ModernMT Comparative Evaluation Study

February 2024

by Kirti Vashee, Tech Evangelist at Translated

based on an independent experiment commissioned to
Achim Ruopp, Polyglot Technology founder

translated.

Summary

translated.

Introduction MT Evaluation for Enterprises

Machine Translation (MT) system evaluation is necessary for enterprises that are considering increasing the use of automated translation to meet the increasing information and communication needs of the global customer. Managers need to understand which MT system is best for their specific use case and language combination, and which MT system will improve the fastest with their data and with the least amount of effort to perform best for the intended use case.

What is the best MT system to use for my specific use case, and this language combination?

Many of the current approaches to making MT system quality assessments are based on unreliable and archaic approaches that provide incomplete and incoherent ratings that are not necessarily useful for business purposes.

The large majority of evaluations done focus on measuring the performance of static MT systems, which are generally designed for generic translation purposes and only measure the performance of the MT system on a single “Test Set” providing a score based on an imperfect measure of syntactic or semantic similarity. If the test set is not closely related to the purpose at hand the results can be misleading and not useful for making intelligent MT system selection decisions. Third-party consultants tend to rank MT systems by these scores (COMET, BLEU, TER, ChrF) even though these metrics only provide a very rough idea of how a system may perform in production use. The truth is that these scores only provide a very rough idea of how MT systems might perform on your actual use case requirements. Also, these measurements may only be true for an instant and should not be considered absolute persistent truth. The results could change every month with new test sets and very small numerical differences between systems are often meaningless

What MT system will improve the fastest with my unique data and need the least amount of effort, to perform the best for my intended use case?

ModernMT is a system designed to be easily tuned to an innumerable number of use cases and generally every user would provide additional data to tune the MT system to perform better on their specific and unique requirements. This is effortless and requires no special technical skill.

ModernMT is a unique MT system amongst the MT systems that are available today, as it is the first dynamic and continuously improving MT system.

It requires no additional effort to use enterprise translation memories (TMs) to tune the MT system to use enterprise context and subject domain. This integration is seamless and transparent and allows a ModernMT system to improve on an hourly basis if active corrective feedback is being provided. As the TM expands with corrections of MT output, the emerging MT system output will continue to improve.

Thus, many MT evaluation approaches that assume that MT systems are essentially static will overlook the dynamic and continuing improvement aspects of a system like ModernMT.

This report is an attempt to provide a more informed and accurate picture of how ModernMT compares to leading public MT systems, with no special effort made by users other than providing access to TM resources for the ModernMT engine to use if they do indeed contain relevant and useful segments.

About The Evaluation Process Used

- The **original comparison** and evaluation presented in this report was done by Achim Ruopp of Polyglot Technology in April 2023 with ModernMT V6 comparing it to public MT system output produced at that same time. In the original evaluation, Adaptive ModernMT outperformed the public MT alternatives on COMET scores in all the languages tested. This experiment was commissioned to Mr Ruopp by Translated to evaluate the incremental benefit of adaptive machine translation over custom systems.

- This **current evaluation** updates the output from all the systems being compared and uses the output from the significantly improved ModernMT V7.

- **In all the comparisons done by Polyglot Technology in the original study, there is no "training" step involved for any of the MT systems that are being compared.**

- **The Adaptive ModernMT is provided access to the TM of a prior segment after it has attempted to translate the segment.** This simulates the active MT output correction process by adding this segment to the referenced TM and provides insight into how easily and quickly the ModernMT system output quality improves.

- Static or **Generic MT systems tend be used without modification by the large majority of users**, but in contrast, every user of an Adaptive MT system like ModernMT will naturally adapt and tune the system to their specific requirements and focus as the technological complexity of doing this has been eliminated.

The Static MT Experience

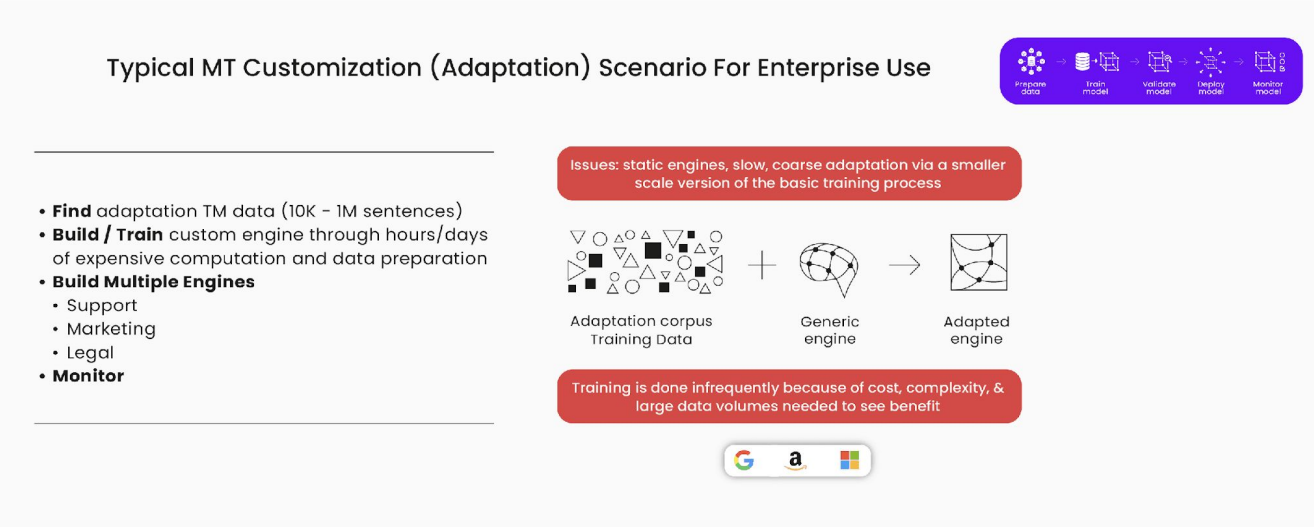
While some Generic (Static) MT systems like Google, Microsoft, and Amazon can also be tuned with customer data (through a process called “training”) there are usually additional costs for this and additional complexity involved. This customization also typically requires large amounts of TM, to see any beneficial impact, relatively much larger volumes of TM than a system like ModernMT requires. The unique approach of ModernMT to optimize and tune the engine at the sentence level greatly reduces the need for large volumes of required TM to tune the MT system to the desired enterprise domain.

The typical MT customization process using static engines is described below. **The customization effort and process is a scaled-down version of the generic engine development process.** Typically, it requires the collection and incorporation of enterprise translation memory relevant to the use case into the generic model via a scaled-down "training process."

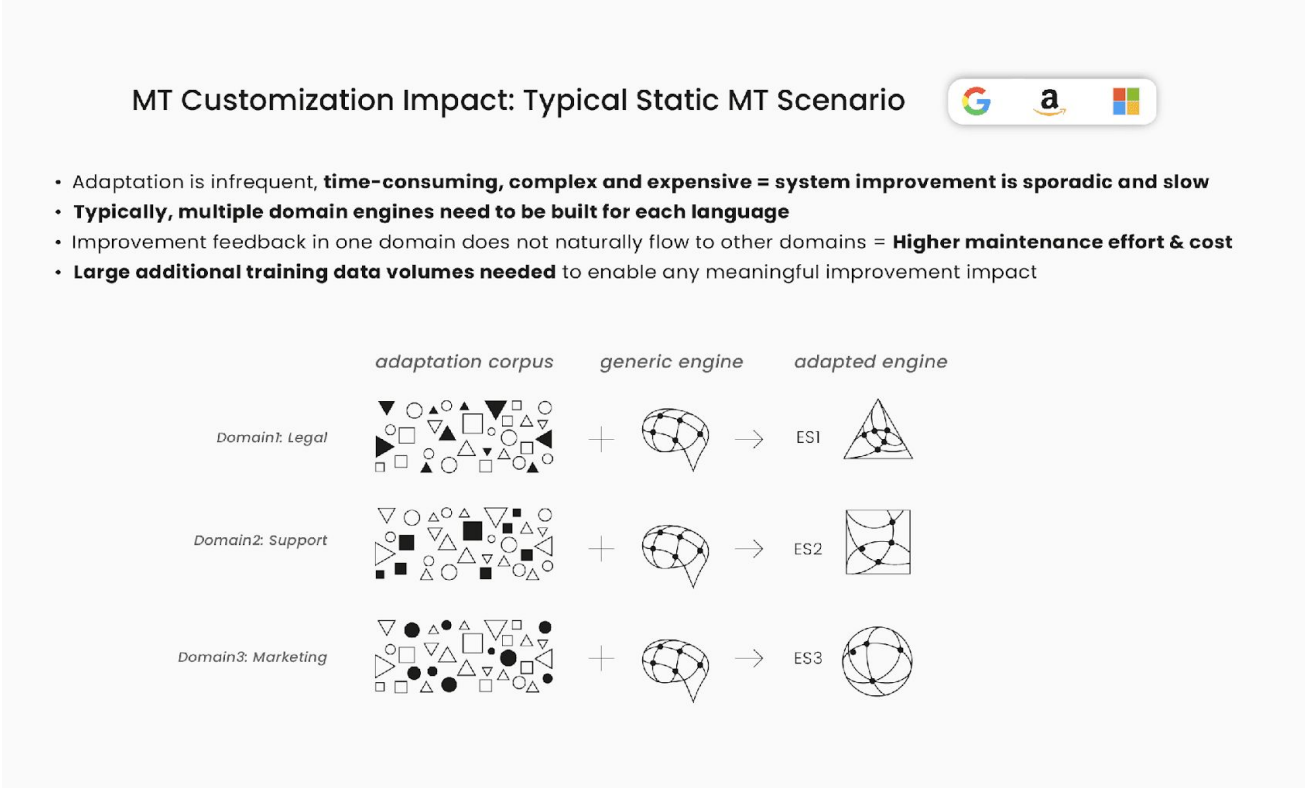
This effort results in limited or coarse optimization if sufficient training data resources are available. The optimization is considered coarse because the training data available to perform the optimization is typically minuscule compared to the base data used in the generic engine.

There is little value in training an engine with limited data as there would be no difference in performance from the generic baseline.

Thus, many attempts to use MT in professional settings face data scarcity problems. Limited data availability limits and reduces the potential impact of adaptation. To further complicate matters, it is usually necessary to build separate engines for each different use case, e.g., customer support, marketing, and legal would all be optimized separately.



Since many global enterprises have multiple product lines and businesses that cross multiple domains (TVs, semiconductors, PCs, home appliances) this will often result in a large number of MT engines needed to cover global business needs. As a result, **it is often necessary to manage and maintain many MT engines.** This management burden is often not understood at the outset when localization teams embark on their MT journey. This complexity also creates a lot of room for error and misalignment as data alignment can easily get out of sync over time.



Over time, many enterprise MT initiatives can be characterized by several problems that are common to users of these static MT systems. These problems are summarized below in order of frequency and importance:

1. **Ongoing scarcity of training data:** Static models require a lot of data to drive improvements. **There is little value in retraining a model until new or corrective data volumes reach critical levels.**
2. **Tedious MTPE experience:** Post-editors must repeatedly correct the same errors because these MT engines do not regularly improve, often leading to worker dissatisfaction.
3. **MT model management overhead and complexity:** There are too many models to manage and maintain, which can lead to misalignment errors.
4. **Communication issues:** Typically, between the MT development team and localization team members and translators, who have very different views of the overall process.
5. **Context insensitivity:** Sentence- and document-level context is typically missing from these custom models.

The **generic (static) MT approach makes sense for large ad-supported portals** where the majority (99%+) of the millions of users will use the MT systems “as-is” without attempting modification or customization.

