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

Application of context vector by attention in automatic evaluation based on word embedding

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

機械翻訳分野において自動的にシステム訳を評価する自動評価は不可欠なものとなっている。現在、ディープラーニング技術に基づく機械翻訳の性能向上は著しく、それに伴い自動評価にも機械翻訳の進展に追随するための性能向上が求められている。文脈処理を可能とするニューラル翻訳が主流となっている現在においては、BLEU [1] に代表される表層レベルの自動評価法のみでは様々な言語現象に対応することが困難である。しかし、これまでに数多くの自動評価法が研究されているにも関わらず、BLEU が今もなお自動評価法のデファクトスタンダードとして利用されている。BLEU には一長一短があることはよく知られているが、他の自動評価法が BLEU に対して明らかなアドバンテージを持っていると感じさせるまでには至っていないことも事実である。

このような現状において、著者らはより良い自動評価法の構築を目的とした研究を行っている。本稿では著者らがこれまでに提案してる、単語分散表現に基づく自動評価法 WE_WPI に対して、文脈情報を利用した新たな自動評価法を提案する。さらに、提案手法を含めた様々な自動評価法を用いて行った WMT20 の評価タスクデータによるメタ評価についても述べる。

2 これまでの提案手法

2.1 IMPACT

これまでに著者らが提案してきた自動評価法をいくつか紹介する。まず、BLEU と同様に表層レベルのみに基づく自動評価法である IMPACT（Intuitive comMon PArts ConTinuum）[2] について述べる。IMPACT はシステム訳と参照訳間において単語を単位とした LCS（Longest Common Subsequence）[3] に基づき共通単語列であるチャンクを求めることでシステム訳に対する評価スコアを算出する。LCS は共通部分列を構成する文字や単語の数を求めるためのアルゴリズムであり、どの文字や単語でチャンクが構成されているかについては明確には示されない。通常は LCS の値が同じであってもその根拠となるチャンクのパターンは数多く存在する。そこで IMPACT では LCS の値が同じチャンクのパターンが複数存在した場合には、チャンクの長さとその位置に基づきパターンを一意に決定する。この処理を再帰的に行うことで、チャンクの長さと出現順を考慮した自動評価を実現している。

2.2 WE_WPI

ニューラルネットワーク技術が自然言語処理分野にもたらした大きな功績の一つとして単語分散表現が挙げられる。Word2Vec のような単語分散表現により単語をベクトル表現に変換することで、単語をこれまでの表層レベルではなく意味レベルで処理できるようになった。WE_WPI（Word Embedding-based automatic MT evaluation using Word Position Information）[4] はその単語分散表現に基づく自動評価法となっている。また、単語分散表現を利用する際には EMD（Earth Mover's Distance）[5] に適用することで文レベルの評

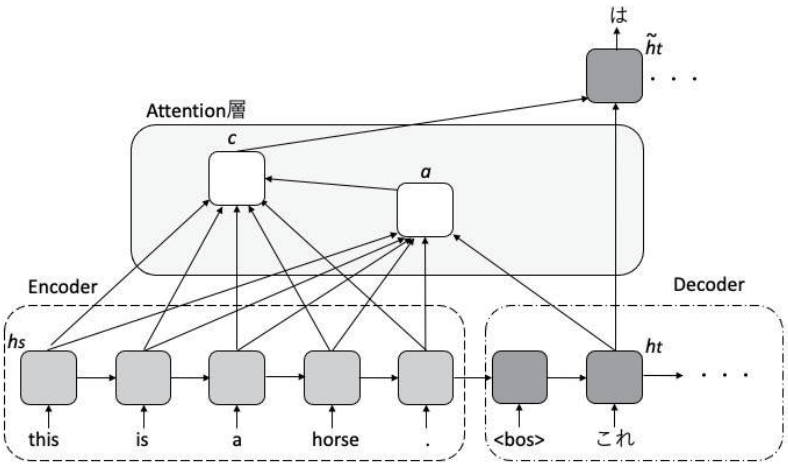

図 1　Attention を用いた Seq-to-Seq モデルの概要図

価スコアを得る。さらに、WE_WPI では語順の違いを評価スコアに反映させるために単語間のコサイン距離に対して単語の出現位置の相対的なずれを負の重みとして付与している。その結果、構成単語が完全一致であっても語順が異なっている場合にはそれに応じて評価スコアが低くなる。また、WE_WPI で使用する単語分散表現は pre-trained のモデルから得られるため、学習を必要としない点も特徴となっている。

