STD: An Automatic Evaluation Metric for Machine Translation Based on Word Embeddings

Pairui Li, Chuan Chen, Wujie Zheng, Yuetang Deng, Fanghua Ye, and Zibin Zheng

***Abstract*—Lexical-based metrics such as BLEU, NIST, and WER have been widely used in machine translation (MT) evaluation. However, these metrics badly represent semantic relationships and impose strict identity matching, leading to moderate correlation with human judgments. In this paper, we propose a novel MT automatic evaluation metric *Semantic Travel Distance* (STD) based on word embeddings. STD incorporates both semantic and lexical features (word embeddings and *n*-gram and word order) into one metric. It measures the semantic distance between the hypothesis and reference by calculating the minimum cumulative cost that the embedded *n*-grams of the hypothesis need to "travel" to reach the embedded *n*-grams of the reference. Experiment results show that STD has a better and more robust performance than a range of state-of-the-art metrics for both the segment-level and system-level evaluation.**

***Index Terms*—Machine translation evaluation, metric, semantic, word embeddings, earth mover's distance, n-gram, word order.**

I. Introduction

MACHINE Translation (MT) evaluation has always been one of the main concerns of researchers in the field of MT. The better evaluation metrics lead to the development of better MT systems [1]. However, MT evaluation is difficult and challenging since natural languages are highly ambiguous and different languages do not always express the same content in the same way [2].

Human evaluation, although highly reliable, encounters inconsistency problems due to both inter- and intra-annotator agreement issues [3]. Moreover, it is extremely expensive and time consuming. To provide more consistent, cheaper and much faster measurements of the performance of MT systems, various automatic evaluation metrics have been proposed in recent years, which assume availability of human translated references and attempt to compare the hypothesis (translation outputs) against the references in different ways.

BLEU [4] has been widely used in MT evaluation due to its ease of implementation, language independence and competitive performance in capturing the translation fluency. It is based on the n-gram matching of the reference and hypothesis. Other metrics such as WER [5], PER [6], NIST [7] and TER [8] have also been commonly used. They mainly focus on the exact matches of the surface words in the output translation. WER, PER and TER compute the edit distance between the hypothesis and reference by calculating the minimum number of editing steps to transform hypothesis to reference; Similar to BLEU, NIST measures the degree of n-gram overlapping between the hypothesis and reference. However, as they mainly focus on the lexical level, they employ strict string matchings between the hypothesis and reference, leading to poor correlation with human judgments. For example, the semantically related words "vacation" and "holiday" and words that differ only by morphological markers, such as "holiday" and "holidays" are considered different words although they have similar meanings. The traditional solution for improving their performance is to use more references, while multiple references are rare and expensive. Besides, these n-gram-based evaluations have proved to be biased towards statistical methods, mainly because they do not allow grammatically-constrained lexical freedom. Moreover, they are mainly suitable for evaluating MT systems' overall performance (system-level evaluation), but unfit for evaluating single translation's quality (segment-level evaluation), which is shown in Table VIII.

To cover the shortages of these lexical-based metrics, in this paper, we introduce a novel metric, which we call the Semantic Travel Distance (STD), for MT evaluation at both system and segment levels. Our approach utilizes word embeddings [9]–[11], which have played an important role in Natural Language Processing (NLP) and Machine Translation (MT). [9] proves that distances between embedded word vectors are to some degree semantically meaningful. To capture the semantics of the translation, we construct a novel document representation and a semantic distance matrix based on n-gram embeddings. More specifically, n-gram embeddings are generated by concatenating the embedding vectors of n words. Besides, to capture word order information in the translation output, we construct a n-gram order distance matrix. The final n-gram distance matrix is the weighted mean of the semantic distance matrix and n-gram order distance matrix. Then, we measure the dissimilarity

Manuscript received January 9, 2019; revised April 22, 2019; accepted June 10, 2019. Date of publication June 14, 2019; date of current version June 28, 2019. This work was supported in part by the Technical Architecture Department at WeChat Group of Tencent, in part by the National Natural Science Foundation of China under Grants 11801595 and 61802450, in part by the Natural Science Foundation of Guangdong under Grant 2018A030310076, in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2016ZT06D211, and in part by the CCF Opening Project of Information System. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Taro Watanabe. *(Corresponding author: Chuan Chen.)*

P. Li, C. Chen, F. Ye, and Z. Zheng are with the National Engineering Research Center of Digital Life, School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: lipr5@mail2.sysu.edu.cn; chenchuan@mail.sysu.edu.cn; yefh5@mail2.sysu.edu.cn; zhzibin@mail.sysu.edu.cn).

W. Zheng and Y. Deng are with the Technical Architecture Department, Wechat Group, Tencent, Guangzhou 510433, China (e-mail: wujiezheng@tencent.com; yuetangdeng@tencent.com).

Digital Object Identifier 10.1109/TASLP.2019.2922845

between the hypothesis and reference by calculating the minimum amount of distance that the embedded n-grams of hypothesis need to "travel" to reach the embedded n-grams of reference. The optimization problem underlying STD reduces to a special case of the well-studied Earth Mover's Distance (EMD) [12] transportation problem and we can leverage existing literature on fast specialized solvers [13]. Finally, for system-level evaluation, we compute the weighted average of the unigram STD score and bigram STD score; for segment-level evaluation, we compute the arithmetic average of the unigram STD score and bigram STD score. If more than one reference is available, the given translation is scored against each reference independently, and the best score is reported.

The STD metric has several intriguing properties:

- It evaluates the semantic dissimilarity between the hypothesis and reference rather than strict string matchings;
- It considers word order information by employing n-gram embeddings and n-gram order distance matrix;
- It is derivable for most languages since it only depends on the availability of word embeddings;
- It outperforms a range of state-of-the-art metrics in MT evaluation tasks for both the segment-level and system-level evaluation;
- Its performance can be easily improved by utilizing higher quality word embeddings.

The remainder of this paper is structured as follows. In Section II, we provide a brief review of the related works on MT evaluation metrics. In Section III, we present the proposed model and algorithm for the novel metric – STD. In Section IV, experiment results for MT evaluation are given to demonstrate the superior performance of the proposed metric to other state-of-the-art metrics. Finally, some concluding remarks are given in Section V.

II. Related Works

Due to the disadvantages of manual evaluation, such as time-consuming, expensive, non-tunable, and non-reproducible, automatic evaluation metrics have been widely used for MT [14]. Papineni [4] proposes BLEU as an automatic MT evaluation metric which is based on the n-gram matching of the reference and hypothesis documents. This is still considered as the most reliable metric and it is used extensively in the MT community to evaluate the translation quality. BLEU averages the precision for unigram, bigram and up to 4-gram and employs a length penalty if the hypothesis is shorter than the best matching (in length) reference. Generally, the final score for a translation is the geometric mean of partial scores for up to 4-gram level, which means that if a hypothesis does not have at least one 4-gram in common with the reference, it will receive a score of zero. Since the scores for many weaker or shorter hypotheses are unfairly and artificially levelled down if the longest substrings in common with the reference are shorter than four words, BLEU is not specifically well suited for segment-level scoring. BLEU's smoothed version [15] uses an add-one technique (where we add one to every n-gram count) to deal with this problem, which shows little improvement over the original.

