単語分散表現と文脈ベクトルの組み合わせによるEarth Mover's Distanceに基づく自動評価

Automatic evaluation based on earth mover's distance by combination of word embeddings and contextual representations

北海学園大学大学院工学研究科教授

越前谷　博

1996年北海学園大学大学院工学研究科修士課程修了。博士（工学）。2013年～現在北海学園大学大学院工学研究科教授。機械翻訳の研究に従事。アジア太平洋機械翻訳協会（AAMT）／Japio特許翻訳研究会委員。

echi@lst.hokkai-s-u.ac.jp　TEL 011-841-1161（内線：7863）

1 はじめに

機械翻訳システムのための自動評価法はBLEU[1]を発端として様々な進化を遂げてきた。今もなお広く利用されているBLEUを含めTER[2]などの初期の自動評価法ではMT訳と参照訳の間の表層情報のみを用いているため単語や文の意味的な類似性を評価スコアに反映することは困難である。

その後、ニューラル翻訳が機械翻訳システムにおけるアーキテクチャの主流になると自動評価法においてもニューラルネットワークの技術を用いた研究が盛んに行われるようになった。それらはモデルの種類や利用の観点よりいくつかに分類される。まず、word2vec[3]やfastText[4]などの事前学習済みの単語分散表現モデルを利用した手法[5][6]では単語分散表現モデルを利用することにより単語の意味をとらえた評価が可能となった。また、初期のWMT（Conference on Machine Translation）のデータを用いてファインチューニングしたモデルを利用する手法[7][8]も提案されている。これらは推定モデルを初期のWMTデータを用いてファインチューニングすることで、よりタスクに適応した自動評価法となっている。さらには、ニューラルネットワークモデルをゼロから構築することで対象ドメインに特化した文脈ベクトルを利用する手法[9][10]もある。これらの手法もWMTデータを用いてニューラルネットワークモデルを構築するため、タスクに対する適応性を向上させることができる。しかし、モデルをファインチューニングおよび新たに構築する手法では大規模な学習データが必要となり、評価精度は学習データの量に強く依存してしまう。

そこで、著者らは単語分散表現モデルによる単語ベクトルに加え、MT訳に対応する原文と参照訳のペアのみを学習データとして構築したニューラルネットワークモデルより得られる文脈ベクトルを利用した手法WE_WPI-Attention[11]を提案した。本稿ではこのWE_WPI-Attentionを発展させた新たな自動評価法を提案し、その有効性をメタ評価の結果に基づいて述べる。提案手法では、WE_WPI-Attentionが単語分散表現のみを用いて得られる評価スコアと文脈ベクトルのみを用いて得られる評価スコアをそれぞれ求めたうえで、それらの加重平均を最終的なスコアとしていた。それに対して、提案手法は単語分散表現モデルによる単語ベクトルとニューラルネットワークモデルによる文脈ベクトルを組み合わせた新たなベクトルを最適輸送アルゴリズムであるEarth Mover's Distance（EMD）[12]の特徴量として用いる。その結果、より直接的にベクトルの持つ言語情報を評価スコアに反映させることができ、人手評価との相関の向上が期待できる。

2 提案手法

提案手法は事前学習された単語分散表現モデルからは単語ベクトル、新たに構築されたニューラルネットワークモデルからは文脈ベクトルを取得し、それらを組み合わせたベクトルをEMDの特徴量として用い、評価スコアを得る。

2.1　ニューラルネットワークモデルの学習

ニューラルネットワークモデルの構築について述べる。本稿ではニューラルネットワークモデルの構築はアテンションベースのLSTM [13] を用いて行う。図1にアテンションベースのLSTMを用いたモデルの学習の概要図を示す。