3 新たな提案手法 WE_WPI-Attention

先述した WE_WPI は単語分散表現の利用により単語の表層レベルではなく意味に基づいたシステム訳の評価を可能とした。しかし、文脈情報を明示的には利用していないため文脈を考慮した自動評価法としては不十分である。そこで、提案手法では WE_WPI に対して文脈情報を取り入れることを目的とした改良を行う。文脈情報は時系列データを対象としたニューラルネットワークである Sequence-to-Sequence（以降、省略して Seq-to-Seq と記す）[6] [7] モデルとニューラル翻訳の性能向上のためにその有効性が確認されている Attention [8] を用いることで得る。Seq-to-Seq モデルはニューラル翻訳の代表的なモデルの一つであり、2 つのリカレントニューラルネットワーク（RNN）を組み合わせたものである。また、Attention は任意の単語に注視しながら翻訳を行うためのアーキテクチャであり、ニューラル翻訳の性能向上において有効であることが示されている。提案手法ではこのような Attention を用いた Seq-to-Seq モデルを利用することで文脈ベクトルを取得し、文脈情報として用いる。そして、システム訳と参照訳の文脈ベクトル間のコサイン類似度と WE_WPI の評価スコアの値による加重平均を最終的な評価スコアとする。このような新たな提案手法を以降、本稿では WE_WPI-Attention と記す。

3.1 Attention を用いた Seq-to-Seq モデル

Attention を用いた Seq-to-Seq モデルは RNN を 2 つ組み合わせた Seq-to-Seq モデルに Attention を導入したものである。このモデルはニューラル翻訳の代表的なモデルの一つとなっている。図 1 に Attention を用いた Seq-to-Seq モデルの概要図を示す。

図 1 において Encoder と Decoder はそれぞれ RNN に相当し、2 つの RNN を組み合わせたものが Seq-to-Seq となる。Encoder の入力は原言語文、Decoder の入力は目的言語文及び Encoder の出力ベクトルである。この Seq-to-Seq に Attention 層を取り入れることで原言語文のどの単語に注視しているかの情報を利用しながら目的言語文を生成することが可能となる。図 1 の $\tilde{h}t$ は以下の式（1）で定義される。この $\tilde{h}t$ が最終的に求める値であり、ht を活性化したものである。

$$\tilde{h}_t(t) = \tanh\left(W_c \begin{bmatrix} c(t) \\ h_t(t) \end{bmatrix} + b\right) \quad (1)$$

式（1）の Wc と b はそれぞれ重みとバイアスであり、パラメータとして学習により更新される。また、$c(t)$ は文脈ベクトルと呼ばれ、$[\cdot]$ は $c(t)$ と $ht(t)$ を結合したベクトルを意味する。文脈ベクトル $c(t)$ は以下

