Olena Karpina
ORCID: 0000-0001-9520-074X

UDC 811.111'332.4
DOI: 10.31558/1815-3070.2023.46.8

EVALUATING THE QUALITY OF MACHINE TRANSLATION OUTPUT WITH HTER IN DOMAIN-SPECIFIC TEXTUAL ENVIRONMENT

У статті представлено ґрунтовний аналіз якості машинного перекладу на прикладі двох популярних систем машинного перекладу: Google Translate та DeepL, у контексті англо-української мовної пари. Дослідження зосереджене на оцінюванні результатів машинного перекладу за метрикою HTER у трьох тематичних галузях: публіцистика, технічна документація та юридичні документи. Процесом оцінювання передбачено розподіл редагувань, зроблених людиною, на вставлення, видалення, заміну та перестановку, кожне з яких продемонструвало певні показники. Дослідження заглиблюється в основні причини редагувань, що випливають з граматичних, стилістичних, культурних і термінологічних труднощів. Хоча результати дослідження продемонстрували досить високу продуктивність обох систем, з різницею між машинним перекладом і версією, відредагованою людиною, менше 1 %, дослідження підкреслило постійну потребу у втручанні людини в процес машинного перекладу.

Ключові слова: *HTER, оцінювання якості машинного перекладу, постредагування, відстань редагування, помилка машинного перекладу, Google Translate, DeepL.*

1. Background. The first decades of the 21st century have witnessed the shift in the strategies employed to machine translation (MT) design (Castilho et el. 2017). The development of neural models and their integration with MT engines significantly enhanced the performance of MT systems, due to their sophisticated architecture and training algorithms, which make them capable of learning on their own errors (trial-and-error-approach). However, despite their undeniable benefits, such as rapid learning, domain adaptability, contextual understanding, neural MT systems still make mistakes which require human intervention at the final step of the translation process.

To determine the amount of human effort needed for refinement of MT output, various technics and approaches have been developed, the most widely known are Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit Ordering (METEOR), Translation Edit Rate (TER), and Human-targeted Translation Error Rate (HTER). Bilingual Evaluation Understudy, or Bleu, is an evaluation metrics which calculates the number of overlapping units (n-grams) in MT and human translation. This measure, proposed by K. Papineni et al. (Papineni et al. 2002) is widely used due to its computational efficiency and simplicity. Unlike BLEU, which has certain limitations, relying excessively on n-gram matching and taking no account of the word order, METEOR is a more comprehensive measure of quality evaluation (Banerjee and Lavie 2005). At the core of this metrics, is also the practice of matching identical words and word stems in human-generated reference translation and machine-generated translation output. However, it considers more linguistic aspects, including synonyms and paraphrases.

Translation Edit Rate or TER is an evaluation tool which measures the distance between machine-generated output and one of the human-generated reference translations, calculating the edits, identified in the process of their aligning (Snover et al. 2006) Lower scores imply the higher performance of MT output.

HTER is an improved version of TER, which increases the chances of MT output to be evaluated objectively by creating a new reference based on MT generated hypothesis sentence (For more information, see section Methods: HTER metrics as an evaluation tool).

Numerous studies have evaluated the quality of English-Ukrainian MT output, presenting a comprehensive overview of the errors identified. This can be found in the papers of Ukrainian researchers A. Gudmanian, A. Sitko, and I. Struk, who evaluated the performance of MT systems Google Translate and Pragma on the basis of the errors, which were categorized into 3 types: stylistic, terminological and content-related (Гудманян та ін. 2019). A comparative study of two online services DeepL and Google Translate was conducted in the context of journalistic writing and literary texts (Моісєєва, 2023). In our previous study, we made an attempt to compare the capacity of rule-based and neural MT services in literary translation context (Карпіна 2020). Literary text was utilized as the research material in the work by V. Karaban and A. Karaban who adopted manual approach in evaluating MT performance (Карабан В. та ін. 2021). However, experimental studies which employ HTER metrics for evaluating the MT performance in English-Ukrainian domain-specific textual environment are currently absent in scientific literature.

The goal of the research is to evaluate and compare the effectiveness and limitations of online services Google Translate and DeepL for English-Ukrainian language pair across three topic domains utilizing HTER metrics.

We aim to achieve this goal by accomplishing the following **objectives**:

- to examine the existing literature on the topic of measuring the quality of MT systems and identify the gaps in the evaluation of English-Ukrainian MT performance using HTER;
- to compile a dataset of domain-specific texts, including journalistic writing, technical documentation and legal documents;
- to apply HTER to evaluate the quality of MT generated by Google Translate and DeepL in each subject domain;
- to compare the performance of Google Translate and DeepL based on the edit distance between MT output and human post-edited translation, highlighting their strong and weak points;
- to examine the underlying causes for the identified MT errors.

2. Methods.

2.1. MT systems. In our study we examined the output of two popular MT services: Google Translate and DeepL. Both of them use AI-powered neural models to perform accurate and fluent translation (Wu et al. 2016, Bhardwaj S. et al. 2020). Nonetheless, according to the results of blind tests, conducted by a board of professional translators hired for the experiment by the DeepL team, DeepL outperformed other popular MT services in European language pairs like English-German, English-French

and English-Spanish. The results of this experiment are accessible on the DeepL official website (DeepL: How do we compare to the competition?).