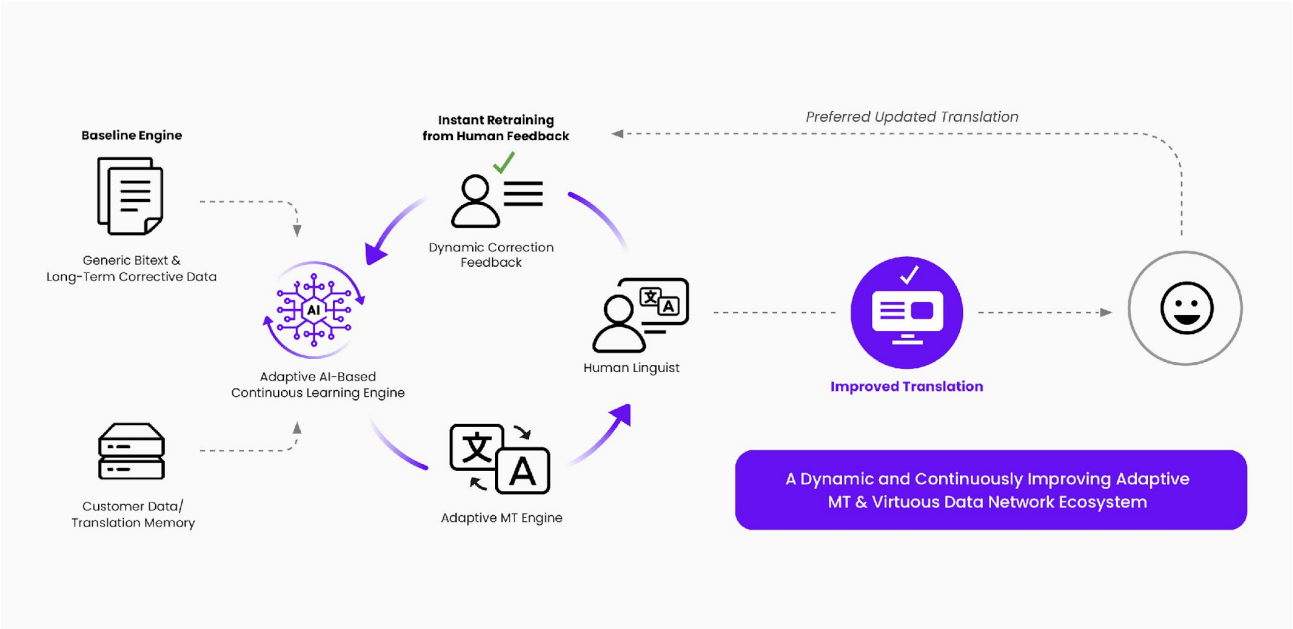
These systems are most often expected to be used without any customization or tuning.

But in the enterprise use setting MT systems need to be able to quickly adapt and improve for the different use cases where multilingual data adds to value to the enterprise mission and customization is almost always a necessity.

The Adaptive MT Experience

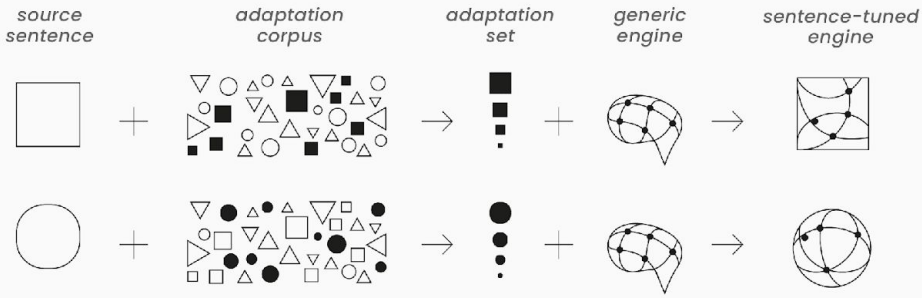
In contrast, **the adaptive MT approach makes more sense for those enterprise and professional translators who almost always attempt to modify the behavior of the generic model to meet the specific and unique needs of a business use case.**

ModernMT is an adaptive MT technology solution designed from the ground up to enable and encourage immediate and continuous adaptation to changing business needs. **It is designed to support and enhance the professional translator's work process and increase translation leverage and productivity.** This is the fundamental difference between an adaptive MT solution like ModernMT and static generic MT systems.



Adaptive MT Adaptation: The ModernMT Scenario

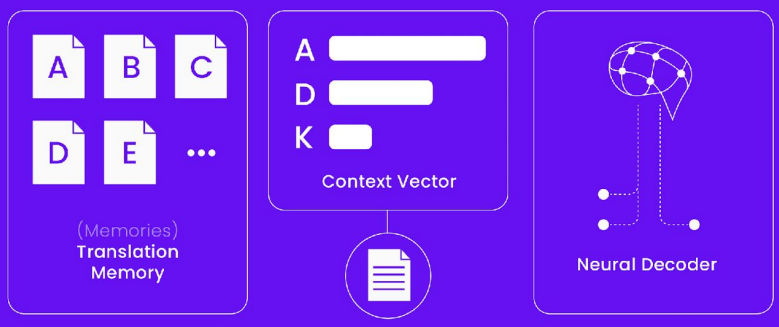
- every sentence is used as a search query in the corpus, and yields a small set of adaptation sentences which are used for micro-training
- **engine gets fine-tuned on the adaptation data set on-the-fly in less than 500 ms**
- **translate with a sentence-specific tuned engine**



While the ModernMT adaptive MT engine also has a baseline generic engine underlying its capabilities, **it is designed to work instantly with any available translation memory resources and to learn instantly from corrective linguistic feedback.**

This is done without any user intervention or action to "train" the system. The user simply points to any available TM and it is used if it is relevant to the translation task at hand. Thus, while many struggle to use MT in an environment where use case requirements are constantly changing, **this adaptive MT system uses memories, corrective feedback, and overall context gathered from both the memories and the overall document.**

ModernMT combines learning from TM, Corrective Feedback, and Context together with the Baseline to provide continuously improving and responsive translation output that adapts instantly to every new use case



This happens seamlessly and transparently with no specific training and technical MT model management actions required from users

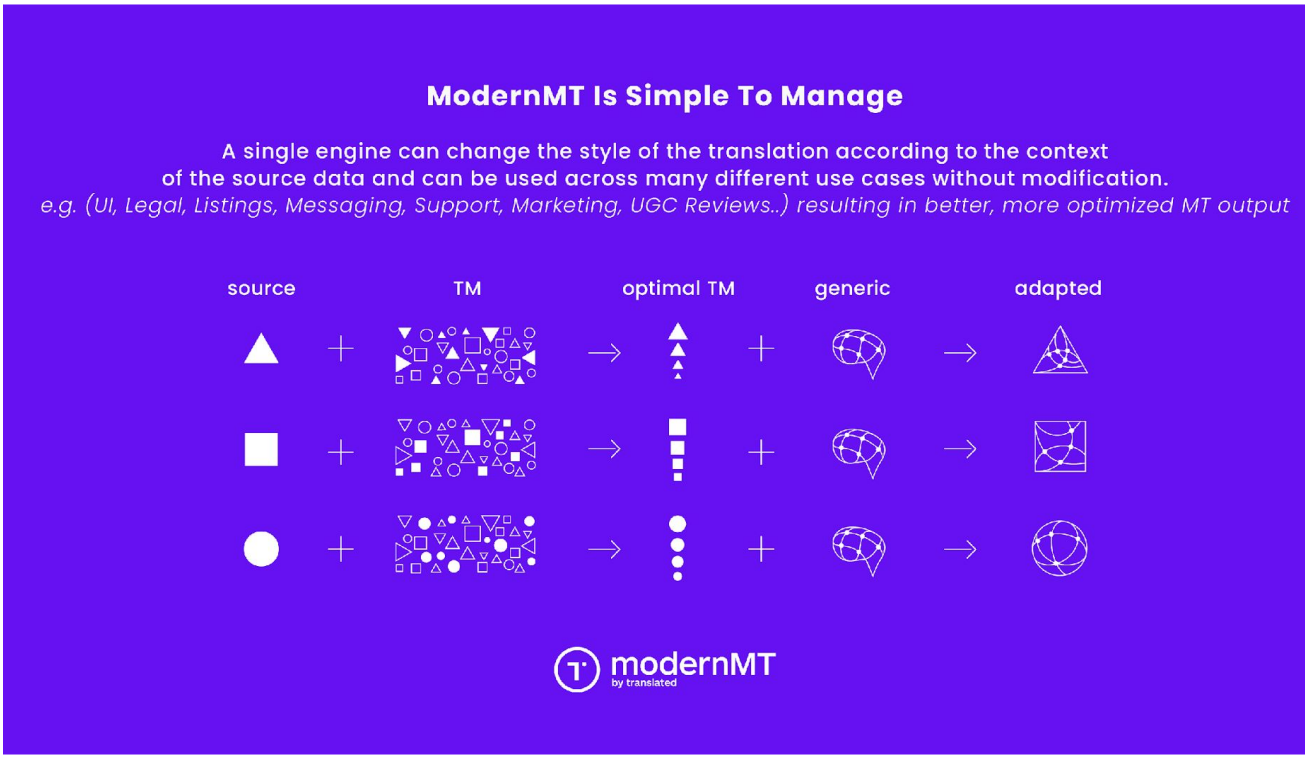
translated.

As the use of MT grows in the enterprise, the benefits of an adaptive MT infrastructure continue to increase over time. The management and maintenance of production MT systems require nothing more than the organization of TM assets and the organized provision of continuous corrective feedback to drive ongoing improvements in system performance.

Thus, content creators and linguistically informed users can be the primary drivers of the ongoing system evolution. Because the underlying continuous improvement process is always active in the background, there is no need for any technology process management by any of these users.

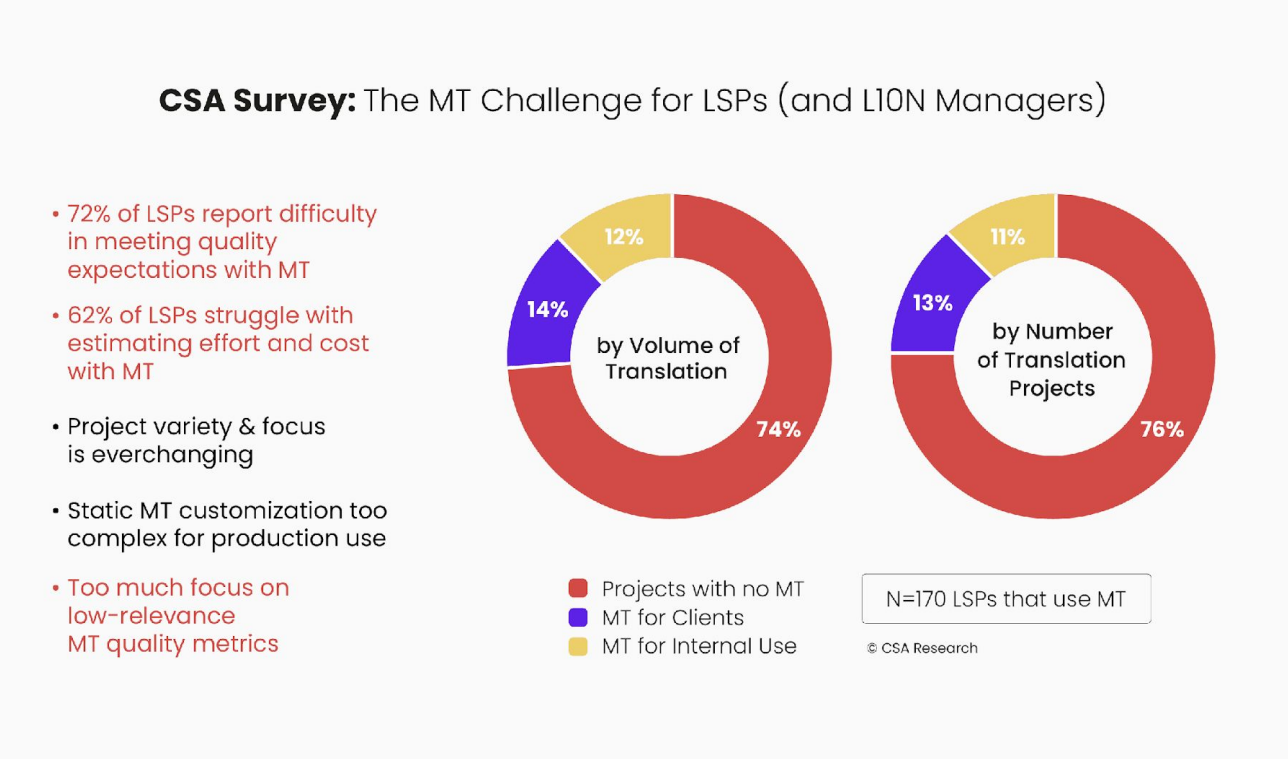
Translation issues that may arise in widespread use, can be quickly identified and corrected by linguists without the need for support from MT technology experts.

New use cases of large-scale deployments can be rapidly deployed by targeting human translation efforts on the most relevant and statistically present content. Adaptive MT technology allows for evolutionary approaches that ensure continuous improvement.



Independent market research points to some key factors that are often overlooked by those attempting to deploy MT in professional and enterprise environments. Surveys conducted by CSA Research and Nimdzi Insights show that most LSPs/Localization Teams in Enterprises struggle to deploy MT in production for three key reasons:

1. **Inability to produce MT output at the required quality levels**. Most often due to a lack of training data needed for meaningful improvement.
2. **Inability to properly estimate the effort and cost of deploying MT in production**.
3. **The ever-changing needs and requirements of different projects** which static MT cannot adapt easily to create a mismatch of skills, data, and competencies.



