タイトル	語学学習支援システムの利用に向けた単語の意味に基づく自動評価法とWMT17の評価タスクを用いたメタ評価
著者	越前谷，博；ECHIZEN’YA，Hiroshi；歌代，崇史；UTASHIRO，Takafumi；田中，洋也；TANAKA，Hiroya；鈴木，聡士；SUZUKI，Soushi；内田，ゆず；UCHIDA，Yuzu；長谷川，大；HASEGAWA，Dai
引用	北海学園大学工学部研究報告(48)：41-51
発行日	2021-01-15

語学学習支援システムの利用に向けた単語の意味に基づく自動評価法とWMT17の評価タスクを用いたメタ評価

越前谷　博*・歌代崇史**・田中洋也***・
鈴木聡士*・内田ゆず*・長谷川大*

Automatic Evaluation based on Word Meaning toward Language Learning Support System and Meta–evaluation using WMT17 Metrics Task

Hiroshi Echizen'ya*, Takafumi Utashiro**, Hiroya Tanaka***,
Soushi Suzuki*, Yuzu Uchida* and Dai Hasegawa*

要　旨

対象言語が日常的に使用されていない環境下での語学学習を支援するための語学学習支援システムの構築を目的とした場合，学習者の語学能力を自動的に評価する処理は重要である．そこで，本報告では語学学習支援システムにおける自動評価の実現に向けた第一段階として機械翻訳における自動評価法の評価精度について述べる．機械翻訳の自動評価では，訳文を評価する際に正解訳を用いてスコアを算出する．本報告では著者が提案した自動評価法WE_WPIについても取り上げる．WE_WPIでは単語の意味を考慮した評価を行うためにニューラルネットにより構築された単語分散表現を使用する．また，訳文を定量的に評価するためにEarth Mover's Distance（EMD）を利用することでスコア化する．本報告ではWMT17における評価タスクで使用されたデータによるメタ評価に基づいて自動評価法の性能評価を行なった．メタ評価の結果，WE_WPIは様々な自動評価法において人手評価との間で高い相関係数を示した.したがって，語学学習支援システムへの適用に向け，大きな期待を抱かせる結果となった．

1　はじめに

語学学習の支援を対象とした語学学習支援システムを実現することは母語以外の言語を学習するために非常に有効と考えられる．本報告における語学学習とは，日本人が日本の学校で英語を学習する場合のように対象言語が日常的に使用されていない環境での学習で

* 北海学園大学工学部
* Faculty of Engineering, Hokkai–Gakuen University
** 北海学園大学経済学部
** Faculty of Economics, Hokkai–Gakuen University
*** 北海学園大学人文学部
*** Faculty of Humanities, Hokkai–Gakuen University

ある［1］．そして，語学学習支援システムは学習者の言語能力を高めることが最大の目的である．さらに語学学習における支援には学習者支援と教師支援に分けられる．学習者支援は学習者に対して文法誤りなどの訂正を行うことが目的となる．その際，語学学習支援システムが単に誤りを自動的に修正してしまうとユーザである学習者に考えさせる機会を奪ってしまうことになり不十分である．したがって，学習者にある程度の認知的負荷を与えることが重要となる．また，教師支援は教材作成の支援，学習者の能力や特徴を分析することが目的となる．特に言語能力評価は様々な評価観点があることから高度に専門的な知識が求められる難しいタスクであるが，能力評価の目的を設定することで間接的に言語能力を測定するための尺度が利用されている．

本報告では教師支援の一つである言語能力の評価に着目する．語学学習支援システムにおける言語能力の自動評価には総合的評価と分析的評価の２種類がある．例えばエッセイの自動採点では総合的評価は一つのエッセイに対して全体スコアを一つ与える．そして，エッセイ自動採点システムは学習者が書いたエッセイに対して，自動的に評価値を付与する．分析的評価では，文法，語彙，内容など複数の観点よりスコアを与える．エッセイの自動採点においては総合的評価が使用されることが多い．エッセイ自動採点システムそのものの評価は，システムと人間のスコアの相関係数を求めることで行われる．多くのシステムの相関係数は0.80から0.85であると報告されている［2，3］．

