All You Need is Source!
A Study on Source-based Quality Estimation for Neural Machine Translation

Jon Cambra Guinea
Welocalize Inc
Frederick, MD, United States
jon.cambra@welocalize.com

Mara Nunziatini
Welocalize Inc
Frederick, MD, United States
mara.nunziatini@welocalize.com

Abstract

Segment-level Quality Estimation (QE) is an increasingly sought-after task in the Machine Translation (MT) industry. In recent years, it has experienced an impressive evolution not only thanks to the implementation of supervised models using source and hypothesis information, but also through the usage of MT probabilities. This work presents a different approach to QE where only the source segment and the Neural MT (NMT) training data is needed, making possible an approximation to translation quality before inference. Our work is based on the idea that NMT quality at a segment level depends on the similarity degree between the source segment to be translated and the engine's training data. The features proposed measuring this aspect of data achieve competitive correlations with MT metrics and human judgment and prove to be advantageous for post-editing (PE) prioritization task with domain adapted engines.

1 Introduction

Quality of Neural Machine Translation (NMT) systems keeps improving and gives humans the ability to translate enormous amounts of segments in a short time. However, raw machine translation is seldom perfect. Therefore, MT in standard localization processes is most of the times followed by some level of human or automated editing aimed at fixing issues in the MT output.

In the translation industry we are witnessing a surge in demand for translation services, as well as increased requests for raw MT (without human review). Often, clients are very concerned about their translation spend, or they do not have time to translate all the content they would like to see translated, therefore more and more of them look for raw machine translation services to get savings and quicker turnaround time. However, depending on the language pairs, use cases and content types involved, raw machine translation for direct consumption (without PE) might not be a good solution.

In an ideal scenario, the quality of the output delivered by MT engines is measured before they are used in production. This exercise is aimed to understand if MT will be helpful for the linguist, or even just to understand if a MT engine training was successful or not.

Typically, the quality of MT translations is measured by comparing how different the MT output is from its reference translation. But how do we measure the quality of MT if we do not have a reference translation? It happens very frequently: imagine that you need to translate a new content type or into a new language pair for which you do not have any reference translation. This conflict led to the recent emergence of Quality Estimation (QE) techniques that try to estimate the quality of a translation when the reference information is not available. In WMT20 QE shared task, state-of-the-art (SOTA) QE models were supervised models trained exclusively on labeled data composed of source segments, the corresponding translations, and human Direct Assessment (DA) (WMT20 QE findings, 2020). The same year, an unsupervised method was proposed to estimate translations (Fomicheva et al., 2020). The paper intro-

duced the idea of using NMT as a glass-box to estimate translation quality by using token level probabilities. From that breakthrough, WMT21 SOTA QE models combined the supervised and unsupervised approaches (WMT21 QE findings, 2021).

Despite the outstanding results of these QE models, we observed that they cannot easily be implemented in production environments for different reasons. Firstly, because a substantial amount of human-labeled data is needed to fine-tune such architectures for a specific language pair and domain. In many cases, it can be problematic to find or generate this type of data without creating some domain shift between the QE training data and the data that has to be estimated in production. This can lead to catastrophic results. Secondly, these models rely on large language models that are expensive to train, store and run. Thirdly, depending on how the adapted NMT models are put in production, they can deprive the owner of NMT probabilities used as features which can also be computationally expensive to extract.

This work tries to elude these challenges by proposing a more "data-centric" direction to estimate NMT quality. Indeed, the importance of data in NMT has been extensively studied in different fields such as domain shift (Wang and Sennrich, 2020), catastrophical forgetting (Goodfellow et al., 2015; Gu and Feng, 2020) and domain robustness (Müller et al., 2020). Hence, it is well known that adapted models will have higher performance on segments from the same domain, or similar to the ones contained in the training data in some aspect. In this perspective, we think that there could be a way to estimate the NMT performance on a segment by checking the source segment and comparing it to the source segments contained in the training data. This work presents two simple techniques to perform this task: a) by measuring the similarity between the segment to be translated and the source segments found in the training data and b) by counting the number of words in the source segment that do not appear in the engine training data (unknown words for the engine).

We evaluate our approach in two steps. Firstly, we create generic engines in three language pairs, and then we adapt each one of them with client-specific data. With these six engines, we translate a set of segments for which the reference is known, score at a segment level those translations using BLEU, chrF3 and COMET, and compute our new source-similarity features. After that, we study and discuss the correlation between these new features and the segment-level MT metrics and human evaluations. Secondly, we focus on the in-domain scenarios to evaluate the impact of using this simple approach as QE metrics to prioritize the segments to be post-edited.

Our main contributions at the end of this study are: (a) a simple, unsupervised and effective approach to estimate the MT quality without checking the reference translation or before producing the translation; (b) an evaluation of how these features correlate with several MT scores and human judgement, both in generic and adapted NMT systems, similarly to previous QE methods; (c) an evaluation of how these features can be used as competitive indicators to prioritize segments to be post-edited. While the study focuses on an unsupervised segment level usage, it opens the door to explain quality changes at a project level and can inspire future architectures for QE models where the source side similarity information could be included.