Given these difficulties, it is worth considering the key requirements for a production-ready MT system. **Why do so many still fail with MT?**

One reason for failure is that many LSPs and localization managers have only used automated metrics to select the "best" MT system for their production needs without having any understanding of how MT engines improve and evolve. Automated MT quality metric scores such as BLEU, Edit Distance, hLepor, and COMET are used to select the "best" MT systems for production work.

These scores are all useful for MT system developers to tune and improve MT systems, but globalization managers who use this approach to select the "best" system may overlook some rather obvious shortcomings of this approach to optimal MT selection.

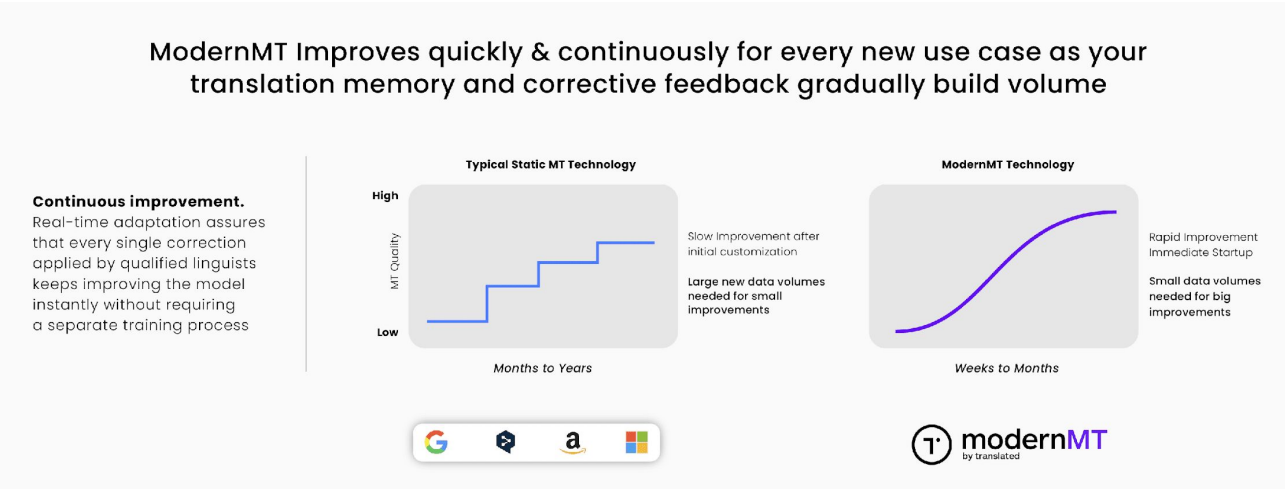
Ideally, the "best" MT system would be determined by a team of competent translators who would run relevant content through the MT system after establishing a structured and repeatable evaluation process. This is slow, expensive, and difficult, even if only a small sample of 250 sentences is evaluated.

Thus, automated measurements (metrics) that attempt to score translation adequacy, fluency, precision, and recall must often be used. They attempt to do what is best done by competent bilingual humans. **These scoring methodologies are always an approximation of what a competent human assessment would determine, and can often be incorrect or misleading, especially with opaque and unrepresentative Test Sets.**

This approach of ranking different MT systems by scores based on opaque and possibly irrelevant reference test sets has several problems. These problems include:

- **These scores do not represent production performance.**
- These scores are **typically obtained on static MT systems and do not capture facts around a system's ability to improve.**
- **These score-based rankings are an OLD snapshot of a constantly changing scene**. If you change the angle or focus, the results will change.
- **Small differences in scores are often meaningless**, and most users would be hard-pressed to explain what these small numerical differences might mean.
- The score is an approximate measure of system performance at a historical point in time and is generally **not a reliable predictor of future performance.**
- **These scores are unable to capture the dynamic evolution typical of an adaptive MT system.**
- Generic, static systems often score higher on these rankings initially but this does not reflect that they are much more difficult to tune and adapt to unique, company-specific requirements.

As a result, **the selection of MT systems for production use based on these score-based rankings can often be suboptimal or simply wrong.** The use of automated metrics to select the "best" MT system is done to manage what is essentially a black-box technology that few understand. NMT system performance can be mysterious and often inscrutable. Scores, as misleading as they may be, can make it easier to justify purchase decisions and show buyer due diligence, even though they produce suboptimal results.



The failure of so many LSPs with MT technology suggests that this approach may not be the best way forward to achieve production-ready and production-grade MT technology.

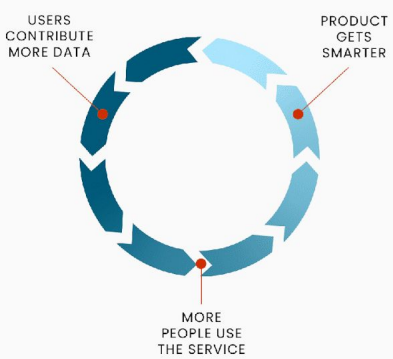
So what criteria are more relevant in the context of identifying production-grade MT technology? The following criteria are much more likely to lead to technology choices that make long-term sense. For example:

- The **speed with which an MT system can be tuned and adapted** to unique corporate content. Systems that require complex training efforts by technology specialists will slow the globalization team's responsiveness.
- The **ease with which the system can be adapted** to unique corporate needs The need to have expensive consulting resources or dedicated MT technology staff on hand and ready to go greatly reduces the agility and responsiveness of the globalization team.
- An **automated and robust MT model improvement process** as corrective feedback and improved data resources are brought to bear.
- The **complexity of MT system management** increases exponentially when multiple vendors are used as they may have different maintenance and optimization procedures. **This suggests that it is better to focus on one or two partners and build expertise through deep engagement.**
- The ability of a system to **enable startup work even if little or no data is available.**
- A **straightforward process to correct any problematic or egregious translation errors**. Many large static systems need large volumes of correction data to override such errors.
- The **availability of expert resources to manage specialized enterprise use cases** and trained human resources (linguists) to help prime and prepare MT systems for large-scale deployment.

It is now common knowledge that machine learning-based AI systems are only as good as the data they use. **One of the keys to long-term success with MT is to build a virtuous data collection system that refines MT performance and ensures continuous improvement.** The existence of such a system would encourage more widespread adoption and enable the enterprise to become multilingual at scale. This would allow the enterprise to break down the barrier of language as a barrier to global business success.

ModernMT is architected for continuous and ongoing improvement that can build long-term translation production leverage enabling the corporation to go multilingual at scale

While initially it may be neccesary to have a significant human corrective input component in production systems, as critical data resources build up the MT system will potentially provide "first draft" translations that could meet or exceed user requirements.



It is easy to assume that all adaptive MT systems employ the same technological strategy. This is not the case and real-time, in-context adaptation can be architected in different ways. In looking more closely at the very few other adaptive MT solutions in the market it is clear that dynamic adaptation can be done using different technological strategies.

However, as more buyers understand that the responsiveness of the MT system matters more than a static COMET score on a random test set, the evaluation strategies will change. It will be more useful to see which systems change most easily with the least amount of effort.

The ModernMT approach to adaptation is to bring the encoding and decoding phases of model deployment much closer together, allowing dynamic and active human-in-the-loop corrective feedback, that is similar to the in-context corrections and prompt modifications we are seeing with large language models. However, it is possible to continuously train ModernMT, as well as modify and adjust inference behavior.

In the future, as Large Language Models (LLMs) become more cost-effective, scalable, secure, and controllable it is possible that they could be used to further enhance SOTA adaptive MT models by improving both core translation quality and output fluency either as stand-alone solutions or more likely as hybrid models that work with MT purpose-focused models that are yet to come.

While LLMs have shown they can perform well in some high-resource languages, the initial evaluations also show that they perform much worse in lower-resource languages. LLMs are not optimized for the translation task, so this is expected. **Since LLMs depend on finding large caches of data in each language this is not a problem that will be solved easily and quickly. The data volumes they need to improve are substantial and often not easily found**.

In contrast, ModernMT just announced support for 200 languages that can all immediately benefit from the continuous improvement infrastructure that underlies the technology, and begin the steady quality improvement process that is described in this article.

However, it is increasingly clear that systems that can improve performance in real-time and respond quickly and efficiently to informed and expert human feedback are very likely to be the preferred approach to solve the challenge of automated language translation at scale.

The Comparative Evaluation Focus

ModernMT output performance is compared in both its static, generic form and its automated adaptation-on-the-fly form against the generic baseline output from Amazon Translate, DeepL, Google Translate, and Microsoft Translator. Beyond out-of-the-box generic translation some of these public MT services allow terminology customization with dictionaries that the user can supply. Some of the services take customization further with customized training with user-provided translation memories (TM). These types of customizations require up-front work selecting, gathering, and cleaning data, running training processes, and continued maintenance and updating of customized systems. The large majority of users, over 95%, of these public systems use these services without any customization, because of the complexity, data, and skill needed to produce better than generic output quality.

ModernMT requires no special action to incorporate and use customer-provided translation memories if these memories are available. The ModernMT system will automatically use data available in translation memories (TM) if it is available, and useful to perform a requested translation request, thus, immediately tuning the generic static system performance to user-supplied translation memory.

Many of the third-party comparative evaluations of MT systems tend to focus ONLY on the performance of generic or static versions of systems as the customization of most MT systems requires large amounts of training data, and is varied enough that comparison is made difficult by the complexity and variance in the customization process.

How well does Adaptive ModernMT work for real-world translation projects? How much better does it translate than generic, non-customized online MT services? We performed an in-depth benchmark translating translation data from the IT domain and scoring the machine translations with well-established automatic metrics.

This report provides an update to an original evaluation done by Achim Ruopp in May 2023 with ModernMT V6 and the prevailing output from the public MT portals at that time. The test and evaluation scripts are identical and applied to the output produced in January 2024.

To ensure that other interested parties may easily replicate, validate, and perform this evaluation we have used a data set that is available in the public domain. The goal of the evaluation was to measure the accuracy and speed of the adaptation of ModernMT to the IT domain and contrast this with generic translations from four major online MT services (Amazon Translate, DeepL, Google Translate, and Microsoft Translator). This is representative of many translation projects in enterprise settings.

The 3D Design, Engineering, and Construction software company Autodesk provides high-quality software UI and documentation translations that were created via post-editing machine translations. This is a great source of evaluation data from the IT domain.

After cleaning and de-duplicating the data, we chose 1000 segments at random from the data for the language pairs:

- US English → German,
- US English → Italian,
- US English → Spanish,
- US English → Brazilian Portuguese, and
- US English → Simplified Chinese

as “Test Sets”.

Adaptive MT with ModernMT is indicated with the identifier **ModernMT** in the results.

In addition, ModernMT can make use of reference TMs that are similar to the translation project at hand. We selected a further 10,000 non-overlapping segments from the Autodesk data for each of the language pairs and evaluated ModernMT in adaptive mode with document context adaptation to the reference TM. Adaptive MT with ModernMT using document context is indicated with the identifier **ModernMT Adaptive** in the results.

Thus, the charts show two different measurements for ModernMT:

1. ModernMT where the correct version of the previous Test Set sentence is added as a reference for all future translations **(ModernMT)**
2. ModernMT Adaptive version with an additional larger TM is referenced to perform each translation **(ModernMT Adaptive).**

Measurement Metrics Used in This Report

The following are the most commonly used metrics to assess MT output quality. While these measurements can be useful it is always best to validate these scores with human assessments to ensure the greatest accuracy.

