Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 13–21
Marseille, 20-25 June 2022

HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation

Serge Gladkoff [1], **Lifeng Han** [2,3]
[1] Logrus Global LLC, Pennsylvania, United States
[2] The University of Manchester, United Kingdom & [3] ADAPT Centre, Dublin City University, Ireland
serge.gladkoff@logrusglobal.com & lifeng.han@{manchester.ac.uk, adaptcentre.ie}

Abstract

Traditional automatic evaluation metrics for machine translation have been widely criticized by linguists due to their low accuracy, lack of transparency, focus on language mechanics rather than semantics, and low agreement with human quality evaluation. Human evaluations in the form of MQM-like scorecards have always been carried out in real industry setting by both clients and translation service providers (TSPs). However, traditional human translation quality evaluations are costly to perform and go into great linguistic detail, raise issues as to inter-rater reliability (IRR) and are not designed to measure quality of worse than premium quality translations. In this work, we introduce **HOPE**, a task-oriented and ***h***uman-centric evaluation framework for machine translation output based ***o***n professional ***p***ost-***e***diting annotations. It contains only a limited number of commonly occurring error types, and uses a scoring model with geometric progression of error penalty points (EPPs) reflecting error severity level to each translation unit. The initial experimental work carried out on English-Russian language pair MT outputs on marketing content type of text from highly technical domain reveals that our evaluation framework is quite effective in reflecting the MT output quality regarding both overall system-level performance and segment-level transparency, and it increases the IRR for error type interpretation. The approach has several key advantages, such as ability to measure and compare less than perfect MT output from different systems, ability to indicate human perception of quality, immediate estimation of the labor effort required to bring MT output to premium quality, low-cost and faster application, as well as higher IRR. Our experimental data is available at `https://github.com/lHan87/HOPE`.

Keywords: Machine Translation Evaluation, Professional Post-Editing, Human Evaluation, Error Classifications

1. Introduction

Recent studies show that human evaluation of machine translation (MT) output quality is the gold standard of translation quality evaluation, since no automated metrics can achieve equally significant results (Freitag et al., 2021; Han et al., 2020b; Han et al., 2021a; Han, 2022). However, existing advanced methods of human quality evaluations, although well developed, such as Multidimensional Quality Metrics (MQM) (Lommel et al., 2014), have the following drawbacks: 1) they are very time- and effort-consuming; 2) they are making it difficult to address the specific needs of MT post-editing, where target quality in many cases is expected to be of what TAUS (the Translation Automation User Society) has coined as "good enough" quality, which is substandard by the premium quality metrics, because they are designed to evaluate a near-premium quality material; 3) they track too many linguistic details that are unnecessary for MT output quality evaluation; 4) do not track MT-specific error types.

For large-scale deployment of MT, a more appropriate quality metric is required which: a) allows for faster learning curve for evaluators to be applied correctly; b) is faster to apply; c) is specifically designed to address less than perfect MT output of "good enough quality"; d) does not track so many unnecessary linguistic details as standard MQM metrics, designed as a tool to measure near-premium quality of human translations.

Devised with these needs and prerequisites in mind, this paper introduces a task-oriented and human-centric evaluation framework named HOPE using professional post-editing annotations for more effective MT evaluation correlating with human judgment [1].

The pilot experiments contain two tasks. Task-I is carried out using English→Russian (EN→RU) language pair in marketing domain with 111 sentence segments, using two MT engines, Google Translator and a customised MT engine. Task-II uses a survey document from business domain containing 671 segments (3,339 words) on the same translation direction but using an alternative NMT engine DeepL. The error types designed as important for post-editing and MT improvement include proper name, impact, required adaptation, terminology, grammar, accuracy, style, and proofreading error. We will explain these error types below in the methodology section. To reflect the severity level of each error, we use a scoring model with geometric progression of error penalty point weights. Error annotation and scoring can be done either without post-editing itself, or during post-editing towards a newly generated post-edited reference translation. Overall, the HOPE evaluation framework provides a human-centric translation quality evaluation of MT output and post-editing.

[1] HOPE has an alternative name: LOGIPEM (LOgrus Global Inverted Post-Editing Metrics) *ref.* `https://logrusglobal.com/`
LH's contribution is done while transferring from ADAPT Centre / Dublin City University to Uni. of Manchester

It is designed specifically to address the specifics of MT output, such as "good enough quality" evaluation tasks, and fully reflect professional post-editing efforts and human perception of translation quality.