2 Related work

QE QE aims to address the problem of evaluating the translation quality of a NMT model when a reference is not available. In recent years, the explosion of multilingual language models like M-BERT (Devlin et al., 2018) or XLM Roberta (Conneau et al., 2019), giving the ability to represent into a single space text from different languages, gave birth to new QE models reaching SOTA results in WMT competitions. In WMT19, a model was presented using cross-lingual sentence embedding information from both source and hypothesis (Zhang and van Genabith, 2019) to learn how to score a translation without a reference. In WMT20, quality estimators like Transquest (Ranasinghe et al., 2020) and COMET as QE (Rei et al., 2020), based on an architecture composed of a multilingual model encoding the source and the hypothesis trained on human-labeled data, outperformed older techniques. In parallel, an unsupervised technique was proposed for QE (Fomicheva et al., 2020). The paper proposed the usage of NMT as a glass-box, which means using the internal states and token level probabilities to reflect the uncertainty of the NMT at inference. This uncertainty revealed consistent correlations with human Direct Assessment (DA).

Therefore, these features are good indicators for QE. A year later, the WMT21 shared task on QE made available data composed of source, translation, and human DA, as well as the resulting glass-box features produced by the NMT model for each translation. As a consequence, the best performing models combined the WMT20 winning architectures with the uncertainty features extracted from the token-level probabilities such as QEMind's (Wang et al., 2021) and Unbabel models (Zerva et al., 2021).

Domain shift in NMT The MT field went from Statistical MT (SMT) to the current NMT models leading to state-of-the-art results in most cases (Stahlberg, 2020). The best performing MT models rely on neural architectures that are trained in two steps. Firstly, the model is trained on large amounts of generic parallel data to get a generic understanding about how to go from the source language to the target language. Secondly, this generic model is fine-tuned with bilingual data from the expected domain before it is used in production. This is what we call domain adaptation.

During this process, due to its neural architecture, NMT suffers from catastrophic forgetting (Goodfellow et al., 2015; Gu and Feng, 2020) which is the process of progressively "forgetting" previous data while strongly fitting to the new in-domain data. The performance on out-of-domain data decreases, while it improves on in-domain data. Therefore, when translating with an adapted model a text different to the in-domain data, the model could fail or produce hallucinations (Müller et al., 2020; Wang and Sennrich, 2020).

3 Source QE for NMT

In this work, we try to extract information that can describe how familiar a segment is to a given engine by comparing each source segment that needs to be translated against all the source segments included in the training data. The two following subsections propose features by transforming the segments into vectors and getting some statistical measurements of the vectors' similarity. These vector similarities are computed with the cosine similarity defined as follows for vectors A and B:

$$\mathrm{sim}(\mathrm{A}, \mathrm{B}) = \frac{A.B}{\|A\| \, \|B\|}$$

With this score we can capture how similar the vectors are. In absolute terms, the values returned are contained in [0,1]: values approaching 1 represent high similarity, while values closer to 0 represent low similarity. For explanatory purposes, we denote S_{train} the set of n_{train} source segments composing the training data and S_{test} the set of n_{test} source segments to translate.

3.1 Bag of words similarity

We create a bag of words (BOW) model for each language pair to transform all segments in $\mathrm{S}_{\mathrm{train}}$ and $\mathrm{S}_{\mathrm{test}}$ into vectors. These vectors are a simplified representation of segments where the features are a bag of words appearing in the document. Hence, the vectors describe how many times a word appears in the encoded segment. For a segment $\mathrm{s}_{\mathrm{train}}$ in $\mathrm{S}_{\mathrm{train}}$ and a segment $\mathrm{s}_{\mathrm{test}}$ in $\mathrm{S}_{\mathrm{test}}$, we denote the corresponding vector representations as $\vec{\mathrm{s}_{\mathrm{train}}}_{\mathrm{bow}}$ and $\vec{\mathrm{s}_{\mathrm{test}}}_{\mathrm{bow}}$. With this information, we compute for every segment in $\mathrm{S}_{\mathrm{test}}$ the following features.

Average BOW similarity This is the arithmetic mean of the cosine similarity between the segment to be translated and all the source segments contained in the training data.

$$\mathrm{avg}_{\mathrm{bow}}(\mathrm{s}_{\mathrm{test}}) = \frac{1}{n_{\mathrm{train}}} \sum_{s \in S_{\mathrm{train}}} \mathrm{sim}(\vec{\mathrm{s}_{\mathrm{test}}}_{\mathrm{bow}}, \vec{\mathrm{s}}_{\mathrm{bow}})$$

With this feature, we try to determine globally how similar the segment is to the full training set. However, this might not be as relevant as we think. Let's picture a scenario where all the segments in $\mathrm{S}_{\mathrm{train}}$ are completely different from the segment to translate except one. If the exception segment is almost identical, we expect that the model probably retained that information and will produce a decent translation by reproducing some similar example.