2.2. Experimental dataset. The dataset consisted of 300 carefully selected sentences belonging to three different domains: journalistic writing, technical documentation and legal documents. The decision to use the dataset of 300 sentences was stipulated by its manageability, diversity and statistical significance. The chosen number of sentences was relatively easy to manage, with utmost precision to evaluation process. Also, the size of the dataset seems sufficient to achieve statistically important results and draw reliable conclusions. Comparing the sentences from three distinctive domains, representing three real-word scenarios, ensures the applicability of the results to a broad context.

Sentences belonging to journalistic writing were predominantly taken from feature articles covering a wide range of topics, such as profiles of famous people, historical events, travel experience, etc. The source of the articles are open-accessed magazines such as Discover magazine, New Scientist, The Guardian, and Wired Magazine, which have established their global position through the long history and high editorial standards, ensuring the credibility of their content. The variety of subjects covering but not limited to science, technology, history, environment etc. aligns with the objectives of our research – to evaluate the performance of the MT services in multi-domain textual environment.

Technical documentation is defined as instructions that tell you how to use a piece of equipment (Cambridge Online Dictionary). In line with this definition, we referred to various user manuals and technical specifications, which serve as primary sources of detailed technical information about wide range of technical products, systems and software including microwave ovens, videocams, air conditioners, electric guitars, computers etc. These documents, often referred to in translation practices and product localization, contain specialized terminology and specific structures peculiar to technical writing. All these considerations make user manuals and technical specifications relevant for comprehensive evaluation of MT quality in technical contexts.

Sentences referring to legal documents were taken from official websites of international organizations, such as UNO, EU, NATO, UNESCO. The selected organisations provide a wide choice of authoritative legal texts, established through their constitutive documents, such as treaties, charters, conventions. These documents are widely used in international diplomacy and policy-making, which justifies their relevance to the chosen category. Official websites of these organisations are publicly accessible, making their content a convenient and credible source for measuring the performance of MT systems within a rich legal framework of various complexity. To enhance the variability of the dataset, we also considered the regulations of international criminal court, which include a set of rules governing its operating.

The length of the sentences under analysis varied from 15 to 30 words. We believe that the selection of sentences within 15-30 word range strikes a balance between short and long sentences. Sentences with the number of words less than 15 could create a wrong impression of the quality of MT services, lacking such linguistic features as

syntactic structures, terminological variability etc. Conversely, sentences exceeding 30 words might be overly complex to fairly measure the systems' performance. By choosing medium-sized sentences, we aim to perform a comprehensive evaluation while avoiding extreme cases that might distort the results of the experiment.

2.3. HTER metrics as an evaluation tool. In order to measure the quality of the sentences in the dataset, we used HTER metrics, which measures the distance between MT output and post-edited human translation. This metrics based on human judgements was offered in respond to the limitations imposed by TER metrics, which did not take into account the phenomenon of semantic equivalence while matching hypothesis (MT output) and reference (human output) sentences (Snover et al. 2006). In contrast to TER, HTER does not only involve identifying the nearest reference sentence from the existing translations, but also presupposes creating new targeted references, generated on the basis of MT output. In other words, the hypothesis sentence is modified to the extent necessary to achieve the required level of accuracy and fluency and be semantically equal to the original.

HTER is defined as a minimum number of edits made by a human annotator to change MT output (referred to as the hypothesis sentence) normalized by the number of words in the resulting human post-edited sentence as (referred to as the reference sentence) (Snover et al. 2006).

$$\text{HTER} = \frac{\#edits}{\#reference\ words}$$

Edits are classified into insertion, deletion, substitution and shift. All edits have equal value, punctuation marks are treated like regular words, and any errors in capitalization are also counted as edits. Consider the example of an English-Ukrainian sentence pair below:

Original sentence: *Despite the difficult and often harrowing backdrop of the last year, the 24-year-old has managed to produce a new 21-track compilation album as DJ Sacred, entitled Dungeon Rap: the Evolution.*

Hypothesis (Google Translate): *Незважаючи на складний і часто жахливий фон минулого року, 24-річний хлопець зумів випустити новий* ***альбом-компіляцію*** *з 21 треку* ***під назвою*** *DJ Sacred під назвою Dungeon Rap: the Evolution.*

Reference sentence (human post-edited version): *Незважаючи на складний і часто жахливий фон минулого року, 24-річний хлопець зумів випустити новий* ***збірник*** *з 21 треку* ***як*** *DJ Sacred під назвою Dungeon Rap: the Evolution.*

In the post-edited version, we corrected a literal translation of the collocation *compilation album* (*альбом-компіляцію*), replacing it with the variant *збірник*. This substitution counts as one edit, regardless of the number of words in the translation unit (we applied the same approach in shifts, where shifting of a phrase had the same edit value as shifting of one word, being regarded as one translation unit).

The other case deals with the distortion of the original sentence semantics, where the term *DJ Sacred* refers to the name of the musician and should be specified with the word *як* instead of *під назвою*, which implies the name of the album. Using the phrase *під назвою* twice in one sentence would be inappropriate from a stylistic perspective.