The rest of the paper is organized as follows: Section 2 introduces related work, Section 3 presents our proposed human-centric HOPE framework, Section 4 carries out our task-oriented experiments, and Section 5 finishes the paper with conclusion and future work.

2. Related Work

In this section, we introduce some related work on post-editing, human assessment methods, task-oriented evaluation, and MT evaluation on English-Russian language pair.

As one of the earliest work on editing distance, (Su et al., 1992) introduced the word error rate (WER) metric into MT evaluation, by calculating the minimum number of editing steps to transform MT output to a reference text. This metric, inspired by Levenshtein Distance (or edit distance), takes word order into account, and the operations include insertion (adding word), deletion (dropping word) and replacement (or substitution, replacing one word with another), the minimum number of editing steps needed to match two sequences.

$$\text{WER} = \frac{\text{substitution+insertion+deletion}}{\text{reference}_{\text{length}}}. \quad (1)$$

One of the weak points of the WER metric is the fact that word ordering is not treated in an effective way. The WER scores are very low when the *word order* of system output translation is "wrong" according to the reference text. In the Levenshtein distance, the mismatches in word order require the deletion and reinsertion of the misplaced words. However, due to the diversity of language expressions, some so-called "wrong" order sentences by WER prove to be *good* translations. To address this problem, the position-independent word error rate (PER) introduced later by (Tillmann et al., 1997) is designed to ignore word order when matching output and reference. Without taking account of the word order, PER counts the number of times that identical words appear in both sentences. Depending on whether the translated sentence is longer or shorter than the reference translation, the rest of the words are either insertions or deletions.

$$\text{PER} = 1 - \frac{\text{correct} - \max(0, \text{output}_{\text{length}} - \text{reference}_{\text{length}})}{\text{reference}_{\text{length}}}. \quad (2)$$

Another way to overcome the excessive penalty on word order in the Levenshtein distance is adding a novel editing step that allows the movement of word sequences from one part of the output to another. This is an editing behavior a human post-editor would do with the cut-and-paste function of a word processor. In this light, (Olive, 2005; Snover et al., 2006) designed the translation edit rate (TER) metric that adds block movement (jumping action) as an editing step. The shift option is performed on a contiguous sequence of words within the output sentence. The TER score is calculated as:

$$\text{TER} = \frac{\#\text{of edit}}{\#\text{of average reference words}} \quad (3)$$

For the edits, the cost of the block movement, any number of continuous words and any distance, is equal to that of the single word operation, such as insertion, deletion and substitution.

TER does not generate a *new reference*. Another metric based on TER is called HTER (human targeted TER), which calculates the minimum of edits to a *new targeted reference* (post-edited translation) (Snover et al., 2006).

There are some evaluation frameworks and platforms that are carried out based on post-editing, such as Aziz et al. (2012), who introduced a tool named PET for post-editing and assessing MT. The aim of PET tool was to facilitate the post-editing of MT output to reach good-enough or publishable quality, and collect post-editing time and "detailed keystroke statistics". However, PET does not give a clear MT system quality comparisons, or their error types and their severity level suggestions.

Regarding human evaluation methods and frameworks, MQM is one of the open-sourced project maintained by group of seasoned experts and professionals (Lommel et al., 2014), initially funded by the European Commission. It has also been adopted by some official MT evaluation shared task challenges such as WMT2021. However, even a subset of the full MQM such as TAUS DQF is too large in size to be adapted to certain practical task oriented evaluations.

Post-editing and translation error categorization using professional translators have been carried out by researchers in assessing neural MT (NMT) outputs in recent studies (Bentivogli et al., 2016; Castilho et al., 2017; Esperança-Rodier and Rossi, 2019; Mutal et al., 2019). For instance, Bentivogli et al. (2016) argued that on a case study of translation quality on English-to-German language pair using the data from IWSLP2015, LSTM based NMT with attention model produces translation output that improves word order in placement of verbs with a large winning margin in comparison to traditional phrase-based statistical MT (PBSMT) model. NMT also produced less morphology and lexical errors than PBSMT in certain degrees. However, as they discussed, NMT still struggles in the aspects of handling long sentences, as well as the correct reordering of "particular linguistic constituents" that needs a deep semantic understanding. However, this mainly focused on three error categories, i.e. morphology, lexical, and word-order. In our method HOPE, we will extend the translation error types into eight commonly occurring ones.