このような状況において，本報告では語学学習支援システムの構築のための第一段階として機械翻訳の自動評価法とそのメタ評価について述べる．自動評価法は機械翻訳システムが生成する訳文を対象に，一つのドキュメントに対してスコアを与えるドキュメント単位での評価と一つの文に対してスコアを与える文単位での評価を行う．その際には著者が提案している自動評価法WE_WPI（**W**ord **E**mbedding – based Automatic MT Evaluation using **W**ord **P**osition **I**nformation）［4］も含める．WE_WPIは単語の意味と語順に着目した自動評価法である．単語の意味を表現するためにニューラルネットにより得られる単語分散表現モデルを用いて，単語の分散表現を得る．また，スコアは訳文と正解文（以降，本報告では参照訳と呼ぶ）を比較することで得るが，その際には単語の分散表現に基づくEarth Mover’s Distance（EMD）［5，6，7］を用いる．また，EMDを用いる場合，文中の語順が反映されないため単語の位置情報をEMDに反映させている．性能評価はWMT17（The Second Conference on Machine Translation）［8，9］の評価タスクによるメタ評価により行う．メタ評価の結果，ドキュメント単位及び文単位のいずれにおいても他手法の中でWE_WPIは上位に位置し，語学学習支援システムへ適用可能性が高いことを確認した．

2　機械翻訳研究における自動評価法

機械翻訳における自動評価法の位置づけに

ついて簡単に述べる．機械翻訳研究におけるブレイクスルーは1990年代以降２度起こったと言える．１度目は1990年代から2000年代における統計的機械翻訳（SMT）[10,11,12]である．この時期は自然言語処理分野全体において機械学習に基づくアプローチが確固たる地位を築き上げていた時期でもあり，機械翻訳研究においてもそうした流れに沿った統計的機械翻訳がブレイクスルーとなった．統計的機械翻訳がブレイクスルーとなり得た大きな要因の一つにオープンソースとして無償で比較的容易にインストール可能であったことが挙げられる．そのことにより，多くの機械翻訳の研究者が統計的機械翻訳を用いた様々な研究を行い，その精度を競い合うこととなった．しかし，改良前の機械翻訳（ベースライン）およびオリジナルのアイデアを盛り込んだ機械翻訳（提案手法）の精度を容易に比較できなければ研究を円滑に進めることは困難である．そのため統計的機械翻訳研究の進展は翻訳精度を迅速に得ることの必要性を増幅させた．そのニーズを満たす役割を果たした技術が自動評価法である．具体的には2002年に提案されたBLEU［13］の登場が自動評価法の存在価値を決定的なものとした．そして，このBLEUに追随してNISTやMETEOR［14］など様々な自動評価法が提案された．著者は2007年にそれまでの自動評価法の問題点を解決する新たな自動評価法としてIMPACT（**I**ntuitive com**M**on **PA**rts **C**on**T**inuum）[15,16,17,18,19,20］を提案した．

自動評価法の重要性は2006年より毎年開催されている機械翻訳に関する様々なタスクを対象としたコンテスト型ワークショップWMT（Workshop on Statistical Machine Translation，2016年よりConference on Machine Translationと名称変更され国際会議として開催されている）において，2008年より評価タスクとして加わっていることからも広く認知されている．

2010年代に入ると統計的機械翻訳に取って代わり，ニューラル機械翻訳（NMT）[21,22,23,24］がブレイクスルーとなり，機械翻訳研究に大きな変革をもたらした．それに伴い，自動評価法においても従来の統計的機械翻訳の評価を目的としたものではなく，ニューラル機械翻訳の評価を目的とした研究が行われるようになった．著者もニューラル機械翻訳に対応するための新たな自動評価法としてWE_WPIを提案している．このように自動評価法の研究は機械翻訳研究を支える技術として共に発展してきた．

2.1 IMPACT

著者はトレーニング及び言語知識を必要とせずに表層情報のみに基づく自動評価法としてIMPACTを提案している．IMPACTは共通部分列（**L**ongest **C**ommon **S**ubsequence：LCS）[25］により訳文と参照訳の間の共通単語列を取得することでスコアを得る．ここでLCSは共通単語列が左右でクロスして出現する場合には，基本的には短い共通単語列の方は完全に無視されてしまうため，語順に極めて厳しいという特徴がある．そこで，IMPACTでは共通単語列の出現位置に関わらず全ての共通単語列をスコアに反映させてい

