Word Embedding-Based Automatic MT Evaluation Metric using Word Position Information

Hiroshi Echizen'ya
Department of Life
Science and Technology
Hokkai-Gakuen University, Japan
echi@lst.hokkai-s-u.ac.jp

Kenji Araki
Graduate School of Information
Science and Technology
Hokkaido University, Japan
araki@ist.hokudai.ac.jp

Eduard Hovy
Language Technologies Institute
Carnegie Mellon University, USA
hovy@cmu.edu

Abstract

We propose a new automatic evaluation metric for machine translation. Our proposed metric is obtained by adjusting the Earth Mover's Distance (EMD) to the evaluation task. The EMD measure is used to obtain the distance between two probability distributions consisting of some signatures having a feature and a weight. We use word embeddings, sentence-level $tf \cdot idf$, and cosine similarity between two word embeddings, respectively, as the features, weight, and the distance between two features. Results show that our proposed metric can evaluate machine translation based on word meaning. Moreover, for distance, cosine similarity and word position information are used to address word-order differences. We designate this metric as **W**ord **E**mbedding-based automatic MT evaluation using **W**ord **P**osition **I**nformation (WE_WPI). A meta-evaluation using WMT16 metrics shared task set indicates that our WE_WPI achieves the highest correlation with human judgment among several representative metrics.

1 Introduction

Recent advances in neural machine translation (NMT) (Sutskever et al., 2014; Luong et al., 2015) are remarkable. Results based on human evaluation have demonstrated that NMT outperforms statistical machine translations significantly (Chiang, 2005; Tufiş and Ceauşu, 2009). The NMT achieved especially high performance in terms of fluency. However, it tends to generate more omission errors than statistical machine translations generate. Unfortunately, it is difficult for automatic evaluation metrics to evaluate outputs with omission errors because those errors are not included as non-match words between the translation and reference. For such cases, the word embedding-based automatic MT evaluation metric, which is based on word position information, is effective.

Various automatic evaluation metrics have been proposed for machine translation, but none is sufficient for NMT. Actually, BLEU (Papineni et al., 2002) is the representative metric based on n-gram matching. Unfortunately, because it is a surface-level metric, it is difficult to address word meaning during evaluation for MT outputs. The word-embedding-based distance measure for document (Kusner et al., 2016) and the word-alignment-based automatic evaluation metric using word embedding (Matsuo et al., 2017) are effective to address word meanings. Nevertheless, they can only ineffectively accommodate word order differences between the translation and reference.

Given those circumstances, a new metric with word embedding-based automatic MT evaluation metric using word position information is proposed in which the evaluation score is obtained by adjusting the Earth Mover's Distance (EMD) (Rubner et al., 1998, 2000) to the evaluation task. The EMD measure represents the distance between two probability distributions. Moreover, the EMD distance is obtained based on a signature consisting of the feature and the weight, and the distance between two features. The feature, weight, and distance must therefore be defined to adjust EMD to the evaluation task.

Proceedings of NAACL-HLT 2019, pages 1874–1883
Minneapolis, Minnesota, June 2 - June 7, 2019.

In our proposed metric, the word embeddings and the sentence-level $tf \cdot idf$ respectively denote the feature and the weight. Consequently, our proposed metric can produce an evaluation based on the word meaning. Moreover, our proposed metric uses word position information in the distance between two word embeddings. The distance is obtained using cosine similarity and the difference of word position between the translation and reference. Results demonstrate that our proposed metric can evaluate translations also considering word order differences. We designate this new metric as **W**ord **E**mbedding-based automatic MT evaluation using **W**ord **P**osition **I**nformation (WE_WPI).

The experimentally obtained results based on the WMT16 metrics shared task (Bojar et al., 2016) demonstrated that our WE_WPI achieves the highest correlation with human judgment among several metrics: BLEU, METEOR (Banerjee and Lavie, 2005), IMPACT (Echizen-ya and Araki, 2007), and RIBES (Isozaki et al., 2010). Moreover, the correlation of WE_WPI is better than that of WE_WPI without word position information (WE). Results therefore confirmed the effectiveness of WE_WPI using word position information.

2 Related Work

Kusner et al. (2016) proposed the Word Mover's Distance (WMD) as a distance measure using word embedding and word alignment. This measure obtains the distance between two documents adjusting EMD to a document. However, it cannot accommodate differences of word order between the translation and reference. Matsuo et al. (2017) also proposed a word-alignment-based automatic evaluation metric using word embeddings for segment-level evaluation. As described in that paper, Maximum Alignment Similarity (MAS) was found to have higher correlation with human evaluation than BLEU for European-to-English, which has similar grammar structures. For Japanese-to-English, which has different grammar structures, Average Alignment Similarity (AAS) showed better correlation with human evaluation than other metrics. However, neither MAS nor AAS uses word position information. Therefore, neither can sufficiently accommodate word order differences. Actually, WE_WPI uses not only the word alignment but also word position information.