COMET Evaluation Results

COMET

Semantic similarity

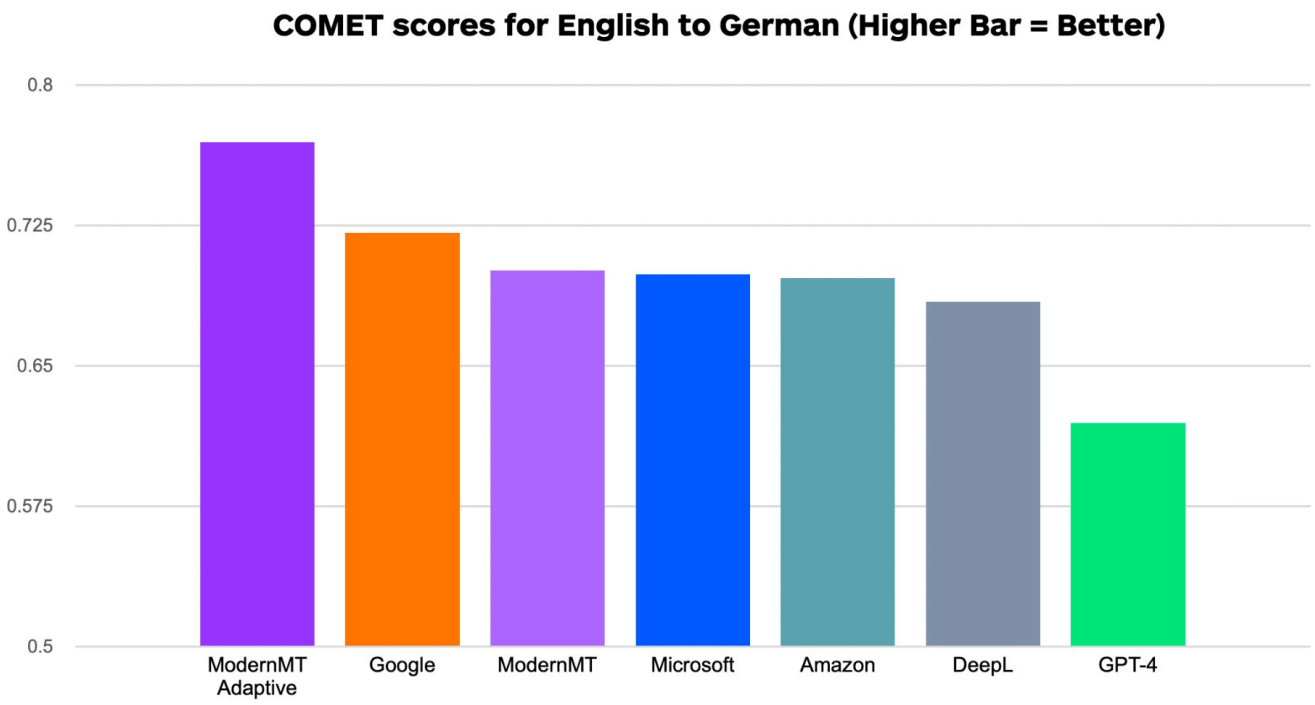
Predicts machine translation quality using information from both the source input and the reference translation. Achieves state-of-the-art levels of correlation with human judgement. May penalize paraphrases/synonyms.

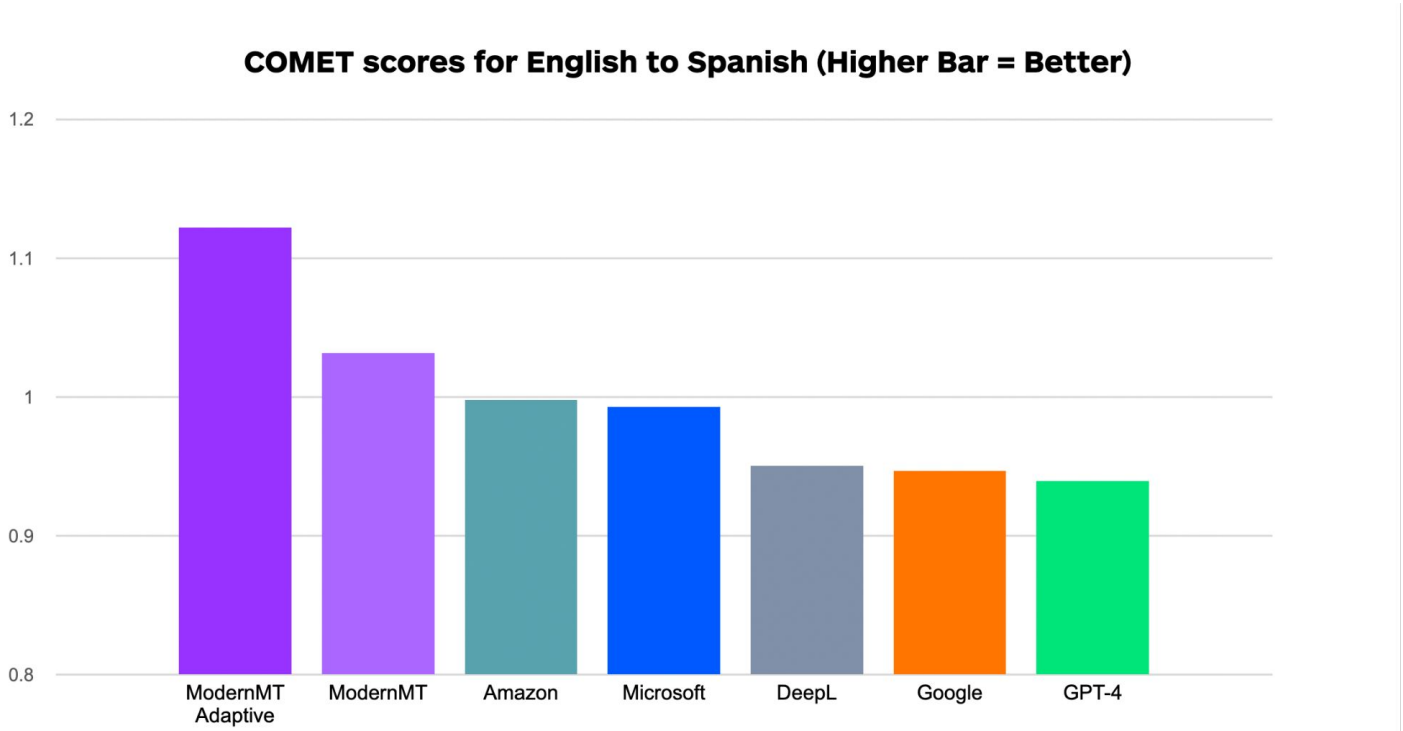
Overview: COMET: A Neural Framework for MT Evaluation

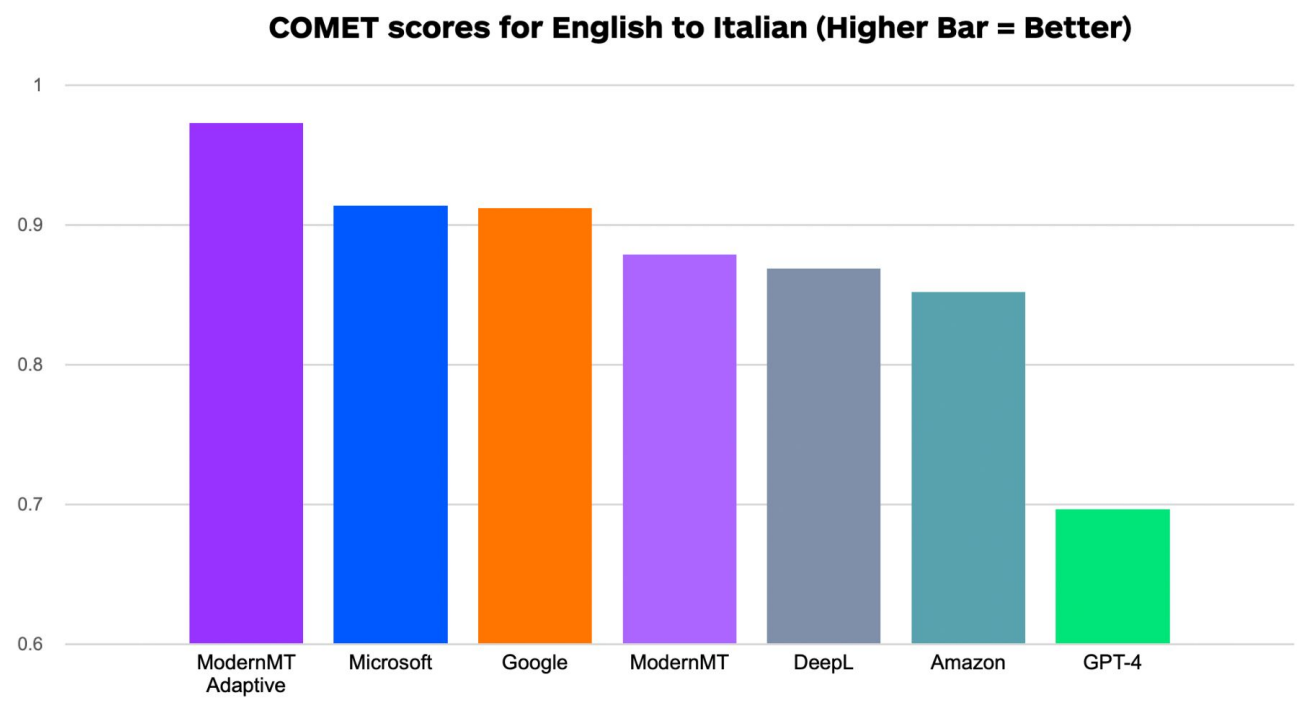
https://aclanthology.org/2020.emnlp-main.213.pdf

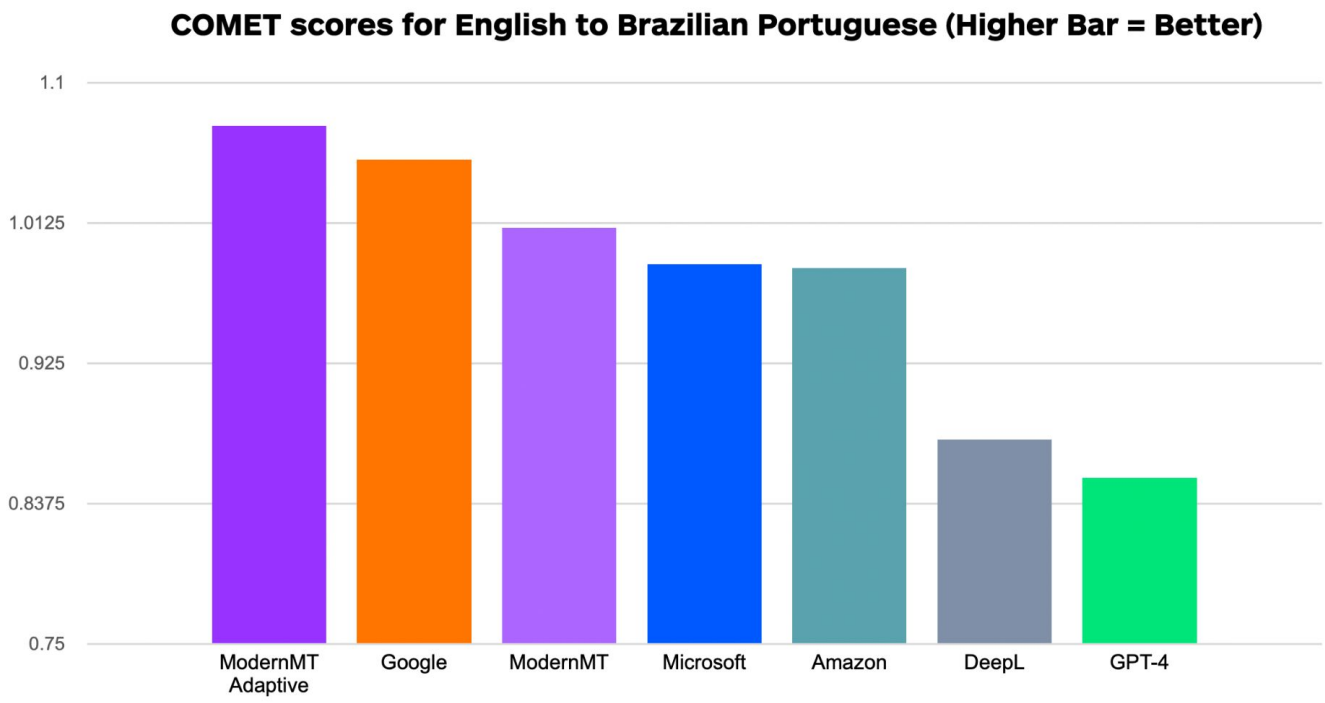
The automatic translation metric COMET has emerged in recent years in academia and industry as the metric that most closely matches human judgments in translation quality. A 2021 Microsoft study established this advantage of COMET across many language pairs. We used the COMET model eamt22-cometinho-da for our evaluation.

