープンデータとなっているため，ワークショップに参加していない場合でも，データを取得し，利用することができる．

そこで，本報告では，このコンテスト型ワークショップ WMT の「自動評価タスク」にて使用され，更に，公開されているデータから近年に開催された 2012 年[2] と 2013 年[3] のデータを用いて行った，提案手法を含めたメタ評価の結果について報告する．メタ評価とは，評価システムの評価を目的として行う性能評価を意味する．複数の自動評価法を含むメタ評価を行うことで，提案手法の特徴や問題点が明らかとなった．

2　自動評価法

我々は従来より，独自の自動評価法を提案[4] し，その有効性について検証してきた．本章では代表的な自動評価法の幾つかを紹介し，その後で提案手法の詳細について述べる．そのことにより，提案手法の特徴とその利点について言及する．

2.1　先行研究

現在までに様々な自動評価法が既に提案されているが，本節では，その代表的なものをいくつか紹介する．

2.1.1　BLEU

最も広く使用されているスタンダードな自動評価法として BLEU（A **Bi**lingual **E**valuation **U**nderstudy）[5] が挙げられる．自動評価の研究が急速に進んだ背景としてこの BLEU の存在は欠かすことはできない．現在では，BLEU よりも高い評価精度を有する自動評価法を考案することが１つの目的となっており，WMT においてもベースラインシステムとして利用されている．

BLEU は n グラム一致率に基づく自動評価法である．以下の式(1)から式(3)に BLEU の計算式を示す．式(1)は n の値を変化させた際の n グラム適合率を示している．$n=1$ の場合には 1-*gram* 適合率，$n=2$ の場合には 2-*gram* 適合率を計算することになる．ここで，適合率とは翻訳文における n グラム一致率である．

式(2)はペナルティを示している．式(1)の n グラム適合率を求める際に問題となるのは，翻訳文が短い場合，過度に高い値を示すことである．例えば，翻訳文が〝the〟や〝a〟だけの場合，明らかに誤った翻訳文であっても参照訳に〝the〟や〝a〟が存在する可能性は高いため，n グラム適合率は高くなる．このような問題を解決するために，翻訳文が短い場合には最終的な評価値にペナルティである式(2)を重み付けとして用いる．具体的には，翻訳文の長さを示す c と参照訳の長さを示す r を比較し，翻訳文の方が長い場合には BP は１となる．即ち，ペナルティを与えない．それに対して，c が r より小さい場合，翻訳文が参照訳に

対して短いため $e^{(1-\frac{r}{c})}$ より得られた値を *BP* として用いる．即ち，*BP* は 1 以下となるため，得られた *n* グラム適合率よりも小さな値となる．このように翻訳文の長さと参照訳の長さの比に応じてペナルティ *BP* の値が決定する．

最終的な評価値は式(3)より得られる．式(3)は *n* の値を変化させた際の各 *n* グラム適合率の相乗平均を示している．使用する *n* グラム適合率の種類としては 1-*gram*，2-*gram*，3-*gram*，そして，4-*gram* の 4 種類，即ち，$N=4$ が適切とされている．また，式(3)の w_n は $1/N$ である．

$$p_n=\frac{\sum_{C\in\{Candidates\}}\sum_{n-gram\in C} Count_{clip}(n\text{-}gram)}{\sum_{C'\in\{Candidates\}}\sum_{n-gram'\in C'} Count(n\text{-}gram')} \tag{1}$$

$$BP=\begin{cases}1 & if \quad c>r\\ e^{(1-\frac{r}{c})} & if \quad c\leqq r\end{cases} \tag{2}$$

$$BLEU=BP\cdot\exp\left(\sum_{n=1}^{N} w_n\log p_n\right) \tag{3}$$

BLEU の評価値は 0.0〜1.0 の範囲であり，値が大きいほど評価が高い．また，BLEU は *n* グラム一致率を求めることで容易に評価値が得られることから高速処理が可能である．一方，BLEU は相乗平均を用いていることから，1-*gram*〜4-*gram* 適合率の値が 1 つでも 0 になると，BLEU の評価値は 0.0 になってしまう．BLEU はドキュメント単位での評価を目的としているため，ドキュメントを評価対象とする場合には問題にはならないが，文単位での評価においては 4-*gram* でマッチする頻度は低くなるため評価値が 0.0 になる可能性が高く，適切な評価が困難となる．したがって，BLEU は文単位の評価には適さないとされている．