Regarding automatic evaluation of English-Russian

(en-ru) MT outputs, the WMT metrics tasks showed that hLEPOR and cushLEPOR achieved cluster-1 performance, on EN→RU MT evaluation in news domain in WMT2013 and WMT2021 metrics shared tasks (Han et al., 2013; Han et al., 2021b; Erofeev et al., 2021). However, there is still an apparent potential to improve the metric's performance at segment level correlation towards professional human judgments.

Overall, these related work present disadvantages such as MQM is too complex and contains many linguistic detail, TER does not correlate well to professional human judgments, HTER based on TER is very abstract in reflecting the translation errors such as how exactly the frequency of "insertion, deletion, substitution" indicates the level of translation quality, and how does such frequency apply to task-oriented assessment, such as for "good enough" situation and post-editing effort? These related work also have short severity scale and normalize error penalty points to closed ranges 0 - 1 or 0 - 100, which make it difficult for human evaluation to present a transparent and tailored analysis of MT output quality.

3. Proposed Models

3.1. Model Design

In designing HOPE we have started from the following premises:

a) We need much fewer error categories than what even the MQM/DQF provides, since in real life scenarios there is seldom time, financial or human resources to dwell deeply into linguistic peculiarities, and in fact users often do not care about subtleties that linguists deem important.

b) At the same time, the HOPE error types should address typical problems of MT output, such as *inadequate* source and the fact that MT output is always *literal*, and in many cases post-editors must spot and correct such things, since they often lead to mistranslations.

c) It is also important that HOPE is scalable by design to cover a wide range of error severity, ranging from very minor errors for near-premium quality to complete garbage, which would be absolutely rejected by traditional metrics, but for MT metrics we need to distinguish the whole range of quality levels some of which are usually unacceptable by conventional metrics designed for premium translations.

d) Ideally, HOPE should allow to measure the MT output and post-editing quality of varying quality in both single output stream and between different streams of text.

3.2. Model Components/Factors

The HOPE specification is based on four pillars: 1) Only the 8 most important error types for the purpose, without error sub-types, 2) New MT-specific error types, 3) Geometric progression of severity levels, and 4) Inverted error score for the translation unit.

The following error types are defined: Impact (**IMP**), Required Adaptation Missing (**RAM**), Terminology (**TRM**), Ungrammatical (**UGR**), Mistranslation (**MIS**), Style (**STL**), Proofreading error (**PRF**), and Proper Name (**PRN**), as shown in Figure 1 with their definitions. Mistranslation is a sub-category of *accuracy* that is defined in MQM typology list [2].

One of the error types, RAM, identifies cases where the source contains an error that has to be corrected, or the target market requires adaptation; both cases go above and beyond the source, which entails an assessment only humans can do.

Another not quite standard error type is "Impact", which is used for inappropriate literal translation impairing the intended impact on the target audience.

Errors of each type can have the following severity differences: (minor, medium, major, severe, critical) with the corresponding values (1, 2, 4, 8, 16).

Error points for each Translation Unit (TU) are added to form the Error Point Penalty (EPP) of the TU (EPPTU) under-study.

$$\text{EPPTU} = \sum_{i} Error_i \times Severity(i) \tag{4}$$

where $Severity(i)$ is the severity level of $Error_i$. Each TU has its own EPPTU not depending on other TUs. Importantly, repeated errors in different TUs are not counted as one error, because MT outputs experience stochastic behavior and errors are not made consistently. One and the same error may repeat itself, but more often is mixed with other instances of a similar error. The system-level score of HOPE is calculated by the sum of overall segment-level EPPTUs:

$$\text{HOPE} = \sum_{TU_j} EPPTU_j = \sum_{i,j} Error_i \times Severity(i) \tag{5}$$

3.3. Deploying HOPE

When doing Translation Quality Evaluation (TQE) with HOPE the evaluator goes from TU to TU and reads the MT output (or post-edited text) comparing it with the source and starts from the most visible error, providing error type code and severity, then goes to the second most visible error, documenting it as well, and in certain rare cases – the third, even if two or three error classifications are usually enough to assess the quality of the translation proposal or post-editing.

The evaluator simply categorizes errors into one of the eight error types and does not spend time on a more detailed classification of minor errors.

3.3.1. Segment/Sentence-Level HOPE

We apply this metric into a sentence level (or segment-level) error severity classification, i.e. (**minor** vs **major**) with the EPPTU score (*1-4* vs *5+*). The benefit

[2] *ref.* `https://themqm.info/typology/` Mistranslation under the terminology of Accuracy.