The number of edits gives the following score:

$$\text{HTER} = \frac{2}{27} = 0{,}074 = 7\ \%$$

We took into account the updates of HTER metrics (Snover, 2009) which allow synonyms and paraphrases. Even if in the process of generating a reference we considered some lexical item more appropriate than the item produced by an MT system in terms of sounding and personal preferences, we tolerated it provided the item did not distort the meaning of the sentence and was stylistically applicable. Additionally, we did not count the edits caused by the agreement of a replaced unit with other sentence members. These adjustments are necessary to normalize the sentence according to the rules of morphology and they are not caused by the deficiency of an MT system.

While performing the translation in MT systems, we put the sentences into necessary contextual environment to minimize the cases of mistranslation due to ambiguity of terms and resolve a possible problem of pronoun coreference.

3. Results and Discussion.

3.1. Quality of performance with the reference to each domain. Although both MT systems demonstrated a good quality of translation in all the domains under observation, – the distance between human post-edited references and MT sentences did not exceed 1 %, – DeepL showcased slightly higher results in all the three domains. Its translation proved to be less literal, more idiomatic and fluent than that produced by Google Translate MT service.

Despite the terminological complexity of technical and legal documentation, journalistic writing turned out to be the most challenging material for translation for both MT engines, with the HTER scores 0,7 % for Google Translate and 0,5 % for DeepL. This can partly be explained by the fact that crafting of legal and technical documentation requires a special approach, which ensures rendering complex information in a clear and concise way. There are special guidelines, such as IBM Style Guide (The IBM Style Guide: Conventions for Writers and Editors), Simplified Technical English (ASD-STE100), IEEE Standard for Systems and Software Engineering – Requirements for Designers and Developers of User Documentation (ISO/IEC/IEEE 26514:2008), the aim of which is to ensure maximum clarity, accuracy and conciseness while avoiding any kind of ambiguity, arising of complex syntax, cultural nuances, terminological inconsistence etc., which in many cases may threaten the well-being and safety of individuals.

Journalistic writing, on the other hand, exhibits the features, which are deliberately avoided in legal and technical writing. Depending on the author's individual style, which is not restricted by any specific guidelines, this type of writing employs idiomatic language, incorporating a number of stylistic devices, idioms, phrasal verbs, colloquialisms, cultural references etc. leading to inaccurate or literal translation by MT engines. Additionally, this type of writing may be marked by a specific style or tone peculiar to each individual publication, which even an AI-powered MT system could struggle to match.

Consider the example below:

Original Sentence: *Queen Nefertiti is best known for the elegant limestone bust signifying her role as the Great Royal Wife of Egyptian pharaoh Akhenaten.*

Hypothesis Sentence: *Цариця Нефертіті найбільш відома завдяки елегантному вапняковому бюсту, який символізує її роль Великої* ***королівської*** *дружини єгипетського фараона Ехнатона.*

Reference Sentence: *Цариця Нефертіті найбільш відома завдяки елегантному вапняковому бюсту, який символізує її роль Великої* ***Царської*** *дружини єгипетського фараона Ехнатона.*

In this example, we spotted the inconsistency of the system in translating the title of Egyptian Queen *Nefertiti.* Having analysed historical and cultural context of the related documents, we discovered that the title of the Egyptian rulers' wives is typically translated into Ukrainian as *цариця,* as well as the power they possessed and the royal family they were part of are referred to as *царська.*

Another example showcases the instances of literal system translation, which required human post-editing:

Original Sentence: *But the location got out, attracting research groups and rich tourists alike, some of whom dove down and took artifacts back to the surface.*

Hypothesis Sentence: *Але це місце* ***вийшло****, приваблюючи дослідницькі групи та багатих туристів, деякі з яких* ***пірнали вниз*** *і* ***повертали*** *артефакти на поверхню.*

Reference Sentence: *Але це місце* ***стало відомим****, приваблюючи дослідницькі групи та багатих туристів, які* ***занурювались*** *і* ***діставали*** *артефакти на поверхню.*

The system misinterpreted the phasal verb *get out* due to its polysemantic nature, while the phrases *dove down* and *took back to the surface* translated into Ukrainian as *пірнали вниз і повертали на поверхню* respectively, sounded somewhat awkward and needed improvement for better fluency.

Nevertheless, the translation of the same sentence by another system (DeepL) proves to be more fluent and idiomatic, though still requiring post-editing.

Hypothesis Sentence: *Але інформація про місцезнаходження корабля просочилася, приваблюючи дослідницькі групи та багатих туристів, деякі з яких пірнали вниз і забирали артефакти* ***назад*** *на поверхню.*

Reference Sentence: *Але інформація про місцезнаходження корабля просочилася, приваблюючи дослідницькі групи та багатих туристів, деякі з* ***них*** *пірнали* ***на дно*** *і* ***піднімали*** *артефакти на поверхню.*

Demonstrating a profound knowledge of the context (there was no mention of the location of the ship in this sentence), the system still fails to achieve the fluency desired for natural sounding of the sentence.