2.1.2 NIST

NIST（**N**ational **I**nstitute of **S**tandards and **T**echnology）[6] もまた BLEU と同様に *n* グラム一致率に基づく自動評価法ではあるが，*n* グラム適合率に対して相互情報量に基づく重みづけを行っている点で異なる．NIST の評価値は式(4)，(5)より得られる．

式(4)は相互情報量による重み付けを *n* グラム適合率に対して行っている．そのため，出現頻度の低い *n-gram* は特徴的な意味のある表現と位置付けられ，*Info* の値，即ち，情報量は多くなる．例えば，2-*gram* において，〝the computer〟と〝the businessman〟の 2 つのフレーズにおいて，〝the〟の出現回数が 10，〝the computer〟の出現回数が 2，そして，〝the businessman〟の出現回数が 5 であるとする．その場合，相互情報量 *Info* の値は〝the computer〟の方が〝the businessman〟よりも高くなる．このように，NIST では，*n* グラム適合率に対して相互情報量に基づく重み付けを行うことにより，意味を考慮した評価法となっている．

そして，式(5)より，この相互情報量に基づく各 *n* グラム適合率の相加平均が NIST の評価値となる．NIST では N の値として一般的に 5 が使用される．また，$\exp\left\{\beta\log^2\left[min\left(\frac{L_{sys}}{\bar{L}_{ref}},1\right)\right]\right\}$ は，ペナルティを意味する．参照訳の長さに対する翻訳文の長さの比と 1 との間の最小値がペナルティの計算に利用される．翻訳文が参照訳よりも長い場合には，ペナルティとしては 1 が選択される．したがって，元々の評価値に対して何の影響も与えない．一方，翻訳文が参照訳よりも短い場合には，参照訳の長さに対する翻訳文の長さの比が 1 未満となるため，その比がペナルティとして選択される．その結果，元々の評価値よりも小さな値となる．

$$Info(w_1\ldots w_n)=\log_2\left(\frac{\text{the \# of occurances of } w_1\ldots w_{n-1}}{\text{the \# of occurances of } w_1\ldots w_n}\right) \tag{4}$$

$$Score=\sum_{n=1}^{N}\frac{\sum_i\left(\sum_{\substack{\text{翻訳文 } i \text{ と参照訳 } i\\ \text{に共通する } w_i\ldots w_n}} Info_i(w_1\ldots w_n)\right)}{\sum_i \text{翻訳文 } i \text{ 中の全 } n\text{-}gram \text{ 数}}\cdot\exp\left\{\beta\log^2\left[min\left(\frac{L_{sys}}{\bar{L}_{ref}},1\right)\right]\right\} \tag{5}$$

NIST の評価値は 0.0〜∞ の範囲で示され，大きな値ほど評価が高い．また，BLEU と同様，*n-gram* を求めるだけで容易に評価値を得られるため，高速処理が可能である．更に，NIST は BLEU と同じくドキュメント単位での評価に特化した自動評価法であり，文単位での評価には適さないとされている．

2.1.3 WER

WER は単語誤り率（**W**ord **E**rror **R**ate）[7] を表

す指標であり，音声認識などでも利用されている．また，WER は編集距離に基づく自動評価法である．編集距離とは置換（substitutions），挿入（insertions），そして，削除（delections）の 3 つの操作に基づいており，翻訳文が参照訳と一致するために，これら 3 つの操作が何回必要かを求めることで得られる．

WER は以下の式(6)より得られる．分母は参照訳の長さ，即ち，参照訳の構成単語数である．また，置換，挿入，そして，削除の 3 つの操作は単語単位で行われる．

$$WER=\frac{substitutions+insertions+deletions}{reference\text{-}length} \tag{6}$$