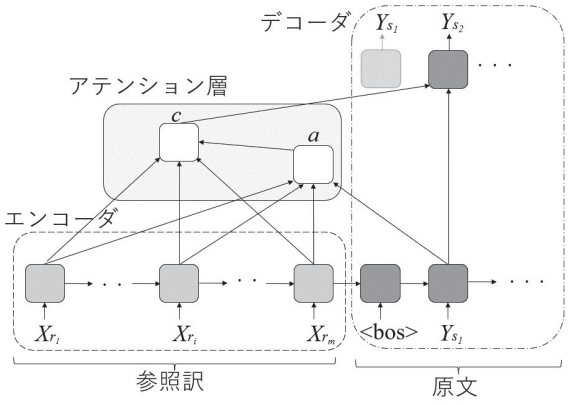

図1　アテンションベースのLSTMを用いた学習の概要図

図1における $\{X_{ri},\ldots,X_{rm}\}$ は参照訳の単語、$\{Y_{s1},Y_{s2},\ldots\}$ は原文の単語を示す。エンコーダには参照訳、デコーダには原文を入力することでLSTMによるエンコーダ・デコーダモデルを学習する。

2.2　評価スコアの計算

学習されたアテンションベースのLSTMモデルのエンコーダにそれぞれ参照訳とMT訳を入力し、その出力を参照訳とMT訳の文脈ベクトルとして用いる。そして、単語分散表現モデルの単語ベクトルと文脈ベクトルを組み合わせ、新たにベクトルを取得する。図2にベクトルの取得とEMDを用いたスコア計算の概要図を示す。

図2の（1）は参照訳 X_r とMT訳 X_h を図1より学習したモデルのエンコーダを用いてベクトル r_i と h_j を取得する場合の概要図である。参照訳は $\{X_{ri},\ldots,X_{rm}\}$ の m 個の単語から構成されている。また、MT訳は $\{X_{hi},\ldots,X_{hn}\}$ の n 個の単語から構成されている。これらの単語を図1より学習したモデルのエンコーダに入力することにより、参照訳においては文脈ベクトル $cr=\{c_{r1},\ldots,c_{rm}\}$、MT訳においては文脈ベクトル $ch=\{c_{h1},\ldots,c_{hn}\}$ が得られる。また、事前学習済みの単語分散表現モデルを利用することで参照訳を構成する単語の個々のベクトルが $s_r=\{s_{r1},\ldots,s_{rm}\}$、MT訳を構成する単語の個々のベクトルが $s_h=\{s_{h1},\ldots,s_{hn}\}$ として得られる。

このようにして得られた参照訳とMT訳それぞれの単語ベクトルと文脈ベクトルを用いて新たなベクトル r_i と h_j を以下の式（1）と（2）より得る。

$$r_{i_k} = exp\left(\alpha \times c_{r_{i_k}}\right) \times s_{r_{i_k}} \qquad (1)$$

$$h_{j_k} = exp\left(\alpha \times c_{h_{j_k}}\right) \times s_{h_{j_k}} \qquad (2)$$

式（1）は参照訳の単語ベクトル s_r と文脈ベクトル c_r を用いたベクトルの計算式、式（2）はMT訳の単語ベクトル s_h と文脈ベクトル c_h を用いたベクトルの計算式

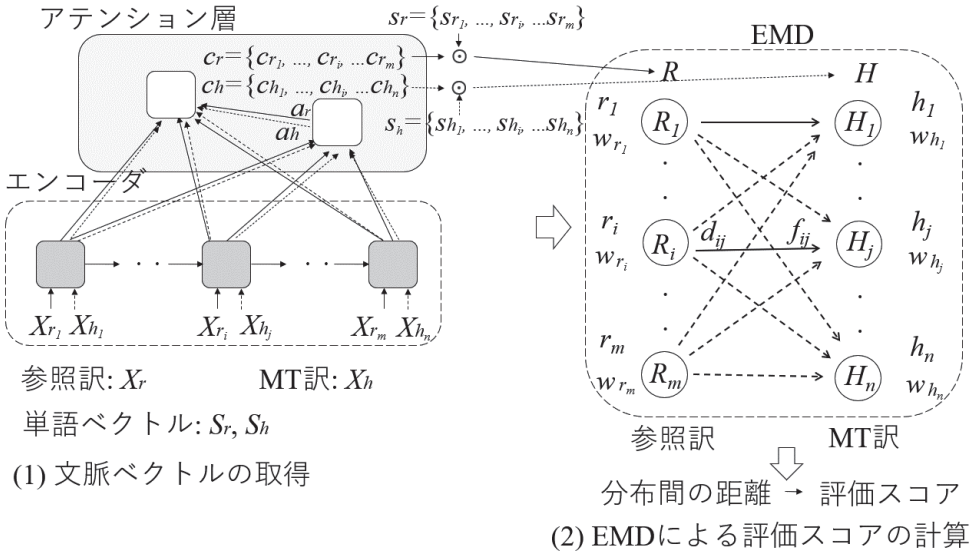

図2　評価スコアの計算の概要図

を示している。式（1）と（2）の α はハイパラメータである。また、k はベクトルの次元数分変化する。したがって、式（1）と（2）の計算はベクトルの要素単位で行われ、全要素に対して適用される。得られたベクトル r_i と h_j は EMD による評価スコアを求める際の特徴量として使用される。

