Article

An Oblivious Approach to Machine Translation Quality Estimation

Itamar Elmakias ^{1,†} and Dan Vilenchik ^{2,*,†}

¹ Department of Industrial Engineering & Management, Ben-Gurion University of the Negev, Beer-Sheva P.O. Box 653, Israel; itamare@post.bgu.ac.il or itamar.elmakias@gmail.com

² School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva P.O. Box 653, Israel

* Correspondence: vilenchi@bgu.ac.il

† These authors contributed equally to this work.

Abstract: Machine translation (MT) is being used by millions of people daily, and therefore evaluating the quality of such systems is an important task. While human expert evaluation of MT output remains the most accurate method, it is not scalable by any means. Automatic procedures that perform the task of Machine Translation Quality Estimation (MT-QE) are typically trained on a large corpus of source–target sentence pairs, which are labeled with human judgment scores. Furthermore, the test set is typically drawn from the same distribution as the train. However, recently, interest in low-resource and unsupervised MT-QE has gained momentum. In this paper, we define and study a further restriction of the unsupervised MT-QE setting that we call *oblivious* MT-QE. Besides having no access to human judgment scores, the algorithm has no access to the test text’s distribution. We propose an oblivious MT-QE system based on a new notion of sentence cohesiveness that we introduce. We tested our system on standard competition datasets for various language pairs. In all cases, the performance of our system was comparable to the performance of the non-oblivious baseline system provided by the competition organizers. Our results suggest that reasonable MT-QE can be carried out even in the restrictive oblivious setting.

Citation: Elmakias, I.; Vilenchik, D. An Oblivious Approach to Machine Translation Quality Estimation. *Mathematics* **2021**, *9*, 2090. <https://doi.org/10.3390/math9172090>

Academic Editor: Fabio Caraffini

Received: 13 August 2021

Accepted: 26 August 2021

Published: 29 August 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: machine translation; unsupervised learning; quality estimation

1. Introduction

Machine translation (MT) is the task of translating text from one natural language to another. Starting in the 1950s, automatic approaches to text translation developed and matured to a level where one can practically use MT. The MT industry started with rule-based MT systems [1], then statistical MT systems [2], and hybrid MT systems [3], and now we are in the era of neural systems [4,5].

As MT is becoming a prevalent mode of translation, its quality is becoming increasingly more critical. The most straightforward option for judging machine translation quality is by human evaluation. The task is performed by experts in translation and linguistics, evaluating the input–output pair from various perspectives such as fluency, adequacy, accuracy, etc.

From a practical point of view, manual evaluation, performed by translation experts, is expensive and takes time. Instead, it is desired to have automatic quick and cheap judgments that approximate human judgment. Machine translation quality evaluation or estimation (MT-QE), refers to an algorithm that produces an evaluation score, which tells the user how good a translation is. The first automatic evaluation methods counted words and sentence-based errors that can be detected automatically. At the same time, general text-level aspects (such as fluency or coherence) were not taken into account. However, in the last decade, new MT evaluation systems were developed to address these aspects.

MT-QE algorithms are used commonly during the development stages of MT systems to measure improvement. They are also used to compare different MT systems.

The “golden standard” metrics in the MT community include BLEU [6], NIST [7], METEOR [8], chrF [9], and TER [10]. Such metrics need reference translations as they compare the MT output with the references and report the comparison scores. If references are available, these metrics can be used to evaluate the output of any number of systems quickly, without the need for human intervention. However, in many situations, references are not available or expensive to obtain. This is, in particular, a problem for less-used languages. The task of reference-less MT-QE is the focus of this work.

Most of the MT-QE algorithms that work without a reference solve the task as a classification problem and are trained in a supervised-learning manner [11–16], to mention a few. The training set typically consists of a large corpus of source–target pairs, along with human judgment scores. In many cases, the test set is sampled from the same training distribution, limiting the evaluation’s generality.

The validation of the algorithm’s score is carried out via correlation coefficients with manual judgment scores. For a meaningful evaluation, a large manual tagged dataset is required (both for testing and training).

When only text is provided for training but no human judgment scores, it is referred to as an unsupervised learning setting. Examples of unsupervised algorithms include the works in [17–24]. This still does not preclude training on the same distribution of text as the test. Thus, the algorithm can use valuable features such as TFIDF (term frequency–inverse document frequency), n -gram statistics, or word embedding tailored to that specific distribution. In fact, in the WMT competitions (WMT is the Workshop on MT which is part of the EMNLP conference), such features are even provided by the organizers.