WER の評価値は 0.0〜∞ の範囲で示され，小さな値ほど評価が高い．即ち，値が小さいほど翻訳文に対する修正操作が必要なく，良質な翻訳文と見なすことができる．また，編集距離はレーベンシュタイン距離[8] に基づき効率的に求めることができるため，高速処理が可能である．更に，編集距離は文単位で算出されるため，ドキュメント単位だけでなく文単位の評価にも有効とされている．WER の特徴としては，左から右に向けて同じ並びで翻訳文と参照訳間の一致単語が存在する場合のみに一致単語は評価値に反映され，一致単語の並びが翻訳文と参照訳間で交差する場合には，評価値には反映されないことから語順に厳しい評価法とされている．

2.1.4 METEOR

METEOR（**M**etric for **E**valuation of **T**ranslation with **E**xplicit **O**rdering）[9] は BLEU や NIST のように翻訳文に対する適合率のみに基づく手法とは異なり，参照訳による再現率も考慮した自動評価法である．METEOR では翻訳文による適合率と参照訳による再現率の F 値を評価値として求める．

以下の式(7)，(8)，そして，(9)より METEOR の評価値は得られる．式(7)における P は適合率，R は再現率を示している．F_{mean} は適合率と再現率による F 値を表している．α はパラメータである．また，式(8)は語順に基づくペナルティである．ch は一致単語の塊であるチャンクの数を示す．また，m は一致単語数を示す．β，γ はそれぞれパラメータである．例えば，一致単語数が 6 で，その全てが翻訳文と参照訳の間で同じ並びかつ連続して出現している場合，ch は 1 となり，m は 6 となる．したがって，Pen の値は $\frac{1}{6}$ より 0.17 となる．それに対して，一致単語数が全て翻訳文と参照訳の間で逆順に出現している場合，ch は 6 となる．したがって，Pen は $\frac{6}{6}$ より 1 となる．この式(8)のペナルティ Pen は式(9)にて（$1-Pen$）として F_{mean} に対する重み付けに使用されるため，Pen が小さいほど式(9)の評価値には影響しない．それに対して，Pen が 1 に近いほど F_{mean} の値は小さくなる．即ち，一致単語の語順が異なるほど，式(9)の（$1-Pen$）は 0 に近づき，F_{mean} の値を小さくする方向に作用する．また，METEOR は 3 つのパラメータ α，β，そして，γ の値が評価精度に大きく影響するとして，言語毎に最適なパラメータ値を設定する必要があるとしている．

$$F_{mean}=\frac{P\cdot R}{\alpha\cdot P+(1-\alpha)\cdot R} \tag{7}$$

$$Pen=\gamma\cdot\left(\frac{ch}{m}\right)^{\beta} \tag{8}$$

$$score=(1-Pen)\cdot F_{mean} \tag{9}$$

METEOR の評価値は 0.0〜1.0 の範囲で示され，大きな値ほど評価が高い．また，METEOR は表層的な単語間の一致だけでなく，オプションとして形態素情報による単語の語形変化や WordNet による類義語に基づく一致も考慮した評価値を算出することができる．しかし，先に述べた自動評価法に比べて，チャンクの抽出処理などが必要なため処理時間は長くなる．一方，METEOR は適合率だけでなく再現率も考慮していることから文単位の評価に適している．

2.1.5 その他の先行研究

上述した自動評価法以外にも様々な自動評価法が提案されている．ROUGE（**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation）[10] シリーズの ROUGE-L は，最長共通部分列（**L**ongest **C**ommon **S**ubsequence，以後，〝LCS〟と呼ぶ．）[11] に基づく手法である．ROUGE-L も評価値は 0.0〜1.0 の範囲で示され，大きな値ほど評価が高いことを意味する．また，LCS の値は WER と同様にレーベンシュタイン距離に基づき効率的に求めることができるため，高速処理が可能である．そして，文単位の評価にも有効である．更に，

ROUGE-L は LCS を用いているため，一致単語の出現順が翻訳文と参照訳で異なる場合，それらの一致単語は評価値には全く反映されず，語順に厳しい評価法と言える．