While journalistic writing posed the greatest challenge for both MT systems, there were slightly varied outcomes when it came to technical and legal documentation. Google Translate performed better in legal documentation environment, with the score of 0,5 %, however, its quality dropped to 0,6 % in technical translation.

Conversely, DeepL exhibited the best outcome in technical translation achieving the score of 0,2 %. However, when it came to the translation of legal documentation, the results were slightly lower – 0,3 %.

The main challenges in translating technical content lie in complex nature of specialised terminology, the presence of various technical abbreviations and acronyms, misunderstanding and misinterpreting of the technical concepts, lack of contextual environment, updates in technical terminology due to constant advancements in technology.

The problem of inadequate interpretation due to possible lack of contextual environment is illustrated in the following example:

Original Sentence: *Do not attempt to operate this oven with the door open since this can result in harmful exposure to microwave energy*.

Hypothesis Sentence: *Не намагайтеся працювати з* ***духовкою*** *з відкритими дверцятами, оскільки це може призвести до шкідливого впливу мікрохвильової енергії.*

Reference Sentence: *Не намагайтеся працювати з* ***піччю*** *з відкритими дверцятами, оскільки це може призвести до шкідливого впливу мікрохвильової енергії.*

The MT engine renders the word oven in its primary meaning – *the part of a cooker with a door, used to bake or roast food* (Cambridge Dictionary), which is typically translated into Ukrainian as *духовка*. The system overlooked the importance of the expression *microwave energy,* which implies an alternate meaning and could result in the accurate translation of the word *oven.*

Noteworthy, DeepL outperformed Google Translate again by demonstrating a better interpretation of the context in this sentence. However, we identified the occurrences of inconsistent translation of the term *oven* in other sentences, translated by DeepL, caused by inadequate context:

Original Sentence: *Remove wire twist-ties and metal handles from paper or plastic containers/ bags before placing them in the oven.*

Hypothesis Sentence: *Зніміть дротяні закрутки та металеві ручки з паперових або пластикових контейнерів/пакетів перед тим, як ставити їх у* ***духовку***.

Reference Sentence: *Зніміть дротяні закрутки та металеві ручки з паперових або пластикових контейнерів/пакетів перед тим, як ставити їх у* ***піч***.

It is worth mentioning that the term *стяжки*, suggested by Google Translate as the translation of the word *twist-ties* better aligns with the context of the sentence. Nonetheless, we retained MT variant, adhering to the principle of semantic equivalence mentioned above. According to this principle, a translation is left unchanged if it does not distort the meaning of the sentence and the fluency is acceptable.

Legal writing exhibits similar challenges of translation as technical writing, the main problem being complex legal terminology, lack of equivalence due to jurisdictional nuances of the legal systems of different countries, literal translation arising of inability to capture the meaning of some specific legal concepts that may not have direct equivalents in the target language.

The specifics of the formal language, imposed by the rules of legal writing occasionally led to the transference of these rules to the target language, resulting in the emergence of constructions uncommon for the latter. Stylistic norms of legal writing incorporate the significant number of passive constructions, used in this type of documentation for the purpose of objectivity and focus on the action. However, such constructions translated literally into Ukrainian showcased the violation of grammatical rules and were post-edited into impersonal constructions, as in the example below:

Original Sentence*: Such person or persons were either hors de combat, or were civilians, medical personnel or religious personnel taking no active part in the hostilities.*

Hypothesis Sentence: *Така особа чи особи **були виключені** з бойових дій, або були цивільними особами, медичним чи релігійним персоналом, які не брали активної участі у бойових діях.*

Reference Sentence: *Так**у** особ**у** чи осіб **було відсторонено** від бойових дій, або вони були цивільними особами, чи належали до медичного чи релігійного персоналу, **який не брав** активної участі у бойових діях.*

This sentence also contains the instances of post-editing which contributes to the fluency of translation: the word *виключені* was replaced by the term *відсторонено*, while the phrase *належали до* was added to ensure smoother translation. Other modifications highlighted in this example were followed by essential adjustments, resulting from the insertion of additional language units and the syntactic alterations. Nonetheless, these changes were excluded from the HTER calculation since they did not arise from deficiencies within the MT systems. The only exception is the phrase *які не брали*, which was modified into *який не брав,* being the case of inaccurate coreference (*персонал не брав*, not *не брали*).

Certain level of terminology inconsistency has also been identified in the course of analysis of MT translation of legal documents. Consider the following examples:

Original Sentence*: The **Registrar** shall have administrative responsibility for the publication of the website of the Court*

Hypothesis Sentence: ***Реєстратор** несе адміністративну відповідальність за публікацію сайту суд.*

Original Sentence*: The **Registrar** shall set the schedule for the elections and inform counsel on the list of counsel by email.*

Hypothesis Sentence: ***Секретар** встановлює графіки виборів і повідомляє про це адвоката список радників електронною поштою.*

The examples illustrate the inconsistency in translating the word *Registrar* within the same context. The system failed to match the title of this position with its Ukrainian equivalent – *секретар судового засідання* and in some context translated it literally – *реєстратор*.

The overall quality of MT translation for three domains calculated according to HTER metrics is presented in Table 1.

Table 1.