One system, DREEM (Chen and Guo, 2015), learns distributed word representations from a neural network model and from distributed sentence representations computed with a recursive autoencoder. Moreover, it uses a penalty based on translation and reference lengths. By contrast, the WE_WPI system specifically examines the difference between the word positions of the translation and reference, not the difference of lengths between the translation and reference. Therefore, it can sufficiently accommodate word order differences. Moreover, it can evaluate the translation efficiently using word embeddings of target languages without requiring large amounts of data or learning time. Our WE_WPI requires no learning of bilingual knowledge or a relation between translation and reference. It needs only a model of word embeddings in advance to apply EMD to the automatic MT evaluation task.

In a non-trained evaluation metric, MEANT 2.0 (Lo, 2017; Bojar et al., 2017) uses a distributional word vector model to evaluate lexical semantic similarity and shallow semantic parses to evaluate structural semantic similarity between the translation and reference. It is a new version of MEANT (Lo and Wu, 2011), which is a non-ensemble and untrained metric. Moreover, MEANT 2.0 - nosrl is a subversion of MEANT 2.0 to evaluate the translation for any output language by removing the dependence on semantic parsers for semantic role labeling (SRL). In that case, phrasal similarity is calculated using n-gram lexical similarities. However, MEANT 2.0 series do not specifically examine the position of each word in the translation and reference. Results show that it is difficult to deal sufficiently with language pairs for which the grammar differs. In WE_WPI, the evaluation score is calculated using the relative difference between the positions of each word in the translation and reference. Therefore, WE_WPI can evaluate translations dealing with word order in languages pairs for which the grammar differs.

3 Word Embedding-Based Automatic Evaluation Metric with Word Position Information (WE_WPI)

3.1 The Earth Mover's Distance (EMD)

3.1.1 Definitions

As described herein, we propose WE_WPI as the automatic MT evaluation metric obtained by adjusting the **E**arth **M**over's **D**istance (EMD) to the

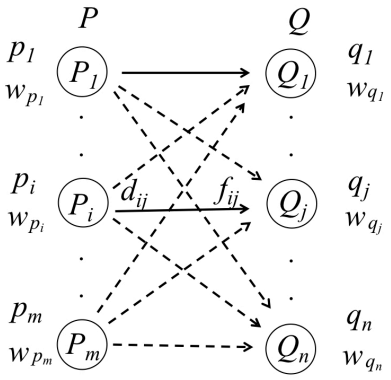

Figure 1: Outline of EMD.

automatic MT evaluation task. First, we describe EMD. Figure 1 depicts an outline of EMD.

In Figure 1, two probability distributions are presented respectively as P and Q. The P and Q consist of some P_i and Q_j, which are the respective signatures. Each signature consists of a feature (*i.e.*, p_i in P_i and q_j in Q_j) and a weight (*i.e.*, w_{p_i} in P_i and w_{q_j} in Q_j). Therefore, two probability distributions P and Q are defined respectively as $P = \{(p_1, w_{p_1})...(p_m, w_{p_m})\}$ and $Q = \{(q_1, w_{q_1})...(q_n, w_{q_n})\}$. Moreover, d_{ij} represents the distance between two features p_i and q_j.

The goal of EMD is to obtain total flow $F = [f_{ij}]$ that minimizes the overall cost from the perspective of a transportation problem. In that case, the overall cost is defined as Eq. (1).

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

Moreover, four constraints are defined for f_{ij}, which is the transportation amount in the transportation problem, to find minimum F as the following Eqs. (2)–(5):

$$f_{ij} \geq 0 \quad 1 \leq i \leq m, \quad 1 \leq j \leq n \quad (2)$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i} \quad 1 \leq i \leq m \quad (3)$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j} \quad 1 \leq j \leq n \quad (4)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right) \quad (5)$$

Constraint (2) shows that each amount of weight f_{ij} is transported only in the direction from signature P_i to signature Q_j to be nonnegative. In Constraint (3), the amount of weight which is supplied from P_i (*i.e.*, $\sum_{j=1}^{n} f_{ij}$) does not exceed w_{p_i}, which is the weight of P_i. Moreover, in Constraint (4), the amount of weight which Q_j receives (*i.e.*, $\sum_{i=1}^{m} f_{ij}$) does not exceed w_{q_j}, which is the weight of Q_j. Finally, the total amount of weight is equal to the weight of the lighter distribution in Constraint (5). In Eqs. (1)–(5), m shows the number of signatures in P; n shows the number of signatures in Q.

The EMD is defined as shown below.

$$EMD(P, Q) = \frac{min(WORK(P, Q, F))}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \quad (6)$$

In Eq. (6), the $min(WORK(P, Q, F))$ is normalized by the minimum amount of work of Eq. (5).

3.1.2 Computing EMD

P_1		P_2		P_3		P_4	
p_1	w_{p_1}	p_2	w_{p_2}	p_3	w_{p_3}	p_4	w_{p_4}
(1,5)	0.6	(5,5)	0.6	(1,1)	0.6	(5,1)	0.6

Table 1: Examples of signatures of P.