また，TER（**T**ranslation **E**rror **R**ate）[12] は WER が編集距離の置換，挿入，そして，削除の 3 つの処理に基づき評価値を算出していたのに対して，シフト処理も加えた 4 つの操作に基づき評価値を算出する．スコアは WER と同様 0.0 以上の値であり，値が小さいほど高い評価となる．また，高速処理も可能であり，文単位の評価にも有効である．これら ROUGE シリーズ，TER は共に一般公開されており，容易に入手及び実行が可能である．

2.2　提案手法

2.1 節で述べた先行研究に対して，我々は多義性のある一致単語を大局的な観点から一意に決定し，かつ，一致単語の語順の違いに柔軟に対応可能な自動評価法を従来より提案している[4]．一致単語の決定を大局的な観点より行うことで，翻訳文と参照訳の間で正しい一致単語を得ることができる．また，一致単語の語順の違いをどの程度，評価値に反映するかをパラメータを用いて柔軟に決定することが可能となる．本節では，前者の処理と後者の処理をそれぞれ「チャンクの決定方法」と「評価値の算出方法」として，提案手法の詳細について述べる．

2.2.1　チャンクの決定方法

提案手法では，始めに一致単語を一意に決定する．一致単語の決定には多義性が存在することから，一致単語を一意に決定することは非常に重要である．例えば，英文においては冠詞の〝the〟や〝a〟は 1 つの文中に複数存在することは少なくない．日本文においても助詞の〝の〟や〝を〟が 1 文中に複数回出現することは少なくない．その際，翻訳文中のどの単語と参照訳中のどの単語を一致単語とすべきなのかには多義性があり，一意に決定することが困難となる．

そこで，提案手法ではそれぞれの一致単語を独立に扱うのではなく，連続して出現する一致単語については，その部分を 1 つの塊，即ち，チャンクを単位として評価値を得る．その結果，一致単語の前後の文脈も考慮した評価を実現できる．チャンクを決定するための具体的な手順は以下のようになる．

始めに，翻訳文と参照訳間において LCS を決定する．例えば，翻訳文として〝a glass guide molded in panel member P made of resin〟，参照訳として〝glass guide of the plastic mounting panel P〟との間でチャンクを決定することを考える．その場合，LCS の値が同じであっても抽出されるチャンクが異なる場合がある．即ち，LCS 経路が複数存在するということである．上述した例では，以下の 2 つの LCS 経路が得られる．ここで，〝[〟と〝]〟の間の単語列は 1 つのチャンクを意味する．

LCS 経路 1
翻訳文：a［glass guide］molded in［panel］member［P］made of the resin
参照訳：［glass guide］of the plastic mounting［panel］［P］
LCS 経路 2
翻訳文：a［glass guide］molded in panel member P made［of the］resin
参照訳：［glass guide］［of the］plastic mounting panel P

この例では，LCS の値は LCS 経路 1，2 共に 4 である．そこで，提案手法では，このように複数の LCS 経路が得られる場合には，以下の式(10)と式(11)に基づき，式(10)の *score* が最大となる LCS 経路を一意に決定する．

$$score = \sum_{c \in c\_num} (length(c)^{\beta} \times pos) \qquad (10)$$

$$pos = \left(1.0 - \left|\frac{c_i}{m} - \frac{c_j}{n}\right|\right) \qquad (11)$$

式(10)の *score* は個々の LCS 経路に対して求める．$length(c)$ は LCS 経路中の個々のチャンクの構成単語数を示す．また，β はチャンクの長さに対する重み付けパラメータであり，1.0 以上を用いる．したがって，*score* は基本的には個々のチャンクが長いほど，大きな値となる．しかし，チャンクの長さだけでは，正しい一致単語を決定できるとは限らないため，翻訳文と参照訳に存在するチャンクの位置も考慮する．例えば，上述した LCS 経路 1 と LCS 経路 2 では，チャンクの長さだけで *score* を求めると LCS 経路 2 が選択される．しかし，LCS 経路 1 の方が，一致単語の抽出

