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	WMT2015データにおける単語の分散表現と単語アライメントを用いた機械翻訳のための自動評価法の有効性について
著者	越前谷，博; ECHIZEN'YA, Hiroshi; 荒木，健治; ARAKI, Kenji
引用	北海学園大学工学部研究報告(46): 175-190
発行日	2019-02-14

Hokkai-Gakuen Organization of Knowledge Ubiquitous through Gaining Archives.

WMT2015データにおける単語の分散表現と単語アライメントを用いた機械翻訳のための自動評価法の有効性について

越前谷　　博*・荒　木　健　治**

Effectiveness of Automatic Evaluation Metrics using Word Embeddings and Word Alignment for Machine Translation in WMT2015 data

Hiroshi ECHIZEN'YA* and Kenji ARAKI**

Abstract

近年，ディープラーニングの技術は自然言語処理においても急速に普及している．単語分散表現に代表されるように意味に基づく言語処理がディープラーニングにより可能となった．本報告では，その単語分散表現を機械翻訳の自動評価に利用する．提案手法では，一般的な自動評価法と同様に機械翻訳システムによる訳文を人手による参照訳と比較し，スコア付けすることで評価を行なう．スコア計算は，Earth Mover's Distanceに基づいて行なう．その際，単語の特徴量に単語分散表現，また，単語の重みには*tf*・*idf*を用いる．また，距離計算にはコサイン距離を用いる．更に，距離行列を作成する際には，訳文と参照訳間の単語アライメントの結果を反映させることで評価精度を向上させる．性能評価実験の結果，単語アライメント及び語順情報を用いることで人手評価との相関が向上し，提案手法の有効性を確認した．

1　はじめに

現在，ディープラーニングの技術は機械翻訳や対話などの自然言語処理研究において革新的な役割を果たしている．機械翻訳分野ではsequence–to–sequenceモデルに基づくシステムが従来主流となっていた統計翻訳［1,2,3］よりも高品質な訳文を生成することが知られている．sequence－to－sequenceモデル［4］はEncoderとDecoderの２つのReccurent Neural Network（RNN）で構成されている．EncoderのRNNは入力文，即ち，単語列をベクトル化する．そし

* 北海学園大学工学部
Faculty 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

て，DecoderのRNNはそのベクトルを受け取り訳文を生成する．このsequence−to−sequenceモデルを用いた機械翻訳を利用することで，より流暢で自然な訳文が生成可能となり，機械翻訳研究は飛躍的に進歩した．このようなディープラーニングを用いた機械翻訳はニューラル機械翻訳［5,6］と呼ばれ，現在，最も盛んに研究されている自然言語処理研究の一つである．

このように機械翻訳分野は劇的な発展を遂げているが，機械翻訳システムの訳文に対する自動評価の精度は十分とはいえない．評価作業が滞るとシステムの開発サイクル全体に悪影響を及ぼすことになる．現在主流となりつつあるニューラル機械翻訳はFluencyの観点での翻訳品質を格段に向上させた．そのため，Fluencyの評価においてはニューラル機械翻訳より得られる訳文に対する十分な差別化は現在の自動評価法では困難である．一方，Adequacyの評価においても，自動評価法のディファクトスタンダートとなっているBLEU［7］を始め現在の自動評価法は十分とはいえない．BLEUは表層レベルのn−gram一致率に基づいているため，単語の意味を扱ったうえでの評価が困難である．したがって，単語の意味を考慮した自動評価法がニューラル機械翻訳のために求められている．このような自動評価の状況において，本報告では単語分散表現を用い，かつ，単語アライメントの結果を用いた，新たな自動評価法を提案する．単語分散表現を用いることにより，単語の意味を考慮することが可能となる．また，単語アライメントを行うことで対応関係にある単語をより正確に反映した自動評価を実現できる．具体的には，Earth Mover's Distance（以降，EMDと呼ぶ）[8,9,10］より訳文と参照訳間の類似度を求める．その際，個々の単語の特徴量には単語分散表現を用いる．また，距離行列の生成の際に単語アライメントにより単語間の対応関係の有無を決定し，対応関係がある場合とない場合とで距離の差別化を図る．更に，対応関係にある単語間においては表層レベルで一致するかどうかと語順情報も考慮したうえで距離を求める．WMT2015［11］の自動評価タスク［12］のデータを用いた性能評価実験の結果，提案手法による評価スコアと人手によるスコアとの相関係数は，単語アライメントを全く考慮しないEMDのみの相関係数よりも高く，提案手法の有効性を確認することができた．また，他の自動評価法とのメタ評価においても，提案手法はMETEOR［13］とほぼ同等な評価精度を示した．