Q_1		Q_2		Q_3	
q_1	w_{q_1}	q_2	w_{q_2}	q_3	w_{q_3}
(2,3)	0.8	(4,3)	0.8	(3,2)	0.8

Table 2: Examples of signatures of Q.

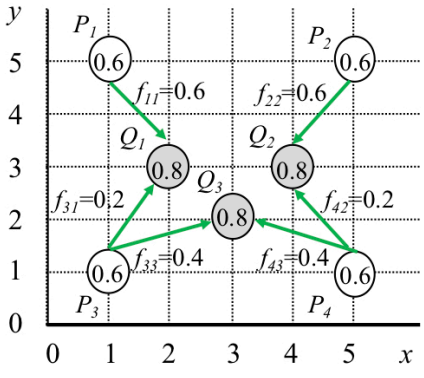

Figure 2: Example of EMD calculation.

We describe the computation of EMD using two probability distributions P and Q in two-dimensional surface. Tables 1 and 2 respectively present the examples of P and Q signatures.

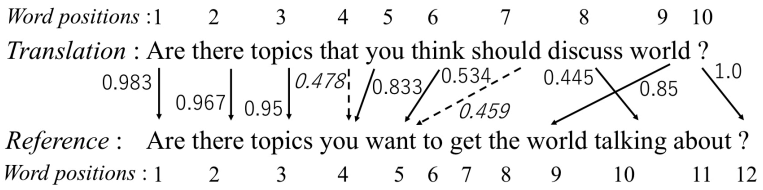

Figure 3: Example of word alignment by WE_WPI.

In Tables 1 and 2, all features p_i and q_j correspond to the coordinate (x, y) of two-dimensional surface.

Figure 2 depicts an example of an EMD calculation based on the signatures in Tables 1 and 2. In Figure 2, the green arrow indicates the amount of weight f_{ij}. All f_{ij} are transported only in the direction from P_i to Q_j according to Constraint (2). In each signature P_i, $\sum_{j=1}^{n} f_{ij}$ does not exceed w_{p_i} by Constraint (3). For example, in P_3, $\sum_{j=1}^{3} f_{3j}$ is 0.6 (=0.2+0.0+0.4). It does not exceed 0.6, which is the weight of P_3. Moreover, in each signature Q_j, $\sum_{i=1}^{m} f_{ij}$ does not exceed w_{q_j} according to Constraint (4). For example, in Q_1, $\sum_{i=1}^{4} f_{i1}$ is 0.8 (=0.6+0.0+0.2+0.0). It does not exceed 0.8, which corresponds to the weight of Q_1. The total amount of weight by $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$ is 2.4. It is equal to 2.4 by $\sum_{i=1}^{m} w_{p_i}$ or 2.4 by $\sum_{j=1}^{n} w_{q_j}$. Therefore, this example of Figure 2 conforms to Constraint (5).

Moreover, the distance between two features is necessary to obtain EMD. When the Euclidean distance is used as the calculation of distance in this example, 2.236 (=$\sqrt{1^2+2^2}$) is obtained as d_{11}, d_{22}, d_{31}, d_{33}, d_{42}, and d_{43}, and other distances are 3.606 (=$\sqrt{2^2+3^2}$) in Figure 2. As a result, 5.366 (=2.236 × (0.6+0.6+0.2+0.4+0.2+0.4)) is obtained as the value of EMD by two probability distributions P and Q in Tables 1 and 2.

We obtain WE_WPI adjusting EMD to the automatic MT evaluation task. Details of application of EMD to WE_WPI are presented in 3.2.2.

3.2 New Automatic MT Evaluation Metric: WE_WPI

3.2.1 Word Alignment using Position Information

For the application of EMD to automatic MT evaluation, we use word alignment results. Word alignment is done using cosine similarity based on word embeddings and the relative difference between the word positions in the translation and reference. In that case, WE_WPI obtains $align\_score$ using Eqs. (7) and (8) presented below.

$$align\_score = cos\_sim(t_i, r_j) \times (1.0 - pos\_inf(T_i, R_j)) \quad (7)$$

$$pos\_inf(T_i, R_j) = \left| \frac{pos(T_i)}{m} - \frac{pos(R_j)}{n} \right| \quad (8)$$

In Eq. (7), t_i and r_j respectively represent the word embeddings of word T_i in the translation and word R_j in the reference. The $cos\_sim(t_i, r_j)$ denotes the cosine similarity between t_i and r_j. Moreover, $pos\_inf(T_i, R_j)$ represents the relative difference between the position of word T_i in the translation and the position of word R_j in the reference. It is defined as Eq. (8). In Eq. (8), $pos(T_i)$ and $pos(R_j)$ respectively denote the positions of word T_i in the translation and word R_j in the reference. Actually, m and n respectively denote the word numbers in the translation and reference. The $pos\_inf(T_i, R_j)$ becomes larger as the relative difference between $pos(T_i)$ and $pos(R_j)$ becomes larger . Therefore, $(1.0 - pos\_inf(T_i, R_j))$ is used as the negative weight for $cos\_sim(t_i, r_j)$. The ranges of $cos\_sim(t_i, r_j)$ and $pos\_inf(T_i, R_j)$ are both 0.0-1.0. Figure 3 depicts an example of word alignment using Eqs. (7) and (8).