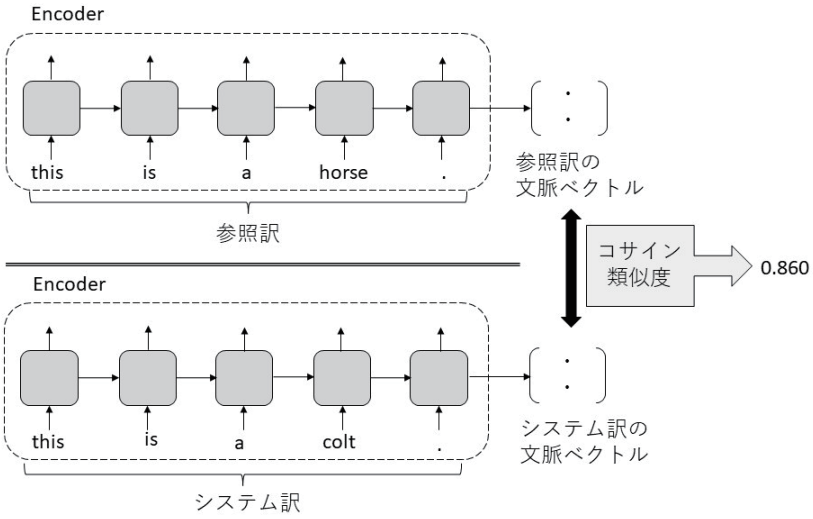

図２　Encoder による文脈ベクトルを用いた類似度計算の概要図

の式（2）より求める。

$$c(t) = \sum_{\tau=1}^{T} a(\tau, t) h_s(\tau) \qquad (2)$$

式（2）は Encoder の各時刻の値がそれぞれどのくらい Decoder に寄与しているかを表すものとなっている。すなわち、$a(\tau, t)$ は各単語に対する重みとして文脈ベクトルに反映される。この重み $a(\tau, t)$ は以下の式（3）より得られる。

$$a(\tau, t) = \mathrm{softmax}\left(g\left(h_s(\tau), h_t(t)\right)\right) \qquad (3)$$

式（3）の $g(\cdot)$ はスコア関数と呼ばれ、いくつか存在するが、今回は内積を用いた。ここで $ht(t)$ は Decoder の任意の単語である。したがって、$a(\tau, t)$ は目的言語文の単語とのスコア関数の値が大きくなる原言語文の単語を求める役割を持つ。また、$a(\tau, t)$ の総和は 1 になるように正規化されており、スコア関数の値が大きい単語についてはその割合が大きくなる。このように Attention を用いることで文脈ベクトルはどの単語が注視されたのかを反映したものとなる。

3.2　文脈ベクトルを用いた類似度計算

提案手法 WE_WPI-Attention では先述した Attention を用いた Seq-to-Seq モデルを学習した後、システム訳と参照訳をそれぞれモデルに与え、Encoder の出力である文脈ベクトルを文脈情報として利用する。図２に Encoder による文脈ベクトルを用いた類似度計算の概要図を示す。

図２の Encoder は同一の学習済みの Attention を用いた Seq-to-Seq モデルの Encoder である。モデルは参照訳と原言語文をペアとして学習させたものである。このモデルに参照訳とシステム訳をそれぞれ Encoder の入力として与え、文脈ベクトルを得る。そして、参照訳の文脈ベクトルとシステム訳の文脈ベクトルとの間でコサイン類似度を求める。

3.3　WE_WPI との組み合わせによる評価スコアの算出

Attention を用いた Seq-to-Seq モデルの Encoder より得られる参照訳とシステム訳の文脈ベクトル間のコサイン類似度及び WE_WPI の評価スコアに対して加重平均を求めることで WE_WPI-Attention の最終的な評価スコアを得る。以下の式（4）はその計算式である。

$$score = \frac{W_1 \cdot \mathrm{cos\_sim} + W_2 \cdot \mathrm{WE\_WPI\ score}}{W_1 + W_2} \qquad (4)$$

式（4）の cos_sim は文脈ベクトル間のコサイン類似度を示す。また、WE_WPI score は WE_WPI の評価スコアを示す。W_1 と W_2 は重みである。文脈ベクトルを得るためのモデルは参照訳と原言語文のみを学習データとしており、小規模かつ極端なものとなっていることを考慮し、cos_sim の重み W_1 の値には 1 を用いた。それに対して、WE_WPI score の重み W_2 の値には 10 を用いた。すなわち、文脈ベクトル間のコサイン