2　先行研究

2.1　BLEU

代表的な自動評価法をいくつか取り上げる．まず，ディファクトスタンダードな自動評価法としてBLEU（A **Bi**lingual **E**valuation **U**nderstudy）［7］が挙げられる．BLEUが提案されたことで自動評価法の研究は大きな注目を集め，現在に至っている．

BLEUは訳文と参照訳間の単語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．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，訳文が短い場合，過度にBLEUのスコアが高くなることである．そのため訳文が短い場合には最終的な評価値に式（2）のペナルティを負の重みとして用いる．最終的なスコアは式（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．通常は$N=4$が適切とされている．また，式（3）のw_nは1/Nである．BLEUスコアは0.0〜1.0の範囲であり，値が大きいほど評価が高い．

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

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

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

2.2 METEOR

METEOR（**M**etric for **E**valuation of **T**ranslation with **E**xplicit **O**rdering）[13] はBLEU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．

以下の式（4），（5），そして，（6）を用いてMETEORのスコアを得る．式（4）におけるPは適合率，Rは再現率を示している．F_{mean}は適合率と再現率を用いたF値を表している．αはパラメータである．また，式（6）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もそれぞれパラメータであり，これらの値は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されている．METEORのスコアもBLEUと同様に0.0〜1.0の範囲を示し，値が大きいほど評価が高い．

$$F_{mean} = \frac{P \times R}{\alpha \times P + (1-\alpha) \times R} \tag{4}$$

$$Pen = \gamma \times \left(\frac{ch}{m}\right)^{\beta} \tag{5}$$

$$score = (1 - Pen) \times F_{mean} \tag{6}$$

2.3 RIBES

RIBES（**R**ank−Based **I**ntutive **B**ilingual **E**valuation **S**core）[14, 15] は日本語と英語間の文法的

な違い，即ち，語順の違いに着目した自動評価法である．また，語順については訳文と参照訳の間の一致単語における順位相関係数を求めることでスコアを得る．以下の式（7）に順位相関係数の計算式を示す．

$$\gamma = \frac{\text{昇順ペア}}{\text{全ペア}} \times 2 - 1 \tag{7}$$

昇順ペアとは単語列を左から右に0,1,・・・とした際に訳文と参照訳の一致単語の並びが訳文と参照訳共に左の単語から右の単語に向かって大きくなる場合，その単語ペアは昇順ペアとなる．逆に，単語の並びが一方は大きくなり，他方が小さくなる場合，その単語ペアは降順ペアとなる．順位相関係数の対象となる単語ペアはすべて一致単語であるため，不一致単語は順位相関係数に反映されないことになる．その結果，BLEUと同様に短い訳文に対しては高いスコアを与えるという問題を引き起こすため以下の式（8）をペナルティとして利用する．

$$\tau = \frac{n}{h} \tag{8}$$

式（8）のnとhはそれぞれ一致単語数と訳文の単語数である．このペナルティを用いたRIBESのスコアは以下の式（9）より得る．

$$RIBES = \frac{\tau + 1}{2} \times P^{\alpha} \tag{9}$$

式（9）の$\alpha(0 \leq \alpha \leq 1)$は重みパラメータである．また，$P$はユニグラム適合率である．RIBESは0.0～1.0の範囲でスコアを出力し，評価が高い程1.0に近い．

2.4　IMPACT

我々は従来より訳文と参照訳の間の共通チャンクの長さと出現順に着目した自動評価法であるIMPACT（**I**ntuitive com**M**on **PA**rts **C**on**T**inum）［16,17,18,19,20,21］を提案している．IMPACTは訳文と参照訳の間の共通チャンクを再帰的に求めることにより語順の違いをスコアに反映している．以下の式（10）に共通チャンクの長さに基づくチャンクのスコアの計算式を示す．

$$Ch_score = \sum_{ch \in ch_num} length\,(ch)^{\beta} \tag{10}$$

式（10）のchは共通チャンクを，$length$（ch）は共通チャンクの構成単語数を示す．ch_numは共通チャンクの数である．そして，βは1.0以上のパラメータである．したがって，訳文と参照訳の間の一致単語が連続して出現すると共通チャンクは長くなりパラメータβがスコアに与える影響が大きくなる．IMPACTではこのCh_scoreを用いて，参照訳を再現できているかを示す再現率及び訳文との一致率を示す適合率をそれぞれ求める．再現率Rと適合率Pの計算式

を式（11）と式（12）にそれぞれ示す．

$$R = \frac{\left(\sum_{i=0}^{RN-1}(\alpha^i \times Ch_score)\right)^{\frac{1}{\beta}}}{m} \tag{11}$$

$$P = \frac{\left(\sum_{i=0}^{RN-1}(\alpha^i \times Ch_score)\right)^{\frac{1}{\beta}}}{n} \tag{12}$$