Maximum BOW similarity Given the previous argument, we hypothesize that we do not need the distance of all the segments since at inference time the NMT model will appeal to the most similar instances of the segment to translate. As a consequence, we will capture the information related to the most similar segment in $\mathrm{S}_{\mathrm{train}}$ by capturing the similarity to it. The feature is defined as follows:

$$\mathrm{max}_{\mathrm{bow}}(\mathrm{s}_{\mathrm{test}}) = \max_{s \in S_{\mathrm{train}}} \mathrm{sim}(\vec{\mathrm{s}_{\mathrm{test}}}_{\mathrm{bow}}, \vec{\mathrm{s}}_{\mathrm{bow}})$$

A limitation of this first group of features is the fact that they rely on a rudimentary transformation as it is BOW modelling. In fact, by definition, this

representation can capture quite well the string or word similarity between two segments. However, this does not constitute an accurate semantic representation.

3.2 Semantic similarity

To describe the semantic relationship between our segments we make use of SOTA models in the semantic textual similarity field, such as sentence transformer models (Reimers, Gurevych, 2019; Reimers, Gurevych, 2020). Thanks to those architectures we transform all segments in S_{train} and S_{test} into sentence embeddings with the 'all-mpnet-base-v2' model which is the mpnet-base model (Song et al., 2020) fine-tuned on a SNLI dataset with more than 1 Billion segment pairs. As a result, the vector representations produced by the model seem to capture the semantic information of the text into a unique space where the distance between two pieces of text is correlated to the semantic similarity. Hence, similar texts are closely represented while different texts have distant representations. As before, we denote $s_{train}\vec{}_{sem}$ and $s_{test}\vec{}_{sem}$ the semantic embedding representation of a segment in S_{train} and S_{test}, and compute the same features as we did with BOW representations:

Average semantic similarity The arithmetic mean of the cosine similarity between s_{test} and every segment in S_{train}

$$avg_{sem}(s_{test}) = \frac{1}{n_{train}} \sum_{s \in S_{train}} sim(s_{test}\vec{}_{sem}, s\vec{}_{sem})$$

Maximum semantic similarity The maximum cosine similarity of s_{test} over all segments in S_{train}

$$max_{sem}(s_{test}) = \max_{s \in S_{train}} sim(s_{test}\vec{}_{sem}, s\vec{}_{sem})$$

3.3 Unknown words

A problem exists with the previous similarity approaches. A segment to be translated can be highly similar to a segment in the training set but with a crucial difference. We illustrate the statement in Table 1. This example presents a segment to be translated which is highly similar to a segment used for engine adaptation. The cosine similarity is 0.95, which is only 0.05 below the score for identical segments (1.00). Both segments share the same structure and same words except for the city name. The city name is responsible for that small difference with a score representing identical segments.

source	The best museums are in **London**.
hyp	Los mejores museos están en **London**.
ref	Los mejores museos están en Londres.
source	The best museums are in Madrid.
ref	Los mejores museos están en Madrid.

Table 1: Example on NMT errors due to unknown words. The first example describes the translation produced by a NMT system. We highlight **in bold the unseen word** in training and **in red the translation error**. The second example corresponds to the most similar segment found in training with a cosine similarity of 0.95

This light difference can be a problem for the NMT model. If the word "London" is not contained in the source side of the training data, the engine will not know how to translate it into Spanish as "Londres" and will certainly produce the untranslated term since it saw that "Madrid" remained untranslated.

Therefore, we create an unk variable to capture the information. For each segment in S_{test}, the unk feature counts the number of unknown words in the segment but not in S_{train}. To do that, we produce from S_{train} the set of words occurring in the dataset which we call $D_{S_{train}}$ and $w_{s_{test}}$ the set of words in a segment s_{test}. The formula below defines how the score is computed.

$$unk(s_{test}) = n(w_{s_{test}}) - n(w_{s_{test}} \cap D_{S_{train}})$$

where n is the operator to count the number of elements, or words in this case, contained in a particular set.

4 Datasets setup

NMT data We call generic data the parallel bilingual pairs used to train the generic engines. All data was extracted from OPUS (Tiedemann, 2012) and contains different domains such as medical, political, scientific or religious among many others. The language pairs involved, and the amount of segment pairs used to train our NMT systems are described in Table 2. The test data used for experiments in Section 5 is obtained from newstest2019 for En-De and News Commentary for En-It. The En-Ko test set was made of segments from multiple domains found in OPUS (Tiedemann, 2012). In-domain data is composed of data provided by an IT security company. More specifically, the content types included in the data are User Interface (UI) and User Assistance (UA). The amount of training data used for

	Generic	In-domain
En-De	11,568,049	181,061
En-It	32,187,643	89,835
En-Ko	17,299,009	173,662

Table 2: Summary table counting the amount of segment pairs used to train NMT systems

the adapted NMT systems can be seen in Table 2. The test data is obtained from documents that were translated and reviewed in the past by human translators, which are not contained in the training data.