類似度の重みを軽くすることでその利用は補助的なものとした。

4 WMT20 評価タスクデータを用いたメタ評価

4.1 評価データと評価方法

著者らのこれまでの提案手法 IMPACT、WE_WPI、そして、新たな提案手法 WE_WPI-Attention を含めた様々な自動評価法によるメタ評価実験を行なった。評価データには WMT20（fifth Conference on Machine Translation）[9] の評価タスクデータを用いた。そこでは英語以外の言語から英語への翻訳（to-English）として 10 言語ペア、英語から英語以外の言語への翻訳（out-of -English）として 8 言語ペアが使用されている。to-English の言語ペアはチェコ語（cs）、ドイツ語（de）、日本語（ja）、ポーランド語（pl）、ロシア語（ru）、タミル語（ta）、中国語（zh）、イヌクティトゥット語（iu）、クメール語（km）、そして、パシュート語（ps）から英語の 10 ペアである。out-of-English においては英語からチェコ語、ドイツ語、日本語、ポーランド語、ロシア語、タミル語、中国語、そして、イヌクティトゥット語の 8 ペアである。また、メ

表 1　to-English のセグメントレベルにおけるメタ評価の結果

	cs-en	de-en	iu-en	ja-en	km-en	pl-en	ps-en	ru-en	ta-en	zh-en
n	14018	16584	8162	15193	3706	21121	3507	14024	12789	62586
WE_WPI-Attention	0.108	0.479	0.217	0.242	0.233	0.096	0.138	0.138	0.227	0.152
WE_WPI	0.102	0.474	0.218	0.238	0.239	0.080	0.134	0.133	0.222	0.151
IMPACT	0.070	0.427	0.188	0.194	0.243	-0.007	0.097	0.009	0.191	0.103
BERT_base-L2	0.103	0.454	0.238	0.263	0.295	0.032	0.159	0.087	0.223	0.141
BERT_large-L2	0.102	0.456	0.251	0.262	0.314	0.044	0.151	0.094	0.245	0.133
BLEURT	0.126	0.456	0.258	0.265	0.327	0.057	0.207	0.093	0.230	0.137
BLEURT-extended	0.127	0.448	0.259	0.271	0.330	0.044	0.161	0.101	0.246	0.137
CharacTER	0.090	0.440	0.214	0.221	0.248	0.023	0.172	0.057	0.138	0.123
chrF	0.086	0.438	0.254	0.242	0.267	0.028	0.144	0.049	0.186	0.132
chrF++	0.090	0.435	0.246	0.245	0.275	0.034	0.145	0.054	0.186	0.130
COMET	0.129	0.485	0.281	0.274	0.298	0.099	0.158	0.156	0.241	0.171
COMET-2R	0.120	0.479	0.257	0.268	0.308	0.098	0.144	0.148	0.253	0.163
COMET-HTER	0.103	0.481	0.198	0.241	0.269	0.080	0.116	0.131	0.227	0.135
COMET-MQM	0.108	0.483	0.215	0.259	0.282	0.080	0.141	0.137	0.227	0.141
COMET-Rank	0.099	0.470	0.188	0.235	0.228	0.073	0.107	0.118	0.199	0.142
EED	0.091	0.440	0.256	0.235	0.271	0.045	0.149	0.053	0.198	0.129
esim	0.110	0.454	0.241	0.239	0.300	0.058	0.147	0.084	0.208	0.138
mBERT-L2	0.119	0.442	0.244	0.251	0.312	0.047	0.151	0.083	0.227	0.133
MEE	0.063	0.402	0.134	0.187	0.206	-0.084	0.078	-0.041	0.114	0.083
parbleu	0.058	0.415	0.167	0.198	0.203	-0.025	0.100	-0.011	0.159	0.095
parchrf++	0.096	0.436	0.232	0.247	0.267	0.027	0.147	0.044	0.184	0.132
paresim-1	0.105	0.464	0.249	0.242	0.292	0.066	0.149	0.089	0.213	0.139
prism	0.143	0.475	0.255	0.272	0.304	0.109	0.165	0.145	0.237	0.167
sentBLEU	0.068	0.413	0.182	0.188	0.226	-0.024	0.096	-0.005	0.162	0.093
SWSS+METEOR	-	-	0.226	0.228	0.264	0.011	0.130	0.048	0.205	0.133
TER	-0.04	0.355	0.021	0.044	0.125	-0.172	-0.036	-0.117	0.046	-0.01
YiSi-0	0.072	0.441	0.261	0.241	0.268	0.035	0.140	0.065	0.183	0.127
YiSi-1	0.117	0.468	0.253	0.277	0.316	0.042	0.147	0.091	0.248	0.146