式（11）のmは参照訳の構成単語数，式（12）のnは訳文の構成単語数を示している．また，式（11）と式（12）の分子は共に語順を考慮した文全体の共通チャンクのスコアを示している．式（10）のCh_scoreは局所的な一致を示すスコアであり，文全体を大局的に捉えたスコアとはなっていない．そこで，訳文が参照訳をどの程度再現しているのかを求めるために，また，訳文と参照訳がどの程度一致しているのかを求めるために式（11）と式（12）をそれぞれ用いる．式（11）と式（12）のαは0.0～1.0のパラメータである．

IMPACTでは訳文と参照訳間の共通チャンク列を共通部分列（**L**ongest **C**ommon **S**ubsequence：LCS）［22］に基づき決定している．しかし，LCSでは仮に２つの共通チャンクの出現順が訳文と参照訳で異なる場合，片方の共通チャンクは無視されることになる．この問題を避けるためにIMPACTでは出現順の異る共通チャンクが存在する場合には，再帰的に共通チャンク列を求め，スコアに反映させている．LCSにより共通チャンク列が決定されるとその共通チャンク列を取り除いたうえで新たにLCSを用いて共通チャンク列を求める．ただし，繰り返し処理により得られる共通チャンク列についてはその回数に応じて負の重みを与える．それが式（11）と式（12）のパラメータαと回数カウンタiである．最初に決定される共通チャンク列についてはカウンタiは０であるためCh_scoreの重みは１となるが，２回目以降の再帰による共通チャンク列に対しては，繰り返し回数が増加するほど負の重みが増加する．式（11）と式（12）のRNは繰り返し処理の終了条件，即ち，共通チャンク列の数を意味する．この繰り返し処理は対象となる訳文と参照訳によって共通チャンク列が異なるため対象となる訳文と参照訳によって動的に決まる．

式（11）と式（12）より再現率と適合率が得られるとそれらを用いて調和平均であるF値を求める．そして，そのF値がIMPACTスコアとなる．計算式を以下の式（13）に示す．

$$IMPACT = \frac{(1+\gamma^2)RP}{R+\gamma^2 P} \tag{13}$$

式（13）のγは$\frac{P}{R}$より得られる．また，IMPACTスコアは0.0～1.0の範囲で出力され，1.0に近いほど評価は高い．

3　提案手法

提案手法は単語の意味と語順を考慮したEMDに基づく自動評価法となっている．始めに単

語分散表現に基いた単語アライメントを行う．次いで，その結果と語順の情報を考慮し，かつ，単語分散表現に基づくEMDより評価スコアを得る．

3.1　単語分散表現に基づく単語アライメント

提案手法では単語アライメントを行うことで対応関係にある単語と対応関係にない単語を差別化する．単語アライメントは単語分散表現に基いて行う．具体的には訳文中の個々の単語を基準として参照訳中の全ての単語との類似度を求める．類似度計算にはコサイン類似度を用いる．単語アライメントの具体例を以下の**図1**に示す．

例えば訳文である“重油/中/の/有害/物質/が/障害/の/原因/で/ある/.”，また，参照訳である“重油/中/に/含ま/れる/有害/物質/が/障害/の/原因/と/なる/.”においては，始めに訳文中の先頭単語“重油”と参照訳中の全ての単語との間のコサイン類似度を求める．そして，コサイン類似度が最も高い単語をアライメント結果とする．**図1**では，参照訳中の“重油”が類似度1.0と最も高いため訳文中の単語“重油”と参照訳の単語“重油”を対応している単語同士とする．

この処理を訳文中の単語の全単語について行う．その結果，“の”と“で”を除いた単語の対応づけが**図1**のように決定される．ここで，単語“の”については訳文中に2つ，参照訳中に1つ存在する．その場合，対応関係が一意に決定できないため単語の対応づけは行わない．また，訳文中の“で”は参照訳中の単語とのコサイン類似度において最大の類似度を持つ単語が複数存在した．その結果，対応関係を一意に決定できずに単語の対応づけは行われない．訳文“重油/中/の/有害/物質/が/障害/の/原因/で/ある/.”と参照訳“重油/中/に/含ま/れる/有害/物質/が/障害/の/原因/と/なる/.”においては，最終的には**図1**のような単語アライメントが得られる．そして，このような単語アライメントの結果を用いてEMDに基づく評価スコアを得る．

3.2　Earth Mover's Distance（EMD）

本節ではEMDの詳細を述べる．EMDは画像検索を目的に2つの分布間の類似度を求めるために提案された［8,9］．その後，言語処理においても情報検索などの分野で利用され，現在は