表1：WMT17の評価タスクデータの多言語から英語方向への訳文を用いたドキュメント単位のメタ評価

	cs–en	de–en	fi–en	lv–en	ru–en	tr–en	zh–en	
n	4	11	6	9	9	10	16	Avg.
Correlation	|*r*|	|*r*|	|*r*|	|*r*|	|*r*|	|*r*|	|*r*|	
AUTODA	0.438	0.959	0.925	0.973	0.907	0.916	0.734	0.836
BEER	0.972	0.960	0.955	0.978	0.936	0.972	0.902	0.954
BLEND	0.968	0.976	0.958	0.979	**0.964**	0.984	0.894	**0.960**
BLEU	0.971	0.923	0.903	0.979	0.912	0.976	0.864	0.933
BLEU2VEC_SEP	0.989	0.936	0.888	0.966	0.907	0.961	0.886	0.933
CDER	0.989	0.930	0.927	0.985	0.922	0.973	0.904	0.947
CHARACTER	0.972	0.974	0.946	0.932	0.958	0.949	0.799	0.933
CHRF	0.939	**0.968**	0.938	0.968	0.952	0.944	0.859	0.938
CHRF++	0.940	0.965	0.927	0.973	0.945	0.960	0.880	0.941
MEANT_2.0	0.926	0.950	0.941	0.970	0.962	0.932	0.838	0.931
MEANT_2.0–NOSRL	0.902	0.936	0.933	0.963	0.960	0.896	0.800	0.913
NGRAM2VEC	0.984	0.935	0.890	0.963	0.907	0.955	0.880	0.931
NIST	1.000	0.931	0.931	0.960	0.912	0.971	0.849	0.936
PER	0.968	0.951	0.896	0.962	0.911	0.932	0.877	0.928
TER	0.989	0.906	0.952	0.971	0.912	0.954	0.847	0.933
TREEAGGREG	0.983	0.920	**0.977**	0.986	0.918	0.987	0.861	0.947
UHH_TSKM	0.996	0.937	0.921	**0.990**	0.914	0.987	0.902	0.950
WER	0.987	0.896	0.948	0.969	0.907	0.925	0.839	0.924
WE_WPI_fastText	0.998	0.965	0.953	0.968	0.945	0.984	**0.908**	**0.960**
IMPACT	**0.999**	0.928	0.934	0.986	0.911	**0.991**	0.899	0.950

る．具体的にはクロスして出現する共通単語列の長い方を優先して，短い方の共通単語列についてはスコアへの影響をパラメータにより小さくなるように制御したうえでスコアに反映させる．その結果，語順に柔軟に対処可能となり，全ての共通単語列をスコアに反映させることが可能となる．

以下の式(1)に個々の共通単語列*ch* のスコアの計算式を示す．式(1)では共通単語列の構成単語数が大きいほどスコアが高くなるようにパラメータβを用いて制御している．βの値は1.0以上である．*length* (*ch*)は各共通単語列*ch* の構成単語数を示し，*ch _num* は文間における共通単語列の数を示している．

$$Ch\_score = \sum_{ch \in ch\_num} length\ (ch\ )^{\beta} \qquad (1)$$

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*R* 及び訳文との一致率を示す適合率*P* を求める．以下の式(2)と式(3)はそれぞれ*R* と*P* を求める計算式である．式(2)と式(3)のαは1.0以下であり，共通単語列のスコアへの影響を制御するためのパラメータである．カウンタ*i* が大きくなるほど共通単語列に対して負の重みとなり*Ch _score* は小さな値となる．*RN* はLCSによる共通部分列の決定処理の処理回数を示している．この値が大きいほどクロスして出現する共通単語列が多いことを意味する．この式(2)と式(3)は全ての共通単語列を用いた式となっている．また，式(2)の*m* は参照訳の構成単語数，式(3)の*n* は訳文の構成単語数を示す．