NMT systems We built MT engines for three language pairs (En-De, En-It and En-Ko) with OpenNMT-tf toolkit (Klein et al., 2020) by training the Transformer architecture (Vaswani et al., 2017) with the generic data described above. Additionally, we adapted those engines with the in-domain data by fine-tuning the final generic model exclusively with in-domain data (Chu et al., 2018).

MT scores In the presence of reference translations or post-edited segments, we automatically score the translations with three different segment-level metrics to have a first view of how our approach correlates with the most commonly-referenced MT metrics in the industry. At a token level, we compute the BLEU score (Papineni et al., 2002) which is extensively used across the industry despite its weakness. At a character level, we use the chrF3 score, which showed high correlations in WMT14 evaluation task (Popovic, 2015). Finally, we also rely on the SOTA metric COMET (Rei et al., 2020) with its last version 'wmt21-comet-mqm'. This metric has been described as the automatic metric which shows the highest correlation with human DA in recent years (Kocmi et al., 2021; Nunziatini, Alfieri, 2021).

Direct Assessment As for Direct Assessment, due to budget constraints, we decided to narrow the experiment to two language pairs which are particularly relevant for us for business reasons: English into Italian and English into German. Three linguists for each language pair performed Direct Assessment on 1,000 machine translated segments. We decided to involve three linguists per language because we believe it is a good compromise between budget restrictions and relevance of the exercise from a statistical point of view. The source segments were randomly selected from projects which were previously translated and reviewed. All segments in this dataset were never seen during training by the domain adapted engines. However, the content type of this dataset is very similar to the domain adapted engine training material content type.

Linguists were provided with detailed evaluation criteria and asked to score Adequacy and Fluency for each segment. For both Adequacy and Fluency, they were asked to provide a score from 1 (lowest) to 5 (highest). In order to get robust scores, fluency and adequacy scores from each annotator were standardized by transforming them into z-scores and averaged across the three linguists.

The linguists involved in this experiment were very familiar with the content type evaluated, as they are the preferred translators for this content type and client. Therefore, close attention was paid to client and domain-specific terminology and segments with little or no context were evaluated considering the context in which those segments would normally appear. Each one of them was allowed plenty of time to complete the exercise, since we understand that scoring the Adequacy and Fluency of 1,000 segments can be tiring and confusing in the long run. In order to make sure that the linguists understood the task correctly, we asked them to start with a small sample and deliver the evaluation, then wait for feedback before proceeding with the biggest sample.

5 First experiment and results

In the following experiment, we describe correlations between the proposed features and the previous MT metrics for generic and domain-adapted systems trained as explained in Section 4. Additionally, we compute the correlations with DA for the in-domain translations with data described in the same Section. Note that, to compute the indicators for the adapted engine, only the in-domain training data is considered to compare the source segments.

5.1 Settings

Benchmarks We use baseline features extracted from previous works in the field to compare the performance of our approach to QE indicators which do not need any training. On the one hand, we make use of **Comet as QE** (Rei et al., 2020), representing a supervised model trained on data from previous WMT competitions. On the other

hand, we compute the **sequence-level translation probability normalized by length** (Fomicheva et al., 2020) defined as **TP**, representing the simplest feature to extract from the NMT model at inference.

5.2 Results

Correlations with MT metrics Table 3 describes Pearson correlation between the proposed MT metrics and the source QE features for generic engine translations. On the one hand, if we compare against the baseline features (TP and $\mathrm{COMET}_{\mathrm{QE}}$), we observe competitive performance in punctual correlations. Indeed, $\mathrm{max}_{\mathrm{sem}}$ provides the best information to estimate COMET above all our proposed approaches for En-It and En-De. For its part, $\mathrm{max}_{\mathrm{bow}}$ correlates with BLEU in En-It but also with COMET in En-De. Additionally, unk can provide information for COMET only in En-De since for the other languages it rarely found segments with unknown word(s). Within this experiment with generic engines, we observe the absence of correlations for $\mathrm{avg}_{\mathrm{bow}}$, $\mathrm{avg}_{\mathrm{sem}}$ and unk when string MT metrics (chrF3, BLEU) are involved. In other words, this table shows that our features strongly correlate with semantic similarity, but not with string similarity between hypothesis and reference. This observation highlights a well-known problem for metrics like BLEU or chrF3: they fail to correctly evaluate the quality of flawless translations which use different terminology or style compared to the reference. It is particularly true in this scenario: because the engine and the test set are generic, we notice that the reference strays away from the source, whereas the model produces more literal translations.