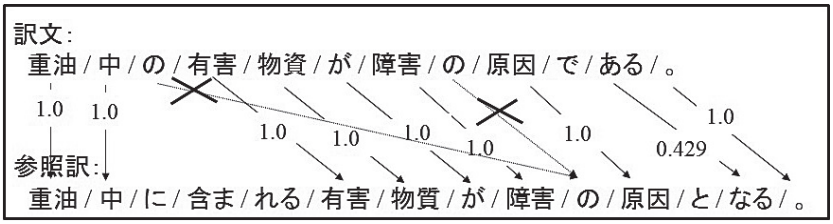

図1：単語アライメントの具体例

ニューラル機械翻訳のフィードバック強化のために利用されている．EMDはある分布から他の分布への輸送問題の最小コストに基いている．以下の式（14）に総輸送コストの計算式を示す．

$$WORK(P,Q,F)=\sum_{i=1}^{m}\sum_{j=1}^{n}d_{ij}f_{ij} \tag{14}$$

２つの分布PとQはそれぞれ$P=(p_1,w_{p_1}),\ldots,(p_m,w_{p_m})$，$Q=(q_1,w_{q_1}),\ldots,(q_n,w_{q_n})$として定義される．$p_i$と$q_j$は特徴量，$w_{p_i}$と$w_{q_j}$は個々の特徴量に対する重みを示す．そして，$d_{ij}$は$p_i$と$q_j$の間の距離である．更に，$f_{ij}$は$p_i$から$q_j$への輸送量を意味する．$F$は総輸送コストの最小値である．

また，EMDでは式（14）の総輸送コストを計算する際，以下の４つの制約条件を満たす必要がある．式（15）は供給地から重要地の一方向にしか輸送されないことを意味している．したがって，逆方向の輸送は行わない．

$$f_{ij}\geq 0,\quad 1\leq i\leq m,\quad 1\leq j\leq n \tag{15}$$

式（16）は供給地iから供給できる容量が供給量w_{p_i}を超えないことを意味している．

$$\sum_{j=1}^{n}f_{ij}\leq w_{p_i},\quad 1\leq i\leq m \tag{16}$$

式（17）は需要地jが受け取れる容量が需要量w_{q_j}以下であることを意味している．

$$\sum_{i=1}^{m}f_{ij}\leq w_{q_j},\quad 1\leq j\leq n \tag{17}$$

式（18）は供給地から移動できる最大総輸送量が供給量の総量と需要量の総量の小さい方であることを意味している．

$$\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}=min\left(\sum_{i=1}^{m}w_{p_i},\ \sum_{j=1}^{n}w_{q_j}\right) \tag{18}$$

EMDは以下の式（19）のように定義される．EMDは２つの分布PとQの間の距離が近いほど出力される値は小さくなる．

$$EMD(P,Q)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}d_{ij}f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}} \tag{19}$$

3.3　単語分散表現及び単語アライメントを用いたEMD

3.3.1　単語の重み付け

EMDを自動評価に適応するにあたり，特徴量，重み，そして，距離を決定する必要がある．提案手法では特徴量に単語分散表現，重み付けには$tf\cdot idf$を利用する．また，距離にはコ

サイン類似度を用いる．距離計算を行う際には単語アライメントの結果に基づいて決定する．本節では単語分散表現に対する重み付けとして用いるtf・idfについて述べる．tf・idfの計算式を以下の式（20）に示す．

$$tf \cdot idf = tf \times \left(\log \frac{N}{df} + 1.0 \right) \tag{20}$$

式（20）を用いることで内容語と機能語の差別化を図っている．tfは任意の文中における対象単語の出現回数である．dfは対象単語が出現する文数である．また，Nは全文数である．対象単語が機能語の場合，数多くの文に出現するため，dfが高い値となり，tf・idfは小さくなる．逆に，内容語の場合，特定の文にしか出現しないためdfは小さな値となり，tf・idfが大きくなる．その結果，内容語と機能語を区別することが可能となる．また，EMDへ適用する際には文中の全単語のtf・idfの総和が1.0になるように正規化する．これは，EMDの値を0.0～1.0のレンジに収めるためである．