Alternative approaches have been designed to address problems with BLEU. [7] proposes NIST metric, which is derived from the BLEU evaluation criterion but differs in one fundamental aspect: instead of n-gram precision, the information gain from each n-gram is taken into account. TER [8] tries to improve the lexical matching process based on the edit distance. ROUGE [16] captures common subsequences. ATEC [17] and PORT [18] utilize the word order information to measure translation quality. METEOR [19] creates an alignment between the reference and hypothesis strings by word-to-word matches. Words can be matched based on their surface forms, stemmed forms, and meanings, which overcomes the shortcoming of strict identity matching in BLEU. After the alignment, METEOR computes the score using a combination of precision, recall, and a penalty to capture word order information. However, the synonymy matching used in METEOR is based on WordNet, which depends on some pre-existing theoretical knowledge, and has poor adaptability to other languages except for English.

Character-based metrics also overcome the issue to some extent. ChrF [20] uses character n-gram F-score for automatic evaluation of MT output. It is simple, language independent and also tokenization independent. BEER [21] also utilizes character n-gram to evaluate lexical accuracy. Besides, it uses hierarchical representations based on PETs (permutation trees) to measure word order. It is a trained metric which depends on the quality of training data to some extent. CharacTer [22] calculates the character level edit distance and performs the shift edit on word level. However, since these metrics are based on character level, they are unsuitable for morphologically poor languages, such as Chinese and Japanese.

Some metrics have been proposed to utilize syntactic information. They usually employ the morphological part-of-speech information, phrase similarity, or sentence structure generated by the linguistic tools such as language parser or chunker. [23] presents three metrics that use syntactic structure and unlabelled dependency information. It uses the subtree kernel introduced in [24] to calculate the similarity between the reference and the candidate translations. [25] proposes a method for obtaining more details about actual translation errors in the generated output by introducing the decomposition of WER and PER over different Part-of-Speech (POS) classes. [26] and [27] incorporate shallow syntactic structures into machine translation metrics. [28] proposes the UHH metric based on sequence and tree kernel functions. Since this group of metrics highly relies on existing syntactic knowledge of languages, it is complicated and impractical especially when faced with a little-known and complex language.

Apart from utilizing syntactic information, there is also a common tendency to capture semantic meanings of translations. One established way is to apply additional linguistic knowledge, such as synonym dictionaries. For example, TER-Plus [29] uses WordNet [30] to compute synonym matches in addition to the four original operations (Insertion, Deletion, Substitution and Shift). [31] uses discourse structure and design discourse-aware similarity measures. MEANT [32] and its variants [33], [34] evaluate translation adequacy by measuring the similarity of the semantic frames and their role fillers between the human reference and machine translations. However, these metrics are inadaptable since they highly depend on available linguistic knowledge, which is difficult to obtain for numerous languages.

Another way to capture semantic meanings is by integrating vector representations especially word embeddings into MT evaluation. [35] uses Latent Semantic Indexing to project sentences as bag-of-words into a low-dimensional continuous space to measure the adequacy on a hypothesis. A monolingual continuous space and a cross-language continuous space have been used to capture the similarity between the hypothesis and reference. With the same idea, [36] proposes a Bayesian Ridge Regressor which uses document-level embeddings as features and METEOR score as target to predict the adequacy of hypothesis. The study of [37] uses vector representation more directly, in which each sentence has been transformed into a vector (they tried 3 kinds of vector representation: one-hot, word embedding and recursive auto-encoder representations). The evaluation score is calculated by the distance between the hypothesis vector and the reference vector, with a length penalty. More recently, [38] combines word embeddings and DBnary [39], a multilingual lexical resource, to enrich METEOR. [40] implements a fuzzy match score for n-gram precisions in the BLEU metric with n-gram word2vec embeddings.

The use of EMD has been pioneered in the computer vision literature [41], [42]. Some publications have studied the EMD approximation for image retrieval applications [43]–[45]. In NLP, [46] proposes the Word Mover's Distance (WMD), which utilizes EMD to measure document dissimilarity. [47] builds EMD on bilingual word embeddings for Machine Translation. To our knowledge, our work is the first to use EMD in MT evaluation.

III. Semantic Travel Distance (STD)

For MT evaluation, our guiding principle is that a good translation should be semantically similar to the references - representing the core meaning of the source utterances. Therefore, STD is proposed to evaluate MT systems by measuring semantic distance between the reference and hypothesis based on word embeddings. Besides, it utilizes n-gram embeddings and n-gram order distance to capture word order features. STD employs the Earth Mover's Distance to compute the minimum cumulative cost that the embedded n-grams of hypothesis need to "travel" to reach the embedded n-grams of reference.

The framework of the proposed STD metric consists of the following three steps:

1) Transforming the reference and the hypothesis documents into vector representations respectively;
2) Constructing a weighted graph based on the representations and word embeddings;
3) Measuring the distance between the reference and hypothesis by utilizing EMD solvers based on the weighted graph.

A. Document Representation

A document can be represented as a bag of weighted n-grams and the weight vector can be considered to be the representation of the document. The two most common ways to represent documents are via a normalized bag-of-words (nBOW) and by their term frequency-inverse document frequency (TF-IDF). However, these two representations are not suitable for document similarity due to their frequent near-orthogonality [48], [49]. Another significant drawback of these representations is that they do not capture the semantics of documents. As a document can be represented as a bag of weighted n-grams, n-grams that are semantically more similar to a document tend to represent it better. Based on this idea, we propose a novel document representation - normalized Similarity representation (nSIM) in the following.

Given a reference document and a hypothesis document in a translation, first we lowercase and tokenize the two documents into two sets of words respectively. Then, we transform the two documents into two sets of n-grams, namely S_{ref} and S_{hyp} respectively. The vocabulary V of two documents can be obtained by $V = S_{ref} \cup S_{hyp}$. Assume we are provided with an n-gram embedding matrix $X \in \mathbb{R}^{d \times n}$ for V, where d denotes the dimension of embedding vectors, n denotes the size of vocabulary, and the i^{th} column x_i represents the embedding of i^{th} n-gram in V. x_i is the concatenation of the embedding vectors (provided with the pretrained word embeddings) of n words in the i^{th} n-gram. For example, the embedding of bigram "eat apple" is generated by concatenating the word embeddings of "eat" and "apple".