1.1. The Oblivious Setting

There is no formal name or distinction between the unsupervised MT-QE setting and the one where the algorithm has no access whatsoever to the text distribution on which its performance is being tested. In particular, no human judgment scores are provided. In this work, we make this distinction and define the latter as the *oblivious learning* setting. We chose the term “oblivious” as it is commonly used in the literature to describe a setting where the algorithm has limited access to the input, e.g., the famous Oblivious Transfer protocol [25] or the Oblivious-Caching algorithms [26].

Oblivious learning makes sense in cases where access to the text’s distribution is impossible or too expensive. Furthermore, the oblivious setting can serve as a benchmark to test the robustness of an MT-QE algorithm to various degrees of noise in the training step. For example, when the unsupervised algorithm [17] was run in oblivious mode by training on general text, its performance degraded by almost 50%.

Another aspect that is related to the oblivious setting is the utilization of a parallel corpus for training the MT-QE algorithm. A parallel corpus consists of two or more monolingual corpora, which are translations of each other. For example, a novel and its translation. To generate such a corpus, corresponding segments, usually sentences or paragraphs, need to be aligned. This should be contrasted with the monolingual setting, where text is available in each language but not coupled or aligned into source–target sentences.

While the oblivious setting does not preclude parallel corpora, the fact that cross-lingual parallel corpora are scarce or non-existent for most of the ~7000 languages spoken on earth makes this additional restriction practical [27]. Thus, we arrive at the research question that will be studied in this paper:

Question: What performance of machine translation quality estimation can be achieved in the oblivious monolingual setting.

1.2. Our Contribution

Our first contribution is to make formal the distinction between unsupervised and oblivious MT-QE. Our second contribution is a new MT-QE algorithm that can be executed in oblivious mode, *ObliQuE* (Oblivious Quality Estimation). The algorithm is inspired by physical systems of particles. We view a sentence as a small system of interacting components (the words represented using word embedding). For each sentence, we compute a *cohesiveness factor*, κ , which reflects the extent to which the meaning of the entire system is a function of the meanings of its constituents. For example, $\kappa(\text{"hot-dog"})$ should be smaller than $\kappa(\text{"dog-house"})$. We take the difference in cohesiveness between source and target sentences as the measure for translation quality. The difference in cohesiveness is some aspect of adequacy, which is commonly understood as the amount of information (meaning) preserved between the reference and the candidate translation.

Our method is compatible with the oblivious setting as it can use word embedding that was trained on a generic text from the relevant source–target language pair (e.g., Wikipedia). Furthermore, our method can use monolingual data, avoiding, for example, cross-lingual word embedding such as used in [28].

We tested our method on standard benchmarks, covering several source–target language pairs: English with Spanish, German and Japanese, and Japanese–Chinese. The performance of our algorithm was better, for example, than the oblivious version of [17], scoring a Spearman rank correlation of 0.37 compared to 0.22–0.28 of [17] on the English–Spanish WMT’12 dataset. Our algorithm came first for the German–English dataset and third in the Russian–English, on data from the WMT’19 third shared QE task [29]. In that task, the baseline provided by the organizers was an unsupervised algorithm, but both supervised and unsupervised algorithms competed. To the best of our knowledge, none were oblivious. Details in Section 4.

To conclude, we positively answer the research question posed above and confirm that a fair estimation of translation quality may be obtained using a general framework that is not tailored to the distribution of the text at hand, which also uses monolingual corpora. Another advantageous aspect of *ObliQuE* is the fact that it is an entirely white-box algorithm. The algorithm, described in Section 2, is straightforward to follow and grasp intuitively and contains no hidden tunable parameters that may impede reproducibility.

2. Methodology

An instance of the MT-QE problem (at the sentence level) is a pair of sentences S (source) and T (target). The output is a score that the algorithm assigns for the quality of the translation $S \rightarrow T$.

Our algorithm first maps the words of S (respectively, T) to vectors using a word embedding, obtained, e.g., via word2vec [30]. The vectors corresponding to the s words of S are stacked into a *sentence matrix* M_S (M_T , respectively), whose rows are the d -dimensional vectors of the word embedding. For example, if S is the sentence “the paper is accepted”, and the word embedding for the i^{th} word w_i in the sentence is the vector $\mathbf{v}_{w_i} = (v_{i1} \ v_{i2} \ \dots \ v_{id})$, then the $4 \times d$ sentence matrix is

$$M_S = \begin{bmatrix} \mathbf{v}_{\text{the}} \\ \mathbf{v}_{\text{paper}} \\ \mathbf{v}_{\text{is}} \\ \mathbf{v}_{\text{accepted}} \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1d} \\ v_{21} & v_{22} & \dots & v_{2d} \\ v_{31} & v_{32} & \dots & v_{3d} \\ v_{41} & v_{42} & \dots & v_{4d} \end{bmatrix}.$$

The two sentence-matrices are the objects from which the MT-QE score will be extracted, using what we call a *cohesiveness* measure.

