Unsupervised Evaluation of Human Translation Quality

Yi Zhou and Danushka Bollegala

Department of Computer Science, University of Liverpool, U.K.

Keywords: Translation Quality, Evaluation of Human Translations, Cross-lingual Word Embeddings, Word Mover's Distance, Bidirectional Minimum Word Mover's Distance.

Abstract: Even though machine translation (MT) systems have reached impressive performances in cross-lingual translation tasks, the quality of MT is still far behind professional human translations (HTs) due to the complexity in natural languages, especially for terminologies in different domains. Therefore, HTs are still widely demanded in practice. However, the quality of HT is also imperfect and vary significantly depending on the experience and knowledge of the translators. Evaluating the quality of HT in an automatic manner has faced many challenges. Although bilingual speakers are able to assess the translation quality, manually checking the accuracy of translations is expensive and time-consuming. In this paper, we propose an unsupervised method to evaluate the quality of HT without requiring any labelled data. We compare a range of methods for automatically grading HTs and observe the Bidirectional Minimum Word Mover's distance (BiMWMD) to produce gradings that correlate well with humans.

1 INTRODUCTION

With the rapid development of international business and multinational companies, there is an increasing demand for translations of user manuals, contracts and various other documents. Even though MT systems have shown promise for automatic translation, they require large parallel corpora for training, which might not be available for resource poor language pairs such as Hindi, Sinhalese, etc. In addition, the performance of MT systems are still far behind professional human translators due to the complex nature of grammar and word usage in languages. The quality of translations generated by MT depends on the distances between the source and target language pairs (Han, 2016). For example, the quality of an English to French translation would be better than an English to Chinese translation, even though both translations are generated from the same MT system (Xu et al., 2018). As a result, HTs are still widely used across numerous industries.

A person's first language, L1, refers to the native language of that person, whereas L2 is a second language spoken by that person. HTs created by L2 speakers can be erroneous due to the different levels of experiences and knowledge of the translators. Often, the quality of translations provided by L2 speakers must be manually verified by professional translators before they can be accepted. A good translation must demonstrate six properties: intelligibility, fidelity, fluency, adequacy, comprehension, and informativeness (Han, 2016). However, manually verifying these properties in an HT is both time consuming and costly. In this paper, we propose an unsupervised method for evaluating the quality of HTs, which addresses this challenging problem.

Translation quality evaluation is a much more complicated task than it might appear in a first glance. Papineni et al. (2002) proposed the bilingual evaluation understudy (BLEU) method to automatically evaluate the quality of MT. They take professional HTs as golden references and consider a better MT should be the one closer to the golden HTs. In contrast, HTs quality evaluation must be done manually because such golden references are not available. People who are familiar with both the source and the target languages are required to evaluate the quality of HTs. The number of such bilingual speakers are limited and might not exist in the case of rare languages. However, human evaluation is time-consuming and not re-usable. MT quality evaluation requires a reference translation. Because of this reason, MT evaluation measures such as BLEU, cannot be used for the purpose of evaluating HTs.

In this paper, we model HT quality evaluation as an unsupervised graph matching problem. Specifi-

Zhou, Y. and Bollegala, D.
Unsupervised Evaluation of Human Translation Quality.
DOI: 10.5220/0008064500550064
In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, pages 55-64
ISBN: 978-989-758-382-7

cally, given a source sentence S and its target translation T, we compare the similarity between the set of words $\{s_1, s_2, \ldots, s_n\}$ in S against the set of words $\{t_1, t_2, \ldots, t_m\}$ in T, using different distance metrics such as Euclidean distance. In this work, we take the advantage of cross-lingual word embeddings between different languages and present a novel approach to automatically evaluate the quality of HTs without accessing to golden references. Our work is inspired by the Word Mover's distance (Kusner et al., 2015), which measures the distance between documents by minimising the cost of transferring embedded words from the source language to the target language. We emphasise that our goal in this paper is *not* to propose a novel MT method nor an evaluation metric for MT. Instead, we consider the problem of automatically detecting high/low quality of human translations, without having any access to reference translations.

Specifically, we report and evaluate different methods for the purpose of unsupervised HT evaluation and compare them against grades given by judges, who are professional translators, for the quality of the HTs using Spearman rank and Pearson correlations. As shown in the experiments, the Bidirectional Minimum Word Mover's distance (BiMWMD) has the strongest correlation with the human assigned grades, indicating that this method is able to accurately detect the low quality and high quality HTs without requiring any human supervision.