Code	Definition	Explanation
IMP	Impact	The translation fails to convey the main thought clearly (even if translation may be literally correct, but proper translation should not be literal in target language, or has poor expression of the main thought).
RAM	Required Adaptation is Missing	Source contains error that has to be corrected, or target market requires substantial adaptation of the source, which translator failed to make. Impact on end user suffers.
TRM	Terminology	Incorrect terminology, inconsistency of translation of entities (forms, sections, etc.)
UGR	Ungrammatical	Translation is ungrammatical - needs to be fixed to convey the meaning properly.
MIS	Mistranslation	Translation distorts the meaning of the source, and presents mistranslation or accuracy error.
STL	Style	Translation has poor style, but is not necessarily ungrammatical or formally incorrect.
PRF	Proofreading error	Linguistic error which does not affect accuracy or meaning transfer, but needs to be fixed.
PRN	Proper Name	A proper name is translated incorrectly.

Figure 1: Error Types in the HOPE Metric.

of such design is that it immediately allows to distill sentences with only minor errors, with EPPTUs 1, 2, 3 and 4, and sentences with major errors (EPPTUs 5 and more). This feature of the metrics allows to see instantly after the annotation how many sentences have only minor errors *vs* those that are no need to edit, and those that contain major errors of any kind.

Since the severity scale is a geometric progression, the sentences with significant errors have much higher EPPTU and are easily distinguishable from TUs with EPPTU 1-4.

One can say that EPPTU 1-4 is precisely what is often meant by "good enough quality" of MT, where budget, time or frequency of visiting the content does not provide for premium quality translation and lower quality is just fine.

The distinction between "unchanged", "good enough" units and "must be fixed" units can serve a corpus (or system) level measure of the MT engine, and can be used to compare engines.

Also, "must be fixed" errors are automatically a recommendation for MT engineers to improve the engine; if a majority of "must be fixed" errors are fixed, the system level quality of the engine will improve significantly.

A high proportion of "must be fixed" errors are also an indication that "MT + PE" (MT plus post-editing) might not be the most efficient process for the source at hand and the MT engine concerned.

3.3.2. Word-Level HOPE

In addition to the segment-level HOPE deployment, we also design a word level HOPE evaluation in its application. The word level HOPE follows the segment-level indicators including "unchanged", "good enough", and "must be fixed". However, the statistics will be reflected at word level, e.g. how many words of the whole document/text belong to each of them three categories. Both segment/sentence-level and word-level HOPE indicators can be used to reflect the overall MT quality in translating the overall material/document. However, they can tell different aspects of the MT systems, e.g. when there are many sentences falling into very different length (short *vs* long sentences).

We leave the selection of either segment-level or word-level HOPE to exact situation where HOPE is deployed. It is always suggested to carry out both these two levels of HOPE evaluation. In our experiments, we will demonstrate both.

4. Experiments

In the first experiment, we demonstrate the application of HOPE evaluation framework at segment-level.

4.1. Task-I: Sentence/Segment-Level HOPE

In this experimental investigation, we are given an MT evaluation task to assess and compare the MT output quality in specific domain from two different MT engines, one is a custom-trained engine and another is stock Google Translate engine. The task is to compare the quality of two engines in a particular domain. We score the EEP value for each MT engine, the ratio of sentences that have no-change, minor, or major error by severity, and the ratio of Mistranslation errors.

4.2. Task-I Setup and Instructions

We are taking for post-editing a collection (file) of 111 strings (sentences) of technical marketing text in CAD/CAM (Computer Aided Design and Computer Aided Manufacturing) domain machine-translated from English to Russian with two different engines, and also a human premium reference translation. Since it is a *marketing* text, it has to be fluent with sufficient impact on the audience, and the specialized

technical nature of the product requires high accuracy of translation and adherence to industry sector terminology.

Post-editing (PE) is not a review of human translation, but a close reading against a source text and the target produced by a machine. It requires training and practice to be efficient and identify a range of different MT mistakes. The overarching aim of PE is to improve the leverage time of a translation project, while maintaining the same quality output.

It must be said that since MT is unable to detect errors in the source and simply relays the source into the target, and especially because modern NMT preserves fluency even at the cost of accuracy for out of domain content, post-editing of modern NMT output is more mentally intense work than revision or review of human translations.

The issue of target quality is important. There are often two opposite PE "modes", such as over-editing vs under-editing.