図２の（2）は EMD による評価スコアの計算の概要図である。EMD は２つの分布間の最適な距離を最適輸送問題として求めるためのアルゴリズムである。ここでは参照訳を分布 R、MT 訳を分布 H とみなし、それぞれの分布を構成する特徴 $\{R_1,...,R_m\}$ と $\{H_1,...,H_n\}$ を単語と位置付けている。参照訳の単語 R_i はすべて特徴量 r_i と重み w_{ri} を持つ。同様に MT 訳の特徴 H_j はすべて特徴量 h_j と重み w_{hj} を持つ。提案手法では、特徴量に式（1）と（2）より得られたベクトル、重みに文レベルの $tf \cdot idf$ を用いる。d_{ij} は R と H の単語間の距離であり、コサイン距離を用いる。また、f_{ij} は輸送量を示しており、最適な f_{ij}、すなわち、式（3）の分子を最小化する f_{ij}^* を求めることが目的となる。EMD は以下の式により得られる。

$$EMD(R, H) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}^*}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \qquad (3)$$

なお、EMD を求める際に輸送量 f_{ij} には以下の４つの制約が加えられる。

$$f_{ij} \geq 0 \ (1 \leq i \leq m, 1 \leq j \leq n) \qquad (4)$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{r_i} \ (1 \leq i \leq m) \qquad (5)$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{h_j} \ (1 \leq j \leq n) \qquad (6)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = min\left(\sum_{i=1}^{m} w_{r_i}, \sum_{j=1}^{n} w_{h_j}\right) \qquad (7)$$

式（4）の制約はすべての荷物が H から R ではなく、R から H に輸送されることを示している。式（5）の制約は R の重み w_{ri} を超える荷物を輸送できないことを示している。式（6）は H の重み w_{hj} を超える荷物を受け入れることができないことを示している。そして、式（7）は全輸送量の上限は R のすべての重みと H のすべての重みの小さい方となることを示している。

提案手法ではこのような EMD において、距離 d_{ij} の計算を単語間のコサイン距離のみを利用するのではなく、対応関係にある単語間の距離と対応関係にない単語間の距離を区別することで語順の違いを EMD のスコアに反映させたものとなっている。提案手法における距離 d_{ij} の計算式を以下の式（8）に示す。

$$d_{ij} = \begin{cases} 1.0 - \cos\_sim(r_i, h_j) \times \exp\left(-pos\_inf\left(X_{r_i}, X_{h_j}\right)\right) \\ (\cos\_sim(r_i, h_j) \text{ が} r_i \text{と} h_{1\sim n} \text{の間の} \cos\_sim \text{の中で最大}) \\ 1.0 \\ (\text{その他}) \end{cases} \qquad (8)$$

式（8）より、参照訳の単語ベクトル r_i と MT 訳の全単語ベクトル $h_{1\sim n}$ 間とのコサイン距離を求め、コサイン距離が最大の単語間のみに対し、対応関係にあるとしてコサイン距離を d_{ij} に反映させる。ただし、参照訳の複数の単語が同一の MT 訳の単語を選択した場合にはコサイン距離の最大のもののみが優先され、それ以外は対応関係が得られなかったとして距離は 1.0 となる。このような場合も含め、対応関係にないと決定された単語間の距離はすべて一律で 1.0 となる。

また、式（8）の $pos\_inf(X_{ri}, X_{hj})$ は参照訳の単語 X_{ri} と MT 訳の単語 X_{hj} の文中の相対的な位置のずれを表し、以下の式（9）より得られる。

$$pos\_inf\left(X_{r_i}, X_{h_j}\right) = \left| \frac{pos(X_{r_i})}{m} - \frac{pos\left(X_{h_j}\right)}{n} \right| \qquad (9)$$