2 RELATED WORK

An HT can be compared against the source text using similarity and distance metrics through cross-lingual word embeddings. Cosine similarity and Euclidean distance have been popularly used for this purpose. Semantic Textual Similarity (STS) systems evaluate the degree of semantic similarity between two sentences. Most of the early work on sentence similarity compute the sentence similarity as the average of the words similarity over the two sentences (Corley and Mihalcea, 2005; Li et al., 2016; Islam and Inkpen, 2008). At SemEval 2012, the supervised systems combining different similarity measures such as lexico-semantic, syntactic and string similarity, using regression models have been proposed (Bär et al., 2012; Šarić et al., 2012). Later, Sultan et al. (2015) propose an unsupervised system based on word alignment. Brychcín and Svoboda (2016) and Tian et al. (2017) model semantic similarity for multilingual and cross-lingual sentence pairs by first translating sentences into the target language using MT, then applying the monolingual STS models. In order to address the problem that human annotated data is limited for resource poor languages, Brychcín (2018) studied linear transformations to map monolingual word embeddings into a common space using bilingual dictionary for cross-lingual semantic similarity.

The distributional hypothesis (Harris, 1954) states that words occurring in the same context tend to have similar meanings. According to the hypothesis, Mikolov et al. (2013) propose distributed Skip-gram and Continuous Bag-of-Words (CBOW) models to learn robust word embeddings from large amount of unstructured texts data. Recent research creating a shared vector space for words across two (bilingual word embeddings) (Artetxe et al., 2017; Chandar A P et al., 2014; Zou et al., 2013) or more (multilingual word embeddings) (Hermann and Blunsom, 2014; Lauly et al., 2014) languages is referred to cross-lingual word embeddings learning. The distance between words from different languages with similar meanings should be close to each other in the shared embedding space (Chen and Cardie, 2018).

The cross-lingual word representations are obtained by training embeddings in different languages independently using monolingual corpora, then map them to a common space through a transformation (Artetxe et al., 2018). Ruder et al. (2017) introduced three different types of alignments in learning cross-lingual word embeddings: word alignment, sentence alignment and document alignment. Word alignment uses bilingual dictionaries with word-by-word translations to learn cross-lingual embeddings (Vulić and Moens, 2015). Sentence alignment requires a parallel corpus (Hermann and Blunsom, 2014; Gouws et al., 2015), which is a collection of texts in one language and the corresponding translations into one or more languages. Document alignment requires document in a comparable corpus across different languages. A comparable corpus contains documents that are not exact parallel translations but convey the same information in different languages (Faruqui and Dyer, 2014; Gouws and Søgaard, 2015).

Several approaches for learning cross-lingual word embeddings have been proposed, which require different types of alignment as supervision. Luong et al. (2015) present the bilingual Skip-Gram model (BiSkip) to learn cross-lingual word embeddings using a parallel corpus (sentence alignment), which can be seen as an extension of the monolingual skip-gram model. The Bilingual Compositional Model (BiCVM) proposed by Hermann and Blunsom (2014) learns cross-lingual word embeddings through sentence alignment. The model leverages the fact that the representations of aligned sentences should be similar. Therefore, semantics can be learned from parallel data. Vulić and Moens (2015) proposed a model to learn cross-lingual word embeddings from non-parallel data. They extend the skip-gram model with negative sampling (SGNS) model and generate cross-lingual word embeddings via a comparable corpus (document alignment).

The method proposed by Kusner et al. (2015) for measuring the distance between two documents is known as the Word Mover's Distance. It considers the distance between documents to be the minimal cost of transforming words from one document to another. However, they take the alignment of each source word to all of the target words to compute the cost of a translation, which is expensive. In this paper, we focus on the sentence alignment and propose the Bidirectional Minimum Word Mover's distance (BiMWMD) method, where we consider the distance between documents to be the cumulative cost of the minimal cost of transferring each source word to the corresponding target word. In addition, our proposed method takes into account the translation flow from both direction (i.e. from the source to the target and from the target to the source).

3 TRANSLATION QUALITY EVALUATION

Our goal is to propose a method, which is able to accurately evaluate the quality of cross-lingual translations, without human supervision. Most translation quality evaluation approaches are based on gold references, which are manually created perfect translations of a source language text to the target language. Our work considers the scenario that there are no such golden references available.

Let us denote the source language by $\mathcal{S}$ and the target language by $\mathcal{T}$. For example, when translating from Japanese to English, $\mathcal{S}$ will be Japanese and $\mathcal{T}$ will be English. Consider a set of words $\mathcal{V}_S$, $\mathcal{V}_T$ respectively in the source and target languages. A cross-lingual word embedding $\boldsymbol{w} \in \mathbb{R}^d$ of a word $w \in \mathcal{V}_S \cup \mathcal{V}_T$ is an embedding that is shared between both $\mathcal{S}$ and $\mathcal{T}$. As already described in Section 2, several methods have been proposed for training accurate cross-lingual word embeddings that we can use for this purpose. Here, we assume the availability of such a set of cross-lingual word embeddings for the source and target languages.