3.3.2　距離計算と距離行列

提案手法では全単語を単語分散表現に変換し，２つの単語分散表現の間の距離を求める．その際，単語アライメントの結果を用いる．また，提案手法では単語分散表現における対義語の関係にある単語間の距離が近くなるという問題を解決するために，単語間の表層情報も距離計算に利用する．以下の式（21）に距離の計算式を示す．

$$d = \begin{cases} 1.0 - simlarity \times pos_info \\ (aligned\ word\ pair\ and\ wd_h = wd_r) \\ 1.0 - similarity^2 \times pos_info \\ (aligned\ word\ pair\ but\ wd_h \neq wd_r) \\ 1.0\,(non - aligne\ word\ pair) \end{cases} \tag{21}$$

$$pos_info = 1.0 - \left| \frac{pos\,(wd_h)}{len\,(h)} - \frac{pos\,(wd_r)}{len\,(r)} \right| \tag{22}$$

式（22）のpos_infoは訳文と参照訳の間の対応関係にある単語間の相対位置のずれを意味する．$pos\,(wd_h)$と$pos\,(wd_r)$はそれぞれ訳文中における対象単語の位置と参照訳中における対象単語の位置である．$len\,(h)$と$len\,(r)$はそれぞれ訳文と参照訳の構成単語数，即ち，文の長さである．pos_infoは相対位置のずれが小さいほど大きな値となり，相対位置のずれが大きほど小さな値となる．そして，このpos_infoを距離計算にて負の重みとして用いる．

式（21）のdは単語間の距離計算式であり，条件によって３通りある．単語間の類似性が高いほどdの値は0.0に近づき，類似性が低いほどdの値は1.0に近づく．また，$similarity$はコサイン類似度を示している．その範囲は0.0～1.0で類似度が高いほど1.0に近い値となる．式

（21）の最初の$1.0-simlarity \times pos_info$は単語アライメントの結果，対応関係が得られ，かつ，単語間の表層も一致している場合の距離計算である．例えば，**図1**の訳文の単語“重油”と参照訳の単語“重油”は対応関係が得られ，かつ表層レベルでも一致しているため，式（21）の最初の式が適用される．したがって，式（22）のpos_infoは0.988となるためdの値は0.002となる．

また，式（21）の中段の$1.0-simlarity^2 \times pos_info$は単語アライメントの結果，意味的な対応関係は存在しているが単語間の表層が不一致の場合の距離計算である．例えば，**図1**の訳文の単語“ある”と参照訳の単語“なる”は単語アライメントの結果より意味的な対応関係は成立しているが表層レベルでは一致していない．そのため式（21）のpos_infの値は0.012となりdの値は0.818となる．単語分散表現を用いた場合，“man”と“woman”のような対義語のコサイン類似度も高くなるという問題がある．そこで，コサイン類似度だけでなく表層情報も利用することで距離計算の精度の向上を図っている．

更に，単語アライメントの結果，対応関係が得られなかった単語間については式（21）の下段の値である$d=1.0$が適用される．これは距離としては最大の値である．式（21）では単語同士が意味的にも表層的にも類似している場合，意味のみが類似している場合，意味的にも表層的にも対応関係が存在しない場合に応じて，距離計算式を分けることで差別化を図っている．更に，提案手法では式（22）の相対位置のずれを距離計算に反映させることで語順の違いを考慮した評価スコアを得ることが可能となる．

次いで，距離計算を用いて距離行列を生成する．**表1**に**図1**の訳文と参照訳間の距離行列を示す．**表1**では単語アライメントより対応関係が得られた単語間の距離を太字で表している．この単語アライメントの結果を利用することでノイズを大幅に削減でき，より高い精度の評価

表1：訳文と参照訳間の距離行

		参照訳													
		重油	中	に	含ま	れる	有害	物質	が	障害	の	原因	と	なる	.
参照訳	重油	**0.002**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	中	1.0	**0.024**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	の	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	有害	1.0	1.0	1.0	1.0	1.0	**0.095**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	物質	1.0	1.0	1.0	1.0	1.0	1.0	**0.083**	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	が	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**0.071**	1.0	1.0	1.0	1.0	1.0	1.0
	障害	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**0.06**	1.0	1.0	1.0	1.0	1.0
	の	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**0.048**	1.0	1.0	1.0	1.0
	原因	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**0.036**	1.0	1.0	1.0
	で	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	ある	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**0.818**	1.0
	.	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**0.0**