The WE_WPI calculates $align\_score$ between a word in the translation and all words in reference. Based on those results, the word with the highest $align\_score$ in the reference is selected as the corresponding word to the word in the translation. In Figure 3, the $align\_score$ between "that" in the translation and "you" in the reference is the highest (*i.e.*, 0.478) among the $align\_score$ between "that" in the translation and all words in the reference. However, it is lower than the $align\_score$ 0.833 between "you" in the translation and "you" in the reference. Therefore, the word which corresponds to "that" in the translation cannot be obtained in the reference. Similarly, the word which

		reference											
		Are	there	topics	you	want	to	get	the	world	talking	about	?
translation	Are	**0.017**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	there	1.0	**0.033**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	topics	1.0	1.0	**0.049**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	that	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	you	1.0	1.0	1.0	**0.154**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	think	1.0	1.0	1.0	1.0	**0.456**	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	should	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	discuss	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**0.555**	1.0	1.0
	world	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**0.139**	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**

Table 3: Distance matrix incorporating translation and reference.

corresponds to "should" in the translation cannot be obtained in the reference.

In contrast, "discuss" in the translation corresponds to "talking" in the reference using $pos\_inf(T_i, R_j)$ of Eq. (8) although "discuss" in the translation corresponds to "topics" in the reference when $(1.0 - pos\_inf(T_i, R_j))$ is not used in Eq. (7) (*i.e.*, $align\_score = cos\_sim(t_i, r_j)$). The 0.477, which is the cos_sim between "discuss" in the translation and "topics" in the reference, is greater than 0.460, which is the cos_sim between "discuss" in the translation and "talking" in the reference. Here, $pos\_inf(T_i, R_j)$ between "discuss" in the translation and "talking" in the reference is 0.033 ($\left|\frac{8}{10} - \frac{10}{12}\right|$). That between "discuss" in the translation and "topics" in the reference is 0.550 ($\left|\frac{8}{10} - \frac{3}{12}\right|$). Consequently, the $align\_score$ of "discuss" in the translation and "talking" in the reference is 0.445 ($0.460 \times (1.0 - 0.033)$). That of "discuss" in the translation and "topics" in the reference is 0.215 ($0.477 \times (1.0 - 0.550)$) using Eq. (7). The WE_WPI can select "talking" in the reference as the corresponding word for "discuss" in the translation using $pos\_inf(T_i, R_j)$. The use of $pos\_inf(T_i, R_j)$ is effective for the correct word alignment.

3.2.2 Adjustment of EMD to the Automatic MT Evaluation Metric

We obtain WE_WPI as new automatic MT evaluation metrics by adjusting EMD to the automatic MT evaluation task. In WE_WPI, the variables P and Q in Figure 1 respectively correspond to a translation T and reference R. Moreover, the features (*i.e.*, p_i and q_j in Figure 1), the weight (*i.e.*, w_{p_i} and w_{q_j} in Figure 1), and distance (*i.e.*, d_{ij} in Figure 1) are required as parameters to adjust EMD to the automatic MT evaluation task. As described herein, we use the word embeddings as features and the sentence-level $tf \cdot idf$ as the weight. The weight definition is presented in Eq. (9) below.

$$w = tf \times \left(\log \frac{N}{df} + 1.0\right) \tag{9}$$

In Eq. (9), tf denotes the appearance frequency of a word in a translation or reference. In addition, df represents the number of sentences in which the word appears in all translations or references. In addition, N is the total number of translations or references. Actually, WE_WPI distinguishes the function word and the content word using Eq. (9). Furthermore, w_{t_i} of the word in the translation and w_{r_i} of the word in the reference by Eq. (9) are normalized respectively using the following Eqs. (10) and (11).

$$\tilde{w}_{t_i} = \frac{w_{t_i}}{\sum_{i=1}^{m} w_{t_i}} \tag{10}$$

$$\tilde{w}_{r_j} = \frac{w_{r_j}}{\sum_{j=1}^{n} w_{r_j}} \tag{11}$$

The dependence of w in Eq. (9) by difference of dataset can be kept to the minimum by normalizing Eqs. (10) and (11). Moreover, we define distance d_{ij}, which is ascertained from the result of the word alignment described in 3.2.1. The d_{ij} is obtained using the following Eq. (12):

$$d_{ij} = \begin{cases} 1.0 - cos\_sim(t_i, r_j) \\ \qquad \times \mathrm{e}^{-pos\_inf(T_i, R_j)} \\ \qquad \text{if } T_i \text{ corresponds to } R_j \\ 1.0 \quad \text{if } T_i \text{ does not correspond to } R_j \end{cases} \tag{12}$$