Let $nSIM \in \mathbb{R}^{1 \times n}$ be the nSIM representation of a hypothesis/reference document, where the i^{th} column $nSIM_i$ is defined as follows:

$$nSIM_i = softmax(SIM_i) = \frac{e^{SIM_i}}{\sum_{i=1}^{n} e^{SIM_i}}, i \in \{1, \ldots, n\},$$

where,

$$SIM_i = \begin{cases} 1, w_i \in S_{doc} \\ sim_{\text{Cosine}}(x_i, X_{doc}) = \frac{x_i \cdot X_{doc}}{\|x_i\| \times \|X_{doc}\|}, w_i \notin S_{doc} \end{cases}, \tag{1}$$

where w_i is the i^{th} n-gram in V, x_i is the n-gram embedding of w_i, and X_{doc} is the centroid embedding of S_{doc}. Specifically, $S_{doc} = S_{hyp}$ for hypothesis representation; $S_{doc} = S_{ref}$ for reference representation. Assume that S_{doc} contains m n-grams. X_{doc} is defined as:

$$X_{doc} = \frac{1}{m} \sum_{i=1}^{n} a \cdot x_i, i \in \{1, \ldots, n\},$$

$$a = \begin{cases} 1, w_i \in S_{doc} \\ 0, w_i \notin S_{doc} \end{cases}. \tag{2}$$

For a better understanding of nSIM representation's superiority to nBOW representation, we take an example for further analysis in Table I. Assume that we have a reference document: "The man will fly to France for a holiday" and a hypothesis document: "The guy is scheduled to France for a vacation". Take unigrams' result as an instance. The nBOW representation and our nSIM representation are shown in Table I. Moreover, the cosine similarity scores between the reference (REF) representation and the hypothesis (HYP) representation are calculated for nBOW and nSIM respectively. For example, since the word "holiday" in the reference document does not appear in the hypothesis document, the nBOW value of "holiday" in the hypothesis is 0.

TABLE I
A TRANSLATION EXAMPLE'S nBOW REPRESENTATION & nSIM REPRESENTATION AND THEIR COSINE SIMILARITIES BETWEEN REFERENCE AND HYPOTHESIS

Reference: The man will fly to France for a holiday.
Hypothesis: The guy is scheduled to France for a vacation.

Representation	Cosine Similarity	Translation	to	will	a	scheduled	for	the	holiday	vacation	man	france	fly	is	guy
nBOW	0.556	REF	0.111	0.111	0.111	0	0.111	0.111	0.111	0	0.111	0.111	0.111	0	0
		HYP	0.111	0	0.111	0.111	0.111	0.111	0	0.111	0	0.111	0	0.111	0.111
nSIM	0.912	REF	0.090	0.090	0.090	0.045	0.090	0.090	0.090	0.050	0.090	0.090	0.090	0.049	0.046
		HYP	0.090	0.048	0.090	0.090	0.090	0.090	0.050	0.090	0.048	0.090	0.045	0.090	0.090

However, even though "holiday" does not appear in the hypothesis, it is still semantically correlated to the hypothesis, which means the word "holiday" can represent the meaning of hypothesis to some extent. To solve this issue, in Equation 1, we measure the semantic similarity between the hypothesis and each word in V if the word does not appear in S_{hyp}. In addition, we employ the softmax function to normalize the representation vector. As illustrated in Table I, the cosine similarity scores of nSIM and nBOW are 0.912 and 0.556 respectively, which means nSIM representation can better utilize semantics of documents since the hypothesis is semantically close to the reference in this example.

Algorithm 1: Algorithm for n-gram Order Distance.

```
Input:
The n-gram list of reference document, $L_{ref}$;
The n-gram list of hypothesis document, $L_{hyp}$;
The n-gram vocabulary of the two documents, $V$.
Output:
N-gram order distance matrix $O$.
[1] Build n-gram order mapping dictionary $d_{ref}$ and
 $d_{hyp}$;
for $i = 0$; $i < len(L_{ref})$; $i++$ do
    $d_{ref}[L_{ref}[i]] = i + 1$
end
for $i = 0$; $i < len(L_{hyp})$; $i++$ do
    $d_{hyp}[L_{hyp}[i]] = i + 1$
end
[2] Initialize $O$ with zeros;
[3] Calculate n-gram order distance;
for $i = 0$; $i < len(V)$; $i++$ do
    for $j = 0$; $j < len(V)$; $j++$ do
        if $(V[i] \in L_{ref})$ & $(V[j] \in L_{hyp})$ &
         $(O[i,j] = 0)$ then
            $O[i,j] = |d_{ref}[V[i]]/len(L_{ref}) -$
             $d_{hyp}[V[j]]/len(L_{hyp})|$
        end
        if $(V[i] \in L_{hyp})$ & $(V[j] \in L_{ref})$ &
         $(O[i,j] = 0)$ then
            $O[i,j] = |d_{hyp}[V[i]]/len(L_{hyp}) -$
             $d_{ref}[V[j]]/len(L_{ref})|$
        end
    end
end
return $O$
```

B. Graph Construction

To achieve the ultimate goal, i.e., to evaluate the translation quality by calculating the distance between the reference and hypothesis, the Earch Mover's Distance (EMD) [12], [50] is adopted to compute the documents' dissimilarity. Given the ground distance (a distance measure between single features) in a feature space, EMD is an algorithm to measure dissimilarity between two multi-dimensional distributions in a feature space, where the ground distance, a distance measure between single features, is given. EMD "moves" this distance from single features to full distributions. Intuitively, given two distributions, one can be thought of as a properly distributed mass of earth in space and the other as a collection of holes in the same space. Then, it measures the minimum amount of work needed to fill the holes with earth. In this case, a unit of work is equivalent to moving a unit of earth by a unit of ground distance. The advantages of EMD [51] over other distribution distance functions are: 1. EMD applies to variable size signatures, which includes fixed size histograms; 2. The cost of moving "earth" in EMD embodies the concept of nearness appropriately, without the quantization problems required by most current methods; 3. EMD allows partial matching in an extremely natural way.

In our context, the distributions are represented by the sets of weighted n-grams, where the weights are obtained from the nSIM representations. A weighted graph is constructed to model the dissimilarity between the reference document and the hypothesis document. And EMD is employed to compute the minimum cost of the weighted graph as the distance between two documents. Given a hypothesis document, a reference document, the vocabulary $V = S_{ref} \cup S_{hyp}$, as well as their nSIM representations, namely $nSIM$ and $nSIM'$, we construct a weighted graph as follows:

- Let $H = \{(w_1, nSIM_1), (w_2, nSIM_2), \ldots, (w_i, nSIM_i)\}$ be the representation of the hypothesis document, w_i denotes an n-gram in vocabulary V and $nSIM_i$ is the weight for w_i, calculated as the i^{th} column of the $nSIM$ representation;
- Let $R = \{(w_1, nSIM'_1), (w_2, nSIM'_2), \ldots, (w_i, nSIM'_i)\}$ be the representation of the reference document, w_i denotes an n-gram in vocabulary V and $nSIM'_i$ is the weight for w_i, calculated as the i^{th} column of the $nSIM'$ representation;
- Let $D = [d_{ij}]$ be the ground distance matrix where d_{ij} is the ground distance between n-gram i and n-gram j. By utilizing n-gram embeddings and n-gram order, D is calculated as the weighted average of the semantic distance matrix $S = [s_{ij}]$ and the n-gram order distance matrix $O = [o_{ij}]$. More specifically, $D = 0.6 * S + 0.4 * O$, where O

is defined in Algorithm 1 and S is defined as:

$$s_{ij} = 1 - \max\left(\frac{x_i \cdot x_j}{\|x_i\| \times \|x_j\|}, 0\right) \tag{3}$$

• Let $G = \{H, R, D\}$ be the weighted graph constructed by H, R and D, where H and R represent an equal number of nodes with different weight and D represents the weight of the edges between H and R.