Let us consider a source language text $S = \boldsymbol{s_1}, \boldsymbol{s_2}, \ldots, \boldsymbol{s_n}$, which is translated to the target language $T = \boldsymbol{t_1}, \boldsymbol{t_2}, \ldots, \boldsymbol{t_m}$ where $\boldsymbol{s}_i \in \mathbf{R}^d$ represents the embedding of i-th word in the source sentence, $\boldsymbol{t}_j \in \mathbf{R}^d$ represents the embedding of j-th word in the target sentence. Here, n and m are the numbers of words in the source and the target texts respectively. Source and target texts could be single or multiple sentences. The methods that we discuss in this paper for evaluating HT quality do not require any sentence-level processing and can be applied to either single sentences or paragraphs that contain multiple sentences.

3.1 Averaged Vector (AV)

Prior work on unsupervised sentence embeddings have found that averaging the word embeddings for the words in a sentence to be a simple but an accurate method for creating sentence embeddings (Arora et al., 2017). Motivated by these prior proposals, we represent both source and target language texts by averaging the cross-lingual word embeddings for the words that appear in each of the texts. We call this the Averaged Vector (AV) method. Specifically, given a source language text $S = \boldsymbol{s_1}, \boldsymbol{s_2}, \ldots, \boldsymbol{s_n}$ and its HT $T = \boldsymbol{t_1}, \boldsymbol{t_2}, \ldots, \boldsymbol{t_m}$, we represent the two texts respectively by embeddings $\bar{\boldsymbol{s}}, \bar{\boldsymbol{t}} \in \mathbb{R}^d$ as given by (1) and (2).

$$\bar{\boldsymbol{s}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s_i} \tag{1}$$

$$\bar{\boldsymbol{t}} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{t_j} \tag{2}$$

After obtaining the vectors for the source sentence and the target sentence, similarity between them can be computed using the cosine similarity between the two vectors $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{t}}$ as given by (3).

$$\begin{aligned}\mathrm{sim}(S,T) &= \cos(\bar{\boldsymbol{s}}, \bar{\boldsymbol{t}}) \\ &= \frac{\bar{\boldsymbol{s}}^\top \bar{\boldsymbol{t}}}{||\bar{\boldsymbol{s}}||\,||\bar{\boldsymbol{t}}||}.\end{aligned} \tag{3}$$

Here, we consider the similarity between S and T as a proxy of the semantic agreement between the source text and its translation, thereby providing a measure of quality. In addition to the simple averaging of word embeddings given in (1) and (2), in our preliminary experiments we implemented tfidf (term frequency inverse document frequency) weighting and SIF (smooth inverse frequency) (Arora et al., 2017) methods for creating sentence embeddings. But for our task of comparing sentences written in different languages, we did not observe any significant improvements in using those weighting methods. Therefore, we decided to use the simple (unweighted) averaging as given in (1) and (2).

3.2 Source-centred Maximum Similarity (SMS)

The AV method described in Section 3.1 can be seen as comparing each word s_i in the source text against each word t_j in the target text. Moreover, it is symmetric in the sense that if we had swapped the source and the target texts, it will return the same similarity score. However, not all words in the source text might be related to all the words in the target text. On the contrary, one word in the source text is often related to only a few words in the target translation. Therefore, we must compare each source word against the most related word in the target translation. For this purpose, we modify the AV method and propose source-centred maximum similarity (SMS) method as described next.

First, we compute the cosine similarity of each embedded word s_i in the source text against all the embedded words $t_1, t_2, \ldots, t_m$ in the target text. Next, the maximum similarity score between s_i and any of $t_1, t_2, \ldots, t_m$ is taken as the score for transforming s_i from the source to target. Finally, we report the averaged similarity score over all the maximal scores as the similarity between S and T as given by (4).

$$\mathrm{sim}(S,T) = \frac{1}{n}\sum_{i=1}^{n} \max_{j=1,\ldots,m} \cos(\boldsymbol{s_i}, \boldsymbol{t_j}) \tag{4}$$

3.3 Target-centred Maximum Similarity (TMS)

TMS is the opposite to the SMS method and is target-centred. This method calculates the cosine similarity of each embedded target word t_j against all the embedded source words $s_1, s_2, \ldots, s_n$. Then, the maximal similarity score is computed as the score of translating each t_j back to a word s_i in the source text. Finally, we take the average score over all the maximal similarity scores of the target words as given by (5).

$$\mathrm{sim}(S,T) = \frac{1}{m}\sum_{i=1}^{m} \max_{i=1,\ldots,n} \cos(\boldsymbol{s_i}, \boldsymbol{t_j}) \tag{5}$$

3.4 Word Mover's Distance (WMD)

WMD is a measure of the distance between documents proposed by Kusner et al. (2015), which is inspired by the Earth Mover's Distance (EMD) (Rubner et al., 2000). WMD can be used to measure the dissimilarity between two text documents. Specifically, it measures the minimum amount of the cost that has to paid for transforming words from a source text S to reach the words in a target text T. By using this metric, we are able to estimate the similarity between a source document and a target document even though they contain no common words.