表 2　out-of-English のセグメントレベルにおけるメタ評価の結果

	en-cs	en-de	en-iu	en-ja	en-pl	en-ru	en-ta	en-zh
n	21121	9339	13159	12830	17689	8330	9087	12652
WE_WPI-Attention	0.477	0.332	0.208	0.503	0.279	0.189	0.374	0.381
WE_WPI	0.477	0.331	-	0.502	0.276	0.192	0.376	0.379
IMPACT	0.457	0.306	0.209	0.486	0.181	0.068	0.525	0.391
BLEURT-extended	0.689	0.447	0.359	0.533	0.430	0.305	0.643	0.460
CharacTER	0.413	0.311	0.309	0.471	0.198	0.143	0.525	0.339
chrF	0.472	0.379	0.344	0.506	0.250	0.153	0.589	0.400
chrF++	0.478	0.367	0.338	0.506	0.255	0.156	0.579	0.388
COMET	0.668	0.468	0.322	0.624	0.462	0.344	0.671	0.432
COMET-2R	0.669	0.463	0.326	0.630	0.445	0.343	0.676	0.434
COMET-HTER	0.665	0.440	0.331	0.601	0.427	0.292	0.640	0.411
COMET-MQM	0.666	0.423	0.313	0.588	0.424	0.281	0.635	0.388
COMET-Rank	0.629	0.379	0.297	0.569	0.388	0.229	0.588	0.380
EED	0.458	0.363	0.361	0.515	0.248	0.155	0.587	0.393
esim	0.469	0.347	0.122	0.522	0.312	0.224	0.599	0.391
mBERT-L2	0.567	0.361	-	0.541	0.350	0.246	0.587	0.432
MEE	0.411	0.289	-0.074	-	0.125	0.027	0.373	-
parbleu	0.460	0.299	0.212	0.052	0.183	0.062	0.340	0.356
parchrf++	0.492	0.355	-	0.527	0.272	0.176	-	0.398
paresim-1	0.475	0.343	0.122	0.510	0.324	0.230	0.599	0.396
prism	0.619	0.447	0.452	0.579	0.414	0.283	0.448	0.397
sentBLEU	0.432	0.303	0.206	0.479	0.153	0.051	0.398	0.396
TER	0.317	0.182	-0.071	-0.591	0.003	-0.121	0.203	-0.36
YiSi-0	0.432	0.349	0.362	0.484	0.233	0.151	0.547	0.319
YiSi-1	0.550	0.427	0.251	0.568	0.349	0.256	0.669	0.463
YiSi-2	0.187	0.296	0.146	0.383	0.115	0.146	0.545	0.152

タ評価はセグメントレベルの観点より相関係数を求めることで行なった。セグメントレベルの相関係数にはケンドール τ を用いた。

4.2　実験結果

表 1 に to-English のセグメントレベルにおけるメタ評価の結果、そして、表 2 に out-of-English のセグメントレベルにおけるメタ評価の結果を示す。なお、WE_WPI-Attention においては Attention を用いた Seq-to-Seq モデルの学習は各言語ペアの参照訳と原言語文を用いて行なった。しかし、文数の多い言語ペアであっても 3,000 未満と小規模であったため全ての文ペアを 10 倍に増やして学習を行なった。また、WE_WPI の en-iu の言語ペアにおいては、イヌクティトゥット語（iu）の pre-trained の単語分散表現モデルが fasttext [10] として提供されていなかったため評価スコアの算出は行っていない。それに伴い、en-iu の WE_WPI-Attention の評価スコアは文脈ベクトル間のコサイン類似度のみを用いて得られた相関係数となっている。表中の数値はケンドール τ を示している。