This problem is overcome in the in-domain experiments presented in Table 4. Indeed, correlations are present to some degree for both string and semantic MT metrics since domain-adapted engines reproduce the style and terminology seen in the training material. Besides, for obvious reasons, the content type itself is not characterized by stylistic flourishes or use of synonyms. In this analysis, $\mathrm{max}_{\mathrm{sem}}$ provides consistent correlations with all the MT metrics for all the language pairs. This indicator computes leading results for string metrics in En-It and En-De, while $\mathrm{max}_{\mathrm{bow}}$ is uncorrelated. Nevertheless, for En-Ko, $\mathrm{max}_{\mathrm{bow}}$ also competes with $\mathrm{max}_{\mathrm{sem}}$. It is also the case for unk, which shows moderate correlations with almost all the metrics for En-It and En-De.

Average features Contrary to our intuition, we notice that $\mathrm{avg}_{\mathrm{bow}}$ and $\mathrm{avg}_{\mathrm{sem}}$ compute low negative correlations with some MT metrics. This means that high-quality translations correspond to source segments with low average similarity to the training set. The assumption is difficult to believe, because it would mean that completely out-of-domain segments are most likely to get high-quality translations than in-domain segments. As a consequence, we decide to drop these features from Table 4.

Correlations with Direct Assessment In Table 4, we also analyze the correlations with Fluency (Fcy) and Adequacy (Adcy) for En-It and En-De. For the first language pair, TP seems to contain the best information to estimate both Adcy and Fcy with medium-high Pearson correlations. This indicator is followed closely by $\mathrm{COMET}_{\mathrm{QE}}$ and our approaches $\mathrm{max}_{\mathrm{sem}}$ and unk, which provide medium-low correlations with these human-labeled metrics. The unk feature outperforms all our proposed approaches for Fcy, while for Adcy $\mathrm{max}_{\mathrm{sem}}$ leads the board. Similarly, for En-De, TP continues to obtain the highest correlations with DA metrics. The second place is shared by $\mathrm{COMET}_{\mathrm{QE}}$ and $\mathrm{max}_{\mathrm{sem}}$, with similar results for both indicators. Furthermore, $\mathrm{max}_{\mathrm{bow}}$ can be ranked after them with low correlations, and unk is only informative for Fcy estimation. Finally, for both language pairs the difference to TP for Fcy is moderate, but we see a larger difference to Adcy, meaning that our approaches are more competitive when measuring Fcy.

We have seen how $\mathrm{max}_{\mathrm{sem}}$ contains competitive information to estimate segment-level quality, even if it does not outperform TP globally in terms of Pearson correlation. However, we observe that our semantic similarity approach has an advantage over features using NMT probabilities in short segments. This type of segments often lack context: this causes uncertainty in NMT as it tends to return low probabilities independently of the accuracy of the translation, while $\mathrm{max}_{\mathrm{sem}}$ is able to indicate better quality if it detects that this segment can be somehow similar to some training instance.

As a final observation, we are aware that the probabilities returned by NMT systems depend on the training and inference data. We could think that our $\mathrm{max}_{\mathrm{sem}}$ and $\mathrm{max}_{\mathrm{bow}}$ indicators are

	En-It			En-De			En-Ko		
	BLEU	**chrF3**	**COMET**	**BLEU**	**chrF3**	**COMET**	**BLEU**	**chrF3**	**COMET**
TP	**0.191**	**0.376**	0.389	**0.200**	**0.297**	0.423	**0.492**	**0.662**	0.440
COMET$_{\mathrm{QE}}$	0.191	0.166	**0.821***	0.053	0.048	**0.824***	0.048	0.004	**0.622***
avg$_{\mathrm{bow}}$	-0.077	0.094	-0.099	-0.029	-0.063	-0.002	-0.021	-0.067	0.037
max$_{\mathrm{bow}}$	**0.123**	0.093	0.042	0.006	0.020	0.168	0.030	0.041	0.015
avg$_{\mathrm{sem}}$	0.048	**-0.133**	-0.152	-0.129	**-0.124**	0.148	0.053	0.043	**-0.198**
max$_{\mathrm{sem}}$	0.027	-0.063	**0.196**	**0.132**	0.032	**0.324**	-0.009	-0.050	0.021
unk	-0.015	-0.044	0.099	-0.010	-0.001	-0.131	-	-	-

Table 3: Pearson correlation table between features and different automatic MT metrics for generic NMT settings. Highest and relevant correlations from all the proposed approaches are in bold; find also in bold the best result between the two baselines. *The correlation is high because COMET and $\mathrm{COMET_{QE}}$ were trained on similar data