Intuitively, cohesiveness measures the extent to which each word supports the meaning of the sentence. For illustration, consider the following three 5-word sentences: *Dear dear dear dear dear* (same word); *Breakfast, dinner, lunch, milk, egg* (same theme); and *Monster, factory, gym, lake, chair* (random themes). Using the word embedding vectors of [31], we

computed the sentence matrix and cohesiveness factor of each sentence. The first sentence scored $\kappa = 1$, and indeed each word fully determines the meaning of the sentence; the second scored $\kappa_2 = 0.46$ (nearly 50%), which reflects the thematic cohesiveness between the words. The third scored $\kappa_3 = 0.22$, roughly $1/\#words = 1/5 = 0.2$, namely, each word contributes uniformly to the meaning of the sentence, which is what one would expect from a random set of words due to symmetry.

We now proceed with the details of the computation of cohesiveness. First, we define the notion of the “main direction” of the sentence S , which is a single vector that captures the semantic meaning of the sentence. A standard way of computing the main direction is by averaging the vectors of the words in the sentence. This choice led to poor performance, and we replaced it with the leading (right) eigenvector of the sentence matrix M_S .

To better understand this choice, let $V = [\mathbf{v}_1, \dots, \mathbf{v}_d]$ be the matrix whose columns are M_S 's right eigenvectors ($\mathbf{v}_1$ is the leading right eigenvector, our candidate for the main direction), U the matrix with left eigenvectors, and let $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m \geq 0$ be M_S 's singular values (we assume that the embedding dimension d is larger than the number of words in the sentence m). The SVD decomposition theorem says that the i^{th} row of M_S (the vector of the i^{th} word w_i) can be written as

$$\mathbf{v}_{w_i} = \sum_{k=1}^m U_{ik} \sigma_k \mathbf{v}_k = U_{i1} \sigma_1 \mathbf{v}_1 + err.$$

That is, the vector of each word in the sentence is composed of the semantic contribution from $\mathbf{v}_1$ (the “main direction” of the sentence) plus an error term, err . In signal processing, the energy in the direction of a certain eigenvector, in our case, $\mathbf{v}_1$, is typically taken to be the ratio between its singular value and the sum of singular values. In our case,

$$\kappa(M) = \frac{\sigma_1}{\sum_{i=1}^m \sigma_i}.$$

That is, the energy of $\mathbf{v}_{w_i}$ in the direction of the sentence (which is represented using $\mathbf{v}_1$) is given by $\kappa(M)$, and this is what we called cohesiveness to begin with. Thus $\kappa(M)$ will be our proxy for cohesiveness.

The Algorithm

Our method, which we call *ObliQuE* (Oblivious Quality Estimation), is described formally below. The procedure receives as input the source sentence S , the target T , word embedding w_S in the source language and w_T in the target one, and an error function $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ which measures the difference in cohesiveness between source and target. Our working hypothesis is that the smaller $\ell(x, y)$, the better the translation. For the evaluation part of this paper we chose $\ell(x, y) = \max\{x, y\} / \min\{x, y\}$.

-
- 1: **procedure** *ObliQuE*(S, T, w_S, w_T, ℓ)
 - 2: Embed S and T using the word-embedding w_S and w_T , respectively.
 - 3: $M_S \leftarrow$ sentence matrix of S
 - 4: $M_T \leftarrow$ sentence matrix of T
 - 5: **return** $\ell(\kappa(M_S), \kappa(M_T))$
-

The algorithm is described for the sentence-level QE task; for the task of document-level QE, the algorithm is applied iteratively on the sentences. The final score is computed as the average of the ℓ -values.

3. Related Work

MT-QE is typically addressed as a supervised machine learning task where the goal is to predict MT quality without relying on reference translation. Traditional feature-based approaches rely on manually designed features obtained from the source and translated sentences, as well as external resources, such as monolingual or parallel corpora [32].