The more the translators develop post-editing skills, the faster they are able to see if a segment is under-edited or over-edited. If a sentence is under-edited, it means it does not comply with customer specifications, e.g., it is not accurate, not clear or contains a glaring error.

Some clients, like in this case, do not want to lower their quality bar. The task definition in this exercise suggests that the quality bar is NOT lowered; in other words, quality standards are the same as with traditional human-only translation process.

If the required quality is the same premium quality as with human translation, then we should not confuse post-editing translator with a "do not over-edit" instruction, which we explain below. The most advanced clients even go as far as to encourage translators to over-edit, setting a goal to achieve performance improvement without quality compromises. Contrary to widespread misconception, translators and editors in both translation and editing are not inclined to edit more than necessary. Translators and editors want to do their job fast and be done with it, and they only edit when they feel that something is not right. When doing post-editing of MT, it is actually important to try to "over-edit", because "do not over-edit" instruction prevents translators (post-editors) from looking into MT errors to identify and fix them correctly. Under-editing consequences are severe, namely the quality of translation memories may quickly and significantly degrade. Some organizations which are doing extensive implementation of MT have the internal guidelines to even instruct translators to intentionally over edit, for the reasons mentioned above, and to maintain the high quality of translation memories going forward, and the quality of translations not to deteriorate, yet preserve the ability of deployment to capture all productivity improvements.

4.3. Challenges from the Evaluation Task

There are some challenges and difficulties in post-editing that we observed not only during this task-oriented post-editing and quality assessment procedure, but in other similar experiments. We list some below. We expect that such challenges may occur in common situations in research and in practice. The description of these challenges may inspire some ideas for further improvements in practice.

1). Lack of context: Sometimes post-edited content may even have a bunch of sentences together, but the general context is missing. For example, the post-editor may have no clue what a particular proper name means, or which module and subsystem the strings are coming from. In some cases, lack of context does not quite allow to understand the meaning of source sentences. The context of a sentence is important, because it takes time to get into the context before you actually do the translation. And if the collection of sentences comes from different documents and have different context, this really slows down the work. Translators work in context much faster, because if you translate separate out of context strings you need to make additional efforts to understand the context of every other string. Often, it's hard to choose a correct term without seeing the wider context.

2). Lack of good legacy data for the translator: If translators are not given a translation memory (TM) to look up a corpus of previous translations, human translators are at disadvantage compared to both MT engine (that has been shown entire training corpus) and traditional workflows where they have access to concordance search in Translation Memory of previously done human translations. Without access to TM it is not possible to look up previous translations, and because of this it's very hard to determine whether the proposal is clumsy but valid legacy translation and therefore may need to be left unchanged, or just some sort of unfortunate "hallucination" of the MT engine, and therefore needs to be edited. We recommend that in real life post-editing jobs TM access is provided to translators, to ease and speed up the decision-making as to whether a translation clumsy to the point of being barely understood has to be corrected or not. In a hurry, leaving all such translations unchanged will degrade translation quality considerably. We would even say that it is not "fair" to human translators (and neither productive, nor beneficial for quality of the end result) to ask them to post-edit MT output trained with huge TM, without access to this TM. If the TM is available, the work of the post-editor becomes much easier. We always use our single source cloud based Memose TM tool, where translators can look up previous translations, and recommend making similar tools available for post-editors and reviewers. Working without access to accumulated knowledge is not the most efficient way, takes additional time (in some cases additional terminological research) and makes the job more difficult. In many cases

such research takes more time. It's like translation before computer aided translation (CAT) tools were invented, — productivity really suffers.

3). Quality of the source: The source text often is malformed, and even contains factual errors, while the work of the human translator is to verify the intended meaning, fix errors in the source and make adaptations. If errors in the source are not fixed, they inevitably make their way into the final translation. Many malformations of the source require correction. In such cases MT output is marked as "stylistic error" because you can't say it that way in the target language, even though translation is not "inaccurate".

4). Quality of proposals (MT outputs): MT proposals are literal by design, and often obscure the source meaning completely.

It also makes it more difficult to find good translations, especially if "do not over-edit" instructions have been given. Such instructions do not really make sense, because this is not a translation of a poem. Translators do not want to edit for the sake of editing, they edit because they need to see how well the intended meaning is transferred, and edit for the better meaning transfer, not because they want to express themselves. Proposals often use similar terms, especially if they have homonyms, without any consistency whatsoever, which is yet another complicating factor. Also, it is more difficult to come up with a good translation, if you are editing a mediocre or mistranslated proposal. Sometimes, it prevents you from even seeing that the proposal is, actually, a mistranslation.