In Eq. (12), $1.0 - cos\_sim(t_i, r_j) \times \mathrm{e}^{-pos\_inf(T_i, R_j)}$ is used as d_{ij} when word T_i in the translation corresponds to word R_j in the reference by the word alignment result. The $pos\_inf(T_i, R_j)$ is obtained by Eq. (8). Here, t_i and r_j respectively correspond to the word embeddings of the words in the translation and reference. The $\mathrm{e}^{-pos\_inf(T_i, R_j)}$ represents the penalty

	cs-en		de-en		fi-en		ro-en		ru-en		tr-en	
Human	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
Systems	6	6	10	10	9	9	7	7	10	10	8	8
mtevalBLEU	.992	.989	.905	.808	.858	.864	.899	.840	.962	.837	.899	.895
METEOR	.995	**.991**	.935	.887	.952	.963	.934	.909	.987	.930	.965	.980
IMPACT	**.997**	.990	.925	.841	.908	.915	.903	.819	.962	.840	.952	.959
RIBES	.995	.990	.948	.891	.894	.901	**.954**	.794	.972	.864	.850	.868
MEANT 2.0 (Lo, 2017)	.989	.990	.947	**.950**	.953	.966	.940	**.946**	.990	**.959**	**.980**	.990
MEANT 2.0 - nosrl	.985	.988	.928	.942	**.969**	**.979**	.917	.930	.984	.958	.978	.987
WE	.986	.976	.918	.903	.954	.963	.885	.884	.989	.938	.976	.991
WE_WPI	.991	.980	**.958**	.927	.955	.957	.919	.877	**.991**	.926	.977	**.993**

Table 4: Absolute Pearson correlation of to-English system-level metric with human assessment variants: RR, standard WMT relative ranking; DA, direct assessment of translation adequacy.

	en-cs		en-de		en-fi		en-ro		en-ru		en-tr	
Human	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
Systems	10		15		13		12		12	12	8	
mtevalBLEU	.968	-	.752	-	.868	-	.897	-	.835	.838	.745	-
METEOR	.960	-	.631	-	.939	-	.873	-	.868	.879	.800	-
IMPACT	**.978**	-	.719	-	.924	-	.911	-	.874	.879	.844	-
RIBES	.968	-	.742	-	**.949**	-	.910	-	.895	.904	.883	-
MEANT 2.0 (Lo, 2017)	-	-	.540	-	-	-	-	-	-	-	-	-
MEANT 2.0 - norsrl	.967	-	.541	-	.902	-	.868	-	**.925**	**.946**	.933	-
WE	.962	-	.609	-	.925	-	.878	-	.899	.910	.930	-
WE_WPI	.967	-	**.780**	-	.931	-	**.917**	-	.914	.923	**.944**	-

Table 5: Absolute Pearson correlation of out-of-English system-level metric with human assessment variants: RR, standard WMT relative ranking; DA, direct assessment of translation adequacy.

to $cos\_sim(t_i, r_j)$ because it becomes smaller as $pos\_inf(T_i, R_j)$ becomes larger. As a result, d_{ij} becomes large when the relative difference between the position of word T_i in the translation and the position of word R_j in the reference (*i.e.*, $pos\_inf(T_i, R_j)$) is large. The d_{ij} by Eq. (12) is 1.0 when word T_i does not correspond to word R_j. Finally, the range of d_{ij} becomes 0.0-1.0.

Moreover, the WE_WPI generates the distance matrix using d_{ij} in Eq. (12). Table 3 presents the distance matrix between the translation "Are there topics that you think should discuss world?" and the reference "Are there topics you want to get the world talking about?" in Figure 3. In Table 3, the bold typeface represents the distance between the two aligned words. The distance matrix using Eq. (12) is effective because it is not influenced by the words which are not aligned between the translation and reference.

The WE_WPI obtains the evaluation score by word embedding, sentence-level $tf \cdot idf$, and the distance matrix based on Eq. (12). The evaluation score of WE_WPI is obtained as Eq. (13).

$$WE\_WPI(T,R) = 1.0 - \frac{min(WORK(T,R,F))}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \quad (13)$$

In that equation, the range of $\frac{min(WORK(T,R,F))}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}$ becomes 0.0-1.0 using the weights normalized by Eqs. (10) and (11). Near 0.0, the distance between T and R is small. However, in the automatic MT evaluation metrics, the score is close to 1.0 when the evaluation for the translation is generally high. Therefore, we obtain WE_WPI by taking the value of $\frac{min(WORK(T,R,F))}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}$ from 1.0. As a result, in between the translation "Are there topics that you think should discuss world?" and the reference "Are there topics you want to get the world talking about?", 0.608 is obtained as the score using Eq. (13).

The WE_WPI can evaluate the translation based on the meanings of words using word embedding. Moreover, it can deal with the word order using the relative difference between the positions of words in the translation and the reference.