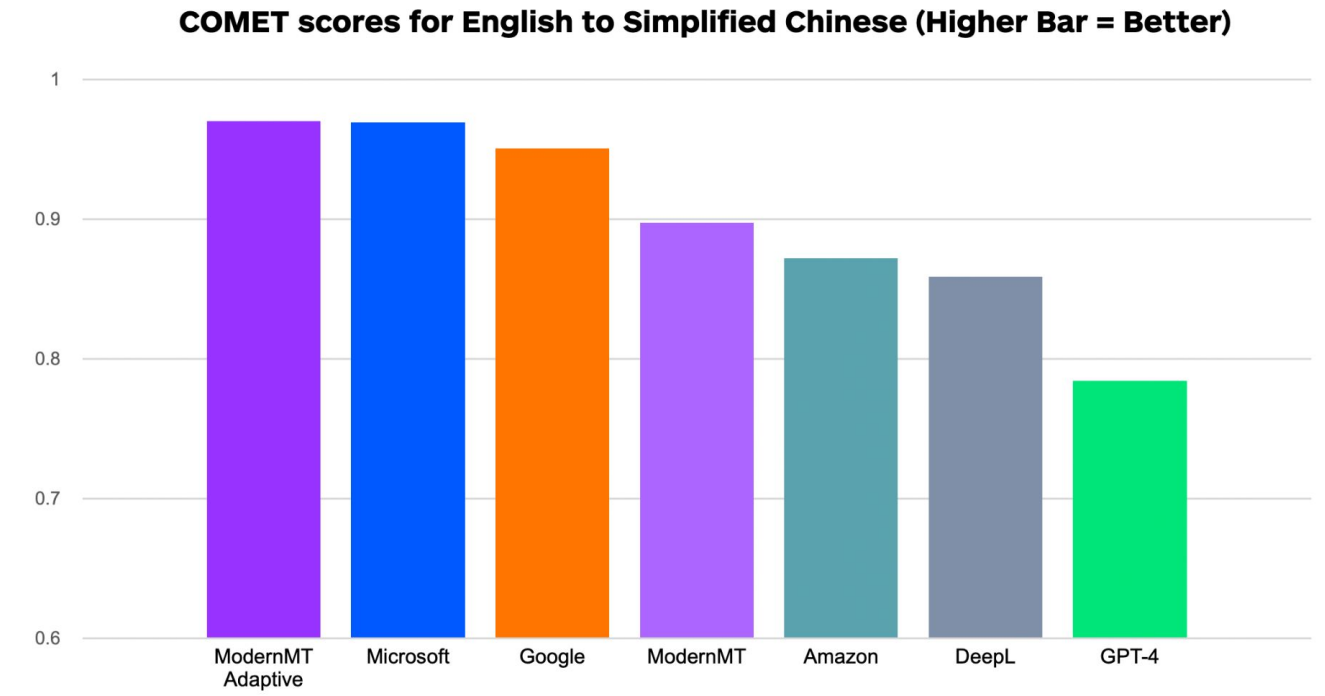
Higher values of COMET indicate better-quality of translations.




COMET scores for English to Spanish (Higher Bar = Better)
1.2
1.1
1
0.9
0.8
ModernMT Adaptive
ModernMT
Amazon
Microsoft
DeepL
Google
GPT-4


COMET scores for English to Italian (Higher Bar = Better)
1
0.9
0.8
0.7
0.6
ModernMT Adaptive
Microsoft
Google
ModernMT
DeepL
Amazon
GPT-4


COMET scores for English to Brazilian Portuguese (Higher Bar = Better)
1.1
1.0125
0.925
0.8375
0.75
ModernMT Adaptive
Google
ModernMT
Microsoft
Amazon
DeepL
GPT-4


COMET scores for English to Simplified Chinese (Higher Bar = Better)
1
0.9
0.8
0.7
0.6
ModernMT Adaptive
Microsoft
Google
ModernMT
Amazon
DeepL
GPT-4

Analysis of COMET Scores

ModernMT outperforms the other generic online MT systems when using an additional reference TM as document context.

This demonstrates that ModernMT is best when using it in typical language industry post-editing processes: iterative post-editing projects for which the language service provider often has existing TMs as reference TMs and when there is a steady inflow of corrective feedback coming to further refine and tune the MT system on customer domain and content.

Using ModernMT avoids the effort and uncertainty of determining whether there is enough ROI in a translation project for building costly custom MT systems, not to mention the maintenance headache a collection of custom MT systems can cause.

SacreBLEU Evaluation Results

SacreBLEU

Syntactic similarity

Compares token-based similarity of the MT output with the reference segment and averages it over the whole corpus. Penalizes omissions and additions. Penalizes paraphrases/synonyms. Penalizes translations of different length.

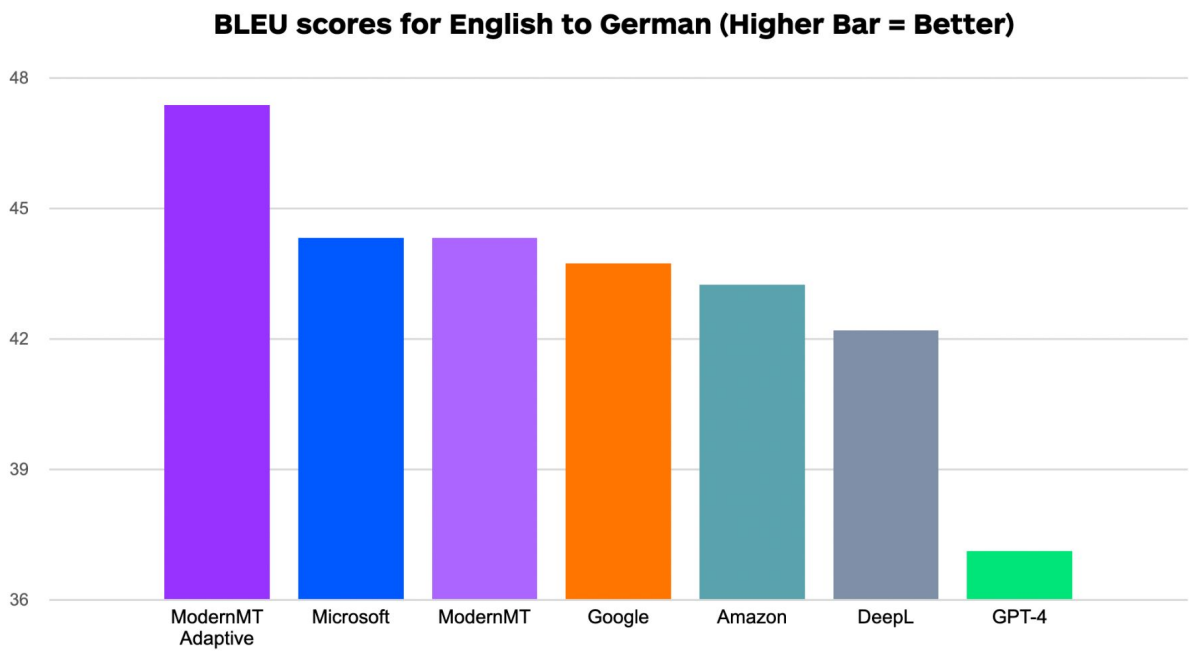
Overview: BLEU: a Method for Automatic Evaluation of Machine Translation

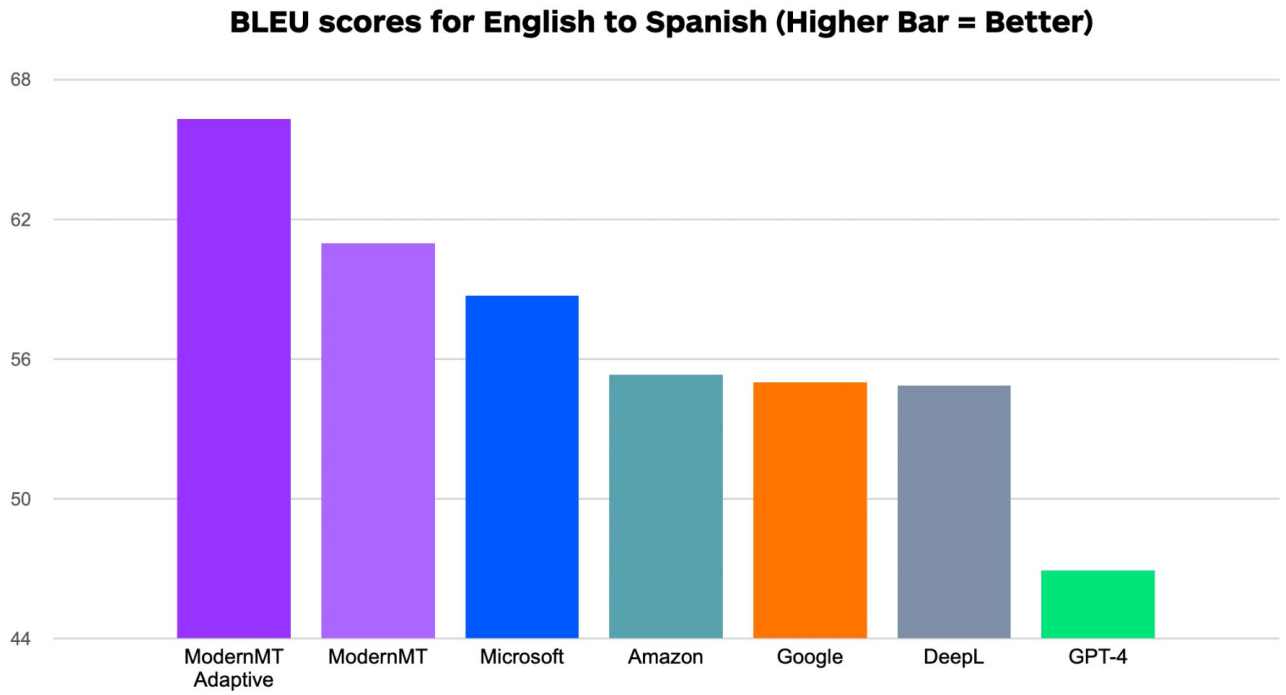
https://aclanthology.org/P02-1040.pdf

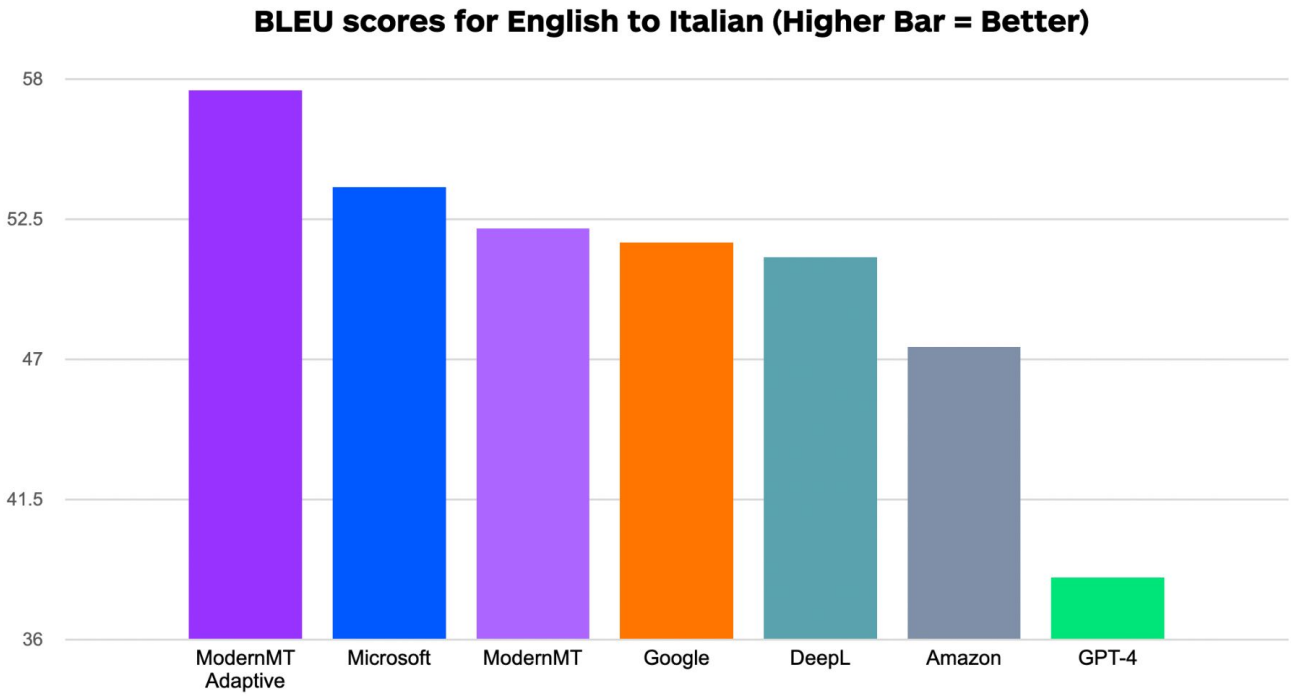
A less technical overview is provided here:

https://blog.modernmt.com/understanding-mt-quality-bleu-scores/

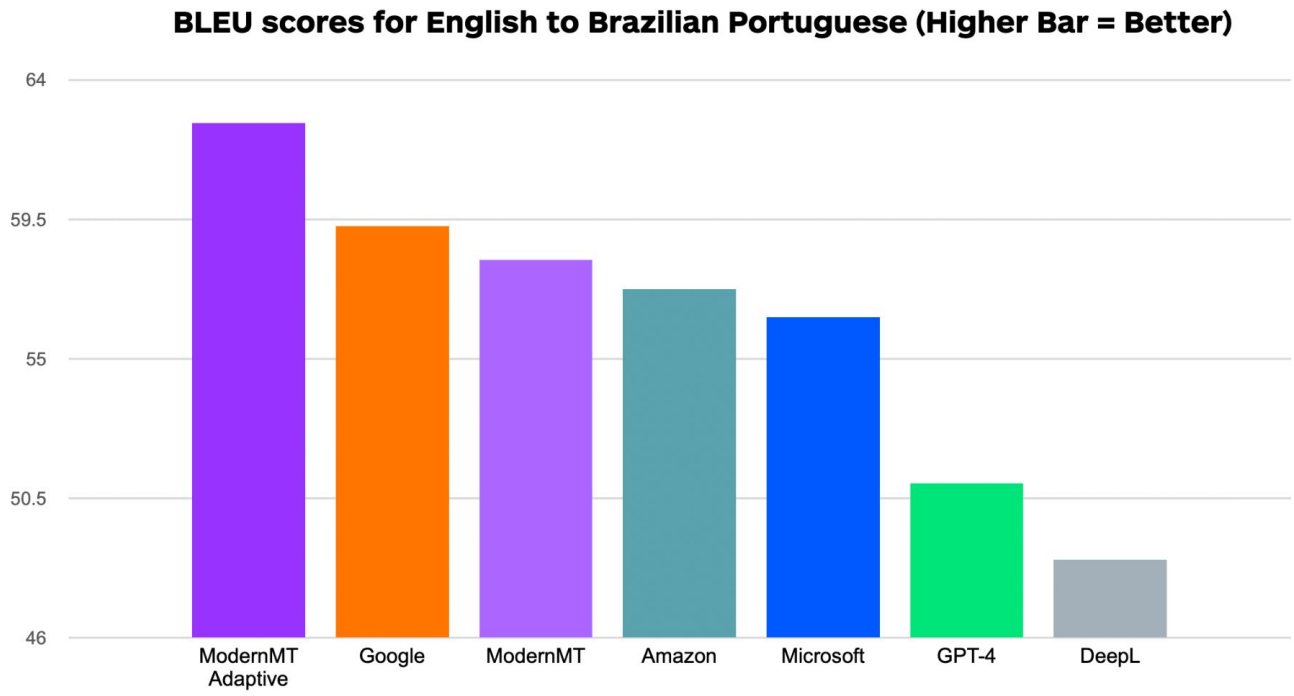
Higher values of SacreBLEU indicate better-quality of translations.

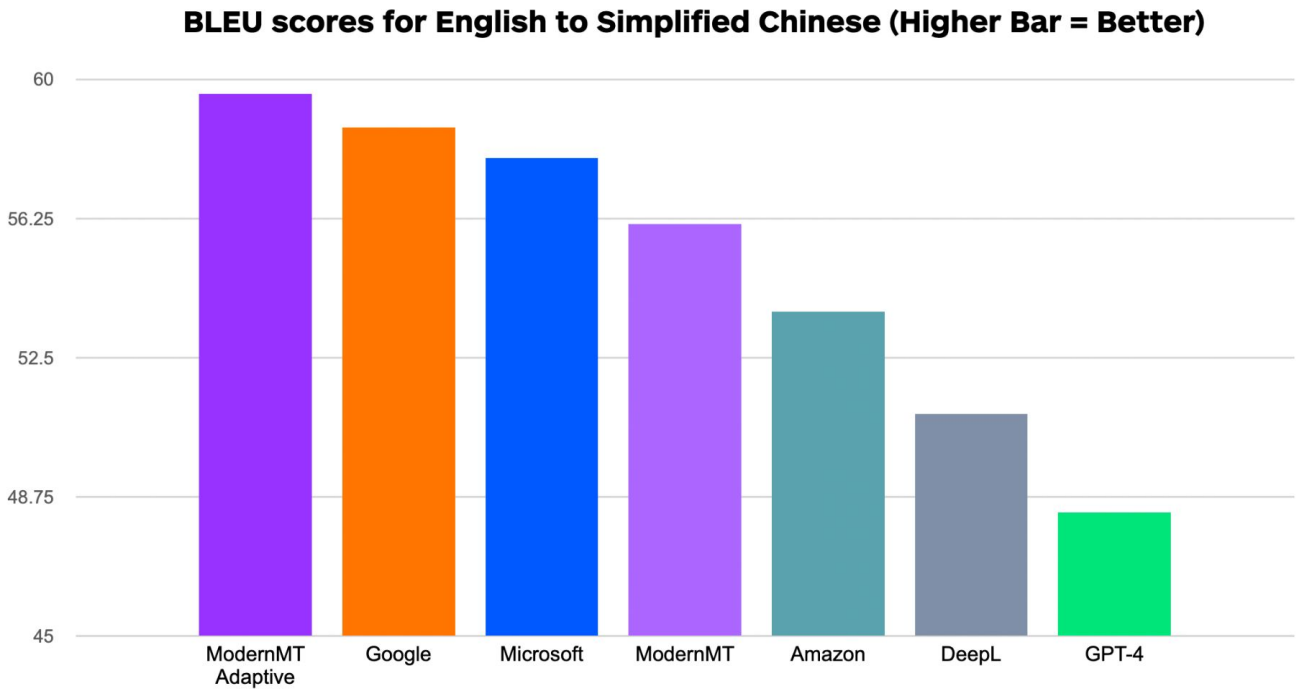



BLEU scores for English to Spanish (Higher Bar = Better)
68
62
56
50
44
ModernMT Adaptive
ModernMT
Microsoft
Amazon
Google
DeepL
GPT-4


BLEU scores for English to Italian (Higher Bar = Better)
58
52.5
47
41.5
36
ModernMT Adaptive
Microsoft
ModernMT
Google
DeepL
Amazon
GPT-4

translated.


BLEU scores for English to Brazilian Portuguese (Higher Bar = Better)
64
59.5
55
50.5
46
ModernMT Adaptive
Google
ModernMT
Amazon
Microsoft
GPT-4
DeepL


BLEU scores for English to Simplified Chinese (Higher Bar = Better)
60
56.25
52.5
48.75
45
ModernMT Adaptive
Google
Microsoft
ModernMT
Amazon
DeepL
GPT-4

Analysis of SacreBLEU Scores

It is interesting to note that the SacreBLEU scores were very consistent with the COMET scores and the same general conclusions can be drawn.

TER Evaluation Results

TER

Syntactic similarity

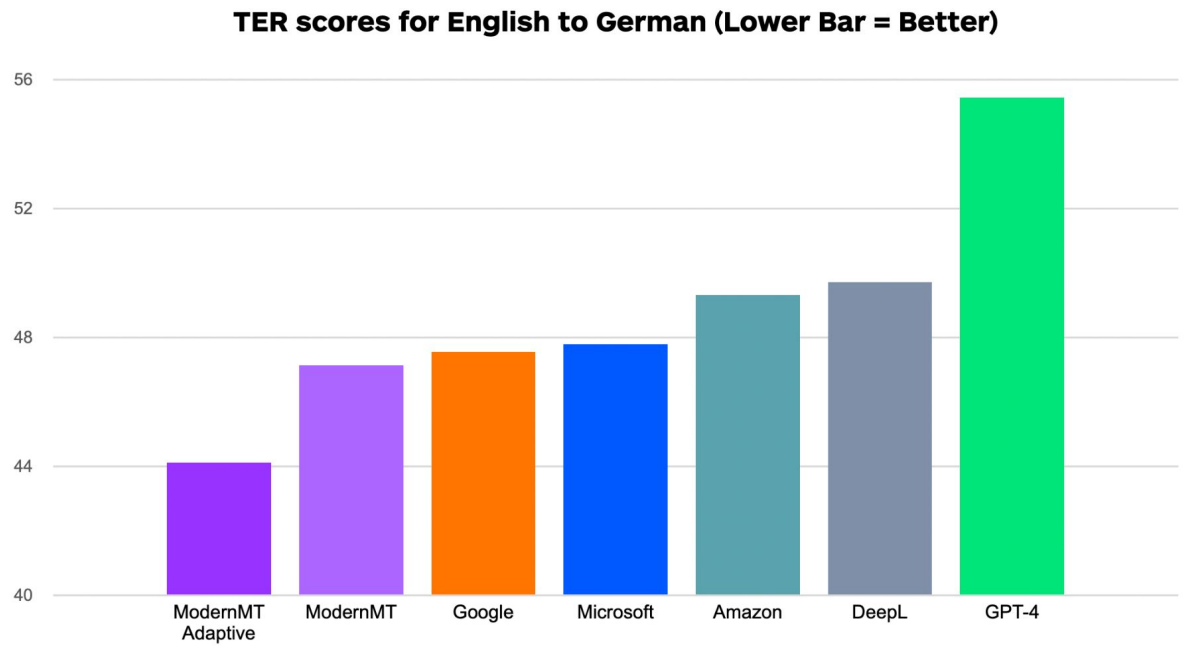
Measures the number of edits (insertions, deletions, shifts, and substitutions) required to transform a machine translation into the reference translation. Penalizes paraphrases/synonyms. Penalizes translations of different length.

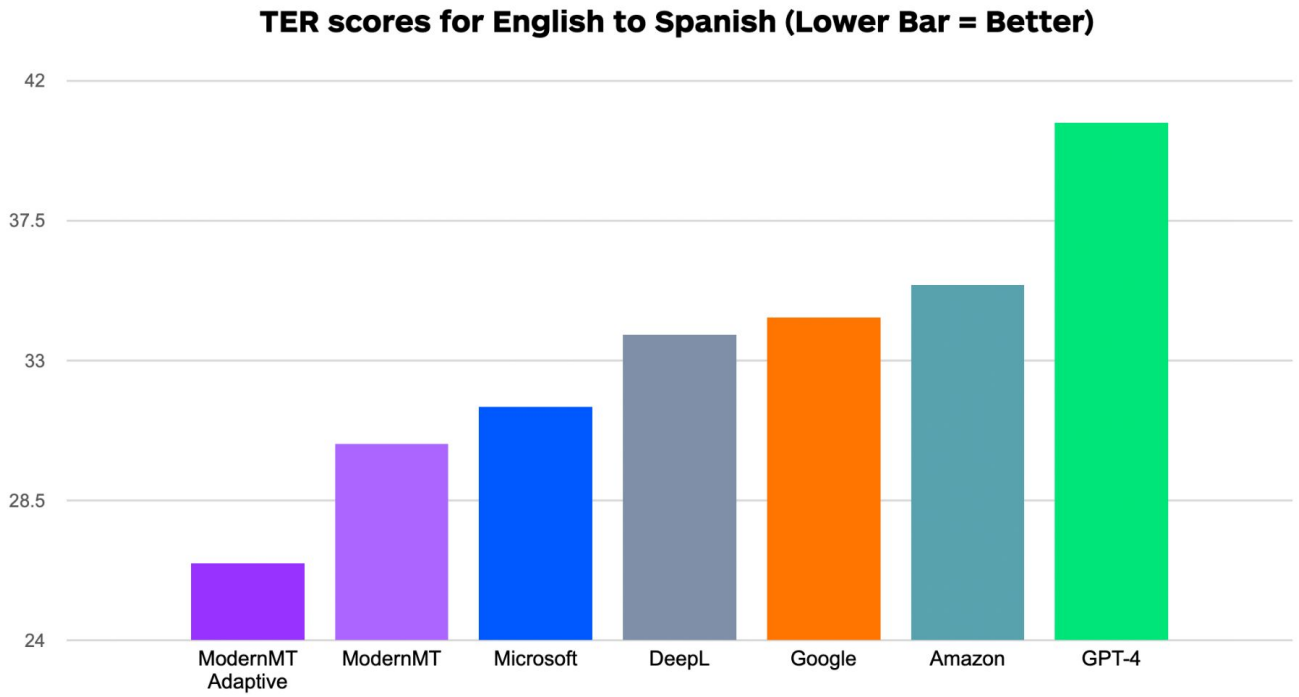
Overview: A Study of Translation Edit Rate with Targeted Human Annotation