C. Distance Calculation

After constructing the weighted graph G, first, we allow each n-gram i in H to "travel" to at least one n-gram in R. Let $F \in \mathbb{R}^{n \times n}$ be a (sparse) flow matrix where $F_{ij} \geq 0$ denotes *how much* of the weight of n-gram i in H travels to n-gram j in R. To transform H entirely into R we ensure that the entire outgoing flow from n-gram i equals its weight $nSIM_i$, i.e. $\sum_j F_{ij} = nSIM_i$. Further, the amount of incoming flow to n-gram j must match $nSIM'_j$, i.e. $\sum_i F_{ij} = nSIM'_j$. Finally, we can define the dissimilarity between the reference document and the hypothesis document as the minimum (weighted) cumulative cost required to move all n-grams from H to R in graph G, i.e. $\sum_{i,j} F_{ij} d_{ij}$.

Formally, the minimum cumulative cost of moving H to R given the constraints is provided by the solution to the following linear program,

$$\begin{aligned} \min_{F \geq 0} \quad & \sum_{i,j=1}^{n} F_{ij} d_{ij}, \\ \text{subject to:} \quad & \sum_{j=1}^{n} F_{ij} = nSIM_i, \quad \forall i \in \{1, \ldots, n\}, \\ & \sum_{i=1}^{n} F_{ij} = nSIM'_i, \quad \forall j \in \{1, \ldots, n\}. \end{aligned} \tag{4}$$

Since the above optimization is a special case of the earth mover's distance, a well studied transportation problem for which specialized solvers have been developed [13], [52]. Note that the solution to the transportation problem is unique. Once the transportation problem is solved, and we have found the optimal flow F^{opt}, the STD score corresponding to n-gram is calculated as:

$$STD_{n-gram} = EMD(H, R) = \sum_{i,j} F_{ij}^{opt} d_{ij}. \tag{5}$$

Formally, the STD scores for segment-level and system-level MT evaluation are defined as follows:

For segment level:

$$STD = 0.5 * STD_{unigram} + 0.5 * STD_{bigram}. \tag{6}$$

For system level:

$$STD = 0.3 * STD_{unigram} + 0.7 * STD_{bigram}. \tag{7}$$

Example 1

Reference: The man will fly to France for a holiday.

Hypothesis: The guy is scheduled to France for a vacation.

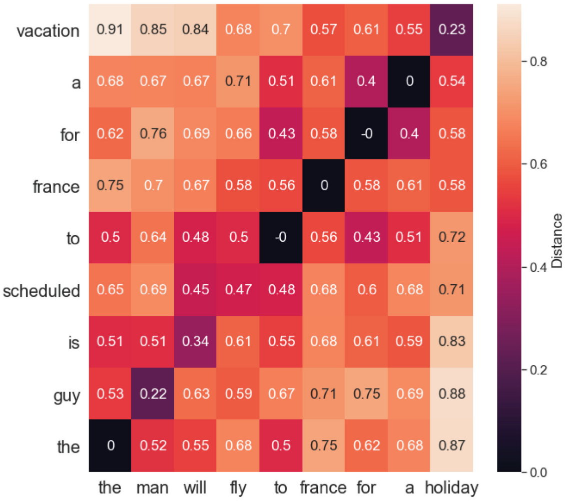

Fig. 1. The heatmap of ground distance matrix D for example 1. The x-axis refers to the reference sentence; the y-axis refers to the hypothesis sentence. The value in the boxes refers to the ground distance between unigrams.

Example 2

Reference: The man will fly to France for a holiday.

Hypothesis: Vacation a for France to scheduled is the guy.

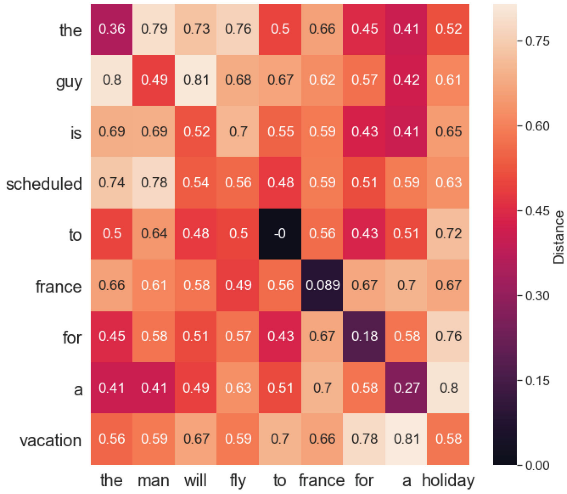

Fig. 2. The heatmap of ground distance matrix D for example 2. The x-axis refers to the reference sentence; the y-axis refers to the hypothesis sentence. The value in the boxes refers to the ground distance between unigrams.

Since STD_{bigram} performs much better than $STD_{unigram}$ for the system-level evaluation, we empirically attach more weight to STD_{bigram} and compute the weighted mean instead of the arithmetic mean. Note that STD score is in the range of [0,1]. As STD measures the distance between the hypothesis and reference, the lower the value of STD is, the better the translation quality is. If more than one reference is available, the given translation is scored against each reference independently, and the best score is reported.

D. Visualization

To better illustrate the STD metric, we present two intuitional examples in Figure 1 and Figure 2, where we plot the heatmap of the ground distance matrix for example 1 and example 2. Note that the x-axis refers to the reference sentence; the y-axis

TABLE II
THE ABBREVIATION FOR LANGUAGE PAIRS

Language pair	Abbreviation
Czech-English	cs-en
German-English	de-en
Finnish-English	fi-en
Russian-English	ru-en
Romanian-English	ro-en

TABLE III
TOOLS USED TO COMPUTE THE COMPARED METRICS IN OUR EXPERIMENT

Metrics	URL
NIST, TER, PER BLEU,sentBLEU,WER	http://github.com/moses-smt/mosesdecoder
WORF1,2,3, CHRF1,2,3	https://github.com/m-popovic/chrF
CharacTer	https://github.com/rwth-i6/CharacTER
BEER	https://github.com/stanojevic/beer

refers to the hypothesis sentence; the value in the boxes refers to the ground distance between unigrams. Two examples have the same reference sentence: “The man will fly to France for a holiday.” The difference is that the hypothesis sentence in example 2 is the reverse of that in example 1. Let’s consider the first example in Figure 1. The ground distance between semantically similar unigrams is small. For instance, the distance between the unigram “vacation” and the unigram “holiday” is 0.23, which is relatively small in this case. However, with the hypothesis of reversed unigram order, example 2 shows different result. For instance, the ground distance between “vacation” and “holiday” becomes 0.58. As we mentioned in Section III-B, the ground distance of STD is the weighted mean of the semantic distance and word order distance. Even though “vacation” and “holiday” are semantically similar, their word order distance is large. By comparing the results between example 1 and example 2, we can see that the STD metric capture not only semantic information, but also word order features.