結果としては適切である．そこで，チャンクの長さだけではなく，チャンクの位置も考慮することが必要と考えられる．提案手法ではこの位置情報を式(11)より求める．

式(11)の *pos* はチャンクの相対的な位置のずれを表している．$\frac{c_i}{m}$ は参照訳中のチャンクの相対位置，$\frac{c_j}{n}$ は翻訳文中のチャンクの相対位置をそれぞれ表している．この 2 つの差の絶対値を求め，絶対値を 1.0 から引くことで相対的な位置のずれを求める．相対位置のずれが大きいほど *pos* の値は小さくなる．また，相対位置のずれが小さいほど *pos* の値は大きくなる．したがって，式(10)に式(11)の *pos* を適用した際には，相対位置のずれが大きいほど，*score* は値はより小さな値となる．

このようにチャンクの長さだけでなく，チャンクの相対的な位置も考慮した式(10)と式(11)を用いることにより，正しい一致単語を有する LCS 経路を選択できる．例えば，上述した LCS 経路 1 と LCS 経路 2 に式(10)と式(11)を適用した場合，LCS 経路 1 の *score* は

$$3.499\Big(=2^{1.2}\times\Big(1.0-\Big|\frac{1}{8}-\frac{2}{12}\Big|\Big)+1^{1.2}\times\Big(1.0-\Big|\frac{7}{8}-\frac{6}{12}\Big|\Big)+1^{1.2}\times\Big(1.0-\Big|\frac{8}{8}-\frac{8}{12}\Big|\Big)\Big),$$

LCS 経路 2 の *score* は

$$3.446\Big(=2^{1.2}\times\Big(1.0-\Big|\frac{1}{8}-\frac{2}{12}\Big|\Big)+2^{1.2}\times\Big(1.0-\Big|\frac{3}{8}-\frac{10}{12}\Big|\Big)\Big)$$

となる．なお，この場合のパラメータ β の値は 1.2 である．この結果より，LCS 経路 1 の *score* は LCS 経路 2 の *score* よりも大きいため，LCS 経路 1 が最適な一致単語を持つ LCS 経路として選択される．

2.2.2 評価値の算出方法

次いで，提案手法では，評価値を算出する．チャンクの決定処理により，複数の LCS 経路，即ち，一致単語の多義性が存在する際には，最適な一致単語を持つ LCS 経路を一意に決定する．そして，その決定された一致単語に基づき評価値を算出する．その算出式を式(12)と式(13)，そして，式(14)に示す．

式(12)の R は参照訳における一致単語に基づく値，即ち，再現率を示している．また，式(13)の P は翻訳文における一致単語に基づく値，即ち，適合率を示している．分子の式は式(12)と式(13)共に同じ式である．*length*(c) はチャンクの長さである．*c_num* はチャンクの数である．上述した LCS 経路 1 においては，チャンクは〝glass guide〞，〝panel”，そして，〝P〞の 3 つである．この 3 つのチャンクのそれぞれの長さに基づき分子の値を計算する．まず，チャンク〝glass guide〞の構成単語数は 2 である．重み付けパラメータ β を 2.0 とすると，このチャンクの値としては $4(=2^{2.0})$ となる．同様に，チャンク〝panel〞と〝P〞についても計算すると，全てのチャンクの値の総和として，$6(=2^{2.0}+1^{2.0}+1^{2.0})$ が得られる．

更に，提案手法では全ての一致単語を評価値に反映させることができるように，このチャンクの決定処理を再帰的に繰り返す．例えば，LCS 経路 1 においては，〝[〞と〝]〞の間に挟まれたチャンクを除いたとしても他のチャンクとして〝of the〞も存在している．しかし，LCS ではチャンクの並びが 2 つの文間で異なる場合には，一致単語として見なされないため，LCS 経路 1 においては〝of the〞はチャンクの対象から外れる．そこで，提案手法では全ての一致単語を評価値に反映可能とするために，決定された一致単語を除いた上で，更にチャンクを決定する．この再帰処理は翻訳文と参照訳間で一致単語が存在しなくなるまで繰り返される．また，再帰処理により抽出されるチャンクは出現順が異なるチャンクであることから，最初に決定されるチャンクとは同一視しない方が適切な場合がある．そこで，提案手法では，出現順が異なるチャンクをどの程度，評価値に反映させるかはパラメータを用いて制御する．