$pos(X_{ri})$ は参照訳の単語 X_{ri} の位置、$pos(X_{hj})$ は MT 訳の単語 X_{hj} の位置を意味する。m は参照訳の構成単語数、n は MT 訳の構成単語数を示す。$pos\_inf\ (X_{ri}, X_{hj})$ は相対的な位置のずれが大きいほど大きな値となり、ずれが小さいほど小さな値となる。したがって、式（8）の対応関係にある単語間の距離計算では相対的な位置のずれが大きいほどコサイン距離 $cos\_sim\ (r_i, h_j)$ に対する重み $exp(\cdot)$ は小さくなり、d_{ij} は大きくなる。逆に相対的な位置のずれが小さいほど $cos\_sim\ (r_i, h_j)$ に対する重み $exp(\cdot)$ は大きくなり、d_{ij} は小さくなる。このような語順の違いを距離計算に導入することは EMD を自然言語処理に適用する際に有効となる [6]。

また、提案手法では EMD の値は 0.0 ～ 1.0 の範囲で得られる。その場合、距離が小さいほど EMD の値も小さくなるため評価は高くなる。したがって、そのまま EMD の値を自動評価法に適用すると、評価が高いほど

スコアは小さくなり、直感とは異なる。そこで以下の式（10）より、評価が高いほどスコアも大きくなるように変換する。

$$score = 1.0 - EMD(R, H) \qquad (10)$$

3 自動評価法のメタ評価実験

3.1 実験方法

メタ評価実験はWMT20（fifth Conference on Machine Translation）[14]の評価タスクデータにある“newstest2020”のデータを用いて行った。評価タスクデータには評価のうえで必要な原文、MT訳、参照訳、そして、人手評価が含まれており、容易にメタ評価実験を行うことができる。

提案手法においては、文脈ベクトルを取得するためにアテンションベースのLSTMによりモデルを学習した。その際の学習データは言語ペア単位ですべての原文と参照訳のペアを用いた。さらに全ペアを10回複製することで学習データを10倍に増加した。その結果、すべての言語ペアにおける最多ペア数はイヌクティトゥット語(iu)と英語(en)の言語ペアの29,710であった。一方、最少ペア数はドイツ語(de)－英語の言語ペアの7,850であった。また、提案手法におけるベクトルの次元数はすべて300次元とした。これは単語分散表現モデルとして用いたfastTextの次元数に合わせたためである。学習時のエポック数はすべての言語ペアにおいて80とした。また、式（1）と（2）で用いたハイパラメータαの値には予備実験の結果に基づき0.1を用いた。したがって、文脈ベクトルよりも単語分散表現モデルの単語ベクトルをより重視したことになる。そして、イヌクティトゥット語についてはfastTextのモデルを取得できなかったため、式（1）のs_{rik}と式（2）のs_{hjk}の値にはすべて1.0を用いた。

メタ評価実験には著者らがこれまでに提案した自動評価法IMPACT[15]、WE_WPI[6]、そして、WE_WPI-Attention[11]も含まれている。IMPACTは表層情報のみに基づく自動評価法であり、チャンクを最長共通部分列に基づき一意に決定し、チャンクの語順を考慮した手法となっている。WE_WPIは事前学習済みの単語分散表現モデルを利用した自動評価法である。その際、EMDの距離計算においては提案手法と同様に語順の情報を用いている。WE_WPIとWE_WPI-Attentionは単語分散表現モデルとしてfastTextを使用しているが、そこにはイヌクティトゥット語の単語分散表現モデルは存在しない。そのため、WE_WPIについては英語-イヌクティトゥット語の評価スコアは求めていない。WE_WPI-Attentionについては文脈ベクトルのみを用いた評価スコアを最終的な評価スコアとした。

メタ評価は各自動評価法が出力する評価スコアと人手評価との相関係数を求めることで行う。その際に得られる評価スコアはシステム単位と文単位の2種類である。そして、評価方法としてシステム単位においてはピアソンの相関係数、文単位においてはケンドールの相関係数を用いている。