Currently, the best performing approaches to QE use NNs to learn useful representations for source and target sentences [14,16,33,34]. A notable example is the Predictor-Estimator (PredEst) model [24], which is based on an encoder–decoder RNN (Recursive NN) architecture (predictor) trained on parallel data for a word prediction task and a unidirectional RNN (estimator) that produces quality estimates by using the context representations generated by the predictor. This method can be run both in supervised and unsupervised modes. Despite achieving good performances, neural-based approaches are resource-heavy and require a significant amount of in-domain parallel corpora and labeled data for training.

Other NN-based algorithms explore internal information from neural models as an indicator of translation quality. They rely on the entropy of attention weights in RNN-based NMT systems [23,35]. However, attention-based indicators perform competitively only when combined with other QE features in a supervised framework.

The few approaches for unsupervised QE, which are not based on NN, are inspired by the work on statistical MT and perform significantly worse than supervised approaches [17,22,36]. For example, Etchegoyhen et al. [36] use lexical translation probabilities from word alignment models and language model probabilities. Their unsupervised approach averages these features to produce the final score. All of these approaches were not tested in an oblivious setting; rather, they computed statistics from the same distribution of the test.

Our approach departs from the NN-based methods in that it is white-box, simple to understand, and has only two parameters (the word embedding and the loss function ℓ). Furthermore, it uses word2vec trained on generic monolingual data and requires no additional training data. It also departs from the statistical approaches as it offers a completely new algorithmic take on the QE problem, which may turn more useful in some settings. As mentioned above, the oblivious version of [17] scored a Spearman rank correlation of 0.22–0.28 compared to our 0.37 on the English-Spanish WMT’12 dataset.

4. Evaluation

In Sections 4.1 and 4.2, we discuss the performance of our method on standard benchmarks that are used in the literature and compare to both supervised and unsupervised algorithms. In Section 4.3, we explore how our algorithm correlates with the results of BLEU [6], the gold standard in the industry. We do that on two datasets that we generated. Finally, we discuss the robustness of *ObliQuE* to the choice of parameters, specifically, the pre-trained vectors.

In all the experiments described in Sections 4.1 and 4.2 we use Google’s word2vec [31] to embed words in English text, and Wikipedia-trained vectors [37] for words that are in other languages. All vectors are 300-dimensional and were trained using the skip-gram architecture with negative sampling.

4.1. Comparing against Supervised Methods

The first batch of tests compares the performance of *ObliQuE* at sentence-level QE against various supervised algorithms. The results are summarized in Table 1. Each column corresponds to a different test set that was made public by previous work. Each test set is on a different pair of languages. The rows of the table describe the results obtained by previous work on that dataset. Four of the six datasets come from the WMT competition; for comparison, we provide the results of the best and baseline supervised systems in that competition. The last row of the table states the IQR (interquartile range) of the results of previous work.

As evident from Table 1, the performance of the supervised baseline (having access to human-judgment annotated data from the same distribution of the train) is on average by merely 25% better than *ObliQuE*. For the last two test sets, De-En and Ru-En from WMT’19, the performance of all algorithms was poor. In this case, our algorithm was much

better than the baseline, and in the De-En case, it was better than the first place in that competition.

Table 1. Correlation with human judgment scores for various language pairs. The bottom three rows refer to scores achieved by other competitors. IQR stands for the interquartile range of the competitors. Complete detail in Section 4.1.

	En-Es WMT'12	En-De WMT'17	Jp-En [38]	Jp-Zh [38]	De-En WMT'19	Ru-En WMT'19
<i>ObliQuE</i>	0.37	0.20	0.31	0.122	0.07	0.064
First place	0.64	0.72	0.52	0.300	0.068	0.089
Baseline	0.58	0.42	0.43	0.125	−0.024	0.022
IQR	0.4–0.6	0.45–0.61	0.43–0.52	0.125–0.3	−0.074–0.022	0.022–0.053

We now proceed to describe in detail the test sets of Table 1. The WMT'12 QE task dataset [39] consists of 442 English–Spanish news texts produced by a phrase-based SMT system called Moses (source in English, target in Spanish). Translations were manually annotated for quality in terms of post-editing effort (1–5 scores). The winner of this MT-QE task is the author of [40] with an algorithm based on SVM and regression trees. The baseline algorithm is SVM trained on 17 features extracted using QUEST++.

The WMT'17 QE task dataset [41] contains English sentences that were translated to German by various MT-systems and ranked by correlation with HTER labels that were computed using TERCOM. We took 479 sentences translated by SYSTRAN.4847 (this system had the largest number of human-scored sentence pairs). The winner of this competition was POSTECH, which is a neural algorithm with predictor–estimator architecture [33,34]. The baseline in this task was again kernel SVM with QUEST features.