	En-It					En-De					En-Ko		
	BLEU	**chrF3**	**COMET**	**Fcy**	**Adcy**	**BLEU**	**chrF3**	**COMET**	**Fcy**	**Adcy**	**BLEU**	**chrF3**	**COMET**
TP	**0.230**	**0.379**	0.349	**0.374**	**0.456**	**0.131**	**0.336**	0.339	**0.217**	**0.343**	**0.344**	**0.531**	0.379
COMET$_{\mathrm{QE}}$	0.199	0.119	**0.646***	0.326	0.312	0.102	0.192	**0.604***	0.193	0.177	0.011	0.026	**0.553***
max$_{\mathrm{bow}}$	0.073	0.055	0.056	0.109	0.127	0.070	0.071	0.170	0.174	0.146	**0.282**	**0.271**	0.163
max$_{\mathrm{sem}}$	**0.241**	**0.161**	0.269	0.246	**0.253**	**0.264**	**0.285**	**0.355**	**0.189**	**0.175**	0.237	0.224	**0.174**
unk(-)	0.138	0.078	**0.374**	**0.282**	0.237	0.139	0.160	0.333	0.156	0.072	0.057	0.065	0.046

Table 4: Pearson correlation between features and different automatic MT metrics and DA scores for domain adapted NMT settings. Highest correlations with all the proposed approaches are in bold; find also in bold the best result between the two baselines.

highly correlated with the averaged probabilities. If we check the Pearson correlation for the domain adapted examples, we observe correlations with TP around 0.3 for $\mathrm{max_{sem}}$ and 0.4 for $\mathrm{max_{bow}}$. Our interpretation of this observation is that the dependence exists. However, this does not imply that the information to estimate quality contained in each indicator is redundant, as it can be seen in the performance difference between $\mathrm{max_{sem}}$ and $\mathrm{max_{bow}}$.

6 Second experiment and results: Post-Editing segment prioritization

Given the previous results showing that $\mathrm{max_{sem}}$ and unk can be considered consistent source QE indicators for domain-adapted engines, we decide to evaluate the impact in a production context where the goal is to maximize the document-level MT quality improvement by performing PE on a small subset of segments only. The following experiment uses both the indicators mentioned above to prioritize the segments to be post-edited, and compares the BLEU performance with other features.

6.1 Settings

We conduct the experiment on En-It and En-De in-domain sets where we have at our disposal, for each source segment, the corresponding hypothesis and reference as well as all the features from the previous experiment along with MT metrics and human annotations. For each QE indicator, we plot the BLEU score after simulating PE on a selected number of segments according to the corresponding indicator. The K percentage of selected segments corresponds to those with the K percent "worse" scores. As an example, if we selected 10% of all segments with $\mathrm{max_{sem}}$, we would post-edit the top 10% segments with lowest similarity scores. On the other hand, if we selected 20% of the segments with unk, we would post-edit the top 20% segments with highest number of unknown words on the source side.

Tested features We test $\mathrm{max_{sem}}$ and unk along with the features used as benchmark in the first experiment: (TP and $\mathrm{COMET_{QE}}$). Additionally, we implement a selection method which combines our two source approaches defined as unk+$\mathrm{max_{sem}}$ which first selects segments based on unk, and once all segments with at least one unknown word have been selected for PE, it uses

$\max_{\text{sem}}$ as the indicator for selection.

Benchmarks In order to understand the performance of the different approaches, we create two benchmarks. On the one hand, a lower benchmark defined as the theoretical random selection where the segments are randomly selected for PE. The values computed for that benchmark are an average approximation of multiple random selections. On the other hand, an upper benchmark described as BLEU selection, where we know beforehand which segments have the worse translations based on BLEU scores. We then use this information to choose the subset of segments to be post-edited. Note that, although this benchmark sets a high standard for the experiment, it can be outperformed when you observe the resulting corpus level BLEU score. This benchmark does not consider segments length which are essential to compute the corpus BLEU as the weighted average of segment BLEU scores.

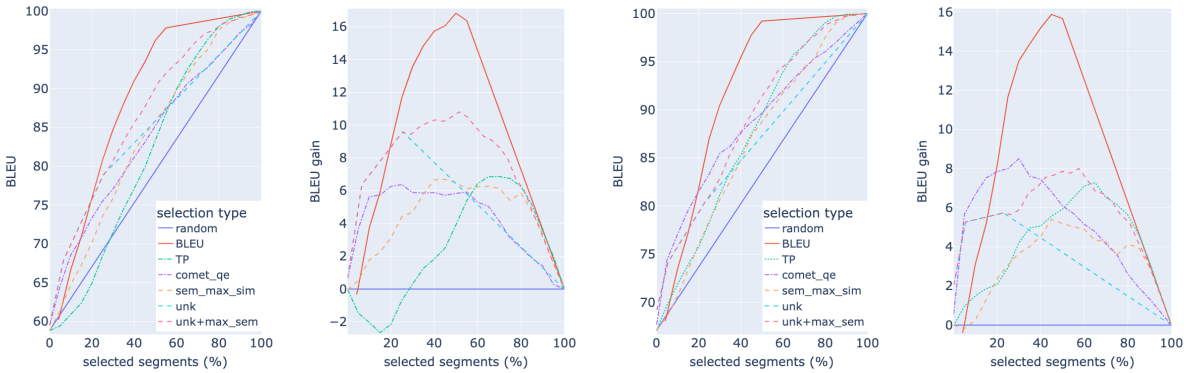
6.2 Results and discussion

The results from this experiment can be seen in Figure 1. Below we comment the results by focusing on the indicators proposed in this paper. The **unk** indicator brings benefit when selecting less than 30% of the segments. In other words, this indicator can help to prioritize segments to post-edit while there are segments with at least one unknown word. When all this type of segments has been post-edited, the remaining ones, with 0 unknown words, can only be selected randomly since the indicator scores them equally. Despite this weakness, we observe that, in the range of interest, the BLEU gain provided by unk surpasses any other indicator except COMET_{QE} for En-It. We can therefore assert that unk is an important feature to select segments to post-edit in the first stages, while segments with unknown words are present.