表1 “to-English”におけるシステム単位のメタ評価結果

Metrics	cs-en	de-en	ja-en	pl-en	ru-en	ta-en	zh-en	iu-en	km-en	ps-en	Avg.
BLEURT-extend	0.771	0.985	0.961	0.551	0.9	0.897	0.945	0.789	0.985	0.942	0.873
COMET	0.783	0.998	0.964	0.591	0.923	0.88	0.952	0.852	0.971	0.941	0.886
esim	0.79	0.998	0.983	0.591	0.928	0.885	0.963	0.807	0.929	0.929	0.8803
OpenKiwi-Bert	0.726	0.989	0.735	0.355	0.862	0.645	0.625	-0.126	0.751	0.753	0.632
OpenKiwi-XLMR	0.76	0.995	0.931	0.442	0.859	0.792	0.905	0.271	0.88	0.865	0.770
prism	0.818	0.998	0.974	0.502	0.908	0.898	0.957	0.833	0.95	0.966	0.8804
YiSi-1	0.832	0.998	0.982	0.543	0.915	0.925	0.961	0.834	0.977	0.953	0.892
IMPACT	0.848	0.996	0.973	0.536	0.934	0.911	0.952	0.714	0.981	0.91	0.876
WE_WPI	0.838	0.998	0.973	0.573	0.939	0.933	0.965	0.776	0.993	0.922	0.8910
WE_WPI-Attention	0.838	0.998	0.971	0.574	0.938	0.933	0.967	0.781	0.992	0.920	0.8912
提案手法	0.849	0.998	0.972	0.541	0.934	0.933	0.964	0.824	0.993	0.929	0.894

表２　“out-of-English”におけるシステム単位のメタ評価結果

Metrics	en-cs	en-de	en-ja	en-pl	en-ru	en-ta	en-zh	en-iu_full	en-iu_news	Avg.
BLEURT-extend	0.989	0.969	0.944	0.982	0.98	0.94	0.928	0.823	0.762	0.924
COMET	0.978	0.972	0.974	0.981	0.925	0.944	0.007	0.86	0.858	0.833
esim	0.908	0.979	0.993	0.969	0.967	0.937	0.972	0.814	0.76	0.922
OpenKiwi-Bert	0.92	0.852	0.363	0.903	0.834	0.846	0.551	0.573	0.808	0.739
OpenKiwi-XLMR	0.972	0.968	0.992	0.957	0.875	0.91	-0.01	0.513	0.68	0.762
prism	0.949	0.958	0.932	0.958	0.724	0.863	0.221	0.957	0.945	0.834
YiSi-1	0.922	0.971	0.969	0.964	0.926	0.973	0.959	0.554	0.523	0.862
IMPACT	0.861	0.932	0.932	0.939	0.961	0.954	0.913	0.405	0.43	0.814
WE_WPI	0.879	0.941	0.964	0.894	0.945	0.936	0.911	-	-	-
WE_WPI-Attention	0.879	0.943	0.966	0.892	0.945	0.932	0.911	0.498	0.655	0.846
提案手法	0.868	0.944	0.964	0.913	0.952	0.935	0.91	0.586	0.695	0.863

表３　“to-English”における文単位のメタ評価結果

Metrics	cs-en	de-en	iu-en	ja-en	km-en	pl-en	ps-en	ru-en	ta-en	zh-en	Avg.
BLEURT-extend	0.127	0.448	0.259	0.271	0.33	0.044	0.161	0.101	0.246	0.137	0.212
COMET	0.129	0.485	0.281	0.274	0.298	0.099	0.158	0.156	0.241	0.171	0.229
esim	0.11	0.454	0.241	0.239	0.3	0.058	0.147	0.084	0.208	0.138	0.198
OpenKiwi-Bert	0.036	0.379	-0.005	0.11	0.168	-0.033	0.076	-0.033	0.118	0.029	0.085
OpenKiwi-XLMR	0.093	0.463	0.056	0.22	0.244	0.059	0.106	0.092	0.188	0.115	0.164
prism	0.143	0.475	0.255	0.272	0.304	0.109	0.165	0.145	0.237	0.167	0.227
YiSi-1	0.117	0.468	0.253	0.277	0.316	0.042	0.147	0.091	0.248	0.146	0.211
IMPACT	0.07	0.427	0.188	0.194	0.243	-0.007	0.097	0.009	0.191	0.103	0.152
WE_WPI	0.102	0.474	0.218	0.238	0.239	0.08	0.134	0.133	0.222	0.151	0.199
WE_WPI-Attention	0.108	0.479	0.217	0.242	0.233	0.096	0.138	0.138	0.227	0.152	0.203
提案手法	0.103	0.475	0.231	0.233	0.235	0.089	0.149	0.134	0.222	0.154	0.2025