5). Quality Specifications: The quality requirements, expectations and the amount of editing greatly depends on the customer specifications, which have to be clearly defined. The customer has to clarify clearly whether he wants to get premium translation, or is fine with "good enough" quality. The instruction "do not over-edit" is not compatible with premium quality requirement. In our HOPE metric "good enough translation" is a proposal with less than 5 penalty points. However, if you look at such sentences "en masse", the text will look significantly clumsier. In many cases customers say that they are okay with "good enough translation", but are dissatisfied at the end of the day, when such translation is delivered. Therefore, thorough discussion on translation grades and requirements is needed, especially for large projects or highly visible content.

Mistranslations pose significant difficulties hiding behind literal translations which look grammatical but do not make sense, if you try to understand the meaning (*ref.* multilingual idiom translation examples from (Valérie Mapelli, 2019) and (Han et al., 2020a)). In this class of problematic sentences all the words at a (superficial) glance seem to be right because MT uses all the words that are used in the training corpus, but in reality the meaning of the source is not making its way into the translation, and terms may be mixed and used incorrectly. Homonyms pose significant problem for NMT. This often happens when the source language "assumes" something and is jargonish, and not quite precise. In fact, sentences that belong to these two cases above take more time to fix than sentences with evident errors, because the post-editor takes time to understand whether fixing is needed, rather than simply quickly correct the proposal. For such situations, the instruction "do not over-edit" slows down post-editing. Very often translation proposals that look fluent are incorrect, and "do not over-edit" instruction will prevent rectifying such situations. We strongly believe that instruction should not be "do not over-edit", because "you must strive to do well, you don't need to make an effort to do things badly, and they will come out without any effort".

4.4. Results and Analyses from Task-I

On average, in such experiments we spend 4 hours on post-editing/annotation work per 100 strings. This does not include a second pass of review/error categories recheck, which we always recommend just to make sure that evaluation is done correctly, and to further improve the Inter-Rater Reliability (IRR). In some cases additional time is required to research the facts that are distorted in the source (not to let through incorrect information).

HOPE evaluation tasks can be done with or without final human reference translation, but in all cases the evaluator classifies errors according to the proposed HOPE quality metrics, and assigns penalty points according to error severity scale. The statistical results of evaluation are summarized in the Figure 2, and obtained quality profiles of System1 and Google Translate on this content are shown on 3. In Figure 2, NOC means "no correction needed" and SEGS means "segment scores" which is the sum of different error type penalty scores.

As can be easily seen from comparing quality profiles of System1 and Google Translate, System1 does pretty good job on this CAD/CAM content, even though it was not trained specifically on this content, but stock Google Translate performs slightly better. This fully reflects the judgment of the evaluator, who shared the overall experience as "Google is slightly better, but not much".

The numbers and diagrams reflect correctly the MT engine quality perceived by the evaluator.

This validates the effectiveness, efficiency, and transparency of our proposed HOPE evaluation framework.

4.5. Task-II: Word-Level HOPE

Here we briefly introduce the experiment we carried out using word-level HOPE evaluation indicators. In this task, we used a survey document from business domain containing 3,339 words (671 segments), where there are many sentences fall into varying length, e.g. very short. For the MT task, we used one of the popular NMT engine DeepL to carry out English-to-Russian translation.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	SRC	System1	NOC	PRN	ACR	STL	TRM	IMP	UGR	PRF	SEGS	Google Translate	NOC	PRN	ACR	STL	TRM	IMP	UGR	PRF	SEGS
	TOTAL by 111 segments		12	32	168	192	235	80	20	8	735		12	22	164	205	207	58	16	6	678
	% of total 111 segments	Segments that do not need editing	11%	4%	23%	26%	32%	11%	3%	1%	6.6		11%	3%	24%	30%	31%	9%	2%	1%	6.1
	TOTAL of segments with scores 1,2,3 and 4		35										45								
	% of segments with scores 1,2,3 and 4	Segments with minor errors (error penalty score <5)	32%										41%								
	TOTAL of segments with scores >4		64										54								
		Segments that need to be edited (error penalty score >5)	58%										49%								

Figure 2: Task-I detailed results of MT engine comparison task: System1 *vs* Google Translate.