The Japanese–English and Japanese–Chinese sentence pairs are taken from [38]. The dataset contains 1676 sentences in Japanese that were obtained from role-playing dialogues of health care providers. The sentences were translated into English and Chinese using their in-house MT system, and quality was graded on a 1–5 scale, reflecting post-edit effort. The QE task was performed by a support vector regression model with a radial basis function (RBF) kernel. The model was trained once with 17 features extracted by QUEST++ (Baseline) and then with additional features extracted from a word-embedding of the sentences (First Place).

4.2. Comparing against Unsupervised Methods

There are very few examples of unsupervised algorithms competing in shared tasks like WMT. This is because the training data in those competitions are published along with human judgment scores, and the goal is of course, to win the competition. Therefore, it makes no sense to give up part of the information. Only recently, attention was drawn to the task of MT through the lens of unsupervised-learning/low-resource languages. Examples include WMT20's first of its kind unsupervised-learning/low-resource competition [27,42].

Nevertheless, the baseline in the third share task of WMT'19 [29] (a baseline is an algorithm provided by the organizers), LASER [28], is an unsupervised algorithm [28], and therefore we can compare against it. Alongside LASER, also supervised-learning algorithms competed in that task. The last two columns of Table 1 show results from that task.

As the test set of that competition was never published, the results that we report are on a sample from the train set that was published by the organizers. We sampled about 200 sentences, which is roughly the size of the test set in that same competition. As evident, *ObliQuE* outperforms LASER by much, and its performance is roughly the same as the best-supervised algorithm.

Also very noticeable is the overall poor performance of all algorithms on that dataset compared to other datasets. This may be attributed to the fact that the QE systems were tested on a variety of MT outputs from different MT systems (unlike the standard case, where all of the test data, as well as train/dev sets, are homogeneous and come from the same MT system). This poor performance provides another motivation for considering the oblivious setting as it allows to estimate the robustness of the algorithm when switching from the same testing-training distribution to a more diverse setting.

4.3. Benchmarking against BLEU

In the second batch of tests, we checked the correspondence between the scores given by *ObliQuE* and the BLEU score, which is the golden standard in MT-QE (except, of course, for human evaluation). We ran the two algorithms at the document level on two very different datasets that we assembled. The first set consisted of 100 online news pieces in English from websites like CNN, NBCNews, and NYTimes. The second dataset consisted of 100 English poems by more than 30 different poets, written in the 19th century. The average number of words per poem was 130 and 196 for a news piece.

BLEU performs evaluation against a reference. The data we collected did not come with a reference. To circumvent this problem, we performed a forward translation into each of German and French and then back to English. We performed this operation independently using three MT systems: Bing, Google, and SDL. All texts and translations are provided as supplementary data. We ended up with 1200 (En,En) pairs: we had 200 original documents in English, and each document went through two agent languages and three MT systems.

We evaluated each (En,En) pair using *ObliQuE* and BLEU, the original English text serving as a reference for BLEU, and recorded that score that each pair received from either algorithm. We then ran a competition, for every pair, between the three MT systems as follows. We fixed an agent language L (L being German or French). Each document d of the 200 original documents resulted in a triplet $\{(d, d_{Bing}^L), (d, d_{Google}^L), (d, d_{SDL}^L)\}$, where d_{Bing}^L stands for the translation of d into L and back to English using Bing, and similarly the other two are defined. We then ran *ObliQuE* and BLEU on every triplet, recording the winning MT for that triplet with respect to BLEU and with respect to *ObliQuE*. Each MT system was ranked by the number of times it won this competition.

The ranking of the three MT systems is detailed in Table 2. It shows that BLEU and *ObliQuE* are aligned in their ranking: Bing > Google > SDL, for German and French, and for both types of documents (poetry and news). The Spearman rank correlation between BLEU and *ObliQuE* was around 0.4 for poetry (across languages), and 0.27 for news via German, and 0.16 for news via French.

Table 2. Evaluation of three MT systems over 100 English poems and 100 news pieces. Each cell is the number of times that the specific MT system got the best score of the three MT systems, according to BLEU or *ObliQuE*. For example, out of 100 English poems that were translated to French and back to English, 47 poems received the highest BLEU score when using Bing for the forward-backward translation; 36 poems received the highest score using Google and 17 poems using SDL.