4.3　考察

表 1 の to-English のセグメントレベルでは WE_WPI-Attention は WE_WPI との比較において、iu-en と km-en の言語ペアを除き高い相関係数を示した。IMPACT に対しても km-en を除き、他の全ての言語ペアにおいて高い相関係数を示した。表 2 の out-of-English のセグメントレベルでは WE_WPI-Attention は WE_WPI に対しては en-de、en-ja、en-pl、そして、en-zh の言語ペアで高い相関係数を示した。IMPACT に対しては、en-cs、en-de、en-ja、en-pl、そして、en-ru の言語ペアにおいて高い相関係数を示した。これらの結果より、WE_WPI に対して Attention を用いた Seq-to-Seq モデルによる文脈ベクトルが有効であるこ

とが明らかとなった。

また、他の自動評価法との比較については表 1 の to-English では WE_WPI-Attention は比較的高い相関係数を示した。全言語ペアの相関係数を対象とした平均を求めた場合、WE_WPI-Attention の平均は 0.203 となり、全 28 の自動評価法において第 9 位であった。それに対して、表 2 の out-of-English では WE_WPI-Attention の相関係数は不十分であった。特に en-ta の相関係数は他の自動評価法に比べて低かった。著者らの提案手法である IMPACT よりも大きく下回っていることから、単語分散表現や文脈ベクトルが有効に機能していないと考えられる。その原因については今後精査する必要がある。

次いで、このようなセグメントレベルのメタ評価の結果における傾向がシステムレベルにおいても同様に見られるかどうかを検証した。表 3 に 3 つの提案手法と表 1 の全言語ペアにおいて WE_WPI-Attention を上回った自動評価法を対象とした to-English におけるシステムレベルの相関係数、表 4 に 3 つの提案手法と表 2 の全言語ペアにおいて WE_WPI-Attention を上回った自動評価法を対象とした out-of-English におけるシステムレベルの相関係数をそれぞれ示す。

なお、相関係数にはピアソンの相関係数を用いた。表 3 より to-English のシステムレベルにおいて WE_WPI-Attention の相関係数は COMET と COMET-2R よりも若干高かった。全言語ペアの相関係数の平均を求めたところ、COMET と COMET-2R の平均がそれぞれ 0.886 と 0.887 であったのに対して WE_WPI-Attention の相関係数の平均は 0.891 であった。したがって、セグメントレベルと同様に to-English においては WE_WPI-Attention は比較的高い相関係数であったといえる。それに対して表 4 の out-of-English では、全言語ペアの相関係数の平均において WE_WPI-Attention は十分とは言えない。en-iu の相関係数が存在しない WE_WPI と mBERT-L2 を除いた場合、WE_WPI-Attention の相関係数の平均値は順位としては第 3 位であった。out-of-English では WE_WPI-Attention に限らず、自動評価法の間で言語ペアによって相関係数の差が大きくなる傾向が見られた。

表 1 から表 4 より、WE_WPI-Attention は to-

表 3　to-English のシステムレベルにおけるメタ評価の結果

	cs-en	de-en	ja-en	pl-en	ru-en	ta-en	zh-en	iu-en	km-en	ps-en
n	12	12	10	14	11	14	16	11	7	6
WE_WPI-Attention	0.838	0.998	0.971	0.574	0.938	0.933	0.967	0.781	0.992	0.920
WE_WPI	0.838	0.998	0.973	0.573	0.939	0.933	0.965	0.776	0.993	0.922
IMPACT	0.848	0.996	0.973	0.536	0.934	0.911	0.952	0.714	0.981	0.910
COMET	0.783	0.998	0.964	0.591	0.923	0.880	0.952	0.852	0.971	0.941
COMET-2R	0.777	0.998	0.964	0.584	0.924	0.881	0.949	0.872	0.970	0.949