4 Experiments

4.1 Experiment Data and Procedure

We conducted evaluation experiments to confirm the effectiveness of WE_WPI. The "newstest2016" set, which is the main test set in WMT16 metrics shared task (Bojar et al., 2016), was used. The script is available at http://www.statmt.org/wmt16/results.html.

	cs-en		de-en		fi-en		ro-en		ru-en		tr-en	
Human	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
♯ Assessments	70k	12k	15k	12k	19k	14k	11k	12k	18k	13k	7k	13k
♯ Translations	8.6k	560	2.4k	560	4.6k	560	2.2k	560	4.7k	560	2.2k	560
Correlation	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r
sentBLEU	.284	.557	.368	.484	.265	.448	.272	.499	.330	.502	.245	.532
METEOR	**.391**	.636	.393	.500	.351	.539	.297	.578	.370	.541	.334	.604
IMPACT	.338	.624	.342	.535	.301	.510	.248	.531	.309	.541	.282	.602
RIBES	.254	.530	.288	.415	.237	.372	.176	.375	.240	.401	.213	.336
MEANT 2.0 (Lo, 2017)	.355	**.674**	.414	.539	**.453**	.510	**.345**	**.607**	**.401**	.535	**.373**	.588
MEANT 2.0 - nosrl	.347	.672	.411	.522	.438	.484	.338	.587	.400	.540	.364	.577
WE	.372	.617	.395	.472	.365	.517	.316	.545	.362	.523	.346	.572
WE_WPI	.387	.649	**.417**	**.548**	.361	**.540**	.308	.555	.371	**.555**	.347	**.625**

Table 6: Segment-level metric results for to-English language pairs with absolute values of correlation coefficients reported for all metrics: correlation of segment-level metric scores with human assessment variants, where τ are official results computed similarly to Kendall's τ and over standard WMT relative ranking (RR) human assessments; r are Pearson correlation coefficients of metric scores with direct assessment (DA) of absolute translation adequacy.

	en-cs		en-de		en-fi		en-ro		en-ru		en-tr	
Human	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
♯ Assessments	118k	-	35k	-	31k	-	7k	-	21k	20k	7k	-
♯ Translations	12.9k	-	6.2k	-	4.1k	-	1.9k	-	6.0k	-	3.0k	-
Correlation	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r
sentBLEU	.223	-	.269	-	.145	-	.171	-	.283	.557	.171	-
METEOR	.245	-	.268	-	.189	-	.177	-	.309	.600	.207	-
IMPACT	.240	-	.263	-	.170	-	.180	-	.297	.609	.231	-
RIBES	.139	-	.188	-	.057	-	.101	-	.206	.442	.153	-
WE	**.359**	-	.347	-	**.360**	-	**.285**	-	**.427**	.625	.336	-
WE_WPI	.352	-	**.371**	-	.357	-	.283	-	.424	**.652**	**.370**	-

Table 7: Segment-level metric results for out-of-English language pairs with the absolute values of correlation coefficients reported for all metrics: absolute correlation of segment-level metric scores with human assessment variants, where τ are official results computed similarly to Kendall's τ and over standard WMT relative ranking (RR) human assessments; r are Pearson correlation coefficients of metric scores with direct assessment (DA) of absolute translation adequacy.

Therefore, we can readily obtain the correlation coefficient between the metrics and human judgments in WMT16 metrics shared task. The WMT16 metrics task includes English paired with Czech, German, Finnish, Romanian, Russian, and Turkish. For all translations, references and scores by human judgment in these language pairs are obtained from the url described above.

For these experiments, we used different automatic MT evaluation metrics for comparison with our WE_WPI: BLEU, METEOR, IMPACT, RIBES, and WE. Here, IMPACT and RIBES, which are surface-based metrics, are effective for language pairs with greatly different word order, such as English and Japanese. In addition, WE is an automatic MT evaluation metric that does not perform word alignment. It uses only $d_{ij} = 1.0 - cos\_sim(t_i, r_j)$ as the d_{ij} of Eq. (12) in the WE_WPI. In both WE and WE_WPI, the word vectors for seven languages (*i.e.*, English, Czech, German, Finnish, Romanian, Russian, and Turkish) were obtained using fastText (Grave et al., 2018).

4.2 Experiment Results and Discussion

Tables 4 and 5 respectively present the correlation coefficient of to-English and out-of-English at the system level. Tables 6 and 7 respectively present the correlation coefficients of to-English and out-of-English at the segment level.