を行うことができる．

3.3.3　評価スコアの計算

EMDは２つの分布の距離を求める手法であるため，分布が類似しているほど値は小さくなり，分布が類似していなければ値は大きくなる．しかし，自動評価法の多くは評価が高いほどスコアは大きく，評価が低いほどスコアは小さくなる．したがって，提案手法でも一般的な自動評価法と同様に評価が高いほど1.0に近く，評価が低い場合には0.0に近くなるように，以下の式（23）を用いて得られたスコアを最終的なスコアとする．

$$\text{提案手法のスコア} = 1.0 - \text{EMD} \tag{23}$$

4　性能評価実験

4.1　実験データ及び実験方法

提案手法の有効性を確認するために他の自動評価法を用いてメタ評価実験を行った．実験に用いたデータはWMT2015の自動評価タスクデータである．WMTデータは翻訳文，参照訳，そして，人手評価も公開されているため数多くの自動評価の研究者に利用されている．今回はWMT2015の自動評価タスクデータの他言語から英語の訳文及び参照訳を用いた．具体的には，フランス語から英語（fr–en），フィンランド語から英語（fi–en），ドイツ語から英語（de–en），チェコ語から英語（cs–en），そして，ロシア語から英語（ro–en）の５種類の訳文と参照訳である．メタ評価では，WMT2015の自動評価タスクに準拠してシステムレベルとセグメントレベルの２つの観点で人手評価との相関係数をそれぞれ求めた．システムレベルの相関係数はピアソンの相関係数，セグメントレベルの相関係数はケンドールの順位相関係数を用いた．

提案手法との比較に用いた自動評価法はBLEU，METEOR，RIBES，IMPACT，EMDのみに基づく手法である．EMDのみに基づく手法と比較することで単語アライメント及び語順情報の有効性を確認することができる．また，EMDのみに基づく手法と提案手法においては単語分散表現としてGoogleNews–vectors–negative300.bin（300次元，語彙数3,000,000）を用いた．これは，新聞記事を学習データとしてword2vec［23］を用いて学習されたモデルである．また，GoogleNews–vectors–negative300.binはストップワードは含まれていない．

4.2　実験結果

メタ評価実験の結果を**表２**と**表３**に示す．**表２**の（）の数値はシステム数，**表３**の（）の数値はセグメント数をそれぞれ示している．表中の“Avg.”はfr–en，fi–en，de–en，cs–en，ru–enの相関係数の平均である．また，提案手法におけるシステムレベルのスコアはセグメントレ

表 2：WMT2015データを用いたシステムレベルの相関係数

Metrics	fr–en (7)	fi–en (14)	de–en (13)	cs–en (16)	ru–en (13)	Avg.
BLEU	0.975	0.929	0.865	0.958	0.851	0.916
METEOR	0.982	0.941	0.960	0.970	0.960	0.962
RIBES	0.975	0.914	0.928	0.953	0.940	0.942
IMPACT	0.986	0.954	0.900	0.980	0.904	0.945
EMDのみ	0.985	0.918	0.920	0.964	0.971	0.952
提案手法	0.990	0.955	0.922	0.989	0.949	0.961

表 3：WMT2015データを用いたセグメントレベルの相関係数

Metrics	fr–en (29770)	fi–en (31577)	de–en (40535)	cs–en (85877)	ru–en (44539)	Avg.
sentBLEU	0.358	0.308	0.360	0.391	0.329	0.349
METEOR	0.380	0.406	0.422	0.439	0.386	0.407
RIBES	0.328	0.266	0.328	0.342	0.307	0.314
IMPACT	0.349	0.307	0.343	0.393	0.317	0.342
EMDのみ	0.365	0.349	0.394	0.411	0.344	0.373
提案手法	0.366	0.342	0.404	0.408	0.348	0.374

ベルのスコアの平均を用いている．

4.3 考察

表 2 のシステムレベルの相関係数では，平均が最も高い値を示したのはMETEORの0.962であった．しかし，提案手法の平均は0.961と大きな差はなく，提案手法はMETEORと同等な評価精度を有している．しかし，**表 3** のセグメントレベルの相関係数においてはMETEORの相関係数が0.407と最も高かった．そして，提案手法の平均はMETEORに次ぐ0.374であった．この結果より，セグメントレベルではMETEORの評価精度が多手法に比べて高いことが確認できる．

一方，EMDのみの手法と提案手法との比較においては，システムレベルでは**表 2** よりEMDのみの手法の平均が0.952であったのに対して提案手法は0.961であった．また，セグメントレベルでは**表 3** よりEMDのみの手法の平均が0.373であったのに対して提案手法では0.374であった．したがって，システムレベルにおいては単語アライメント及び語順情報の利用の有効性を確認することができた．

性能評価実験からシステムレベルにおいては提案手法の有効性を確認できたが，セグメントレベルでは提案手法の有効性は十分とはいえない．そこで，セグメントレベルの評価精度の向上を目的に，提案手法と我々が従来より提案しているIMPACTを組み合わせた“提案手法 with

IMPACT”を提案する.“提案手法 with IMPACT”は以下の式（24）に示すように提案手法のスコアとIMPACTより得られるスコアの平均を評価スコアとする.

$$\text{提案手法 with IMPACTのスコア} = \frac{\text{提案手法のスコア}+\text{IMPACT}}{2.0} \tag{24}$$

“提案手法 with IMPACT”を用いたセグメントレベルの相関係数を**表 4** に示す．また，**表 4** には参考として**表 3** と同様の提案手法の相関係数も示している.