Let us assume that two text documents are represented as normalised bag-of-words vectors and the i-th target word t_i appears $h(t_i)$ times in the target text T. The normalised frequency $f(t_i)$ of t_i is given by (6).

$$f(t_i) = \frac{h(t_i)}{\sum_{j=1}^{m} h(t_j)} \tag{6}$$

Likewise, the normalised frequency, $f(s_j)$ of a word s_j in the source text S is given by (7).

$$f(s_j) = \frac{h(s_j)}{\sum_{i=1}^{n} h(s_i)} \tag{7}$$

Then, the transportation problem can be formally defined as the minimum cumulative cost of moving words from a S to T under the constraints specified in the following linear programme (LP).

$$\text{minimise} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{T}_{ij} c(i,j) \tag{8}$$

$$\text{subject to:} \quad \sum_{j=1}^{m} \mathbf{T}_{ij} = f(s_i), \forall i \in \{1,\ldots,n\} \tag{9}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = f(t_j), \forall j \in \{1,\ldots,m\} \tag{10}$$

$$\mathbf{T} \geq \mathbf{0} \tag{11}$$

Here, $\mathbf{T} \in \mathbb{R}^{n \times m}$ is a nonnegative *flow* matrix that is learnt by the LP, and $c(i,j)$ is the cost of translating (transforming) the word s_i to t_j. We measure this translation cost as the Euclidean distance between the embeddings of s_i and t_j as given by (12).

$$c(i,j) = ||\boldsymbol{s_i} - \boldsymbol{t_j}||_2 \tag{12}$$

Intuitively, if $c(i,j)$ is high for translating s_i to t_j, then the (i,j) element T_{ij} of $\mathbf{T}$ can be set to a small (possibly zero) value to reduce the objective given by (13). The equality constraints given in (9) and (10) specify respectively column and row stochasticity constraints for $\mathbf{T}$. In other words, these equality constrains ensure that the total weights transferred from each source word to the target text, and vice versa are preserved, making $\mathbf{T}$ a *double stochastic* matrix. Note that each source text word s_i is allowed to match against one or more target text words t_j under these constraints.

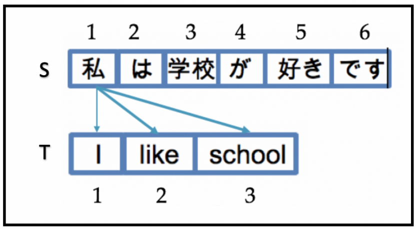

Figure 1: Translating a word in Japanese source (S) text into English target (T) text. The perfect alignment between S and T is $s_1 \to I$, $s_2 \to null$, $s_3 \to school$, $s_4 \to null$, $s_5 \to like$, and $s_6 \to null$. The thin arrow represents the minimum cost translating 私 to *I*. The correct translations are likely to have smaller distances (costs) associated with.

3.5 Bidirectional Minimum WMD (BiMWMD)

WMD method described in Section 3.4 is symmetric in the sense that even if we swap the source and target texts we will get the same score for the translation quality. On the other hand, SMS and TMS methods described respectively in Sections 3.2 and 3.3 are both asymmetric translation quality evaluation methods. Following SMS and TMS methods, we extend WMD method such that it considers the translation quality from the point-of-view of the source text, which we refer to as the *Source-centric Minimum WMD* (SMWMD) and from the point-of-view of the target text, which we refer to as the *Target-centred Minimum WMD* (TMWMD). Finally, we combine the two extensions and propose the Bidirectional Minimum WMD (BiMWMD) to evaluate the translation quality from both point of views. Next, we describe SMWND, TMWMD and BiMWMD methods.

SMWMD: Source-centred Minimum WMD (SMWMD) method considers the translation flow to be from the source sentence to the target sentence. Figure 1 shows an example of measuring distance from a source text S to a target text T. In SMWMD, we measure the minimum cost of translating each source word s_i to any word in T, and consider the sum of costs as the objective function for the LP. Similar to WMD, we denote $\mathbf{T}_{ij} \geq 0$ to be the flow matrix translating s_i to t_j according to the cost $c(i, j)$ given by (12). As we found that the normalised frequencies $f(t_i)$ and $f(s_j)$ have little effect on the results through the experiments, we assign both frequencies to be 1 to simplify the objective function.