In Tables 4–7, RR represents the correlation based on the relative ranking by human judgment to 5 translations at a time. The bold typeface shows the highest correlation coefficient among all correlation coefficients of metrics. Moreover, the coefficients of MEANT 2.0 described in (Lo, 2017) are added to Tables 4–6. Here, WE_WPI achieves the highest correlation with human judgment in Table 5, DA in Table 6, and Table 7. Especially, the correlation coefficients of WE_WPI are high with language pairs for which the grammar differs (*i.e.*, English-to-German (en-de), German-to-English (de-en),

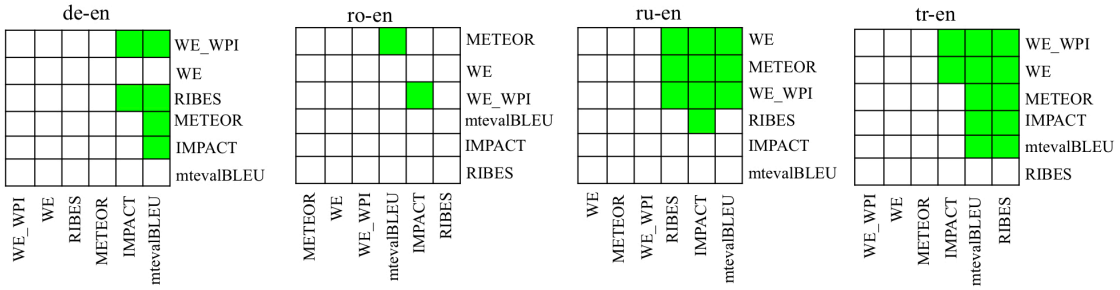

Figure 4: To-English system-level metric significance test of results for human assessment variants, where DA denotes the direct assessment of translation adequacy. Green cells show a significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to the Williams test.

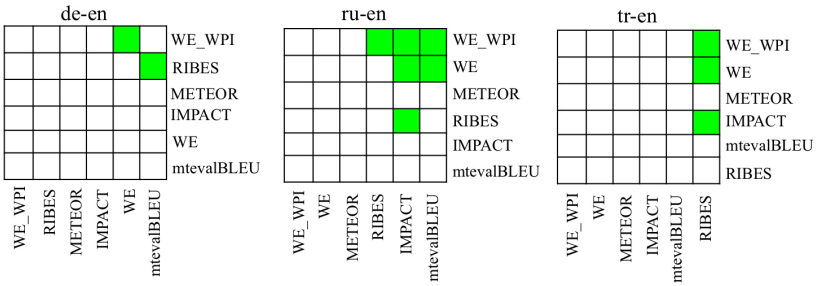

Figure 5: To-English system-level metric significance test of results for human assessment variants, where RR denotes the standard WMT relative ranking for the translation task system only. Green cells show a significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to the Williams test.

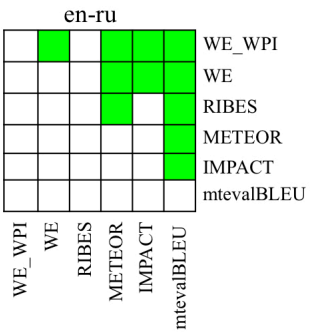

Figure 6: English-to-Russian system-level metric significance test of results for human assessment variants, where DA denotes direct assessment of translation adequacy. Green cells show a significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to the Williams test.

English-to-Turkish (en-tr), and Turkish-to-English (tr-en)). Therefore, the WE_WPI is effective with such language pairs because it uses word position information.

Moreover, we investigated the significance of WE_WPI results and those of other metrics except those of MEANT 2.0 and MEANT 2.0 - nosrl. As described herein, Williams significance test (Williams, 1959) was used to assess differences in dependent correlations. Figures 4–9 present significance test results for every competing pair of metrics, including those of our WE_WPI. However, the language pairs for which significant differences could not be obtained in any competing pair of metrics are excluded from Figures 4–9 (*i.e.*, cs-en and fi-en in Figure 4, cs-en, fi-en and ro-en in Figure 5, en-cs in Figure 7).

In Figures 4–9, green cells signify that the metric shows significant difference from other metrics with 95% or greater confidence. Results demonstrated that our WE_WPI yielded significantly different results among metrics. Particularly, WE_WPI was found to have significantly better results than those of WE at the segment level, as shown in Figures 8 and 9. This particular result demonstrates that the word position information in WE_WPI is effective for segment-level evaluation.

Moreover, WE_WPI does not need much time to calculate the scores described in 3.2.2. However, it takes time to calculate $tf \cdot idf$ of words and to change the surface-level words to the word vectors. It is efficient to calculate $tf \cdot idf$ of all words in the translations and references, and to extract the word vectors, which correspond to the words in the translations and references, from the fastText models in advance.

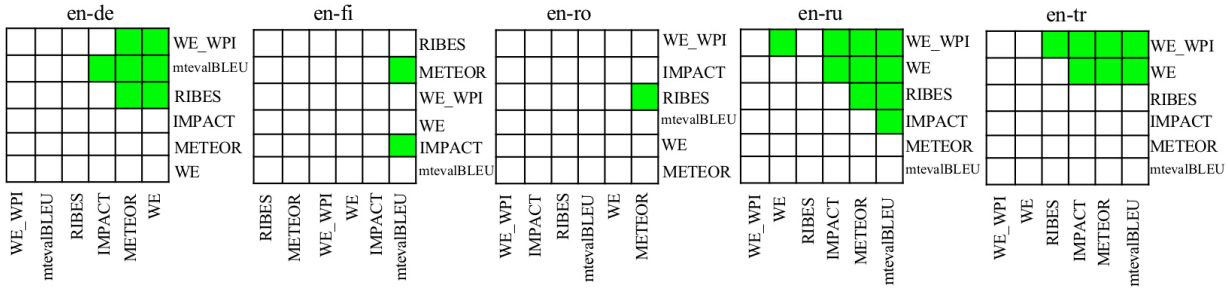

Figure 7: Out-of-English system-level metric significance test of results for human assessment variants, where RR denotes the standard WMT relative ranking for translation task system only. Green cells show a significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to the Williams test.