IV. RESULTS

We evaluate STD on two machine translation evaluation datasets. We first describe each dataset and a range of classic and state-of-the-art metrics. We then compare the performance of STD and the competing metrics on these datasets. Finally, we examine how different word embeddings affect the performance of STD. Python code for the STD metric is available at https://github.com/Lipairui/Semantic-Travel-Distance.

A. Experiment Setup

We evaluated our metric STD on two metric task datasets: WMT-15 [53] and WMT-16 [54]. Each dataset contains four to-English language pairs: Czech, German, Finnish and Russian to English in WMT-15; Romanian, German, Finnish and Russian to English in WMT-16. The abbreviation for language pairs used in the experiment are described in Table II. Using Direct Assessment (DA) [55] as “golden truths”, we employed absolute value of Pearson Correlation Coefficient (r) to measure the correlation between metrics and human judgement at both system level and segment level. The Pearson Correlation Coefficient is as follows:

For system level:

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}}, \quad (8)$$

where H_i is the human judgment score of the i^{th} system in a given translation direction, M_i is the corresponding score as predicted by a given metric. $\overline{H}$ and $\overline{M}$ are their means respectively.

For segment level:

$$r = \frac{\sum_{i=1}^{n}(H'_i - \overline{H'})(M'_i - \overline{M'})}{\sqrt{\sum_{i=1}^{n}(H'_i - \overline{H'})^2}\sqrt{\sum_{i=1}^{n}(M'_i - \overline{M'})^2}}, \quad (9)$$

where H'_i is the human judgment score of the i^{th} translation data (reference-hypothesis document pair) in a given translation direction, M'_i is the corresponding score as predicted by a given metric. $\overline{H'}$ and $\overline{M'}$ are their means respectively.

The word embedding used in our STD implementation is the freely-available *fastText* word embedding[1] [11], which has 2 million word vectors trained on Common Crawl (600B tokens). Besides, we utilize PyEMD[2] to calculate the earth mover’s distance, which is a Python wrapper for Ofir Pele and Michael Werman’s implementation of the Earth Mover’s Distance [13], [56].

For a better understanding of the general performance of our metric, we compare STD with a range of state-of-the-art metrics: sentBLEU, WordF1,2,3, ChrF1,2,3, CharacTer and BEER at the segment level. As for system-level evaluation, we employ BLEU, NIST, TER, PER WER, WordF1,2,3, ChrF1,2,3, CharacTer and BEER. To show the results more clearly, we highlight the optimal and suboptimal performance in bold and underline respectively. The tools we use to compute are described in Table III. A brief introduction of the metrics used in our experiment is listed below:

– BLEU [4]: This metric is based on the n-gram matching of the reference and the hypothesis, which averages the precision for unigram, bigram and to 4-gram and applies a length penalty if the generated sentence is shorter than the best matching (in length) reference translation.

– sentBLEU [16], [57]: This metric is BLEU’s smoothed version for the segment-level evaluation, which uses an add-one technique based on BLEU metric.

– NIST [7]: This metric is derived from the BLEU evaluation criterion but differs in one fundamental aspect: instead of n-gram precision, the information gain from each n-gram is taken into account.

– TER [8]: This metric is defined as the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references.

[1] https://fasttext.cc/docs/en/english-vectors.html
[2] https://github.com/wmayner/pyemd

TABLE IV
SEGMENT-LEVEL ABSOLUTE PEARSON CORRELATION COEFFICIENT BETWEEN AUTOMATIC METRICS AND DA HUMAN ASSESSMENT ON THE **WMT-15** DATASET; THE OPTIMAL AND SUBOPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINE RESPECTIVELY

Language pair	**cs-en**	**de-en**	**fi-en**	**ru-en**	Average
Dataset size	500	500	500	500	–
sentBLEU	0.474	0.528	0.527	0.534	0.516
WordF1	0.526	0.554	0.549	0.578	0.552
WordF2	0.537	0.557	0.563	0.592	0.562
WordF3	0.539	0.556	0.565	0.594	0.563
ChrF1	0.539	0.598	0.564	0.596	0.574
ChrF2	0.551	0.604	0.573	0.610	0.584
ChrF3	0.552	0.603	0.572	0.611	0.585
BEER	0.547	0.590	0.586	0.621	0.586
CharacTer	0.534	**0.613**	0.583	0.617	0.587
STD	**0.577**	0.603	**0.652**	**0.644**	**0.619**

TABLE V
SEGMENT-LEVEL ABSOLUTE PEARSON CORRELATION COEFFICIENTS BETWEEN AUTOMATIC METRICS AND DA HUMAN ASSESSMENT ON THE **WMT-16** DATASET; THE OPTIMAL AND SUBOPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINE RESPECTIVELY

Language pair	**fi-en**	**ru-en**	**ro-en**	**de-en**	Average
Dataset size	560	560	560	560	–
WordF1	0.435	0.497	0.508	0.464	0.476
sentBLEU	0.448	0.502	0.499	0.484	0.483
WordF2	0.445	0.503	0.522	0.471	0.485
WordF3	0.447	0.504	0.525	0.473	0.487
ChrF1	0.454	0.522	0.570	0.452	0.499
BEER	0.462	0.533	0.551	0.471	0.504
ChrF2	0.457	0.534	0.581	0.469	0.510
ChrF3	0.455	0.535	**0.582**	0.472	0.511
CharacTer	0.470	0.516	0.549	**0.545**	0.520
STD	**0.536**	**0.543**	0.567	0.521	**0.542**

– **PER** [6]: This metric computes the post-editing distance between the hypothesis and the reference translation.

– **WER** [5]: This metric counts and aggregates three kinds of errors (substitution, deletion, insertion), normalized by the length of the reference.

– **CharacTer** [22]: This metric calculates the character level edit distance and performs the shift edit on word level. Different from TER's strict matching criterion, a hypothesis word is considered to match a reference word and could be shifted, if the edit distance between them is lower than a threshold.

– **WordF1,2,3** [58]: This metric calculates a simple F-score combination of the precision and recall of word n-grams of maximal length 4 with different setting for the β parameter ($\beta = 1$, 2, or 3).

– **ChrF1,2,3** [58]: This metric uses character n-gram F-score of character n-grams of maximal length 6 with different setting for the β parameter ($\beta = 1$, 2, or 3).

– **BEER** [21]: This metric utilizes character n-gram to evaluate lexical accuracy. Besides, it uses hierarchical representations based on PETs (permutation trees) to measure word order.