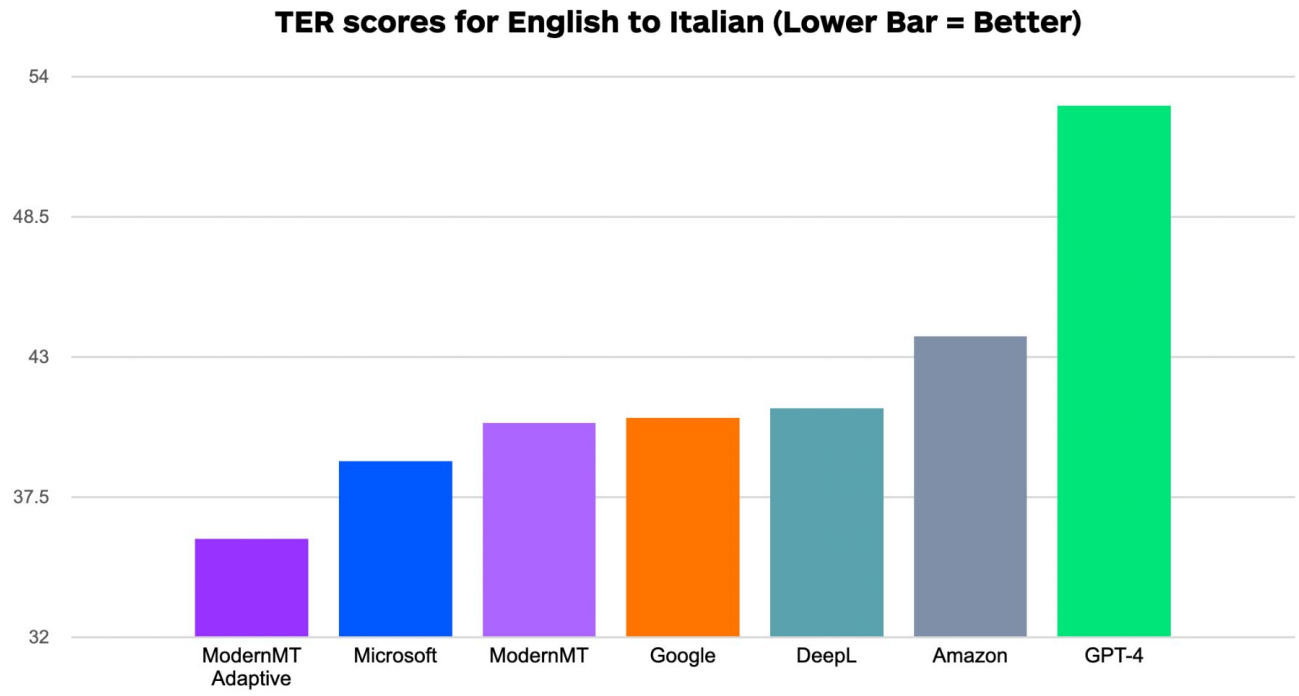
https://aclanthology.org/2006.amta-papers.25.pdf

When analyzing machine translation for post-editing we are interested in the effort a post-editor has to put in to produce high-quality post-edited translations from the machine translation drafts. The best way to do this is to measure the time post-editors spend on the task. However, due to the distributed nature of translation, this data is often hard to come by. The second-best metric is to measure the text editing effort. A well-established metric for text editing efforts is the translation edit rate (also called translation error rate) - TER.

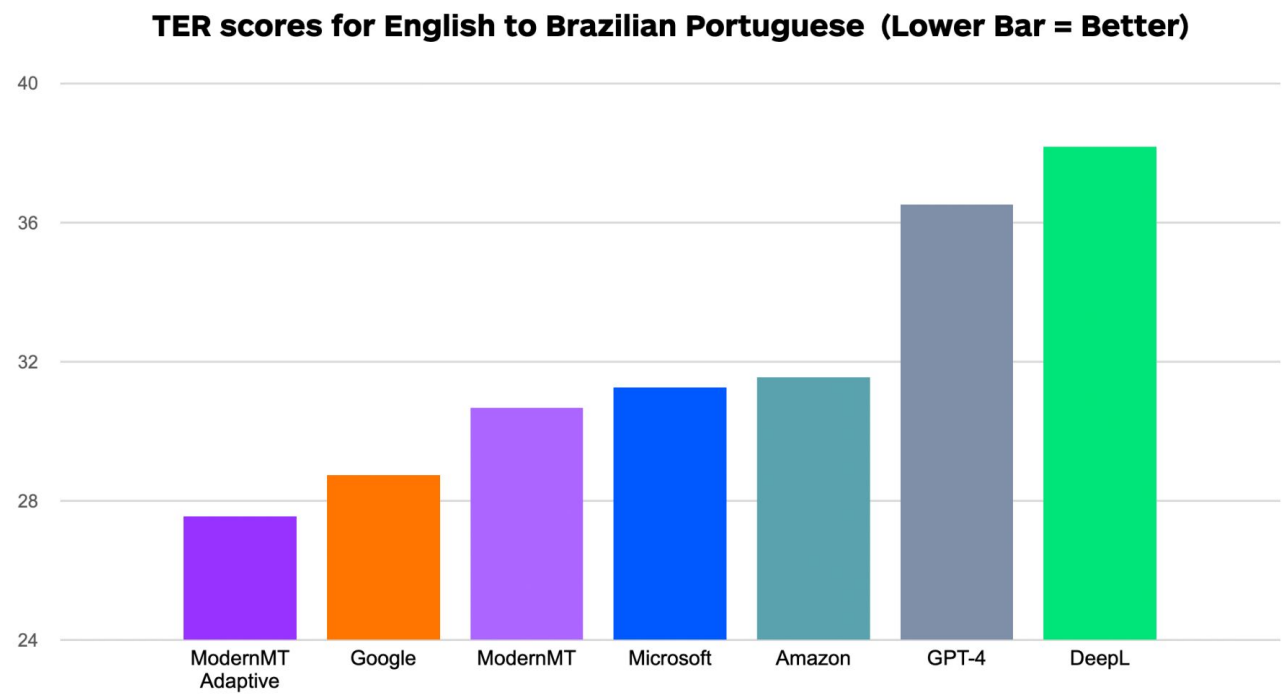
Lower values of TER indicate lower editing effort and thus better translations.

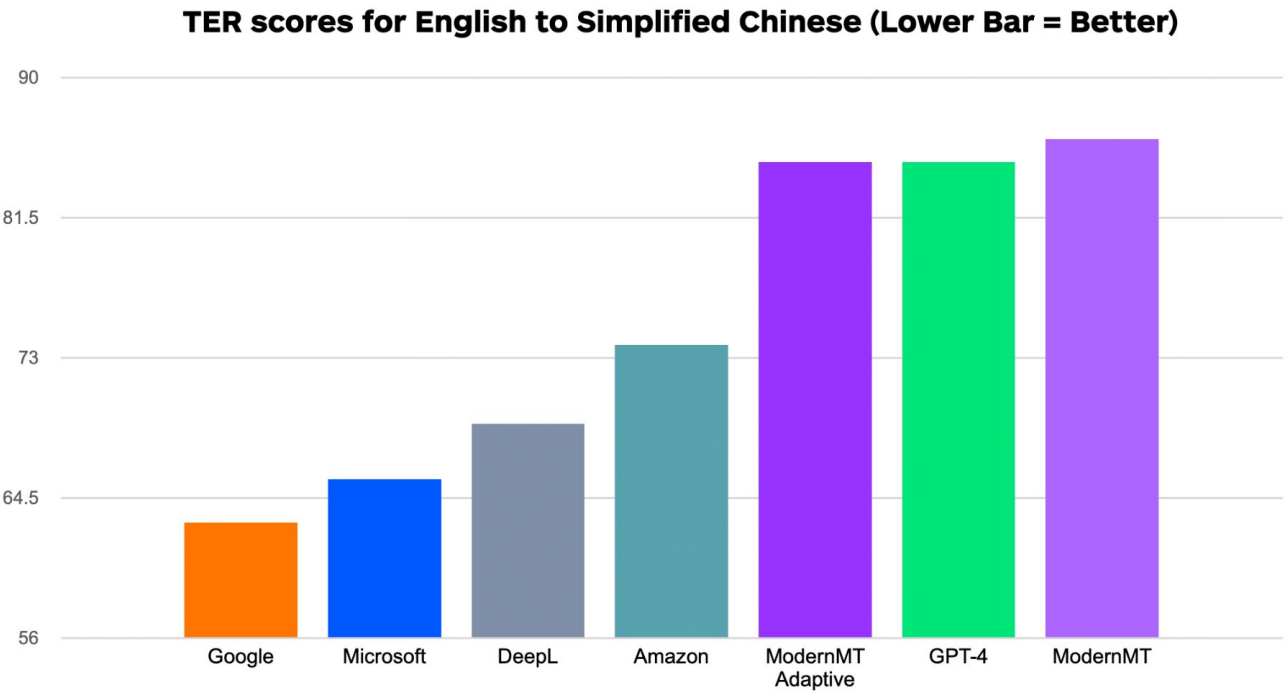



TER scores for English to Spanish (Lower Bar = Better)
42
37.5
33
28.5
24
ModernMT Adaptive
ModernMT
Microsoft
DeepL
Google
Amazon
GPT-4


TER scores for English to Italian (Lower Bar = Better)
54
48.5
43
37.5
32
ModernMT Adaptive
Microsoft
ModernMT
Google
DeepL
Amazon
GPT-4

translated.


TER scores for English to Brazilian Portuguese (Lower Bar = Better)
40
36
32
28
24
ModernMT Adaptive
Google
ModernMT
Microsoft
Amazon
GPT-4
DeepL


TER scores for English to Simplified Chinese (Lower Bar = Better)
90
81.5
73
64.5
56
Google
Microsoft
DeepL
Amazon
ModernMT Adaptive
GPT-4
ModernMT

translated.

Analysis of TER Scores

ModernMT scores significantly worse on TER than COMET. Looking at the data more closely we determine that this has most likely to do with tokenization - **TER is more sensitive to whitespace differences than COMET.** Google Translate more often outputs correct spaces around Latin script words in the reference than ModernMT.

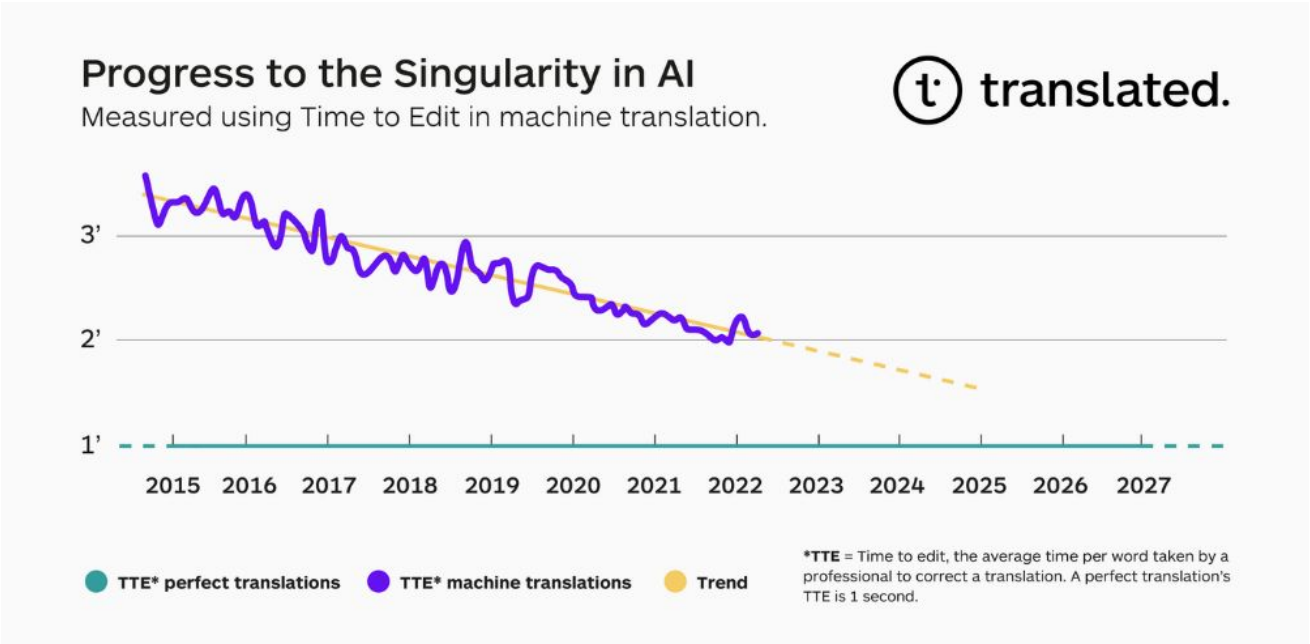
This difference in tokenization, however, does not affect the performance of ModernMT in actual production use settings as the COMET and SacreBLEU scores show. Older versions of BLEU score calculations also have this hyper-sensitivity to minor tokenization differences.

Longer-Term Implications of Continuous Improvement

Translated makes extensive use of MT in their production translation work and has found that TTE is a much better proxy for MT quality than measures like Edit Distance, COMET, or BLEU. They have found that rather than using these automated score-based metrics, **it is more accurate and reliable to use a measurement of the actual cognitive effort extended by professional translators during the performance of production work.**

Consistent scoring and quality measurement are challenging in the production setting because this is greatly influenced by varying content types, translator competence, and changing turnaround time expectations. A decade of careful monitoring of the production use of MT has yielded the data shown below. Translators were not coerced to use MT and it was only used when it was useful.