式(12)と式(13)の分子の式にある $\sum_{i=0}^{RN-1}$ がチャンクの抽出における再帰処理を示している．*RN* はチャンクの決定処理の回数である．LCS 経路 1 の場合には，最初のチャンクの決定処理で〝glass guide〞，〝panel”，そして，〝P〞が抽出され，2 度目の決定処理で〝of the〞が抽出されるため，*RN* は 2 である．また，α は再帰処理により決定される一致単語をどの程度評価値に反映させるかを制御するためのパラメータであり，1.0 以下の値を取る．例えば，LCS 経路 1 では，パラメータ α の値を 0.5 とするとチャンク〝glass guide〞，〝panel”，そして，〝P〞に対しては，$i=0$ であるため $1(=0.5^0)$ がチャンクの値の総和 6 に対して重

み付けされる．この場合，チャンクの値の総和に対しては影響を与えないということになる．次いで，再帰処理により決定されたチャンク〝of the〟においては，パラメータ β の値を 2.0 とすると，チャンクの値としては 4($=2^{2.0}$) となる．そして，このチャンクに対する出現順に対する重み付けは，$i=1$ より 0.5($=0.5^1$) となるため，チャンクの値としては 2($=0.5\times4$) となる．これで全ての一致単語の抽出が終了となるため，最初のチャンクの値 6 と再帰処理によるチャンクの値 2 を足すことで，式(12)と式(13)の分子の値は最終的に 8 となる．このようにパラメータ α を用いることで，出現の異なる一致単語をどの程度反映させるかを制御可能となる．1.0 の場合には，一致単語の出現順を問わない評価値が得られ，0.0 に近い値ほど出現順に厳密な評価値が得られることになる．

更に，式(12)と式(13)の分子の計算式により得られたチャンクの値に対して，参照訳の長さ m と翻訳文の長さ n を用いて正規化することで，それぞれ式(12)の R と式(13)の P を算出する．例えば，LCS 経路 1 の場合，R と P はそれぞれ $0.354\left(=\sqrt{\frac{8}{8^{2.0}}}\right)$ と $0.236\left(=\sqrt{\frac{8}{12^{2.0}}}\right)$ になる．そして，提案手法における最終的な評価値が式(14)により得られる．ここで，式(14)の γ は P/R より得られる．例えば，LCS 経路 1 では γ の値は $0.667\left(=\frac{0.236}{0.354}\right)$ となる．したがって，最終的な評価値 $AE\ score$ は

$$0.264\left(=\frac{(1+0.667^2)\times0.354\times0.236}{0.354+0.667^2\times0.236}\right)$$

となる．

$$R=\left(\frac{\sum_{i=0}^{RN-1}(\alpha^i\sum_{c\in c\_num}length(c)^\beta)}{m^\beta}\right)^{\frac{1}{\beta}} \quad (12)$$

$$P=\left(\frac{\sum_{i=0}^{RN-1}(\alpha^i\sum_{c\in c\_num}length(c)^\beta)}{n^\beta}\right)^{\frac{1}{\beta}} \quad (13)$$

$$AE\ score=\frac{(1+\gamma^2)RP}{R+\gamma^2P} \quad (14)$$

このように提案手法では，翻訳文と参照訳間の適切な一致単語を決定し，その上で一致単語の語順の違いをどの程度評価値に反映させるかをパラメータにより制御する．したがって，従来の自動評価法に比べ，より厳密な評価が可能になると考えられる．

更に，提案手法では処理時間の短縮のための最適化[13] を行っている．提案手法では，全ての LCS 経路を求め，その中から適切な一致単語を持つ LCS 経路を選択するため，LCS 経路の数が多くなると処理時間が増加する．この問題を解決するために，LCS 経路の効率的な探索のための最適化を行っている．