The performance of $\mathbf{max_{sem}}$ can be described in two ranges: [0%,40%] and [40%,100%]. In the first range, $\max_{\text{sem}}$ is between the two worst indicators. In fact, in En-De it only outperforms TP, while for En-It our proposed feature provides the lowest improvement closely behind TP. Additionally, there is a common trend in this range for both language pairs where the gain provided over the baseline monotonously increases. In the second range, the BLEU gain provided by the indicator remains globally constant at the maximum value reached in the first range and has a competitive performance compared to other indicators using hypothesis information: for En-De it outperforms COMET_{QE} and competes with TP, while for En-It it is better than COMET_{QE}, but behind TP. Finally, the heuristic indicator $\mathbf{unk{+}max_{sem}}$ can be seen as the best technique to approach the upper benchmark. For En-De, the method consistently outperforms all the other indicators for every selected amount. For En-It, the BLEU gain provided by this method is only outperformed by COMET_{QE} for the first 40% segments. Above that threshold, our combined approach outperforms any other indicator. These results are not a surprise given the previous observations made on each source QE indicator. In the first range, the poor performance of $\max_{\text{sem}}$ is compensated with the benefits from unk. While in the second range, our approach wins thanks to the advantages given by $\max_{\text{sem}}$, leading to high and constant BLEU gain.

7 Business Implementations

There are many scenarios in which this feature could be useful in production, for a Language Service Provider. While we briefly mentioned quite a few ideas in this paper, we would now like to focus on the implementation that we believe would bring the biggest benefit to the client. If we used $\text{unk}{+}\max_{\text{sem}}$ to identify a fixed amount of mostly challenging segments, by looking at the source only, and decided to post-edit only this sample of potentially incorrect segments, the client could get a dramatic improvement in the quality of the content translated with a little effort. By knowing the budget of the client for translating a document, we could estimate the number of words that can be post-edited with that budget and the extent of the improvement we could get. Let's assume that the client has budget (or time) only to post-edit the 10% of the document. In a traditional scenario, the client would probably rather have raw MT on everything, prioritize post-editing only on those part of the content (if any) that get more visibility, or even worse, post-edit only some randomly selected chunks of text. Conversely, by using these indicators, we could aim at performing post-editing only on the top 10% segments that we know have a higher probability of containing issues. Similarly, if the client has no fixed budget or turnaround time, but is trying anyway to save as much money and



(a) En-De BLEU evolution (b) En-De BLEU gain over random selection (c) En-It BLEU evolution (d) En-It BLEU gain over random selection

Figure 1: PE selection strategy comparison showing: competitive results for unk on the first 30-40% of selected segments for PE, and $\max_{\text{sem}}$ for larger selections; superiority of unk+$\max_{\text{sem}}$ for En-De and competitive results for En-It.

time as possible, we could recommend that they do PE only on that percentage of segments which could increase BLEU. This estimation would help them publish their content more quickly, because part of it would not need any human intervention and would enable linguists to focus only on what really needs to be fixed. Also, while there might of course still be errors in the MT output that do not get reviewed, this approach gives clients with budgetary constraints a focused way to spend and some certainty that the worst segments will not reach the reader.

8 Conclusions

In this paper, we offered a new approach to unsupervised segment-level QE for NMT systems by only evaluating the source segment with the help of NMT training data. By using sentence transformers and bag of words methods, we transformed all the segments into vectors and computed the maximum semantic and string similarity. These scores, along with a feature counting the number of unknown words for the NMT system, seem to contain relevant information for estimating the translation quality at a character, token, semantic, fluency and adequacy levels before producing the translation. The results were comparable to other QE techniques using NMT hypothesis or probabilities.

Moreover, we analyzed how the different indicators can heuristically help prioritize segments for PE. On the one hand, the unknown words count is an insightful indicator to select the very first segments to prioritize by choosing segments with one or more unknown words. On the other hand, the maximum semantic similarity is advantageous when the PE task can be applied to more than 40% of the segments. As a result, the combination of both indicators to select segments for PE led to the highest BLEU gains above all the QE indicators in most data selection settings.