	French				German			
	Poetry		News		Poetry		News	
	BLEU	ObliQuE	BLEU	ObliQuE	BLEU	ObliQuE	BLEU	ObliQuE
Bing	47%	45%	65%	58%	56%	67%	86%	70%
Google	36%	35%	34%	29%	26%	25%	8%	16%
SDL	17%	20%	1%	13 %	18%	8%	6%	14%

4.4. Robustness

In this section, we evaluate the robustness of our method. Recall that we run our method in oblivious mode. In other words, it has no access to the distribution of the test set during train time. Our method uses pre-trained word embedding, which required text for training. It is, therefore, natural to ask how the choice of text to train word2vec affects the performance of the algorithm. The training of a word2vec embedding further involves fixing several parameters like the window size, the dimension of the embedding, or whether to use negative sampling or not. These parameters are explained in [30].

In this work, we use pre-trained vectors, and therefore we can check robustness for varying parameters of existing versions of word2vec. The two parameters we checked robustness for are whether negative sampling was used and which text was used.

In the tests of Section 4.1, we used a word2vec embedding trained on Google news text with negative sampling. This section replaces it with two word2vec versions, both trained on Wikipedia text, one with negative sampling and one without. The non-English text was embedded by training on Wikipedia text with negative sampling. We checked what happens when using Wikipedia but without negative sampling (we call the two options WikiNeg and WikiNorm). Table 3 shows the correlation with human judgment scores when *ObliQuE* is parameterized with the various combinations. The first line of Table 3 corresponds to the results summarized in Table 1.

As evident from the first two rows of Table 3, our method is robust to changes in the source of text that was used for training the embedding. Specifically, as long as the embedding used for the source and target languages was trained with negative sampling, the results are pretty much the same, regardless of the actual text used (Google news or Wikipedia entries). On the other hand, when using negative sampling only for one language (last two rows of Table 3), the performance is poor.

Table 3. Repeating the tests described in Table 1 with various pre-trained vector combinations. Each row corresponds to a specific pair of versions of a word2vec embedding. The text source and the negative sampling flag are embedded in the name of the version. The left-hand-side (lhs) version was used for the source language and the rhs for the target.

<i>ObliQuE</i>	En-Es (WMT'12)	En-De (WMT'17)	Jp-Zh [38]
GoogleNeg-WikiNeg	0.37	0.20	0.31
WikiNeg-WikiNeg	0.34	0.20	0.24
GoogleNeg-WikiNorm	−0.01	0.05	−0.01
WikiNeg-WikiNorm	0.16	−0.02	0.07

5. Discussion and Limitation

Machine translation is being used by millions of people daily, and therefore evaluating the quality of MT systems is an important task. While human evaluation of MT output remains crucial to look for ideas to improve MT systems still further, it is not scalable by any means. MT evaluation offers a cheap and fast alternative.

The standard pipeline of training an MT and MT-QE systems relies on a large bilingual corpus of source–target pairs, along with a human judgment. However, to this date, there is no machine translation available for most of the approximately 7000 languages spoken on the planet Earth due to scarcity of large bi-lingual corpora for training. Therefore, methods of unsupervised machine translation (and quality estimation) are important for alleviating this problem. This aspect of low-resource MT and MT-QE is an emerging field. For example, only in WMT 2020 was there a first such shared task.

We proposed an even stricter version of unsupervised MT-QE, which we called oblivious MT-QE. In the oblivious setting, besides having no scored pairs of source–target sentences, the algorithm has no access to source–target pairs from the distribution of the text on which its performance is then tested. We showed that despite such a restrictive setting,

a competitive performance of MT-QE could be achieved. We compared the performance of our oblivious algorithm to high-resource supervised learning MT-QE systems and concluded that performance degrades but remains competitive.

Our aim in this work was not to design the best MT-QE system but rather to understand if there are “universal signals” in language that can be harnessed to the task of MT-QE. The oblivious MT-QE setting allowed us to answer this question affirmatively by presenting an algorithm that performs “blind” MT-QE quite successfully.

One limitation of this current work is the use of word2vec, which is a “static” vectorization approach. The same written words, despite having different meanings, are always vectorized the same. It may be that context-based vectorization types (e.g., BERT) would be a better choice. We intend to explore this direction in future research.

Another direction for future research is to add cohesiveness as an additional feature to an existing MT-QE system, supervised or not, and check to what extent it will boost its performance.

Finally, an interesting question left for future research is whether sentence cohesiveness could be used during the training of an MT system in order to improve its quality.