3　性能評価実験

3.1　実験データ

本報告では，WMT2012 と WMT2013 を用い，提案手法を含む複数の自動評価法に対するメタ評価を行った．実験データは WMT2012[14] と WMT2013[15] の「自動評価タスク」で使用されたデータを入手して用いた．WMT2012 データは，チェコ語(cs)と英語(en)間，ドイツ語(de)と英語間，スペイン語(es)と英語間，そして，フランス語(fr)と英語間で双方向でのデータが存在するが，今回は最も広く使用されている英語の翻訳データに限定したメタ評価を行った．また，WMT データには翻訳文に対応する参照訳と人手評価も含まれている．人手評価は 5 つの機械翻訳システムが出力した翻訳文 5 文を評価者が比較し，それぞれに対して 5 段階評価を与えたものである．WMT2013 データにおいては，WMT2012 データの言語に加え，ロシア語(ru)と英語間の双方向のデータも存在する．

3.2　実験方法

実験の手順について述べる．まず，翻訳文と参照訳を入力データとして自動評価システムより評価値を得た．本実験で用いた自動評価法はドキュメント単位においては，BLEU（ver.13a）[16]，NIST（ver.13a）[16]，METEOR（ver1.4）[17]，そして，提案手法[18] である．また，文単位においては，METEOR（ver1.4）と提案手法を用いた．2.1 で述べたように，BLEU と NIST についてはドキュメント単位のみの評価を対象にしているため，文単位に対しては使用しなかった．次いで，各自動評価法において，評価値と人手による参照訳との間の相関を求めた．参照訳の数は各翻訳文に対して 1 つである．ドキュメント単位の相関について

表 1 ：WMT2012 を用いたドキュメント単位におけるスピアマンの順位相関係数

Metrics	cz-en (6 systems)	de-en (16 systems)	es-en (12 systems)	fr-en (15 systems)	Avg.	Rank
提案手法	0.886	0.676	0.958	0.807	0.832	3
BLEU	0.886	0.674	0.958	0.796	0.828	4
NIST	0.943	0.700	0.944	0.779	0.841	2
METEOR	0.943	0.841	0.979	0.818	0.895	1

表 2 ：WMT2012 を用いた文単位におけるケンドールの順位相関係数

Metrics	cz-en (11,155 sentences)	de-en (12,042 sentences)	es-en (9,880 sentences)	fr-en (11,682 sentences)	Avg.	Rank
提案手法	0.189	0.207	0.208	0.226	0.207	2
METEOR	0.223	0.279	0.248	0.243	0.248	1

は，スピアマンの順位相関係数，文単位の相関については，ケンドールの順位相関係数をそれぞれ求めた．相関係数の利用方法については WMT と同様である．

なお，提案手法における自動評価システムのパラメータ α と β の値には，予備実験に基づき最適と考えられる 0.1 と 1.2 をそれぞれ用いた．

3.3 実験結果と考察

表 1 には WMT2012 を用いたドキュメント単位におけるスピアマンの順位相関係数，表 2 には WMT2012 を用いた文単位におけるケンドールの順位相関係数を示す．そして，表 3 には WMT 2013 を用いたドキュメント単位におけるスピアマンの順位相関係数，表 4 には WMT2013 を用いた文単位におけるケンドールの順位相関係数を示す．表中の項目〝Rank〟は各言語毎の相関係数の平均〝Avg.〟に対して順位付したものである．

表 1 と表 3 より，ドキュメント単位においては提案手法のランキング結果はそれぞれ 2 位と 3 位であり，全自動評価法の中で中間的な位置付けであった．相関係数は 0.8 を上回っており，十分ではないが人手評価との相関はあると考えられる．言語毎の相関係数を見ると，提案手法においては，表 1 より〝de-en〟の相関係数が最も低く，他手法との比較においても相関係数は低い．この結果は，語順の大きく異なる言語間の翻訳の場合，提案手法が十分に機能していないことが考えられる．表 3 においても，提案手法では，〝ru-en〟を除き〝de-en〟の相関係数は他の言語に比べ，低い相関係数を示している．この問題を解決するためのには，式(12)，式(13)における語順の影響を制御するためのパラメータ α に対して適切な値を調査することなどが考えられる．また，表 3 では，〝ru-en〟の相関係数が他の言語の相関係数に比べ，非常に低い．しかし，これは他手法においても同様の傾向であり，提案手法だけの問題とは言えない．即ち，〝ru-en〟の評価精度の向上は自動評価法全般の問題と考えられ，今後精査する必要がある．