表 2：WMT17の評価タスクデータの英語から多言語方向への訳文を用いたドキュメント単位のメタ評価

	en−cs	en−de	en−fi	en−lv	en−ru	en−tr	en−zh	
n	14	16	12	17	9	8	11	Avg.
Correlation	|*r*|	|*r*|	|*r*|	|*r*|	|*r*|	|*r*|	|*r*|	
AUTODA	0.975	0.603	0.879	0.729	0.850	0.601	0.976	0.802
AUTODA−TECTO	0.969	−	−	−	−	−	−	−
BEER	0.970	0.842	0.976	0.930	0.944	0.980	0.914	0.937
BLEND	−	−	−	−	0.953	−	−	−
BLEU	0.956	0.804	0.920	0.866	0.898	0.924	0.981	0.907
BLEU2VEC_SEP	0.963	0.810	0.942	0.859	0.903	0.911	−	−
CDER	0.968	0.813	0.965	0.930	0.924	0.957	**0.983**	0.934
CHARACTER	**0.981**	**0.938**	0.972	0.897	0.939	0.975	0.933	0.948
CHRF	0.976	0.863	**0.981**	0.955	0.950	0.991	0.976	**0.956**
CHRF+	0.976	0.855	0.980	0.956	0.948	0.988	−	−
CHRF++	0.974	0.852	0.979	0.956	0.945	0.986	0.976	0.953
MEANT_2.0	−	0.858	−	−	−	−	0.956	−
MEANT_2.0−NOSRL	0.976	0.770	0.972	0.959	**0.957**	0.991	0.943	0.938
NGRAM2VEC	−	−	0.940	0.862	−	−	−	−
NIST	0.962	0.769	0.957	0.935	0.920	0.986	0.976	0.929
PER	0.954	0.687	0.949	0.851	0.887	0.963	0.934	0.889
TER	0.955	0.796	0.961	0.909	0.933	0.967	0.970	0.927
TREEAGGREG	0.947	0.773	0.965	0.927	0.921	0.983	0.938	0.922
WER	0.954	0.802	0.960	0.906	0.934	0.956	0.954	0.924
WE_WPI_fastText	0.969	0.844	0.980	**0.966**	0.956	**0.996**	0.956	0.952
IMPACT	0.945	0.800	0.966	0.937	0.927	0.973	0.984	0.933

$$R = \frac{\left(\sum_{i=0}^{RN-1}(\alpha^i \times Ch\_score)\right)^{\frac{1}{\beta}}}{m} \quad (2)$$

$$P = \frac{\left(\sum_{i=0}^{RN-1}(\alpha^i \times Ch\_score)\right)^{\frac{1}{\beta}}}{n} \quad (3)$$

そして，式(2)と式(3)より得られるRとPの調和平均を求めることで，最終的なスコアを得る．その計算式を以下の式(4)に示す．

$$IMPACT = \frac{(1+\gamma^2)RP}{R+\gamma^2 P} \quad (4)$$

式(4)のγは$\frac{P}{R}$より得られる．また，IMPACTスコアは0.0から1.0の範囲で出力され，1.0に近いほど評価は高くなる．

3　WE_WPI

WE_WPIでは単語アライメントとスコア計算の２つのステップよりスコアを得る．単語アライメントは単語の分散表現に基づき訳文と参照訳との間で対応関係にある単語ペアを決定する処理である．スコア計算はその結果を単語の語順情報として利用することでEMDに基づきスコアを計算する処理である．

また，EMDを自動評価法に適用する際には，３つのパラメータを定義する必要がある．EMDは輸送問題の最適解を求めるアルゴリズムであり，２つの分布間の距離を計算する．それぞれの分布は複数の特徴量から構成されており，WE_WPIでは２つの分布を構成する特徴量を訳文と参照訳の構成単語の分散表現に対応させている．また，個々の特徴量は輸送問題の観点から荷物に相当する重み

表3：WMT17の評価タスクデータの多言語から英語方向への訳文を用いた文単位のメタ評価

Human Evaluation	cs–en DA	de–en DA	fi–en DA	lv–en DA	ru–en DA	tr–en DA	zh–en DA	Avg.
n	560	560	560	560	560	560	560	
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	
AUTODA	0.499	0.543	0.673	0.533	0.584	0.625	0.583	0.577
BEER	0.511	0.530	0.681	0.515	0.577	0.600	0.582	0.571
BLEND	**0.594**	**0.571**	**0.733**	0.577	**0.622**	**0.671**	**0.661**	**0.633**
BLEU2VEC_SEP	0.439	0.429	0.590	0.386	0.489	0.529	0.526	0.484
CHRF	0.514	0.531	0.671	0.525	0.599	0.607	0.591	0.577
CHRF++	0.523	0.534	0.678	0.520	0.588	0.614	0.593	0.579
MEANT_2.0	0.578	0.565	0.687	**0.586**	0.607	0.596	0.639	0.608
MEANT_2.0–NOSRL	0.566	0.564	0.682	0.573	0.591	0.582	0.630	0.598
NGRAM2VEC	0.436	0.435	0.582	0.383	0.490	0.538	0.520	0.483
SENTBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.481
TREEAGGREG	0.486	0.526	0.638	0.446	0.555	0.571	0.535	0.537
UHH_TSKM	0.507	0.479	0.600	0.394	0.465	0.478	0.477	0.486
WE_WPI_fastText	0.547	0.493	0.692	0.550	0.564	0.602	0.589	0.577
WE_WPI_fastText_BERT	0.533	0.514	0.710	0.539	0.575	0.632	0.623	0.589
IMPACT	0.494	0.489	0.630	0.469	0.505	0.581	0.570	0.534