B. Machine Translation Evaluation

TABLE VI
SYSTEM-LEVEL ABSOLUTE PEARSON CORRELATION COEFFICIENT BETWEEN AUTOMATIC METRICS AND DA HUMAN ASSESSMENT ON THE **WMT-15** DATASET; THE OPTIMAL AND SUBOPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINE RESPECTIVELY

Language pair	**cs-en**	**de-en**	**fi-en**	**ru-en**	Average
System number	16	13	14	13	–
WER	0.888	0.884	0.853	0.895	0.880
TER	0.907	0.890	0.872	0.907	0.894
BLEU	0.957	0.865	0.929	0.851	0.901
PER	0.963	0.864	0.871	0.931	0.921
NIST	0.973	0.901	0.894	0.910	0.920
WordF1	0.987	0.877	0.961	0.887	0.928
WordF3	0.981	0.877	0.969	0.896	0.931
WordF2	0.983	0.877	0.968	0.894	0.931
CharacTer	0.965	**0.981**	0.928	0.881	0.939
ChrF3	0.980	0.948	0.915	0.936	0.945
ChrF2	0.983	0.945	0.925	0.942	0.949
ChrF1	**0.989**	0.930	0.950	0.946	0.954
BEER	0.987	0.933	0.955	0.958	0.958
STD	0.982	0.946	**0.972**	**0.987**	**0.972**

TABLE VII
SYSTEM-LEVEL ABSOLUTE PEARSON CORRELATION COEFFICIENTS BETWEEN AUTOMATIC METRICS AND DA HUMAN ASSESSMENT ON THE **WMT-16** DATASET; THE OPTIMAL AND SUBOPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINE RESPECTIVELY

Language pair	**fi-en**	**ru-en**	**ro-en**	**de-en**	Average
System number	9	10	7	10	–
PER	0.767	0.887	0.748	0.730	0.783
WER	0.768	0.837	0.762	0.822	0.797
WordF1	0.808	0.852	0.804	0.780	0.804
WordF2	0.806	0.831	0.815	0.786	0.809
WordF3	0.803	0.833	0.818	0.787	0.810
TER	0.846	0.847	0.793	0.834	0.830
BLEU	0.864	0.837	0.840	0.808	0.837
NIST	0.929	0.854	0.807	0.801	0.848
BEER	0.972	0.901	0.852	0.879	0.901
ChrF1	**0.980**	0.898	0.865	0.868	0.903
ChrF2	0.967	0.918	0.886	0.893	0.916
CharacTer	0.927	0.930	0.883	0.929	0.917
ChrF3	0.958	0.923	0.892	0.902	0.919
STD	0.975	**0.942**	**0.893**	**0.937**	**0.937**

For the segment-level evaluation, results in Tables IV and V show that STD obtains the best performance among a range of state-of-the-art metrics on average. More specifically, STD has the best performance on Czech-English (cs-en), Finnish-English (fi-en) and Russian-English (ru-en) translations on the WMT15 dataset. As for the result on the WMT16 dataset, STD obtains the best performance on Finnish-English (fi-en) and Russian-English (ru-en) translations as well as the suboptimal result on German-English (de-en) translation. Since STD shows an optimal/suboptimal performance on most language pairs, it is more robust than other metrics. On the average, STD shows a higher correlation with human judgements than other metrics for the segment-level evaluation.

For the system-level evaluation, results in Tables VI and VII show that, on average, STD also performs the best among a range of state-of-the-art metrics. More specifically, on the WMT15 dataset, STD obtains the best performance on Finnish-English (fi-en) and Russian-English (ru-en) translations. As for the result on the WMT16 dataset, STD shows the best performance on Russian-English (ru-en), Romanian-English (ro-en)

TABLE VIII
Translation Examples Evaluated With sentBLEU and $STD^*(1-STD)$. R Denotes the Reference Sentence; H_1, H_2 and H_3 Denote Different Hypothesis Sentences

ID	Sentence	sentBLEU	STD^*
R	Dettori soaks up the cheers as he brings in Predilection after victory	–	–
H_1	Dai Tuli brought the Predilection after the victory, causing the crowd to cheer.	0.135	0.883
H_2	Dai Tuli received the Predilection after winning, causing the audience cheers.	0.156	0.869
H_3	After winning the winner, Doodle led the Predilation, causing the whole field.	0.104	0.855

TABLE IX
Introduction of the Pretrained Word Embedding Models Used in Our Experiment

Metric	Training algorithm	Corpus	Dimension size	Token size	Vocabulary size	Web page
STD_{w300}	*word2vec*	Google News	300	100B	3M	https://code.google.com/archive/p/word2vec/
STD_{fw300}	*fastText*	Wikipedia	300	16B	1M	https://fasttext.cc/docs/en/english-vectors.html
STD_{fc300}		Common Crawl	300	600B	2M	
STD_{g25}	*GloVe*	Twitter	25	2B	1.2M	https://nlp.stanford.edu/projects/glove/
STD_{g50}		Twitter	50	2B	1.2M	
STD_{g100}		Twitter	100	2B	1.2M	
STD_{g200}		Twitter	200	2B	1.2M	
STD_{g300}		Common Crawl	300	840B	2.2M	

and German-English (de-en) translations as well as the suboptimal result on Finnish-English (fi-en) translation. STD shows a great performance on different language pairs, which proves its high robustness. These results demonstrate that STD outperforms other metrics for the system-level evaluation.

From the above results, we can conclude that STD has a better and more robust performance than other state-of-the-art metrics for both the segment-level and system-level evaluations. In addition, we can see that word-level based metrics, such as BLEU, NIST, WER, PER, TER and WordF, have relatively lower correlation with human judgements than character-based metrics like CharacTer, ChrF and BEER. Furthermore, the results show that a method like STD that utilizes flexible semantic matching has better performance than strict lexical matching. By utilizing word embeddings and n-gram order, STD shows promising performance for MT evaluation.

To better understand why STD can achieve better correlation with human judgments than lexical-based metrics, we select, in Table VIII, some interesting examples for further analysis. R denotes the reference sentence and H_1, H_2 and H_3 denote different hypothesis sentences. We calculate the sentBLEU score and the STD score for each translation example. Note that for sentBLEU, the higher the score is, the better the translation quality is while STD is the opposite. To present the result more intuitively, we compute the STD score as $1 - STD$, namely STD^*.

According to the human judgments, H_1 is better than H_2 and H_2 is better than H_3 in this example. Besides, three hypotheses are semantically similar to the translation reference. In this case, sentBLEU assigns the scores of 0.135, 0.156 and 0.104 to H_1, H_2 and H_3 respectively. In contrast, STD associates H_1 with 0.883, H_2 with 0.869 and H_3 with 0.855. It's obvious that STD better matches human's judgement.

C. *Word Embeddings*

As STD is based on word embeddings, we examine how different word embeddings influence the metric's performance on the WMT16 dataset. Apart from the aforementioned freely-available *fastText* model, we also utlize seven other freely-available models pretrained on diverse corpus with different algorithms including *fastText* [11], *word2vec* [9] and *Glove* [10]. Table IX shows the detailed information of the pertrained models used in our experiment. We compare the STD metric of different word embeddings with BLEU/sentBLEU, BEER and CharacTer for both the system-level and segment-level evaluations. The optimal and suboptimal results are highlighted in bold and underline respectively. And the STD model STD_{fc300}, which is used in the previous experiment, is also highlighted.