Author Contributions: Conceptualization D.V.; methodology D.V. and I.E.; software I.E.; writing D.V. and I.E.; funding acquisition D.V.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the ISF grant number 1388/16.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Forcada, M.; Ginestí-Rosell, M.; Nordfalk, J.; O'Regan, J.; Ortiz-Rojas, S.; Pérez-Ortiz, J.; Sánchez-Martínez, F.; Ramírez-Sánchez, G.; Tyers, F. Apertium: A free/open-source platform for rule-based machine translation. *Mach. Transl.* **2011**, *25*, 127–144. [\[CrossRef\]](#)
- Koehn, P.; Och, F.J.; Marcu, D. Statistical Phrase-Based Translation. In Proceedings of the NAACL '03—2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—Volume 1, Edmonton, AB, Canada, 27 May–1 June 2003; Association for Computational Linguistics: Stroudsburg, PA, USA, 2003; pp. 48–54. [\[CrossRef\]](#)
- Costa-jussà, M.R.; Fonollosa, J.A. Latest trends in hybrid machine translation and its applications. *Comput. Speech Lang.* **2015**, *32*, 3–10. [\[CrossRef\]](#)
- Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.
- Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
- Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th ACL, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
- Doddington, G. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In Proceedings of the Second International Conference on Human Language Technology Research (HLT), San Diego, CA, USA, 27–27 March 2002; pp. 138–145.
- Lavie, A.; Agarwal, A. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007; pp. 228–231.
- Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisboa, Portugal, 17–18 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 392–395. [\[CrossRef\]](#)
- Snoover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In Proceedings of Association for Machine Translation in the Americas, Cambridge, MA, USA, August 2006; pp. 223–231.
- Kreutzer, J.; Schamoni, S.; Riezler, S. QUality Estimation from ScraTCH (QUETCH): Deep Learning for Word-level Translation Quality Estimation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisboa, Portugal, 17–18 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 316–322. [\[CrossRef\]](#)
- Martins, A.F.T.; Junczys-Dowmunt, M.; Kepler, F.N.; Astudillo, R.; Hokamp, C.; Grundkiewicz, R. Pushing the Limits of Translation Quality Estimation. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 205–218. [\[CrossRef\]](#)