HTER results for Google Translate and DeepL translation in three topic domains

Sentence Domain	Google Translate		DeepL	
	Score	**%**	**Score**	**%**
Journalistic	0,007	0,7	0,005	0,5
Technical	0,006	0,6	0,002	0,2
Legal	0,005	0,5	0,003	0,3
Overall	**0, 006**	**0,6**	**0,003**	**0,3**

3.2 Types of edits with the reference to each MT system and domain. According to HTER metrics, all MT errors were classified by the type of edit performed by the human annotator, which makes up 4 types of edits: insertion, deletion, substitution and shift.

In all the three domains, regardless the MT engine used, the predominant type of edits was substitution. It accounted for 63,58 % and 59,17 % for journalistic writing for (Google Translate and DeepL respectively); 61,4 % and 74,58 % for technical writing; and 79 % and 80 % for legal documents.

The primary causes of most substitution instances were ambiguity, insufficient knowledge of terminology, inadequate pronoun coreference and subject-predicate agreement, russian loanwords, which are no longer used in Ukrainian, contextual misinterpretation, lack of accuracy and idiomaticity of translation.

In the following example, the system demonstrated insufficient knowledge of computer terminology referring to document formatting and visual design.

Original Sentence: *Paper or originals of A4, B5 or LT size can be placed either in a portrait direction or in a landscape direction.*

Hypothesis Sentence: *Папір або оригінали формату А4, В5 або LT можна розміщувати в портретному або альбомному* ***напрямку****.*

Reference Sentence: *Папір або оригінали формату А4, В5 або LT можна розміщувати як в портретній, так і в альбомній* ***орієнтації****.*

The terms *portrait* and *landscape direction* were translated word-for-word, whereas the accurate translation should be *портретна та альбомна орієнтація****.***

The second most frequent type of edits is insertion. This tendency holds true across all the three domains translated by the Google translate MT system: 22,54 % for journalistic writing, 21,1 % for technical writing and 13 % for legal writing. A slightly different outcome was observed in DeepL performance, which positioned insertion as the second most frequent post-editing technique in legal and technical documentation, however, yielded deletion as the second most frequent edit in journalist writing, the score for deletion being 20,83 %, whereas insertion stood only at 17,5 %.

The analysis of experimental dataset unveiled diverse reasons for insertion. These encompass cultural and terminology gaps, lack of information required for conveying the intended meaning, restructuring of the sentences in order to improve the natural

flow, adaptation of grammatical constructions to grammatical and stylistic norms of the target language, which required extra linguistic units.

Deletion occurred in the process of post-editing for various reasons. The primary cause for deletion was the presence of extraneous words resulting from the literal translation of the MT engine. Consider the example:

Original Sentence: *"She developed a very impressive peeling algorithm, you know, she's faster than humans, about three times faster," says Brecht.*

Hypothesis Sentence: *«Вона розробила **дуже** вражаючий алгоритм очищення, знаєте, вона швидша за людей, приблизно в три рази швидше», – каже Брехт.*

Reference Sentence: *«Вона розробила вражаючий алгоритм очищення, уявляєте, вона швидша за людей, приблизно втричі швидша», – каже Брехт.*

In this sentence, the MT of the phrase *a very impressive* as *дуже вражаючий* consists of a redundant linguistic unit *дуже***.** The semantic structure of the adjective *вражаючий* already incorporates a semantic component of intensity, making any other intensifying elements sound unnecessary as they break the natural flow of the sentence and contradict to the stylistic norms of the target language.

Shift as a type of edits demonstrated the lowest frequency in all the three domains for both MT systems with the results 2,31 % and 2,5 % in journalistic writing (for Google Translate and DeepL respectively); 7,02 % and 3,39 % for technical writing, and 1 % and 3,64 % for legal writing.

Most instances of shift occurred when the MT output replicated the syntactic structure of the source language, leading to inaccurate and hardly readable constructions in the target language. This can be observed in the following example:

Original Sentence: *The duty roster of legal officers of the Chambers shall be maintained by the Presidency and made available to the Registry.*

Hypothesis Sentence: *Список чергувань юридичних працівників Палат веде Президія та надається до Секретаріату.*

Reference Sentence: *Президія веде список чергувань юридичних працівників Палат та надає його Секретаріату.*

The shift employed in the reference sentence contributes to naturalness of sounding and ensures correct grammatical structuring. The insertion of the word *його* clarifies that it was the list available to the Registry, which enhances the clarity of translation.

The tables and figures below provide a general overview of the number of occurrences of each type of edits across the three domains in Google Translate and DeepL output.

Table 2

Google Translate: Distribution of Edit Types.

Type of edits	**Journalistic**		**Technical**		**Legal**	
	#	%	#	%	#	%
Insertion	39	22,54	24	21,1	13	13

Deletion	20	11,56	12	10,52	7	7
Substitution	110	63,58	70	61,4	79	79
Shift	4	2,31	8	7,02	1	1
Total # edits	**173**	**100**	**114**	**100**	**100**	**100**

Table 3

DeepL: Distribution of Edit Types.