表 2 と表 4 の文単位においては，提案手法，METEOR 共に相関係数は非常に低いものであった．特に表 4 より WMT2013 の提案手法における相関係数は 0.2 を下回った．文単位の評価は，全ての自動評価法にとって最も深刻な問題であり，今後積極的にその解決に向けて取り組まなければならない．

4 まとめ

本報告では，毎年開催されるコンテスト型ワークショップ WMT が公開している WMT2012 と WMT2013 のデータを用いて行った，提案手法を含む自動評価法のメタ評価の結果について述べた．メタ評価の結果，提案手法の評価精度のランキング結果は他手法に比べて最上位に位置するものではなかった．しかし，特定の言語の相関係数が低くなっていることが要因であることから，今後はパラメータのチューニングを行うことで言語毎の過度なばらつきを抑え，ランキング結果の改善を図る．

更に，今後は自動評価全般の問題である，文単位の評価精度の向上のための研究を進める予定で

表 3 ：WMT2013 を用いたドキュメント単位におけるスピアマンの順位相関係数

Metrics	cz-en (11 systems)	de-en (17 systems)	es-en (12 systems)	fr-en (13 systems)
提案手法	0.909	0.909	0.937	0.934
BLEU	0.945	0.897	0.853	0.951
NIST	0.900	0.828	0.804	0.786
METEOR	0.982	0.946	0.923	0.967

Metrics	ru-en (19 systems)	Avg.	Rank
提案手法	0.721	0.882	2
BLEU	0.614	0.852	3
NIST	0.465	0.757	4
METEOR	0.889	0.941	1

表 4 ：WMT2013 を用いた文単位におけるケンドールの順位相関係数

Metrics	cz-en (85,469 sentences)	de-en (128,668 sentences)	es-en (67,832 sentences)	fr-en (80,741 sentences)
提案手法	0.148	0.167	0.176	0.142
METEOR	0.222	0.236	0.241	0.194

Metrics	ru-en (151,422 sentences)	Avg.	Rank
提案手法	0.123	0.151	2
METEOR	0.226	0.224	1

ある．

参考文献

［1］P. Koehn. 2010. Statistical Machine Translation. *Cambridge University Press*.

［2］C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut and L. Specia. 2012. Findings of the 2012 Joint Workshop on Statistical Machine Translation. Proceedings of the 7th Workshop on Statistical Machine Translation. pp.10–51.

［3］O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz M. Post, R. Soricut and L. Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. Proceedings of the Eighth Workshop on Statistical Machine Translation. pp.1–44.

［4］H. Echizen-ya and K. 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.

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

［6］G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Proceedings of the second International Conference on Human Language Technology Research. pp.138–145.

［7］G. Leusch, N. Ueffing and H. Ney. 2003. A Novel String-to-String Distance Measure With Applications to Machine Translation Evaluation. Proceedings of the 9th Machine Translation Summit (MT Summit). pp.311–318.

［8］T. Komori and S. Katagiri. 1992. GPD Training of Dynamic Programming-based Speech Recognizers. *Journal of the Acoustical Society of Japan (E) 13(6)*. pp.341–349.

［9］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.

[10] Chin-Yew Lin and Franz Josef Och. 2004. Auto-

matic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL). pp.606-613.

[11] D. S. Hirschberg. 1975. A Linear Space Algorithm for Computing Maximal Common Subsequences. *Communications of the ACM*. Volume 10 Issue 6. pp. 341-343.

[12] M. Snover, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas (AMTA). pp. 223-231.

[13] H. Echizen'ya, K. Araki and E. Hovy. 2012. 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.

[14] http://www.statmt.org/wmt12/results.html.

[15] http://www.statmt.org/wmt13/results.html.

[16] http://www.itl.nist.gov/iad/mig//tools/.

[17] http://www.cs.cmu.edu/~alavie/METEOR/index.html#Download.

[18] http://www.lst.hokkai-s-u.ac.jp/~echi/impact.html.