Our work opens the door to new perspectives in QE. Firstly, we know that the source QE features presented are just a small sample of many other indicators that could be computed to compare a source segment to the NMT training data. Nevertheless, this paper highlights the importance of looking back to the training data to evaluate how easily and accurately a segment can be translated by a NMT system. Consequently, as it happened with glass-box features in the last WMT QE task, we think that future research on QE supervised models should incorporate these features or any other information that compares the data to be translated against the engine training data.

References

Chaojun Wang and Rico Sennrich. 2020. On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online.

Chu, Chenhui and Wang, Rui. 2018. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*

Conneau, Alexis and Khandelwal, Kartikay and Goyal, Naman and Chaudhary, Vishrav and Wenzek, Guillaume and Guzmán, Francisco and Grave, Edouard and Ott, Myle and Zettlemoyer, Luke and Stoyanov, Veselin. 2019. Unsupervised cross-lingual

representation learning at scale. *arXiv preprint arXiv:1911.02116*

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*

Fomicheva, Marina and Sun, Shuo and Yankovskaya, Lisa and Blain, Frédéric and Guzmán, Francisco and Fishel, Mark and Aletras, Nikolaos and Chaudhary, Vishrav and Specia, Lucia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics* 8, 539-555, MIT Press

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville and Yoshua Bengio. 2015 An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. *arXiv preprint arXiv:1312.6211*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* Istanbul, Turkey. European Language Resources Association (ELRA).

Klein, Guillaume and Hernandez, François and Nguyen, Vincent and Senellart, Jean. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 102-109

Kocmi, Tom and Federmann, Christian and Grundkiewicz, Roman and Junczys-Dowmunt, Marcin and Matsushita, Hitokazu and Menezes, Arul. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv:2107.10821*

Libovickỳ, Jindřich and Rosa, Rudolf and Fraser, Alexander. 2019. How language-neutral is multilingual BERT? *arXiv preprint arXiv:1911.03310*

Mathias Müller, Annette Rios and Rico Sennrich. 2020. Domain Robustness in Neural Machine Translation. *arXiv preprint arXiv:1911.03109*.

Nunziatini, Mara and Alfieri, Andrea. 2021 A Synthesis of Human and Machine: Correlating "New" Automatic Evaluation Metrics with Human Assessments. *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, "440-465" Virtual, *Association for Machine Translation in the Americas*, *https://aclanthology.org/2021.mtsummit-up.29"*.

Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311-318.

Popović, Maja 2015. chrF: character n-gram F-score for automatic MT evaluation *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392-395.

Ranasinghe, Tharindu and Orasan, Constantin and Mitkov, Ruslan. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. *arXiv preprint arXiv:2011.01536*

Rei, Ricardo and Stewart, Craig and Farinha, Ana C and Lavie, Alon. 2020. COMET: A neural framework for MT evaluation *arXiv preprint arXiv:2009.09025*

Reimers, Nils and Gurevych, Iryna. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics https://arxiv.org/abs/1908.10084*

Reimers, Nils and Gurevych, Iryna. 2020 Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, *Association for Computational Linguistics*.

Shuhao Gu and Yang Feng. 2020. Investigating Catastrophic Forgetting During Continual Training for Neural Machine Translation. *arXiv preprint arXiv:2011.00678*.

Song, Kaitao and Tan, Xu and Qin, Tao and Lu, Jianfeng and Liu, Tie-Yan. 2020. Mpnet: Masked and permuted pre-training for language understanding *Advances in Neural Information Processing Systems* 16857-16867

Specia, Lucia and Blain, Frédéric and Fomicheva, Marina and Fonseca, Erick and Chaudhary, Vishrav and Guzmán, Francisco and Martins, André FT. 2020. Findings of the WMT 2020 shared task on quality estimation. *Association for Computational Linguistics*

Specia, Lucia and Blain, Frédéric and Fomicheva, Marina and Zerva, Chrysoula and Li, Zhenhao and Chaudhary, Vishrav and Martins, André 2021. Findings of the WMT 2021 shared task on quality estimation *Association for Computational Linguistics*

Stahlberg, Felix. 2020. Neural Machine Translation: A Review and Survey.

Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia. 2017. Attention is all you need. *Advances in neural information processing systems*, volume 30

Wang, Jiayi and Wang, Ke and Chen, Boxing and Zhao, Yu and Luo, Weihua and Zhang, Yuqi. 2021. QEMind: Alibaba's Submission to the WMT21 Quality Estimation Shared Task. *arXiv preprint arXiv:2112.14890*

Zerva, Chrysoula and van Stigt, Daan and Rei, Ricardo and Farinha, Ana C and Ramos, Pedro and de Souza, José GC and Glushkova, Taisiya and Vera, Miguel and Kepler, Fabio and Martins, André FT. 2021. Ist-unbabel 2021 submission for the quality estimation shared task. *Proceedings of the Sixth Conference on Machine Translation*, 961-972.

Zhang, Jingyi and van Genabith, Josef. 2020. Translation Quality Estimation by Jointly Learning to Score and Rank. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2592-2598.