Type of edits	**Journalistic**		**Technical**		**Legal**	
	#	%	#	%	#	%
Insertion	21	17,5	7	11,86	5	9,09
Deletion	25	20,83	6	10,17	4	7,27
Substitution	71	59,17	44	74,58	44	80
Shift	3	2,5	2	3,39	2	3,64
Total # edits	**120**	**100**	**59**	**100**	**55**	**100**

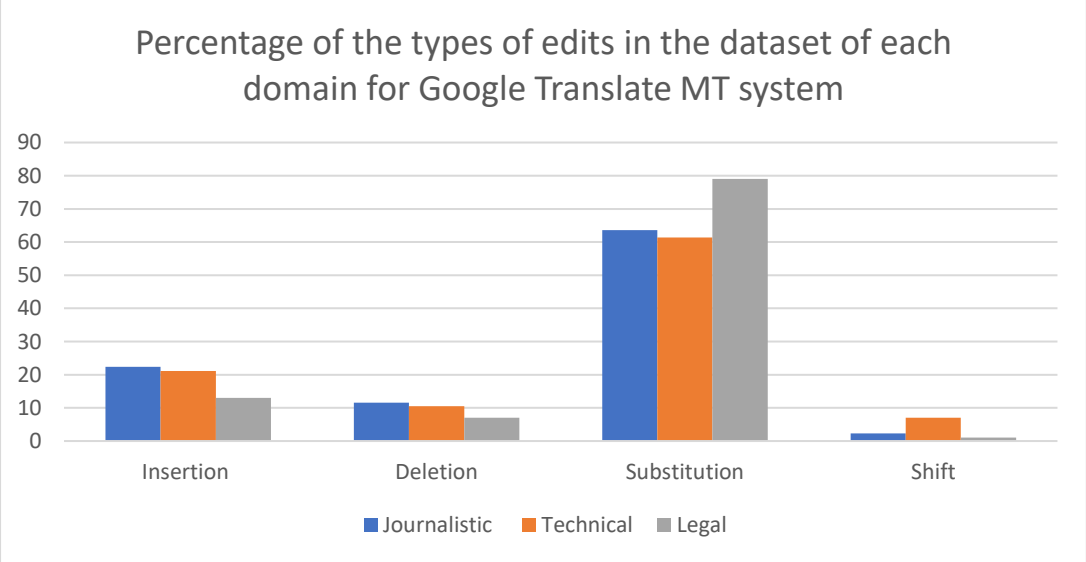

Fig.1

Percentage of the types of edits in the dataset of each domain for Google Translate MT system

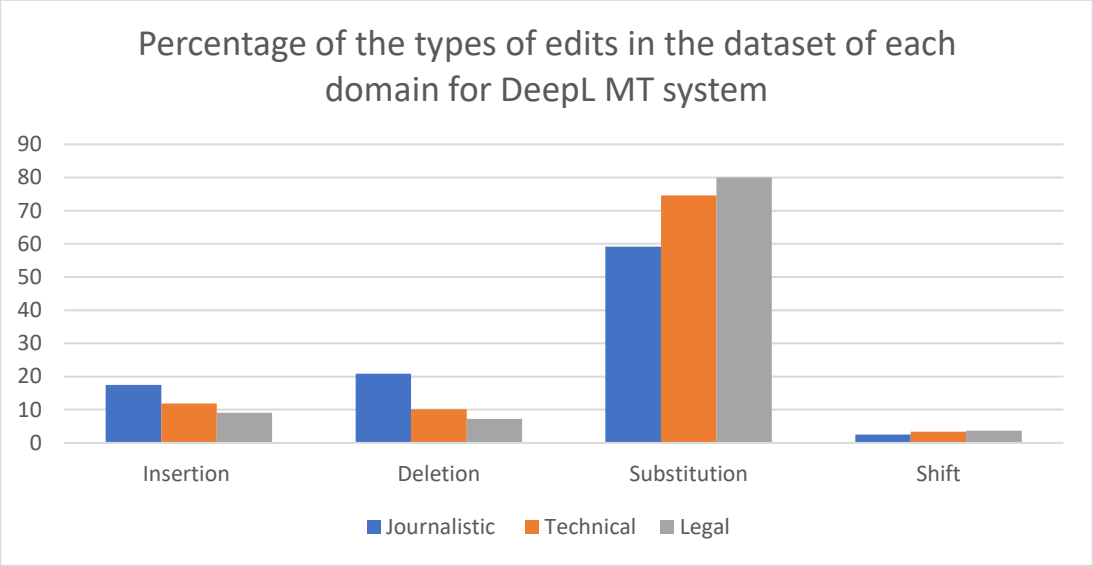

Fig. 2

Percentage of the types of edits in the dataset of each domain for DeepL MT system

4. Conclusions. The MT services Google Translate and DeepL have demonstrated a high performance, translating across all the three domains: journalistic writing, technical documentation and legal documents. The edit distance between MT outcome and human post-edited version has not exceeded 1 %. This implies the idea of their applicability in a variety of contexts, making them quick and effective tools for providing translation with minimum human efforts. Despite the high evaluation results, the study has emphasized the need for human post-editing. Irrespective of the domain of translation, both MT systems revealed instances of inaccurate translation stemming from the disregard for grammatical and stylistic norms of Ukrainian language. Multiple occurrences of russianisms – calqued russian words, passive constructions, uncommon for Ukrainian language, as well as instances of low readability and awkward phrasing have been identified. It suggests that human intervention is still crucial for ensuring high-quality translation, especially in culture-sensitive context.