Then, the optimisation problem can be written as follows:

$$\text{minimise} \quad \sum_{i=1}^{n} \min_{j=1,\ldots,m} \mathbf{T}_{ij} c(i,j) \tag{13}$$

$$\text{subject to:} \quad \sum\nolimits_{j=1}^{m} \mathbf{T}_{ij} = 1, \forall i \in \{1,\ldots,n\} \tag{14}$$

$$\sum\nolimits_{i=1}^{n} \mathbf{T}_{ij} = 1, \forall j \in \{1,\ldots,m\} \tag{15}$$

$$\mathbf{T} \geq \mathbf{0} \tag{16}$$

To simplify the objective function in (13), we use y_i to replace $\mathbf{T}_{ij}c(i,j)$, where $\mathbf{T}_{ij}c(i,j)$ is the actual cost of transforming words from one document to another and y_i is an upper bound on $\mathbf{T}_{ij}c(i,j)$. Let us denote the actual objective by TC given by (17) and its upper bound by Y given by (18).

$$TC(S,T) = \sum_{i=1}^{n}\sum_{j=1}^{m} T_{ij} c(i,j) \tag{17}$$

$$Y(S,T) = \sum_{i=1}^{n} y_i \tag{18}$$

Using y_i, we can rewrite the previous optimisation problem as an LP as follows:

$$\text{minimise} \quad \sum_{i=1}^{n} y_i \tag{19}$$

$$\text{subject to:} \quad \mathbf{T}_{ij} c(i,j) \leq y_i \tag{20}$$

$$\sum_{j=1}^{m} \mathbf{T}_{ij} = 1, \forall i \in \{1,\ldots,n\} \tag{21}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = 1, \forall j \in \{1\ldots,m\} \tag{22}$$

$$\mathbf{T} \geq \mathbf{0} \tag{23}$$

We collectively denote the minimum translation cost for translating a source text S into a target text T given by solving the LP above as, $\mathrm{SMWMD}(S,T)$, which can be either $TC(S,T)$ or $Y(S,T)$. During our experiments, we will study the difference between the actual objective (TC) and its upper bound (Y) for the purpose of predicting the quality of HT.

TMWMD: An accurate translation of a given source text must not only correctly translate the information contained in the source text but it must also not add new information that was not present in the original source text into the translation. One simple way to verify this is to back translate the target text to the source and measure their semantic distance. For this purpose, we modify the WMD objective, in the same manner as we did for SMWMD but pivoting on the target text instead of the source text. We refer to this approach as the Target-centred Minimum WMD (TMWMD).

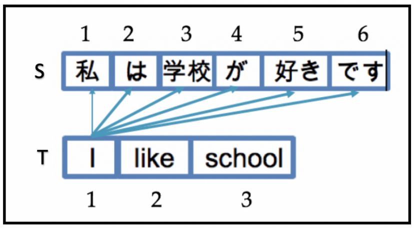

Figure 2: Translating a word in Japanese source (S) text into English target (T) text. The perfect alignment between S and T is $s_1 \rightarrow I$, $s_2 \rightarrow null$, $s_3 \rightarrow school$, $s_4 \rightarrow null$, $s_5 \rightarrow like$, and $s_6 \rightarrow null$. The light arrow represents the minimum cost alignment.

Specifically, we define the distance between S and T as the minimal cumulative distance required to move all words from the target text T to the source text S. An example is give in Figure 2, where the target word I is being compared against all the words in the source (indicated by arrows) and the closet Japanese translation s_1 is mapped by a thin arrow. TMWMD is the solution to the following LP:

$$\text{minimise} \quad \sum_{j=1}^{m} y_j \tag{24}$$

$$\text{subject to: } \mathbf{T}_{ij} c_(i,j) \leq y_j \tag{25}$$

$$\sum_{j=1}^{m} \mathbf{T}_{ij} = 1, \forall i \in \{1,\ldots,n\} \tag{26}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = 1, \forall j \in \{1 \ldots, m\} \tag{27}$$

$$\mathbf{T} \geq \mathbf{0} \tag{28}$$

Note that TMWMD is the mirror image of SMWMD in the sense that by swapping S and T we can obtain the LP for SMWMD.

Likewise SMWMD, TMWMD can also be computed using the actual objective ($TC(S,T)$) or the upper bound ($Y(S,T)$). We collectively denote these two variants as $\mathrm{TMWMD}(S,T)$.