表４　“out-of-English”における文単位のメタ評価結果

Metrics	en-cs	en-de	en-iu	en-ja	en-pl	en-ru	en-ta	en-zh	Avg.
BLEURT-extend	0.689	0.447	0.359	0.533	0.43	0.305	0.643	0.46	0.483
COMET	0.668	0.468	0.322	0.624	0.462	0.344	0.671	0.432	0.499
esim	0.469	0.347	0.122	0.522	0.312	0.224	0.599	0.391	0.373
OpenKiwi-Bert	0.262	0.168	-0.115	-0.529	0.153	0.164	0.169	0.077	0.044
OpenKiwi-XLMR	0.607	0.369	0.06	0.553	0.347	0.279	0.604	0.377	0.400
prism	0.619	0.447	0.452	0.579	0.414	0.283	0.448	0.397	0.455
YiSi-1	0.55	0.427	0.251	0.568	0.349	0.256	0.669	0.463	0.442
IMPACT	0.457	0.306	0.209	0.486	0.181	0.068	0.525	0.391	0.328
WE_WPI	0.477	0.331	-	0.502	0.276	0.192	0.376	0.379	-
WE_WPI-Attention	0.477	0.332	0.208	0.503	0.279	0.189	0.374	0.381	0.343
提案手法	0.471	0.339	0.25	0.5	0.28	0.184	0.367	0.381	0.347

3.2　メタ評価実験の結果

表 1 に “to-English（MT 訳が英語）” におけるシステム単位のメタ評価結果、表 2 には “out-of-English（MT 訳が英語以外）” におけるシステム単位のメタ評価結果を示す。また、表 3 に “to-English” における文単位のメタ評価結果、表 4 には “out-of-English” における文単位のメタ評価結果を示す。表中の “Avg.” は全言語ペアの相関係数の平均である。

3.3　考察

表 1 の “to-English” におけるシステム単位のメタ評価結果より、提案手法の“Avg.”が最も高い値を示した。言語ペアにおいては “cs-en” と “km-en” の相関係数が他の自動評価法に対して比較的高い値を示した。表 2 の “out-of-English” におけるシステム単位のメタ評価結果においては提案手法の “Avg.” は 3 番目に位置している。提案手法と WE_WPI-Attention の “en_iu_full” と “en_iu_news” を比較すると、提案手法の相関係数が大きく向上している。イヌクティトゥット語については単語ベクトルを利用できないため文脈ベクトルのみを用いて評価スコアを求めている。その際、WE_WPI-Attention では参照訳と MT 訳の文脈ベクトル間のコサイン距離のみを用いている。それに対して、提案手法では EMD を利用する際に語順も考慮したスコア計算となっている。したがって、提案手法では語順情報の利用がより効果的に働いたと考えられる。

表 3 の “to-English” における文単位のメタ評価結果では、提案手法の “Avg.” の順位は 6 番目であった。WE_WPI-Attention との比較においてわずかではあるが、低い値となった。表 4 の “out-of-English” の文単位のメタ評価結果においては、“Avg.” の順位は 7 番目であった。著者らが従来から提案してきた自動評価法との比較では最も高い値であったが、他の自動評価法との比較では “Avg.” の値はかなり低くなっている。WE_WPI-Attention との比較においてはやはり言語ペア “en-iu” の相関係数は大きく向上している。

表 1 から表 4 より提案手法は “out-of-English” の文単位を除くと他の自動評価法と比べて遜色のない性能を示しているといえる。文単位のメタ評価においては COMET [8] が最も高い性能を示しており、この結果は文献 [16] においても言及されている。COMET はファインチューニングしたモデルを利用していることから提案手法のアプローチとは異なるが提案手法のさらなる性能向上が必要である。