Future research could consider expanding the evaluation domains to a wider spectrum of contexts, i. e. literary text, healthcare literature, financial documentation etc. Such perspective could enhance the practicality and real-world value of the research, addressing specific needs and demands of MT users. Beyond domain expansion, future perspectives could also include cross-language comparison of MT performance, including additional pair of languages in the experimental evaluation scope. Measuring MT performance in multilingual contextual environment could enable a more comprehensive understanding of language-specific challenges of each MT service and provide MT users with the detailed information leading to the right choice of a translation tool.

Reference

1. Гудманян, А., Сітко, А., Струк, І. «Функціонально-прагматична адекватність машинного перекладу публіцистичних текстів». [В] *Науковий журнал Львівського державного університету безпеки життєдіяльності «Львівський філологічний часопис»*: зб. наук. праць 5. Львів, 2019: 48–54.
   [Hudmanyan A., Sitko A., Struk I. «Funktsional'no-prahmatychna adekvatnist' mashynnoho perekladu publitsystychnykh tekstiv». [V] *Naukovyy zhurnal L'vivs'koho derzhavnoho universytetu bezpeky zhyttyediyal'nosti «L'vivs'kyy filolohichnyy chasopys»*: zb. nauk. prats' 5. L'viv, 2019: 48–54.]
2. Карабан, В. І., and А. В., Карабан. «Чи настає вже ера художнього машинного перекладу?(контекстуальні помилки машинного перекладача Deepl)». [В] *Мова і культура,* 2021: 438–445.
   [Karaban, V. I., and A. V., Karaban. «Chy nastaye vzhe era khudozhn'oho mashynnoho perekladu? (kontekstual'ni pomylky mashynnoho perekladacha DeepL)». [V] *Mova i kul'tura*, 2021: 438–445.
3. Карпіна, Олена «Компаративний аналіз літературного й машинного перекладів (на матеріалі фрагментів роману С. Плат "The Bell Jar")». [В] *Актуальні питання іноземної філології* : наук. журн. / редкол. І. П. Біскуб (гол. редактор) та ін. Луцьк: Східноєвроп. нац. ун-т ім. Лесі Українки 3, 2020: 94–101.
   [Karpina, Olena «Komparatyvnyy analiz i mashynnoho perekladiv ((na materiali frahmentiv romanu S. Plat "The Bell Jar")». [V] *Aktual'ni pytannya inozemnoyi filolohiyi* : nauk. zhurn. / redkol. I. P. Biskub (hol. Redactor ta in. Luts'k: Skhidnoyevrop. nats. un-t im. Lesi Ukrayinky 3, 2020: 94–101.]
4. Моісєєва, Наталія, Ольга, Дзикович, and Аліна, Штанько. «Машинний переклад: порівняння результатів та аналіз помилок DeepL та Google Translate». [В] *Advanced Linguistics* 11, 2023: 78–82.
5. [Moisyeyeva, Nataliya, Ol'ha, Dzykovych, and Alina, Shtan'ko. «Mashynnyy pereklad: porivnyannya rezul'tativ ta analiz pomylok DeepL ta Google Translate». [V] *Advanced Linguistics* 11, 2023: 78–82.]
6. ASD Simplified Technical English Specification ASD-STE100. URL: https://www.asd-ste100.org/ (29.08.2023)
7. Bhardwaj, Sh., Hermelo, D. A., Langlais, Ph., Bernier-Colborne, G., Goutte, C., and Simard, M... "Human or Neural Translation?". [V] *In Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain. International Committee on Computational Linguistics, 2020: 6553–6564. URL: https://aclanthology.org/2020.coling-main.576.pdf (9.09.2023)
8. Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., Waya, A. "Is Neural Machine Translation the New State of the Art?". [V] *The Prague Bulletin of Mathematical Linguistics* 108(108), 2017:109–120. DOI: 10.1515/pralin-2017-0013
9. How does DeepL work? URL: https://www.deepl.com/en/blog/how-does-deepl-work
10. ISO/IEC/IEEE 26514:2008 (IEEE Standard for Systems and Software Engineering - Requirements for Designers and Developers of User Documentation). URL: https://www.iso.org/standard/43073.html (29.08.2023)
11. Banerjee, S. and Lavie, A. [V] "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". *Proceedings of the ACL 2005 Workshop on Intrinsic and*