BiMWMD: SMWMD and TMWMD are evaluating the translation quality in one direction only. If the translation cost from source to target as well as from target to source are both small, then it is an indication of a higher quality translation. To quantitatively capture this idea, we propose Bidirectional Minimum Word Mover Distance (BiMWMD) as a translation quality evaluation measure. BiMWMD is defined by (29) and is the sum of optimal translation costs returned individually by SMWMD and TMWMD.

$$\mathrm{BiMWMD}(S,T) = \mathrm{SMWMD}(S,T) + \mathrm{TMWMD}(S,T) \tag{29}$$

From the definition (29), it follows that BiMWMD is a symmetric translation quality measure, similar to WMD. However, BiMWMD and WMD are solving different LPs, hence returning different translation quality predictions. Specifically, in WMD the minimal cumulative cost for translating each word in the source text S to all the words in the target text T is returned as the objective. On the other hand, BiMWMD solves two independent LPs, each considering only a single direction (SMWMD considers the case of translating from S to T, whereas TMWMD considers the case of translating from T to S). As we later see in Section 4.4, BiMWMD shows a higher degree of correlation with human ratings for translation quality than WMD.

4 EXPERIMENTS

In this section, we evaluate the different translation quality measures described in Section 3. For this purpose, we annotated a translation dataset as described in Section 4.1 and use correlation against human grades as the evaluation criteria. Experimental results are discussed in Section 4.4.

4.1 Dataset

For evaluating the different translation quality measures described in Section 3, we created a dataset by selecting 1030 sentences from Japanese user manuals on Digital cameras. We then asked a group of 50 human translators to translate the selected Japanese sentences to English. The human translators are all native Japanese speakers who have studied English as a foreign language. The human translators for this experiment were recruited using a crowd-sourcing platform that is operational in Japan. The human translators have different levels of experience in translating technical documents ranging broadly from very experienced translators to beginners. We believe this would give us a broad spectrum of human translations for evaluation purposes. Each Japanese sentence was assigned to one of the human translators in the pool of human translators and was asked to write a single English translation.

Next, we randomly selected 130 pairs of Japanese-English translated sentence pairs and asked four humans, who are bilingual speakers of Japanese and English and are professionally qualified translators with over 10 years of experience in translating technical documents, to rate each of the quality of the selected translation pairs. We call these four humans as *judges* to distinguish them from the pool of human translators who wrote the translations. Specifically, we asked each of the four judges to grade a translated sentence pair by assigning one of the following four grades.

Grade 1 Quality Translations: A perfect translation. There are no further modifications required. The translation pair is scored in a range of $0.76 - 1.00$.

Grade 2 Quality Translations: A good translation. Some words are incorrectly translated but the overall meaning can be understood. The translation pair is scored in a range of $0.51 - 0.75$.

Grade 3 Quality Translations: A bad translation. There are more incorrectly translated words than correctly translated words in the translation. The translation pair is scored in a range of $0.26 - 0.50$.

Grade 4 Quality Translations: Requires re-translation. The translation cannot be comprehend or conveys a significantly different meaning to the source sentence. The translation pair is scored in a range of $0.00 - 0.25$.

The average of the grades assigned by the four judges to a translated sentence pair is considered as its final grade.

4.2 Cross-lingual Word Embeddings

All of the translation quality measures we proposed in Section 3 require cross-lingual word embeddings. To create cross-lingual word embeddings between Japanese and English languages in an unsupervised manner, we align publicly available monolingual word embeddings. Specifically, we first use the monolingual word embeddings, which are trained on Wikipedia and Common Crawl using fastText (Grave et al., 2018). Because our dataset contains Japanese and English words, we train two separate monolingual word embedding sets for Japanese and English. Next, we use the unsupervised adversarial training methods proposed by Conneau et al. (2017) and implemented in MUSE[1] to align the Japanese and English word embedding spaces, without requiring any bilingual dictionaries or parallel/comparable corpora. Although it is possible to further improve the performance of this cross-lingual alignment using bilingual lexical resources, by not depending on any such resources we are able to realistically estimate the performance of the different methods we proposed in Section 3 when such resources are not available.

4.3 Evaluation Measures

Recall that our goal in this work is to predict the quality of human translations without having access to any reference translations. Therefore, we would like to know whether the translation quality scores returned by the different methods we proposed in Section 3 are correlating with the grades given by the human judges to the human translations in the dataset we created in Section 4.1. To evaluate the level of agreement between the grades and the translation quality scores, we compute the Pearson and Spearman rank correlation coefficients between these two sets of numbers. Pearson correlation coefficient measures the linear relationship between two variables, whereas Spearman correlation coefficient considers only the relative ordering.