表 4 よりIMPACTのスコアを提案手法のスコアと組み合わせることでセグメントレベルの相関係数の平均は0.374から0.395に向上した.“提案手法 with IMPACT”の相関係数はMETEORの0.407を上回るものではないが，提案手法に対し，大幅にその差を縮めることができた．したがって，提案手法とIMPACTの組み合わせ方を更に考慮することでセグメントレベルの相関係数をより向上できる可能性があると考えられる.

次いで，システムレベルにおける人手評価と提案手法の具体的なスコアについて述べる．**表 5** にfr–enの人手評価と提案手法及びEMDのみの手法のスコアを示す．表中の（）は**表 2** におけるfr–enの相関係数である.

表 5 より人手評価のスコアではシステムBが２位，システムCが３位であるのに対し，提案手法とEMDのみの手法では共にシステムBが３位，システムCが２位と逆転している．しかし，それ以外は全て人手評価と同順位である．そのため提案手法，EMDのみの手法共にfr–enの相関係数はそれぞれ0.990，0.985と高い相関係数を示した．提案手法の相関係数が高い理由は，システムレベルの相関係数にピアソンの相関係数を用いているため人手評価のスコアと絶

表 4：WMT2015データを用いた提案手法と“提案手法with IMPACT”のセグメントレベルの相関係数

Metrics	fr–en (29770)	fi–en (31577)	de–en (40535)	cs–en (85877)	ru–en (44539)	Avg.
提案手法	0.366	0.342	0.404	0.408	0.348	0.374
提案手法with IMPACT	0.391	0.365	0.414	0.440	0.363	0.395

表 5：fr–enにおけるシステムレベルの人手評価と提案手法及びEMDのみの手法のスコアの比較

MT systems	人手評価のスコア	人手評価の順位	提案手法のスコア(0.990)	提案手法の順位	EMDのみの手法のスコア(0.985)	EMDのみの手法の順位
システムA	0.498	1	0.528	1	0.717	1
システムB	0.446	2	0.523	3	0.714	3
システムC	0.415	3	0.526	2	0.715	2
システムD	0.275	4	0.519	4	0.708	4
システムE	0.223	5	0.516	5	0.707	5
システムF	−0.423	6	0.475	6	0.677	6
システムG	−0.434	7	0.443	7	0.656	7

対値の差がより小さい提案手法に効果的だったためと考えられる．

次いで提案手法におけるセグメントレベルの人手評価とEMDのみの手法及び“提案手法 with IMPACT”のスコアの具体例を**表 6** に示す．**表 6** はフランス語から英語とドイツ語から英語の翻訳を行った場合の自動評価の具体例である．フランス語から英語への翻訳の例では，機械翻訳システム 1 は“a tax system that promotes high incomes and pensioners.”と翻訳し，機械翻訳システム 2 は“a tax system that favors high earners and pensioners.”と翻訳した．参照訳は“a tax system which favors high incomes and people of independent means.”である．これらの訳文に対する人手評価はそれぞれ 3 と 1 であった．EMDのみの手法ではそれぞれ0.612と0.628を評価スコアとして付与した．また，“提案手法with IMPACT”ではそれぞれ0.442と0.441を評価スコアとして付与した.“提案手法 with IMPACT”はわずかながら機械翻訳システム 2 よりも機械翻訳システム 1 の訳文に対して高いスコアを与えている．この結果は人手評価と一致している．

しかし，その逆にEMDのみの手法による評価スコアの方が人手評価との相関が高い場合もあった．**表 6** のドイツ語から英語の翻訳においては，機械翻訳システム 1 は“its revenue was $2.57 billion in 2013, an increase of 13 percent compared to 2012 design.”，機械翻訳システム 2 は“its revenues were 2013 in 2,57 billion dollars to 2012 compared with an increase of 13%.”と翻訳した．また，参照訳は“it had revenues of $2.57 billion in 2013, up 13 percent from 2012.”である．これらの訳文に対する人手評価はそれぞれ 3 と 5 であった．EMDのみの手法による評価スコアは機械翻訳システム 1 の訳文に対しては0.587，機械翻訳システム 2 に対しては0.588と