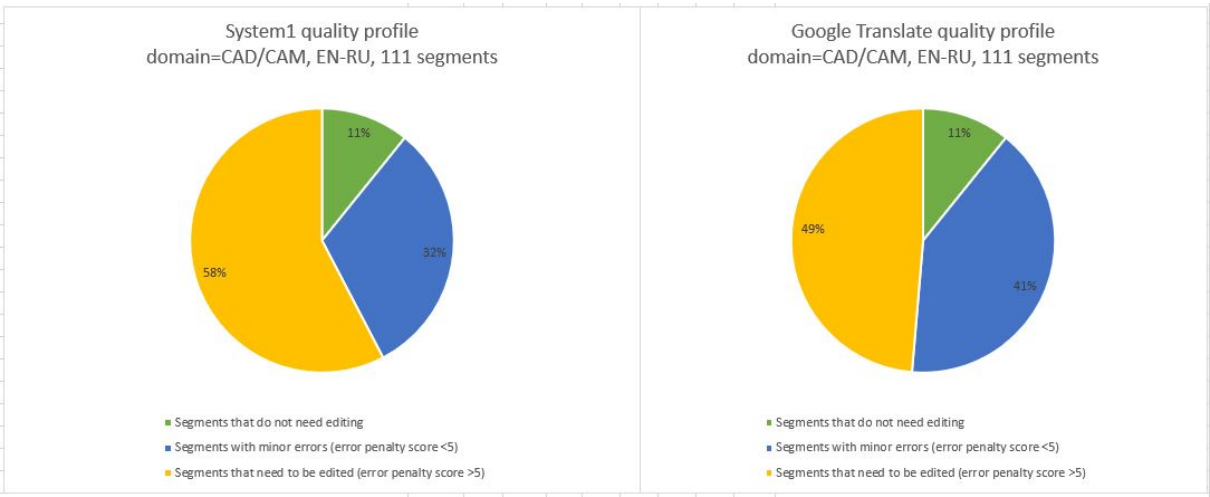

Figure 3: Task-I HOPE quality indicators: comparison of System1 *vs* Google Translate.

The evaluation comparison using segment-level vs word-level count via HOPE is shown in Figure 4 and 5.

5. Discussion and Conclusion

Overall, the proposed metric HOPE is much faster and simpler to apply than any standard MQM-based metric, including Dynamic Quality Framework (DQF) launched by TAUS [3] , yet it can be seen as an MQM implementation tailored specifically for evaluating MT output. For evaluation of MT output it is just as accurate and even more informative than standard MQM, such as DQF, and it provides additional immediate and valuable insights about the post-editing effort; it is much more adapted to the specific purpose of assessing MT and post-editing; and it is much more precise and specific than holistic rubric scores. HOPE allows to see breakdown of the estimated post-editing effort by its complexity from both segment-level and word-level. In its current form, HOPE allows to correctly and clearly quantify the "feeling" that professional post-editing translators have about the text, and provides factual numerical data instead of opinions like "in general this MT output is good, but there are certain things to be fixed", etc. It also allows to compare engines and their performance on various domains and content types in a fast, reliable and very clear, concise and transparent, quantifiable way.

Finally, according to the confidence estimation research on translation quality evaluation by (Gladkoff et al., 2021), our current sample size for MT quality assessment especially Task-II using 671 segments (3,000+ words) is large enough to support the confident conclusion drawn from the results. Regarding Task-I, it would be more accurate if we extend the sample size a bit more, e.g. double the current sample size, even though this does not affect the demonstration of the effectiveness of the evaluation framework and methodology. We leave this into our future work, as well as doing more experiments on different systems, language pairs, domains and content types[4].

6. Acknowledgment

We thank Renat Bikmatov for participation and the valuable feedback on the paper. We thank the anonymous reviewers for very valuable comments and suggestions on our work. Lifeng Han's contribution was partially funded by The University of Manchester and ADAPT Research Centre: the ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

[3] https://www.taus.net/think-tank/news/press-release *ref.* 2012/06 "TAUS LAUNCHES DYNAMIC QUALITY EVALUATION FRAMEWORK"

[4] Our experimental data for Task-I is available at https://github.com/lHan87/HOPE; Task-II involves commercial data that is confidential at this stage.

	Unchanged	Minor	Major
Total Segments (671)	278	275	118
Percent of Segments	41%	41%	18%
Wordcount (3339)	691	1718	930
Percent of Wordcount	21%	51%	28%

Figure 4: Quality Indicators via Segment *vs* Word Level HOPE: number counts and percentage.