13. Martins, A.F.T.; Astudillo, R.; Hokamp, C.; Kepler, F. Unbabel's Participation in the WMT16 Word-Level Translation Quality Estimation Shared Task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 806–811. [\[CrossRef\]](#)
14. Wang, J.; Fan, K.; Li, B.; Zhou, F.; Chen, B.; Shi, Y.; Si, L. Alibaba Submission for WMT18 Quality Estimation Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*; Association for Computational Linguistics: Belgium, Brussels, 2018; pp. 809–815. [\[CrossRef\]](#)
15. Specia, L.; Paetzold, G.; Scarton, C. Multi-level Translation Quality Prediction with QuEst++. In *Proceedings of the ACL-IJCNLP 2015 System Demonstrations*, Beijing, China, 26–31 July 2015; Association for Computational Linguistics and the Asian Federation of Natural Language Processing: Beijing, China, 2015; pp. 115–120. [\[CrossRef\]](#)
16. Kepler, F.; Trénous, J.; Treviso, M.; Vera, M.; Martins, A.F.T. OpenKiwi: An Open Source Framework for Quality Estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 117–122. [\[CrossRef\]](#)
17. Moreau, E.; Vogel, C. Quality Estimation: An experimental study using unsupervised similarity measures. In *Proceedings of the Seventh Workshop on Statistical Machine Translation, WMT@NAACL-HLT*, Montreal, QC, Canada, 7–8 June 2012; pp. 120–126.
18. Banchs, R.; Li, H. AM-FM: A Semantic Framework for Translation Quality Assessment. In *Proceedings of the 49th ACL*, Portland, OR, USA, 19–24 June 2011; Volume 2, pp. 153–158.
19. Banchs, R.E.; D'Haro, L.F.; Li, H. Adequacy-Fluency Metrics: Evaluating MT in the Continuous Space Model Framework. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 472–482. [\[CrossRef\]](#)
20. D'Haro, L.; Banchs, R.; Hori, C.; Li, H. Automatic Evaluation of End-to-End Dialog Systems with Adequacy-Fluency Metrics. *Comput. Speech Lang.* **2018**, *55*. [\[CrossRef\]](#)
21. Yankovskaya, E.; Tättar, A.; Fishel, M. Quality Estimation and Translation Metrics via Pre-trained Word and Sentence Embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 101–105. [\[CrossRef\]](#)
22. Popovic, M. Morpheme- and POS-based IBM1 scores and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, QC, Canada, 7–8 June 2012; Association for Computational Linguistics: Montreal, QC, Canada, 2012; pp. 133–137.
23. Yankovskaya, E.; Tättar, A.; Fishel, M. Quality Estimation with Force-Decoded Attention and Cross-lingual Embeddings. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*; Association for Computational Linguistics: Belgium, Brussels, 2018; pp. 816–821. [\[CrossRef\]](#)
24. Kim, H.; Jung, H.Y.; Kwon, H.; Lee, J.H.; Na, S.H. Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2017**, *17*. [\[CrossRef\]](#)
25. Rabin, M.O. *How To Exchange Secrets with Oblivious Transfer*; Technical Report TR-81; Aiken Computation Lab, Harvard University: Cambridge, MA, USA, 1981.
26. Frigo, M.; Leiserson, C.E.; Prokop, H.; Ramachandran, S. Cache-Oblivious Algorithms. *ACM Trans. Algorithms* **2012**, *8*, 4, 22. [\[CrossRef\]](#)
27. Ahmadnia, B.; Dorr, B.J. Augmenting Neural Machine Translation through Round-Trip Training Approach. *Open Comput. Sci.* **2019**, *9*, 268–278. [\[CrossRef\]](#)
28. Artetxe, M.; Schwenk, H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 597–610. [\[CrossRef\]](#)
29. Fonseca, E.; Yankovskaya, L.; Martins, A.F.T.; Fishel, M.; Federmann, C. Findings of the WMT 2019 Shared Tasks on Quality Estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 1–10. [\[CrossRef\]](#)
30. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems NIPS'13*, Harrahs and Harveys, Lake Tahoe, CA, USA, 5–8 December 2013; Volume 2, pp. 3111–3119.
31. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. GoogleNews-vectors-negative300.bin.gz—Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
32. Specia, L.; Cancedda, N.; Dymetman, M.; Turchi, M.; Cristianini, N. *Estimating the Sentence-Level Quality of Machine Translation Systems*; EAMT: Barcelona, Spain, 2009; pp. 28–35.
33. Kim, H.; Lee, J.H.; Na, S.H. Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, 7–8 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 562–568. [\[CrossRef\]](#)
34. Kim, H.; Lee, J.H. Recurrent Neural Network based Translation Quality Estimation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 787–792. [\[CrossRef\]](#)
35. Rikters, M.; Fishel, M.; Bojar, O. Visualizing Neural Machine Translation Attention and Confidence. *Prague Bull. Math. Linguist.* **2017**, *109*, 39–50. [\[CrossRef\]](#)

-
36. Etchegoyhen, T.; Martínez Garcia, E.; Azpeitia, A. Supervised and Unsupervised Minimalist Quality Estimators: Vicomtech's Participation in the WMT 2018 Quality Estimation Task. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels, 31 October–1 November 2018; Association for Computational Linguistics: Belgium, Brussels, 2018; pp. 782–787. [[CrossRef](#)]
 37. Park, K. Pre-Trained Word Vectors of 30+ Languages. 2017. Available online: <https://github.com/Kyubyong/wordvectors> (accessed on 1 August 2021).
 38. Fujita, A.; Sumita, E. Japanese to English/Chinese/Korean Datasets for Translation Quality Estimation and Automatic Post-Editing. In Proceedings of the 4th Workshop on Asian Translation, Taipei, Taiwan, 27 November–1 December 2017; pp. 79–88.
 39. Callison-Burch, C.; Koehn, P.; Monz, C.; Post, M.; Soricut, R.; Specia, L. Findings of the 2012 Workshop on Statistical Machine Translation. In Proceedings of the WMT@NAACL-HLT, Montreal, QC, Canada, 7–8 June 2012; pp. 10–51.
 40. Soricut, R.; Bach, N.; Wang, Z. The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task. In Proceedings of the Seventh Workshop on Statistical Machine Translation, Montreal, QC, Canada, 7–8 June 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 145–151.
 41. Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huang, S.; Huck, M.; Koehn, P.; Liu, Q.; Logacheva, V.; et al. Findings of the 2017 Conference on Machine Translation (WMT17). In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7–8 September 2017; pp. 169–214.
 42. Farajian, M.A.; Lopes, A.V.; Martins, A.F.T.; Maruf, S.; Haffari, G. Findings of the WMT 2020 Shared Task on Chat Translation. In Proceedings of the Fifth Conference on Machine Translation, Association for Computational Linguistics, Online, 3 June 2020; pp. 65–75.