を有している．EMDではこの重みを他方の分布の特徴量に分配するために分割され，それが輸送量となる．作業量は輸送量と特徴量間の距離の積として定義される．EMDは作業量全体を最小化するためのアルゴリズムである．WE_WPIでは重みには文レベルの*tf*・*idf*，距離計算にはコサイン距離を用いている．さらに語順の違いを反映させるために距離計算には単語の出現位置の相対的なズレを用いている．

以下の式(5)に重みを得るための*tf*・*idf*の計算式を示す．ここで*tf*は文中の任意の単語の出現頻度を示す．また，*df*は任意の単語が出現する文の数を示す．*N*は全ての文数である．式(5)の*tf*・*idf*を用いることで内容語と機能語を差別化する．

$$tf \cdot idf = tf \times \left(\log \frac{N}{df} + 1.0 \right) \tag{5}$$

以下の式(6)に単語間の距離計算式を示す．*cos_sim*はコサイン距離，$pos_in\,f(T_i, R_j)$は訳文の単語T_iと参照訳の単語R_jの出現位置の相対的なズレを示しており，それは式(7)より求める．対応関係にない単語ペアの値は距離の最大値である1.0とし，対応関係にある単語ペアの場合は意味的に近くかつ出現位置の相対的なズレが少ないほど値は小さくなる．

$$d = \begin{cases} 1.0 - cos\_sim \times e^{-pos\_inf(T_i, R_j)} \\ (if\ T_i\ corresponds\ to\ R_j) \\ 1.0 (if\ T_i\ does\ not\ correspond\ to\ R_j) \end{cases} \tag{6}$$

$$pos\_in\,f(T_i, R_j) = \left| \frac{pos(T_i)}{m} - \frac{pos(R_j)}{n} \right| \tag{7}$$

ここでEMDは距離計算であり，距離が近いほど値は小さくなる．そのため自動評価法のスコアとしてEMDの値をそのまま用いると評価が高いほど値は小さくなり，反比例の

表 4：WMT17の評価タスクデータの英語から多言語方向への訳文を用いた文単位のメタ評価

	en–cs	en–de	en–fi	en–lv	en–ru	en–tr	en–zh	
Human Evaluation	DARR	DARR	DARR	DARR	DA	DARR	DA	Avg.
n	32,810	3,227	3,270	3,456	560	247	560	
Correlation	τ	τ	τ	τ	$\|r\|$	τ	$\|r\|$	
AUTODA	0.041	0.099	0.204	0.130	0.511	0.409	0.609	0.286
AUTODA–TECTO	0.336	–	–	–	–	–	–	–
BEER	0.398	0.336	**0.557**	0.420	0.569	0.490	0.622	0.485
BLEND	–	–	–	–	0.578	–	–	–
BLEU2VEC_SEP	0.305	0.313	0.503	0.315	0.472	0.425	–	–
CHRF	0.367	0.336	0.503	0.420	0.605	0.466	0.608	0.472
CHRF+	0.377	0.325	0.514	0.421	0.609	0.474	–	–
CHRF++	0.368	0.328	0.484	0.417	0.604	0.466	0.602	0.467
MEANT_2.0	–	0.350	–	–	–	–	0.727	–
MEANT_2.0–NOSRL	0.395	0.324	0.565	0.425	**0.636**	0.482	0.705	0.505
NGRAM2VEC	–	–	0.486	0.317	–	–	–	–
SENTBLEU	0.274	0.269	0.446	0.259	0.468	0.377	0.642	0.391
TREEAGGREG	0.361	0.305	0.509	0.383	0.535	0.441	0.566	0.443
WE_WPI_fastText	0.399	0.322	0.552	0.407	0.524	0.490	0.754	0.493
WE_WPI_fastText_BERT	**0.404**	**0.359**	0.552	**0.432**	0.546	**0.498**	**0.766**	**0.508**
IMPACT	0.292	0.264	0.473	0.304	0.500	0.433	0.719	0.426