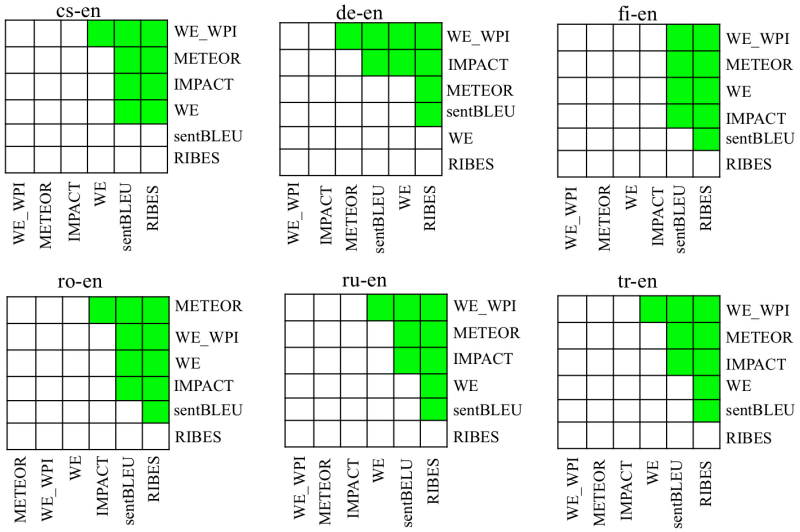

Figure 8: To-English segment-level metric significance test of results for human assessment variants, where DA denotes direct assessment of translation adequacy. Green cells show marked benefits obtained with the metric in a given row over the metric in a given column according to the Williams test.

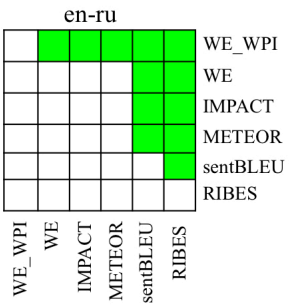

Figure 9: English-to-Russian segment-level metric significance test of results for human assessment variants, where DA denotes direct assessment of translation adequacy: green cells show marked benefits obtained with the metric in a given row over the metric in a given column according to the Williams test.

5 Conclusion

As described herein, we proposed WE_WPI as a new automatic MT evaluation metric. It produces an evaluation based on the meanings of words using word embedding. Moreover, it can accommodate word-order differences. Evaluation experiments demonstrated that our WE_WPI obtains the highest correlation with human judgments among several representative metrics in language pairs for which the grammar differs, and demonstrated that it is significantly better than other metrics at segment-level evaluation.

Our future work will improve WE_WPI to obtain high-quality evaluation scores in combination with other metrics. We will conduct evaluation experiments using various data. Moreover, we will use WE_WPI to improve NMT quality. For instance, WE_WPI can be used easily in Minimum Risk Training (MRT) (Shen et al., 2016), which minimizes the expected loss on the training data.

Acknowledgments

This work was partially supported by grants from Hokkai-Gakuen University.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Result of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation (WMT17)*, pages 489–513.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Result of the wmt16 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 199–231.

Boxing Chen and Hongyu Guo. 2015. Representation based translation evaluation metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 150–155.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 263–270.

Hiroshi Echizen-ya and Kenji Araki. 2007. Automatic evaluation of machine translation based on recursive acquisition of an intuitive common parts continuum. In *Proceedings of the Eleventh Machine Translation Summit*, pages 151–158.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2016. From word embeddings to document distance. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 957–966.

Chi–kiu Lo. 2017. Meant 2.0: Accurate semantic mt evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation (WMT17)*, pages 589–597.

Chi–kiu Lo and Dekai Wu. 2011. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, pages 220–229.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Junki Matsuo, Mamoru Komachi, and Katsuhito Sudoh. 2017. Word-alignment-based segment-level machine translation evaluation using word embeddings. arXiv:1704.00380. Version 1.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Gubas. 1998. A metric for distributions with applications to image databases. In *Proceedings of ICCV*, pages 59–66.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Gubas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(12):99–121.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1683–1692.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Neural Information Processing Systems*, arXiv:1409.3215. Version 3.

Dan Tufiş and Alexandru Ceauşu. 2009. Factored phrase-based statistical machine translation. In *Proceedings of the 5th Conference Speech Technology and Human-Computer Dialogue*, pages 1–7.

Evan James Williams. 1959. Regression analysis. 14. Wiley New York.