表 4　out-of-English のシステムレベルにおけるメタ評価の結果

	en-cs	en-de	en-ja	en-pl	en-ru	en-ta	en-zh	en-iu_full	en-iu_news
n	12	14	11	14	9	15	12	11	11
WE_WPI-Attention	0.879	0.943	0.966	0.892	0.945	0.932	0.911	0.498	0.655
WE_WPI	0.879	0.941	0.964	0.894	0.945	0.936	0.911	-	-
IMPACT	0.861	0.932	0.932	0.939	0.961	0.954	0.913	0.405	0.430
BLEURT-extended	0.989	0.969	0.944	0.982	0.980	0.940	0.928	0.823	0.762
COMET	0.978	0.972	0.974	0.981	0.925	0.944	0.007	0.860	0.858
COMET-2R	0.983	0.972	0.986	0.982	0.872	0.959	-0.066	0.848	0.867
COMET-HTER	0.976	0.951	0.989	0.974	0.803	0.925	-0.073	0.900	0.888
COMET-MQM	0.974	0.881	0.974	0.967	0.788	0.910	0.084	0.870	0.867
mBERT-L2	0.946	0.970	0.977	0.976	0.946	0.944	0.934	-	-
prism	0.949	0.958	0.932	0.958	0.724	0.863	0.221	0.957	0.945
YiSi-1	0.922	0.971	0.969	0.964	0.926	0.973	0.959	0.554	0.523

English の方が out-of-English よりもセグメントレベル、システムレベル共に高い性能を示し、他手法との比較においても WE_WPI-Attention が上位に位置していることを確認できた。WE_WPI-Attention のアプローチ面での特徴としては文脈ベクトルを取得するために学習が必要となるが、学習に用いるデータは原言語文とそれに対応する参照訳のみと小規模であるため、多くの学習時間を要しない。また、Attention を用いた Seq-to-Seq モデルは参照訳に極端に偏ったモデルとなり、参照訳に強く依存すると考えられる。しかし、式（4）において WE_WPI 側の重みを大きくしているため、そのことによる影響は抑制されていると考えられる。

5 おわりに

本稿では著者らがこれまでに提案している単語分散表現に基づく自動評価法 WE_WPI の評価精度向上を目的として、文脈ベクトルを利用した新たな自動評価法を提案した。提案手法 WE_WPI-Attention を含めた WMT20 によるメタ評価実験の結果、WE_WPI に比べて WE_WPI-Attention はセグメントレベルの相関係数において高い値を示したことから文脈ベクトルの有効性が確認された。また、他の自動評価法との比較では to-English のセグメントレベルのメタ評価において、全言語ペアの平均値が全 28 の自動評価法の中で 9 番目であったことから、提案手法が比較的上位に位置していることを確認した。

今後は英語がシステム訳となる to-English だけでなく、様々な言語がシステム訳となる out-of-English のメタ評価においても提案手法がより高い評価精度を示すように改良を進める予定である。そして、今後も評価精度のより高い自動評価法の実現に向けた研究を継続していく。

参考文献

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

[2] Hiroshi Echizen-ya, and Kenji Araki, 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, 2007.

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

[4] Hiroshi Echizen' ya, Kenji Araki, and Eduard Hovy, 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, 2019.

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

[6] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1724-1734, 2014.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, Sequence to Sequence Learning with Neural Networks, Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS 2014), pp.3104-3112, 2014.
[8] D. Bahdanau, K. Cho, and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, 3rd International Conference on Learning Representations (ICLR), 2015.
[9] Nitika Mathur, Johnny Tian-Zheng Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar, Results of the WMT20 Metrics Shared Task, Proceedings of the 5th Conference on Machine Translation (WMT), pp.688-725, 2020.
[10] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, Volume 5, pp.135-146, 2017.