関係となる．そこで，以下の式(8)を用いることで値が高いほど評価も高くなるように変換する．

$$WE\_WPI = 1.0 - EMD \tag{8}$$

4 性能評価実験

4.1 実験データ及び実験方法

実験データにはWMT17の評価タスクで使用された訳文，参照訳，そして，人手評価を用いた．WMT17では言語ペアとしては英語とチェコ語，英語とドイツ語，英語とフィンランド語，英語とラトビア語，英語とロシア語，英語とトルコ語，そして，英語と中国語の7つが使用されている．したがって，これらの言語ペアにおいて双方向で翻訳した訳文と参照訳を用いて自動評価法はスコアを算出する．自動評価法が出力するスコアはドキュメント単位と文単位の2種類である．そして，メタ評価は自動評価法より得られたスコアと人手評価値との間でドキュメント単位と文単位それぞれ相関係数を求めることで行う．また，データはニュースに関する内容となっている．メタ評価の対象となる自動評価法にはWMT17で示されている自動評価法に加え，著者が提案している自動評価法IMPACTとWE_WPIも用いた．WE_WPI_fastTextは単語の分散表現を取得するためのモデルとしてfastText［26］モデルを用いている．

4.2 実験結果

表 1 から**表 4** にメタ評価の結果を示す．**表 1** は多言語から英語方向への訳文を用いたドキュメント単位のメタ評価の結果，**表 2** は英

語から多言語方向への訳文を用いたドキュメント単位のメタ評価の結果，**表 3** は多言語から英語方向への訳文を用いた文単位のメタ評価の結果，そして，**表 4** は英語から多言語方向への訳文を用いた文単位のメタ評価の結果である．表中のcsはチェコ語，deはドイツ語，fiはフィンランド語，lvはラトビア語，ruはロシア語，trはトルコ語，zhは中国語，そして，enは英語を示す．“Avg.”は 7 つの言語ペアの相関係数の平均を示す．太字は自動評価法の中で最も高い相関係数であったことを示す．**表 1** と**表 2** の n はドキュメント数，**表 3** と**表 4** の n は文数を示す．また，$|r|$ はピアソンの相関係数の絶対値，τ はケンドール τ を示す．**表 3** と**表 4** の“Human Evaluation”における“DA”は人手評価として絶対評価，また，“DARR”は絶対評価を相対評価に変換した値を人手評価として用いたことを示す．

4.3 考察

表 1 より多言語から英語方向への訳文を用いたドキュメント単位ではWE_WPI_fastTextとBLEND［27］の“Avg.”が最も高かった．ドキュメント単位においてはいずれの自動評価法も高い相関係数を示しており，信頼性は高いと考えられる．**表 2** より英語から多言語方向への訳文を用いたドキュメント単位ではCHRFの“Avg.”が0.956と最も高かった．しかし，WE_WPI_fastTextの“Avg.”は0.952であり，CHRFに次いで高かった．したがって，WE_WPI_fastTextはドキュメント単位において安定して高い相関係数を示すことが確認された．

表 3 と**表 4** の文単位のメタ評価を行うにあたり，自動評価法WE_WPIに関してはWE_WPI_fastTextとWE_WPI_fastText_BERTの 2 種類を用いた．文単位の相関係数はドキュメント単位の相関係数に比べて低く，評価精度は十分とは言えない．そこで，本報告では文単位の評価精度の向上を目的にWE_WPI_fastText_BERTを新たに加えた．WE_WPI_fastText_BERTは単語の分散表現を得る際に用いるモデルとしてfastTextのモデルだけではなく，BERT［28］のモデルも用いている．BERTは文脈情報を反映させた単語の分散表現を得ることが可能である．WE_WPI_fastText_BERTでは具体的にはfastTextモデルより得られる単語の分散表現を用いたコサイン距離とBERTモデルより得られる単語の分散表現を用いたコサイン距離の積を式(6)の cos_sim とした．fastTextモデルもBERTモデルも共に事前学習されたモデルを使用しているため容易に用いることが可能である．