また、提案手法が “to-English” において安定した性能を示したことから、利用した単語分散表現モデルとの関係についても調査が必要である。英語の単語分散表現モデルが他の言語の単語分散表現モデルに比べて提案手法により効果的であった可能性があるため、単語分散表現モデルに対する検証も必要と考えられる。

4　おわりに

本稿では、事前学習済みの単語分散表現モデルから取得した単語ベクトルと MT 訳に対応する原文と参照訳のペアのみを学習データとして構築したアテンションベースの LSTM モデルのエンコーダから取得した文脈ベクトルを組み合わせ、それを EMD の特徴量としてスコア計算を行う新たな自動評価法を提案し、WMT20 の評価データによるメタ評価を行った。その結果、“out-of-English” の文単位のメタ評価以外においては提案手法は WMT20 の評価タスクにおける SOTA モデルと遜色ない性能を示した。

しかしながら、その精度は十分とはいえないため、さらなる精度向上のための改良が必要である。提案手法の一つの利点は文脈ベクトルを取得するために構築したアテンションベースの LSTM モデルに必要な学習データを原文とその参照訳のみとしている点にあるが、今後は原文と参照訳に出現する情報のみを利用したデータ・オーギュメンテーションに取り組む予定である。それにより学習モデルの性能を向上させ、より有効な文脈ベクトルが取得されることで提案手法の性能向上をもたらすと考えられる。

参考文献

[1] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proc. 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318（2002）.

[2] Snover, M., Dorr, B., Schwartz, R., Micciulla,

L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Proc. 7th Conference of the Association for Machine Translation in the Americas, pp. 223-231 (2006) .

[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26, 3111-3119 (2013) .

[4] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 5, pp.135-146 (2017) .

[5] Lo, C.: YiSi - A Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In: Proc. Fourth Conference on Machine Translation, Volume 2: Shared Task Papers, pp. 507-513 (2019) .

[6] Echizen-ya, H., Araki, K., Hovy, E.: Word Embedding-Based Automatic MT Evaluation Metric using Word Position Information. In: Proc. 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1874-1883 (2019) .

[7] Sellam, T., Pu, A., Chung, H.W., Gehrmann, S., Tan, Q., Freitag, M., Das, D., Parikh, A.P.: Learning to Evaluate Translation Beyond English: BLEURT Submissions to the WMT Metrics 2020 Shared Task. In: Proc. 5th Conference on Machine Translation, pp. 921-927 (2020) .

[8] Rei, R., Stewart, C., Farinha, A.C., Lavie, A.: COMET: A Neural Framework for MT Evaluation. In: Proc. 2020 Conference on Empirical Methods in Natural Language Processing, pp. 2685-2702 (2020) .

[9] Mathur, N., Baldwin, T., Cohn, T.: Putting Evaluation in Context: Contextual Embeddings improve Machine Translation Evaluation. In: Proc. 57th Annual Meeting of the Association for Computational Linguistics, pp. 2799-2808 (2019) .

[10] Thompson, B., Post, M.: Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In: Proc. 2020 Conference on Empirical Methods in Natural Language Processing, pp. 90-121 (2020) .

[11] 越前谷博：単語分散表現に基づく自動評価法における Attention による文脈ベクトルの利用 . Japio YEAR BOOK 2021, pp. 286-293 (2021) .

[12] Rubner, Y., Tomasi, C., Guibas, L.J.: A Metric for Distributions with Applications to Image Databases. In: Proc. 1998 IEEE International Conference on Computer Vision, pp. 59-66 (1998) .

[13] Bahdanau, D., Cho, K., Bengio,Y.: Neural Machine Translation by Jointly Learning to Align and Translate. In: Proc. Third International Conference on Learning Representations (2015) .

[14] Mathur, N., Wei, J.T.-Z., Freitag, M., Ma, Q., Bojar, O.: Results of the WMT20 Metrics Shared Task. In: Proc. 5th Conference on Machine Translation, pp. 688-725 (2020) .

[15] Hiroshi Echizen-ya, and Kenji Araki.: Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, In: Proc. of the Eleventh Machine Translation Summit, pp.151-158, (2007) .

[16] Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., Menezes, A.: To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics

for Machine Translation. In: Proc. Sixth Conference on Machine Translation, pp. 478-494 (2021) .