Extrinsic Evaluation Measures for MT and/or Summarization: 2005 URL: https://aclanthology.org/W05-0909/ (29.08.2023)
12. Cambridge dictionary URL: https://dictionary.cambridge.org/dictionary/english/oven (9.09.2023)
13. How do we compare to the competition? URL: https://www.deepl.com/en/quality.html
14. Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. "Bleu: a Method for Automatic Evaluation of Machine Translation". [V] *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* 2002 URL: https://aclanthology.org/P02-1040.pdf (29.08.2023)
15. Snover, Matthew G. et al. "TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate". [V] *Machine Translation* 23 (2), September 2009: 117–127. DOI: 10.1007/s10590-009-9062-9
16. Snover, Matthew, et al. A Study of Translation Edit Rate with Targeted Human Annotation. *In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers.* Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas 2006: 223–231.
17. The IBM Style Guide: Conventions for Writers and Editors. URL: https://ptgmedia.pearsoncmg.com/images/9780132101301/samplepages/0132101300.pdf (29.08.2023)
18. Wu Y., Schuster M., Chen Zh., Le Quoc V., Norouzi M., Macherey W., Krikun M., Cao Yu., Gao Q., Macherey K., Klingner J., Shah A., Johnson M., Liu X., Kaiser L., Gouws S., Kato Y., Kudo T., Kazawa H., Stevens K., Kurian G., Patil N., Wang W., Young C., Smith J., Riesa J., Rudnick A., Vinyals O., Corrado G., Hughes M., and Dean, J. "Google's neural machine translation system: Bridging the gap between human and machine translation". arXiv: 1609.08144v2 [cs.CL] 2016. https://doi.org/10.48550/arXiv.1609.08144

List of sources

1. Discover magazine URL: https://www.discovermagazine.com/
2. European Convention on Human Rights. European Court of Human Rights URL: https://www.echr.coe.int/documents/d/echr/convention_eng
3. European Council URL: https://www.consilium.europa.eu/en/council-eu/presidency-council-eu/
4. New Scientist. Weekly Magazine URL: https://www.newscientist.com/
5. Regulations of the Office of the Prosecutor. Published by the International Criminal Court 2011 URL: https://www.icc-cpi.int/sites/default/files/Publications/Regulations-of-the-Office-of-the-Prosecutor.pdf
6. SafeManuals URL: https://safe-manuals.com/
7. The Guardian URL: https://www.theguardian.com/
8. The North Atlantic Treaty. The North Atlantic Treaty Organization URL: https://www.nato.int/cps/en/natohq/official_texts_17120.htm
9. UNESCO. Constitution URL: https://www.unesco.org/en/legal-affairs/constitution?hub=66535
10. United Nations Charter (full text) URL: https://www.un.org/en/about-us/un-charter
11. Wired Magazine URL: https://www.wired.co.uk/

EVALUATING THE QUALITY OF MACHINE TRANSLATION OUTPUT WITH HTER IN DOMAIN-SPECIFIC TEXTUAL ENVIRONMENT

Olena Karpina
Applied Linguistics Department, Lesya Ukrainka Volyn National University, Lutsk, Ukraine.

Abstract

Background: The implementation of neural networks in MT systems design has greatly challenged the existence of human translation. The emergence of translating models which adopt mechanisms of translation, imitating the work of the human brain, aroused high expectations of

immediate breakthrough. However, despite significant improvements in accuracy and fluency of AI-powered MT systems, human assistance remains essential in the translation process.

Purpose: The Purpose of the research is to evaluate and compare the effectiveness and limitations of free online services Google Translate and DeepL for English-Ukrainian language pair across three topic domains utilizing HTER metrics.

Results: Google Translate and DeepL demonstrated rather high level of performance, with the edit distance less than 1 % in each of the three domains. Nonetheless, it is still early to talk about self-sufficient MT systems which can operate completely without human assistance. The main causes for MT translation errors were identified as terminological issues, including wrong translation equivalent and terminology inconsistency, contextual issues, stemming in from the inability to interpret a wider context; accuracy errors due to the gap between grammatical systems of the source and target languages, fluency concerns, and various cultural and stylistic discrepancies.

Discussion: The most challenging input for both MT systems appeared journalistic writing, with the HTER scores 0,7 % for Google Translate and 0,5 % for DeepL (the percentage indicates edit distance between MT and human post-edited translation). The errors made by MT systems are rooted in the stylistic features of this genre of writing, bearing traits of the author's individual style, including idioms, phrasal verbs, stylistic figures. In technical writing, DeepL performed considerably better, with the edit distance just 0,2 %, while Google Translate exhibited the most favorable performance within the legal textual environment, demonstrating the edit distance of 0,5 %, whereas in technical writing the outcome was slightly worse – 0,6 %. DeepL, having outperformed Google Translate in all experimental domains, exhibited the edit distance of 0,3 % in technical writing.

Concerning the types of edits, categorized according to HTER metrics into insertion, deletion, substitution and shift, the most frequent edit employed by human post-editors was substitution, accounting for roughly over a half of all edits made during the post-editing process. Notably, in legal writing its score raised to 79 % for Google Translate and 80 % for DeepL, which can be explained by terminological inappropriacy and structural challenges due to distinct syntactic rules of the source and target languages. The least frequent edit was shift, its value did not exceed 4 % for all experimental domains.

Keywords: HTER, MT quality evaluation, post-editing, edit distance, MT error, Google Translate, DeepL.

Vitae

Olena Karpina PhD in Philology (Germanic Languages), Associate Professor of Applied Linguistics Department, Lesya Ukrainka Volyn National University.

The scope of scientific interests covers translation studies, linguistics of emotion, lexical semantics, communicative linguistics.

Correspondence: karpina@vnu.edu.ua

Надійшла до редакції 04 вересня 2023 року
Рекомендована до друку 01 жовтня 2023 року