表 3 より多言語から英語方向への訳文を用いた文単位ではBLENDが最も高い相関係数を示した．BLENDは既存の複数の自動評価法によるアンサンブル学習により訳文の評価を行う手法である．有効な自動評価法の組み合わせを形式化するためにSVM回帰（SVR）を用いて学習を行う．また，その際には人手評価として絶対評価（DA）を使用している．一般的にトレーニングベースの自動評価法はノントレーニングベースの自動評価法に比べ，高い相関係数が得られる．しかし，トレーニングベースの自動評価法は学習処理を

必要とするため利便性の観点ではノントレーニングベースの自動評価法に対して劣ると考えられる．

表 4 より英語から多言語方向への訳文を用いた文単位ではWE_WPI_fastText_BERTが最も高い相関係数を示した．“en–cs”，“en–de”，“en–lv”，“en–tr”，そして，“Avg.”において最も高い値を示した．**表 3** と**表 4** の文単位のメタ評価の結果より，WE_WPI_fastText_BERTはWE_WPI_fastTextに比べ高い評価精度を示した．これはBERTモデルによる単語の分散表現の使用が有効であったことを示している．

表 1 から**表 4** の全てのメタ評価結果においてIMPACTよりもWE_WPIの評価精度は高かった．IMPACTは表層情報に基づく自動評価法であり，語形変化や同義語に追随することができないという問題点がある．それに対して，WE_WPIは単語の意味に相当する分散表現を使用しているため，語形変化や同義語に追随することが可能である．その結果，特に**表 3** と**表 4** の文単位のメタ評価結果において，単語の意味を考慮したWE_WPIは表層情報のみに基づくIMPACTよりも高い評価精度が得られたと考えられる．

5　まとめ

本報告では，語学学習支援システムにおける言語能力の自動評価に着目し，機械翻訳における自動評価法を取り上げ，その評価精度について述べた．その際には，WMT17の評価タスクを用いて行なったメタ評価の結果に基づき述べた．メタ評価の結果，表層情報に基づく自動評価法よりも単語の意味に基づく自動評価法が高い評価精度を示すことを確認した．さらに，単語の意味に相当する分散表現を用いる際には，複数のモデルから得た単語の分散表現を利用することで評価精度を向上できることを確認した．

今後は自動評価法WE_WPIの評価精度を向上させるための検討及び改良を行う．さらに語学学習支援システムにおいて自動評価法を利用可能にするための研究を進める予定である．

謝辞

本研究は，令和元年度学術研究助成費（総合研究）の助成を受けたものである．

References

［1］奥村学監修，永田亮著．2017．語学学習支援のための言語処理（自然言語処理シリーズ11）．コロン社．

［2］Keith, T.Z. 2003. Validity of Automated Essay Scoring Systems. *Routledge*. pp.147–167.

［3］石岡恒憲．2004．記述式テストにおける自動採点システムの最新動向．**行動計量学**，*Vol.31, No.2*. pp.67–87.

［4］Hiroshi Echizen’ya, Kenji Araki, Eduard Hovy. 2019. Word Embedding – Based Automatic MT Evaluation Metric using Word Position Information. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp.1874–1883.

［5］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.

［6］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.

[7] 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.

[8] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Phillip Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). Proceedings of the Conference on Machine Translation (WMT). Volume 2: Shared Task Papers. pp.169−214.

[9] Ondřej Bojar, Yvette Graham and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. Proceedings of the Conference on Machine Translation (WMT). Volume 2: Shared Task Papers. pp.489−513.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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 the 3rd Workshop on Patent Translation pp.9−16.

[17] 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.

[18] Hiroshi Echizen'ya, Kenji Araki and Eduard Hovy. 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.

[19] 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 Processing. pp.230−236.

[20] 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.

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

[22] Minh−Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals and Wojciech Zaremba. 2015. Addressing the Rare Word Problem in Neural Ma-

chine 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.

[23] 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.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. *Attention Is All You Need*. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). pp.6000–6010.

[25] A. Apostolico and C. Guerra. 1987. The Longest Common Subsequence Problem Revisited. *Algorithmica, Volume 2, issue 1 – 4.* pp. 315 – 336, Springer.

[26] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics, Volume 5.* pp.135–146.

[27] Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a Novel Combined MT Metric Based on Direct Assessment – CASICT–DCU submission to WMT17 Metrics Task. *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers.* pp.598–603.

[28] Jacob Devlin, Ming – Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* pp.4171–4186.