Results in Table X and Table XI show that word embeddings have a significant influence on STD's performance for both the segment-level and system-level evaluations. The STD based on *GloVe* embeddings shows relatively lower correlation with human judgements except for STD_{g300}, which performs the best for the system-level evaluation. Besides, the STD based on *fastText* and *word2vec* embeddings, such as STD_{fw300}, STD_{fc300} and STD_{w300}, have better performance than other metrics,

TABLE X
Segment-Level Absolute Pearson Correlation Coefficient Between Automatic Metrics and DA Human Assessment on the **WMT-16** Dataset; the Optimal and Suboptimal Results Are Highlighted in Bold and Underline Respectively

Language pair Dataset size	**fi-en** 560	**ru-en** 560	**ro-en** 560	**de-en** 560	Average –
STD_{g200}	0.178	0.206	0.214	0.237	0.208
STD_{g100}	0.203	0.253	0.245	0.273	0.244
STD_{g50}	0.227	0.253	0.255	0.301	0.259
STD_{g25}	0.272	0.386	0.358	0.364	0.345
STD_{g300}	0.506	0.476	0.508	0.427	0.479
sentBLEU	0.448	0.502	0.499	0.484	0.483
BEER	0.462	0.533	0.551	0.471	0.504
CharacTer	0.470	0.516	0.549	**0.545**	0.520
STD_{fw300}	0.529	0.522	**0.575**	0.499	0.531
STD_{w300}	**0.555**	<u>0.534</u>	0.521	0.519	<u>0.532</u>
$\boldsymbol{STD_{fc300}}$	<u>0.536</u>	**0.543**	<u>0.567</u>	<u>0.521</u>	**0.542**

TABLE XI
SYSTEM-LEVEL ABSOLUTE PEARSON CORRELATION COEFFICIENT BETWEEN AUTOMATIC METRICS AND DA HUMAN ASSESSMENT ON THE **WMT-16** DATASET; THE OPTIMAL AND SUBOPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINE RESPECTIVELY

Language pair	**fi-en**	**ru-en**	**ro-en**	**de-en**	Average
System number	9	10	7	10	–
STD_{g200}	0.839	0.831	0.744	0.786	0.800
$STDg100$	0.861	0.803	0.766	0.819	0.812
STD_{g25}	0.801	0.814	0.810	0.833	0.815
STD_{g50}	0.922	0.750	0.772	0.825	0.817
BLEU	0.864	0.837	0.840	0.808	0.837
BEER	0.972	0.901	0.852	0.879	0.901
CharacTer	0.927	0.930	0.883	0.929	0.917
STD_{w300}	**0.981**	0.913	0.879	0.901	0.919
$\boldsymbol{STD_{fc300}}$	$\underline{0.975}$	0.942	0.893	0.937	0.937
STD_{fw300}	0.959	$\underline{0.955}$	$\underline{0.903}$	$\underline{0.952}$	$\underline{0.942}$
STD_{g300}	0.953	**0.975**	**0.935**	**0.971**	**0.959**

which might ascribe to their larger size of dimension and training corpus. In general, the larger the training vocabulary size of the model is, the better the model performs, which is in line with those of Mikolov *et al.* [59], that in general more data (as opposed to simply relevant data) creates better embeddings. By incorporating more data and optimizing the training algorithm, we can create better word embeddings and promote the performance of STD.

V. CONCLUSION

In this paper, we propose a novel MT evaluation metric STD based on word embeddings, which measures the semantic distance between the hypothesis and reference rather than strict string matchings. It also incorporates word order features by employing n-gram order distance. Experiments on two recent WMT metric tasks indicate that STD has a better performance than a range of state-of-the-art metrics for both the segment-level and system-level evaluations. Besides, its performance can be easily improved by utilizing higher quality word embeddings. Moreover, it only depends on the availability of word embeddings, which should be available, or at least derivable for most languages.

In future work, we consider to improve the STD metric from the following three aspects. First, we can utilize higher quality word embeddings, such as the recently proposed *ELMo* [60] and *BERT* models [61], which are proved to be powerful and promising. Second, we consider to transform the STD metric from a word-level based metric into a character-level based one since character-level based metrics have shown better performance. Third, incorporating syntactic features into our metric, such as Part-of-Speech (POS) and dependency information, may lead to higher correlation with human judgements.

REFERENCES

[1] C. Liu, D. Dahlmeier, and H. T. Ng, "Better evaluation metrics lead to better machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 375–384.

[2] H. Somers, *Computers and Translation: A Translator's Guide*, vol. 35. Amsterdam, The Netherlands: John Benjamins Publ., 2003.

[3] J. Blatz *et al.*, "Confidence estimation for machine translation," in *Proc. 20th Int. Conf. Comput. Linguistics*, 2004, p. 315.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[5] K.-Y. Su, M.-W. Wu, and J.-S. Chang, "A new quantitative quality measure for machine translation systems," in *Proc. 14th Conf. Comput. Linguistics*, 1992, vol. 2, pp. 433–439.

[6] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf, "Accelerated DP based search for statistical translation," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, 1997, pp. 2667–2670.

[7] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proc. 2nd Int. Conf. Human Lang. Technol. Res.*, 2002, pp. 138–145.

[8] M. Snover, N. Madnani, B. J. Dorr, and R. Schwartz, "Fluency, adequacy, or HTER?: Exploring different human judgments with a tunable MT metric," in *Proc. 4th Workshop Statist. Mach. Transl.*, 2009, pp. 259–268.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[10] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017.

[12] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proc. 6th Int. Conf. Comput. Vision*, 1998, pp. 59–66.

[13] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *Proc. 12th Int. Conf. Comput. Vision*, 2009, vol. 9, pp. 460–467.

[14] L. Han, "Machine translation evaluation resources and methods: A survey," 2016, arXiv:1605.04515.

[15] B. Chen and C. Cherry, "A systematic comparison of smoothing techniques for sentence-level BLEU," in *Proc. 9th Workshop Statist. Mach. Transl.*, 2014, pp. 362–367.

[16] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics*, 2004, p. 605.

[17] B. Wong and C. Kit, "ATEC: Automatic evaluation of machine translation via word choice and word order," *Mach. Transl.*, vol. 23, no. 2/3, pp. 141–155, 2009.

[18] B. Chen, R. Kuhn, and S. Larkin, "PORT: A precision-order-recall MT evaluation metric for tuning," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Long Papers*, 2012, vol. 1, pp. 930–939.

[19] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.

[20] M. Popović, "chrF: Character n-gram F-score for automatic MT evaluation," in *Proc. 10th Workshop Statist. Mach. Transl.*, 2015, pp. 392–395.

[21] M. Stanojević and K. Sima'an, "Evaluating MT systems with beer," *Prague Bull. Math. Linguistics*, vol. 104, no. 1, pp. 17–26, 2015.