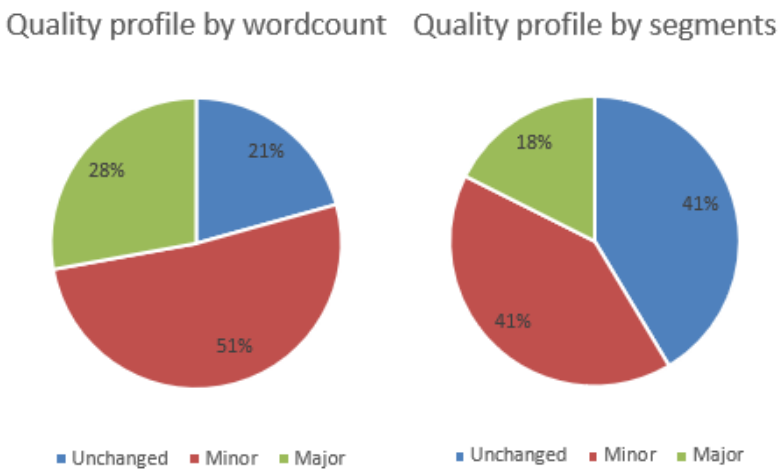

Figure 5: HOPE quality indicators via word-level (left) *vs* segment-level (right) in percentage.

7. Bibliographical References

Aziz, W., Castilho, S., and Specia, L. (2012). PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November. Association for Computational Linguistics.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, Y., Lohar, P., Way, A., Valerio, A., Miceli Barone, A. V., and Gialama, M. (2017). A comparative quality evaluation of pbsmt and nmt using professional translators. In *MT Summit 2017*, 09.

Erofeev, G., Sorokina, I., Han, L., and Gladkoff, S. (2021). cushLEPOR uses LABSE distilled knowledge to improve correlation with human translations. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 421–439, Virtual, August. Association for Machine Translation in the Americas.

Esperança-Rodier, E. and Rossi, C. (2019). Time is everything: A comparative study of human evaluation of smt vs. nmt. In *ranslating and the computer 41*.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *arXiv e-prints*, page arXiv:2104.14478, April.

Gladkoff, S., Sorokina, I., Han, L., and Alekseeva, A. (2021). Measuring uncertainty in translation quality evaluation (TQE). *CoRR*, abs/2111.07699.

Han, L., Wong, D. F., Chao, L. S., He, L., Lu, Y., Xing, J., and Zeng, X. (2013). Language-independent model for machine translation evaluation with reinforced factors. In *Machine Translation Summit XIV*, pages 215–222. International Association for Machine Translation.

Han, L., Jones, G., and Smeaton, A. (2020a). AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online, December. Association for Computational Linguistics.

Han, L., Jones, G., and Smeaton, A. (2020b). MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2970–2979, Marseille, France, May. European Language Resources Association.

Han, L., Smeaton, A., and Jones, G. (2021a). Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online, May. Association for Computational Linguistics.

Han, L., Sorokina, I., Erofeev, G., and Gladkoff, S. (2021b). cushLEPOR: customising hLEPOR metric using optuna for higher agreement with human judgments or pre-trained language model labse. In *Proceedings of Six Conference on Machine Translation (WMT2021), In Press*. Association for Computational Linguistics.

Han, L. (2022). *An Investigation into Multi-Word Expressions in Machine Translation*. PhD thesis, Dublin City University. https://doras.dcu.ie/26559/.

Lommel, A., Burchardt, A., and Uszkoreit, H. (2014). Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463, 12.

Mutal, J., Volkart, L., Bouillon, P., Girletti, S., and Estrella, P. (2019). Differences between SMT and NMT output - a translators' point of view. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 75–81, Varna, Bulgaria, September. Incoma Ltd., Shoumen, Bulgaria.

Olive, J. (2005). Global autonomous language exploitation (gale). In *DARPA/IPTO Proposer Information Pamphlet*.

Snover, M., Dorr, B. J., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceeding of AMTA*.

Su, K.-Y., Ming-Wen, W., and Jing-Shin, C. (1992). A new quantitative quality measure for machine translation systems. In *Proceeding of COLING*.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *Proceeding of EUROSPEECH*.

8. Language Resource References

Valérie Mapelli. (2019). *Chinese-English Database of Proverbs and Idioms (Chengyu) Lexical Conceptual*. ELRA, ISLRN 506-728-933-717-0.