表 6：セグメントレベルの人手評価とEMDのみの手法及び“提案手法 with IMPACT”のスコアの比較

原文（フランス語）			参照訳（英語）		
une fiscalité qui favorise les hauts revenus et les rentiers.			a tax system which favors high incomes and people of independent means.		
機械翻訳システム 1			機械翻訳システム 2		
a tax system that promotes high incomes and pensioners.			a tax system that favors high earners and pensioners.		
人手評価	EMDのみ	提案手法 with IMPACT	人手評価	EMDのみ	提案手法 with IMPACT
3	0.612	0.442	1	0.628	0.441

原文（ドイツ語）			参照訳（英語）		
seine einnahmen lagen 2013 bei 2,57 milliarden $, ein anstieg von 13 prozent im vergleich zu 2012.			it had revenues of $2.57 billion in 2013, up 13 percent from 2012.		
機械翻訳システム 1			機械翻訳システム 2		
its revenue was $2.57 billion in 2013, an increase of 13 percent compared to 2012 design.			its revenues were 2013 in 2,57 billion dollars to 2012 compared with an increase of 13%.		
人手評価	EMDのみ	提案手法 with IMPACT	人手評価	EMDのみ	提案手法 with IMPACT
3	0.587	0.415	5	0.588	0.232

なり，人手評価と同様に機械翻訳システム１の訳文より機械翻訳システム２の訳文の方が評価が高かった．それに対して，“提案手法 with IMPACT”では機械翻訳システム１の訳文に対しては0.415，機械翻訳システム２の訳文に対しては0.232となり，機械翻訳システム２の訳文より機械翻訳システム１の訳文に対して高いスコアを与えている．この結果は人手評価とは異なっており，“提案手法 with IMPACT”の評価が不十分であることを示している．

自動評価法では一般的にセグメントレベルの評価精度はシステムレベルの評価精度に比べて低く，そのことが大きな課題となっているが，今回のメタ評価実験においても**表３**に示すようにどの自動評価法でもセグメントレベルの相関係数は低い結果となった．したがって，セグメントレベルの評価精度を向上させるためのより一層の取り組みが必要である．

５　まとめ

本報告では，WMT2015の自動評価タスクを用いて，提案手法の有効性について述べた．提案手法では単語分散表現と単語アライメントの結果をEMDに反映させたうえで自動評価を行う．その結果，システムレベルの相関係数はMETEORとほぼ同等の値を得ることができた．更に，セグメントレベルの相関係数を向上させるために，我々が従来より提案している自動評価法IMPACTを提案手法と組み合わせた自動評価法を提案した．その結果，セグメントレベルの相関係数を大幅に向上させることができた．しかし，セグメントレベルの相関係数は十分とはいえず，今後の課題である．

今後はセグメントレベルの相関係数の向上のために単語アライメントの精度を向上させる予定である．また，語順の違いを単語間の距離計算に利用する際により適切な方法を検討する予定である．更に，様々なデータを用いたメタ評価を行う予定である．

謝辞

本研究は，平成29年度北海学園学術研究助成金（一般研究）の助成を受けたものである．

REFERENCES

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

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

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

［４］Ilya Sutskever, Oriol Vinyals and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *Neural Information Processing Systems.*

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

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

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

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

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

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

[11] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Phillip Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. Proceedings of the Tenth Workshop on Statistical Machine Translation. pp.1−46.

[12] Miloš Stanojević, Amir Kamran, Philipp Koehn and Ondřej Bojar. 2015. Results of the WMT 15 Metrics Shared Task. Proceedings of the Tenth Workshop on Statistical Machine Translation. pp.256−273.

[13] 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.

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

[15] 平尾努，磯崎秀樹，須藤克仁，Kevin Duh，塚田元，永田昌明．2014．語順の相関に基づく機械翻訳の自動評価法．自然言語処理 *Vol.21*，*No.3*．pp.421−444.

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

[17] Hiroshi Echizen−ya, Terumasa Ehara, Sayori Shimohata, Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro and Noriko Kando. 2009. Meta−Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR−7. Proceedings of the3rd Workshop on Patent Translation pp.9−16.

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

[19] Hiroshi Echizen'ya, Kenji Araki and Eduard Hovy. Optimization for Efficient Determination of Chunk in Automatic Evaluation for Machine Translation. Proceedings of the1th International Workshop on Optimization Techniques for Human Language Technology. pp.17−30.

[20] Hiroshi Echizen'ya, Kenji Araki and Eduard Hovy. 2013. Automatic Evaluation Metric for Machine Translation that is Independent of Sentence Length. Proceedings of the 9th Recent Advances in Natural Language Process-

ing. pp.230–236.

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

[22] A. Apostolico and C. Guerra. 1987. The Longest Common Subsequence Problem Revisited. *Algorithmica, Volume2, issue 1–4.* pp.315–336, Springer.

[23] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, 2013. Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at International Conference on Learning Representaions.