[22] W. Wang, J.-T. Peter, H. Rosendahl, and H. Ney, "Character: Translation edit rate on character level," in *Proc. 1st Conf. Mach. Transl., Shared Task Papers*, 2016, vol. 2, pp. 505–510.

[23] D. Liu and D. Gildea, "Syntactic features for evaluation of machine translation," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 25–32.

[24] M. Collins and N. Duffy, "Convolution kernels for natural language," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 625–632.

[25] M. Popović and H. Ney, "Word error rates: Decomposition over POS classes and applications for error analysis," in *Proc. 2nd Workshop Statist. Mach. Transl.*, 2007, pp. 48–55.

[26] Y. S. Chan and H. T. Ng, "MAXSIM: A maximum similarity metric for machine translation evaluation," in *Proc. 46th Annu. Meeting Assoc. Computat. Linguistics*, 2008, pp. 55–62.

[27] Z. Zhu, D. Bernhard, and I. Gurevych, "A monolingual tree-based translation model for sentence simplification," in *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 1353–1361.

[28] M. Duma and W. Menzel, "UHH submission to the WMT17 quality estimation shared task," in *Proc. 2nd Conf. Mach. Transl.*, 2017, pp. 556–561.

[29] M. G. Snover, N. Madnani, B. Dorr, and R. Schwartz, "TER-Plus: Paraphrase, semantic, and alignment enhancements to translation edit rate," *Mach. Transl.*, vol. 23, no. 2/3, pp. 117–127, 2009.

[30] C. Fellbaum, *WordNet: An Electronic Lexical Database* (Language, Speech, and Communication). Cambridge, MA, USA: MIT Press, 1998.

[31] F. Guzmán, S. Joty, L. Màrquez, and P. Nakov, “Using discourse structure improves machine translation evaluation,” in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Long Papers*, 2014, vol. 1, pp. 687–698.

[32] C.-k. Lo and D. Wu, “MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames,” in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Lang. Technol.*, 2011, vol. 1, pp. 220–229.

[33] C.-k. Lo, M. Beloucif, M. Saers, and D. Wu, “XMEANT: Better semantic MT evaluation without reference translations,” in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Short Papers*, 2014, vol. 2, pp. 765–771.

[34] C.-k. Lo, P. Dowling, and D. Wu, “Improving evaluation and optimization of MT systems against meant,” in *Proc. 10th Workshop Statist. Mach. Transl.*, 2015, pp. 434–441.

[35] R. E. Banchs, L. F. D’Haro, and H. Li, “Adequacy-fluency metrics: Evaluating MT in the continuous space model framework,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 3, pp. 472–482, Mar. 2015.

[36] M. Vela and L. Tan, “Predicting machine translation adequacy with document embeddings,” in *Proc. 10th Workshop Statist. Mach. Transl.*, 2015, pp. 402–410.

[37] B. Chen and H. Guo, “Representation based translation evaluation metrics,” in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics/7th Int. Joint Conf. Natural Lang. Process., Short Papers*, 2015, vol. 2, pp. 150–155.

[38] C. Servan, A. Bérard, Z. Elloumi, H. Blanchon, and L. Besacier, “Word2Vec vs DBnary: Augmenting METEOR using vector representations or lexical resources?” in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers*, 2016, pp. 1159–1168.

[39] G. Sérasset, “DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF,” *Semantic Web*, vol. 6, no. 4, pp. 355–361, 2015.

[40] A. Tättar and M. Fishel, “Bleu2vec: The painfully familiar metric on continuous vector space steroids,” in *Proc. 2nd Conf. Mach. Transl.*, 2017, pp. 619–622.

[41] S. E. Robertson *et al.*, “Okapi at TREC-3,” in *Proc. 3rd Text Retrieval Conf.*, 1995, p. 109.

[42] Z. Ren, J. Yuan, and Z. Zhang, “Robust hand gesture recognition based on finger-earth mover’s distance with a commodity depth camera,” in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1093–1096.

[43] K. Grauman and T. Darrell, “Fast contour matching using approximate earth mover’s distance,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2004, vol. 1, p. I.

[44] S. Shirdhonkar and D. W. Jacobs, “Approximate earth mover’s distance in linear time,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2008, pp. 1–8.

[45] E. Levina and P. Bickel, “The earth mover’s distance is the mallows distance: Some insights from statistics,” in *Proc. 8th IEEE Int. Conf. Comput. Vision*, 2001, vol. 2, pp. 251–256.

[46] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, “From word embeddings to document distances,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 957–966.

[47] M. Zhang, Y. Liu, H. Luan, M. Sun, T. Izuha, and J. Hao, “Building earth mover’s distance on bilingual word embeddings for machine translation,” in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2870–2876.

[48] B. Schölkopf, J. Weston, E. Eskin, C. Leslie, and W. S. Noble, “A kernel approach for learning from almost orthogonal patterns,” in *Proc. Eur. Conf. Mach. Learn.*, 2002, pp. 511–528.

[49] D. Greene and P. Cunningham, “Practical solutions to the problem of diagonal dominance in kernel document clustering,” in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 377–384.

[50] G. Monge, “Mémoire sur la théorie des déblais et des remblais,” *Histoire de l’Académie Royale des Sciences de Paris*, vol. s1-14, pp. 139–143, 1781.

[51] Y. Rubner, C. Tomasi, and L. J. Guibas, “The earth mover’s distance as a metric for image retrieval,” *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[52] H. Ling and K. Okada, “An efficient earth mover’s distance algorithm for robust histogram comparison,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 840–853, May 2007.

[53] M. Stanojević, A. Kamran, P. Koehn, and O. Bojar, “Results of the WMT15 metrics shared task,” in *Proc. 10th Workshop Statist. Mach. Transl.*, 2015, pp. 256–273.

[54] O. Bojar, Y. Graham, A. Kamran, and M. Stanojević, “Results of the WMT16 metrics shared task,” in *Proc. 1st Conf. Mach. Transl., Shared Task Papers*, 2016, vol. 2, pp. 199–231.

[55] Y. Graham, T. Baldwin, and N. Mathur, “Accurate evaluation of segment-level machine translation metrics,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2015, pp. 1183–1191.

[56] O. Pele and M. Werman, “A linear time histogram metric for improved sift matching,” in *Proc. Eur. Conf. Comput. Vision*, Oct. 2008, pp. 495–508.

[57] J. Gao and X. He, “Training MRF-based phrase translation models using gradient ascent,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2013, pp. 450–459.

[58] M. Popović, “chrF deconstructed: Beta parameters and n-gram weights,” in *Proc. 1st Conf. Mach. Transl., Shared Task Papers*, 2016, vol. 2, pp. 499–504.

[59] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, arXiv:1301.3781.

[60] M. E. Peters *et al.*, “Deep contextualized word representations,” in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol., Long Papers*, 2018, pp. 2227–2237.

[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

Authors’ photographs and biographies not available at the time of publication.