4.4 Results

In Table 1, we compare the different HT quality evaluation measures described in Section 3. Recall that some methods return similarity scores (AV, SMS, TMS), whereas others return distances (WMD, SMWMD, TMWMD, BiMWMD). To compare both similarities and distances equally, we convert the distances to similarities for each method by

$$1 - \frac{\text{distance}}{\text{maximum distance}}.$$

We use the interior-point method to solve the LPs in all cases. A higher degree of correlation with the grades assigned by the judges for the translations indicate a reliable quality prediction measure. From Table 1, we see that averaging the word embeddings for creating text/sentence embeddings and then measuring their cosine similarity (AV) provides a low-level of correlation. Comparing SMS and TMS methods, we see that centering on the target provides a higher degree of correlation than by centering on the source. A similar trend can be observed when comparing SMWMD and TMWMD. In fact, SMWMD returns negative correlation values for both Spearman and Pearson correlation coefficients. We see that BiMWMD returns the best correlation scores against judges' grades among all methods compared in Table 1. This result shows that it is important to consider both directions of the translations to more accurately estimate the quality of a human translation.

To study the effect of various parameters and settings associated with the BiMWMD method, we evaluate it under different configurations. Specifically, to analyse the effect of normalising word embeddings, we consider three settings: ℓ_1 normalisation, ℓ_2 normalisation and no normalisation (**No**). To decide be-

[1] https://github.com/facebookresearch/MUSE

Japanese	English	Grade	Score
カメラをテレビに接続するための映像と音声用のケーブル。映像と音声信号を送信する。	A cable used to connect the camera to a standard definition television or video device, supplying both audio and video signals	0.76	0.255
写真撮影用レンズの絞りは、微調整できるようにするために複数枚の板（絞り羽根）を重ね合わせて作られている。6 枚羽根絞りの場合、6 枚の羽根で 6 角形ができており、このような絞りを虹彩絞りという。	A type of stop mechanism with multiple overlapping blades offering fine control over aperture. The six blades create a hexagonal opening referred to as an iris diaphragm.	0.54	0.212
各部名称。いわゆる「撮影設定変更ボタン」。ボタンを押すと、液晶モニターに撮影に関する情報が表示される。	Camera part; the “shooting information”button, used to display shooting settings in the monitor. The term in parentheses is represented by an icon in the manuals except when it appears in the list of camera parts.	0.43	0.233
カメラを縦に構えて撮影すること	Of images: Taller than it is wide (a.k.a. portrait orientation). “Tall”images are taken with the camera in “tall”or “portrait”orientation.	0.23	0

Figure 3: Scores assigned by BiMWMD method to translation pairs graded by the human judges. We have scaled both BiMWMD scores and judges' grades to [0,1] range for the ease of comparison.

Table 1: Performance of the different HT quality evaluation methods. The best correlations are in bold.

Method	Spearman r	Pearson ρ
AV	0.2628	0.3076
SMS	0.1505	0.3224
TMS	0.4576	0.4851
WMD	0.3953	0.5003
SMWMD	-0.3928	-0.2328
TMWMD	0.4164	0.4199
BiMWMD	**0.5895**	**0.5296**

tween actual objective $TC(S,T)$ (given by (17)) vs. its upper bound $Y(S,T)$ (given by (18)), we consider each of the two values separately as the value returned by BiMWMD and measure the correlation against judges' grades. The row and column stochasticity constraints adds a large number of equality constraints to the LPs. Adding both row and column stochasticity constraints at the same time often makes the LP infeasible. To relax the constraints and to empirically study the significance of the row and column stochasticity constraints, we run BiMWMD with row stochasticity constraints only (denoted by **Row**) vs. column stochasticity constraints only (denoted by **Column**). All possible combinations of these configurations are evaluated in Table 2.

Table 2: Different settings for the BiMWMD method. Normalisation of word embeddings: ℓ_1, ℓ_2 and unnormalised (**No**), **Row** and **Column** denote using only row or column stochasticity constraints in the LP. Moreover, we can consider the actual objective (**TC**) or its upper bound (**Y**) as the value of BiMWMD.

Method	Spearman r	Pearson ρ
ℓ_2+Y+Row	0.0033	0.1764
ℓ_2+Y+Column	**0.5895**	**0.5296**
ℓ_2+TC+Row	-0.2510	-0.0767
ℓ_2+TC+Column	-0.2510	-0.0711
ℓ_1+Y+Row	0.3136	0.1834
ℓ_1+Y+Column	0.5721	0.5125
ℓ_1+TC+Row	-0.2510	-0.0730
ℓ_1+TC+Column	-0.2510	-0.0687
No+Y+Row	-0.0728	0.0284
No+Y+Column	0.5146	0.4878
No+TC+Row	-0.2073	-0.0395
No+TC+Column	-0.2053	-0.0414

From Table 2, we see that the best performance is obtained with ℓ_2 normalised cross-lingual word embeddings. We see that column stochasticity constraints are more important than the row stochasticity constraints. Moreover, using the value of the upper bound ($Y(S,T)$) as BiMWMD is more accurate than the actual objective ($TC(S,T)$). Recall that the flow matrix **T** has nm number of parameters. The number of parameters increases with lengths of source and target texts. Therefore, it is possible to set most of those nm elements to zero to minimise the actual objective, thereby satisfying the inequality $T_{ij}c(i,j) \leq y_j$ in LP. Therefore, the sum of upper bounds $\sum_j y_j$, which is the objective minimised by the reformed LP, is a better proxy as BiMWMD.

A good measure for predicting the quality of HTs must be able to distinguish low quality HTs from high quality HTs. If we can automatically decide whether a particular HT is of low quality without another human having to read it, then it is possible to prioritise such low quality HTs for retranslation or to be verified by a human in charge of quality control. This is particularly useful when we have a large number of translations to verify and would like to check the ones which are most likely to be incorrect. To understand the scores assigned by BiMWMD to translations of different grades, in Figure 3, we randomly select translation pairs with different grades and show the scores predicted by BiMWMD, which was the best method among the methods proposed in Section 3. We see that high scores are assigned by BiMWMD method for translations that are also rated as high quality by the human judges, whereas low scores are assigned to translations that are considered to be of low quality by the judges.

5 CONCLUSION

We proposed different methods for automatically predicting the quality of human translations, without having access to any gold standard reference translations. In particular, we proposed a broad range of measures covering both symmetric and asymmetric measures. Our experimental results show that Bidirectional Minimum Word Mover Distance method is in particular demonstrates a high degree of correlation with grades assigned by a group of judges to a collection of human done translations. In future, we plan to evaluate this method for other language pairs and integrate it into a translation quality assurance system.

REFERENCES

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proc. of ICLR*.

Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proc. of ACL*, pages 451–462.

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proc. of ACL*, pages 789–798.

Bär, D., Biemann, C., Gurevych, I., and Zesch, T. (2012). Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proc. of SemEval*, pages 435–440.

Brychcín, T. (2018). Linear transformations for cross-lingual semantic textual similarity. *arXiv:1807.04172*.

Brychcín, T. and Svoboda, L. (2016). Uwb at semeval-2016 task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In *Proc. of SemEval*, pages 588–594.

Chandar A P, S., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V. C., and Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In *Proc. of NIPS*, pages 1853–1861.

Chen, X. and Cardie, C. (2018). Unsupervised multilingual word embeddings. pages 261–270.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv:1710.04087v3*.

Corley, C. and Mihalcea, R. (2005). Measuring the semantic similarity of texts. In *Proc. of ACL Workshop*, pages 13–18.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proc. of EACL*, pages 462–471.

Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In *Proc. of ICML*.

Gouws, S. and Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proc. of NAACL HLT*, pages 1386–1390.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proc. of LREC*, pages 3483–3487.

Han, L. (2016). Machine translation evaluation resources and methods: A survey. *arXiv*.

Harris, Z. S. (1954). Distributional structure. *Word*, pages 146–162.

Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. In *Proc. of ACL*, pages 58–68.

Islam, A. and Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(2):10.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *Proc. of ICML*, pages 957–966.

Lauly, S., Boulanger, A., and Larochelle, H. (2014). Learning multilingual word representations using a bag-of-words autoencoder. *arXiv*.

Li, Y., McLean, D., Bandar, Z. A., Crockett, K., et al. (2016). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge & Data Engineering*, pages 1138–1150.

Luong, T., Pham, H., and Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proc. of VSMNLP Workshop*, pages 151–159.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proc. of ICLR*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Prof. of ACL*, pages 311–318.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, pages 99–12.

Ruder, S., Vulić, I., and Søgaard, A. (2017). A survey of cross-lingual word embedding models. *arXiv*.

Šarić, F., Glavaš, G., Karan, M., Šnajder, J., and Bašić, B. D. (2012). Takelab: Systems for measuring semantic text similarity. In *Proc. of SemEval*, pages 441–448. Association for Computational Linguistics.

Sultan, M. A., Bethard, S., and Sumner, T. (2015). Dls$@$ cu: Sentence similarity from word alignment and semantic vector composition. In *Proc. of SemEval*, pages 148–153.

Tian, J., Zhou, Z., Lan, M., and Wu, Y. (2017). Ecnu at semeval-2017 task 1: Leverage kernel-based traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proc. of SemEval*, pages 191–197.

Vulić, I. and Moens, M.-F. (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proc. of IJCNLP*, pages 719–725.

Xu, R., Yang, Y., Otani, N., and Wu, Y. (2018). Unsupervised cross-lingual transfer of word embedding spaces. In *Proc. of EMNLP*, pages 2465–2474.

Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proc. of EMNLP*, pages 1393–1398.