The measurement used to describe ongoing progress with MT is Time To Edit (TTE). **This is a measurement made during routine production translation work and represents the time required by the world's highest-performing professional translators to check and correct MT-suggested translations.**



The data are compelling because of the following reasons:

- The **sheer scale of the measurements across actual production work** is described in the link above. The chart focuses on measurements across 2 billion edits where long-term performance data was available.
- The chart represents what has been **observed over seven years, across multiple languages, measuring the experience of professional translators** making about 2 billion segment edits under real-life production deadlines and delivery expectations.
- **Over 130,000 carefully selected professional translators contributed to the summary measurements shown on the chart.**
- **The segments used in the measurements are all "no TM match" segments as this represents the primary challenge in the professional use of MT.**
- The broader ModernMT experience also shows that **highly optimized MT systems for large enterprise clients are already outperforming the sample shown** above which represents the most difficult use case of no TM match.
- **A very definite linear trend shows that if the rate of progress continues as shown, it MAY be possible to produce MT segments that are as good as those produced by professional translators within this decade.** This is the point of singularity at which the time top professionals spend checking a translation produced by the MT is not different from the time spent checking a translation produced by their professional colleagues which may or may not require editing.

It is important to understand that the productivity progress shown here is highly dependent on the superior architecture of the underlying ModernMT technology which learns dynamically, and continuously, and improves daily based on ongoing corrective feedback from expert translators. ModernMT output has thus continued to steadily improve over time. It is also highly dependent on the operational efficiency of the overall translation production infrastructure at Translated SRL.

The **virtuous data improvement cycle that is created by engaged expert translators** providing regular corrective feedback provides the right kind of data to drive ongoing improvements in MT output quality. **This improvement rate is not easily replicated by public MT engines and periodic bulk customization processes that are typical in the industry.**

The corrective input is professional peer revision during the translation process - and this expert human input "has control," and guides the ongoing improvement of the MT, not vice versa. **While overall data, computing, and algorithms are critical technological foundations to ongoing success, expert feedback has a substantial impact on the performance improvements seen in MT output quality.**

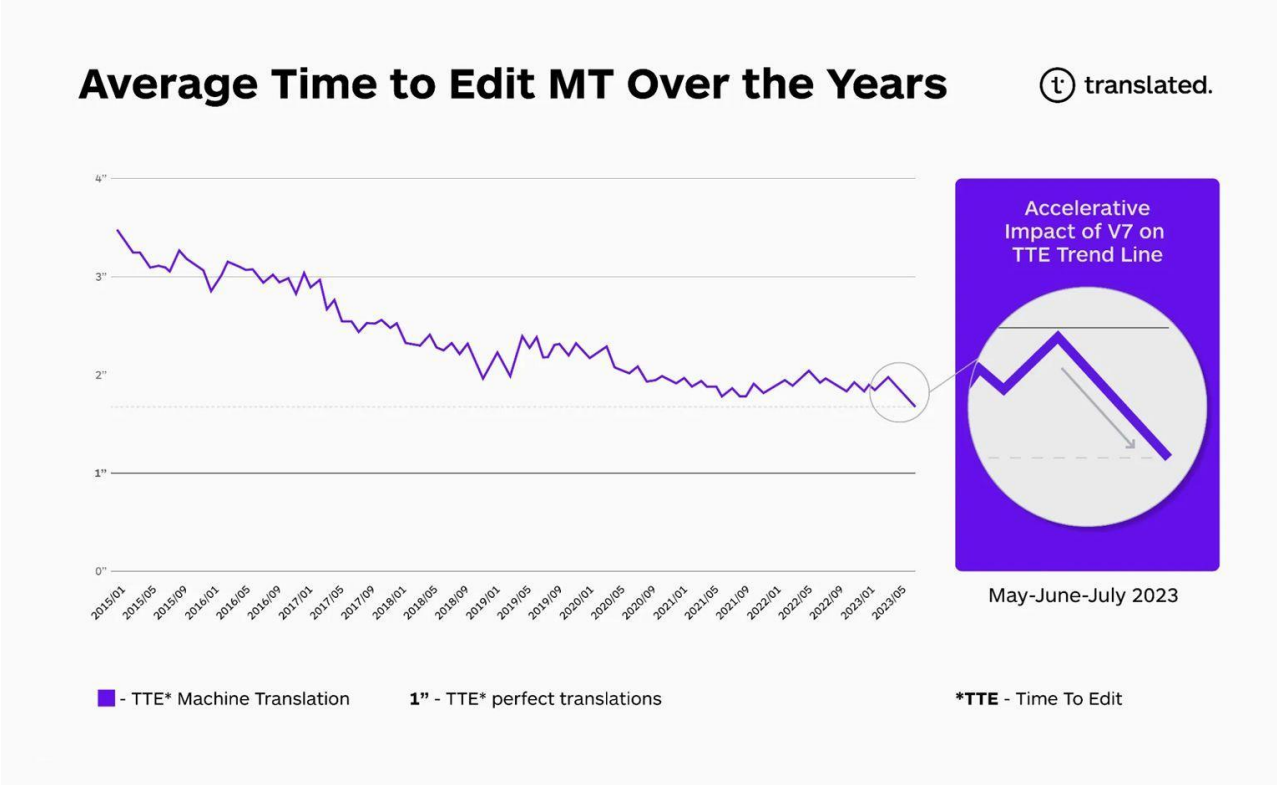
The final quality of translations delivered to customers is measured by a metric called EPT (Errors per thousand words) which in most cases is 5 or even as low as 2 when two rounds of human review are used. **The EPT rating provides a customer-validated objective measure of quality that is respected in the industry, even for purely human translation work when no MT is used.**

There is a strong, symbiotic, and mutually beneficial relationship between the MT and the engaged expert translators who work with the technology. The process is quite different from typical clean-up-the-mess PEMT projects with poorly customized static models where the feedback loop is virtually non-existent, and where the MT systems barely improve even with large volumes of post-edited data.

Responsive, Continuously Improving MT Drives Engagement from Expert Translators Who See Immediate Benefit During the Work Process

The combined effect of all the improvements and innovations introduced in ModernMT V7 has had a significant impact on the overall performance and capabilities of ModernMT.

The MT quality is now considered to be 45% to 60% better than the previous version according to human evaluations.



These improvements have greatly reduced the Time to Edit (TTE) for MT suggestions. At the end of July 2023, the aggregate TTE measured across tens of thousands of samples showed a 20% reduction, reaching a record low of 1.74 seconds. This milestone indicates an acceleration towards singularity in translation, a trend further supported by preliminary TTE data collected continuously since the 1.74 seconds record was established.

The Problem with Industry Standard Automated Metrics for MT Quality Assessment

It has become fashionable in the last few years to use automated MT quality measurement scores like BLEU, Edit Distance, hLepor, and COMET as a basis to select the “best” MT systems for production work. And some companies use different MT systems for different languages in an attempt to maximize MT contributions to production translation needs. **These scores are all useful for MT system developers to tune and improve MT systems, however, globalization managers who use this approach may overlook some rather obvious shortcomings of this approach for MT selection purposes.**

Here is a summary listing of the shortcomings of this best-MT-based-on-scores approach:

1. These scores are typically based on **measurements of static systems**. The score is ONLY meaningful on a certain day with a certain test set and actual MT performance may be quite different from what the static score might suggest. **The score is a measurement of a historical point and is generally not a reliable predictor of future performance.**

2. Most enterprises need to adapt the system to their specific content/domain and thus **the ability of a system to rapidly, easily, and efficiently adapt to enterprise content is usually much more important** than any score on a given day.

3. **These scores do not and cannot factor in the daily performance improvements that would be typical of an adaptive, dynamically, and continuously improving system like ModernMT, which would most likely score higher every day it was actively used and provided with corrective feedback. Thus, they are of very limited value with such a system.**

4. These scores can vary significantly with the test set that is used to generate the score and scores can vary significantly as test sets are changed. **The cost of generating robust and relevant test sets often compromises the testing process as the test process can be gamed.**

5. Most of these scores are only **based on small test sets with only 500 or so sentences** and the actual experience in production use on customer data could vary dramatically from what a score based on a tiny sample might suggest.

6. Averaged over many millions of segments, **TTE gives an accurate quality estimate with low variance and is a more reliable indicator of quality issues in production MT use.** Machine translation researchers have had to rely on automated score-based quality estimates such as the edit distance, or reference-based quality scores like COMET and BLEU to get quick and dirty MT quality estimates because they have not yet had the opportunity to work with such large (millions of sentences) quantities of data collected and monitored in production settings.

7. As enterprise use of MT evolves the needs and the expected capabilities of the system will also change and thus static scores become less and less relevant to the demands of changing needs.

8. Also, such a score does not incorporate the importance of overall business requirements in an enterprise use scenario where other workflow-related, integration, and process-related factors may actually be much more important than small differences in scores.

9. Leading-edge research presented at EMNLP 2022 and similar conferences provide evidence that **COMET-optimized system rankings frequently do not match what "gold-standard" human assessments would suggest as optimal. Properly done human assessments are always more reliable in almost every area of NLP.** The TTE measurements described above inherently allow us to capture human cognition impact and quality assessment at a massive scale in a way that no score or QE metric can today.

10. Different MT systems respond to adaptation and customization efforts in different ways. The benefit or lack thereof from these efforts can vary greatly from system to system especially when a system is designed to primarily be a generic system. **Adaptive MT systems like ModernMT are designed from the outset to be tuned easily and quickly with small amounts of data to fit a wide range of unique enterprise use cases**. ModernMT is almost never used without some adaptation effort, unlike generic public MT systems like Google MT which are primarily used in a default generic mode.

A “single point quality score” based on publicly sourced sentences is simply not representative of the dynamically changing, customized, and modified potential of an active and evolving enterprise adaptive MT system that is designed to be continuously adapted to unique customer use case requirements.

When it is necessary to compare two MT systems in a buyer selection and evaluation process, double-blind A/B human evaluations on actual client content would probably produce the most accurate and useful results that are also better understood by the executive and purchasing management.

Additionally, MT systems are not static: the models are constantly being improved and evolving, and what was true yesterday in quality comparisons may not be true tomorrow. For these reasons, understanding how the **data, algorithms, and human processes** around the technology interact is usually more important than any static score-based comparison snapshot. A more detailed discussion of the MT system comparison issues is provided here.

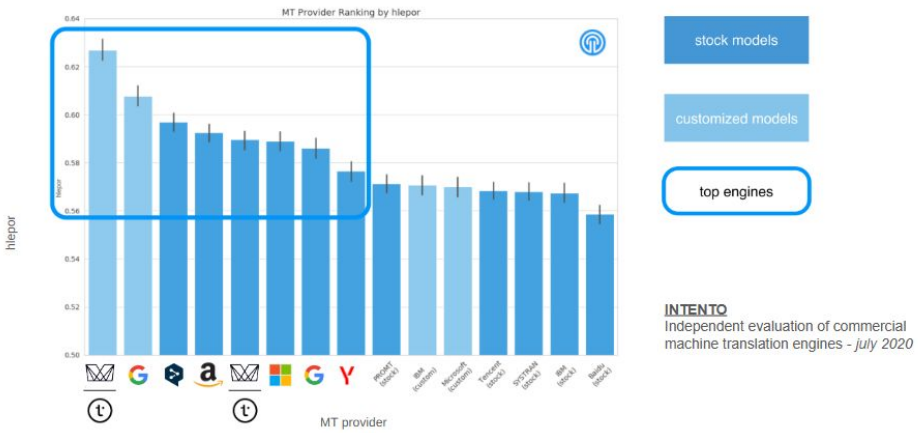
Conducting accurate and consistent comparative testing of MT systems is difficult with either automated metrics or human assessments. We are aware that the industry struggles in its communications about translation quality with buyers. Both are easy to do badly and difficult to do well. **However, in most cases, properly done human A/B tests will yield much more accurate results than automated metrics.**

Other Independent MT System Evaluations

1. One excellent example of how an MT system might perform with your unique data can be shown by **how a system performs with brand new data where we have a high degree certainty that the model has not seen**.
2. In the summer of 2020, this kind of evaluation was possible when in the early days of the COVID pandemic all the major MT systems were provided with COVID-related translation memory and terminology and then evaluated using automatic metrics to understand which systems improved the most on translating new COVID-related material. In July 2020, TAUS and Intento conducted a research project where they compared all the major MT systems which were all provided with the same COVID-related training corpus called the “Corona Crisis Corpora”. At this point, all the MT vendors were still gathering corpus to enable their MT systems to translate material related to the pandemic at the highest possible accuracy.
3. ModernMT demonstrated its ability to learn new specialized domains quickly in the tests that were conducted and where it was the best performer across the engines that were given the Corona Crisis Corpora.

Average hLEPOR Scores in German

Custom NMT Evaluation with COVID data (2020)



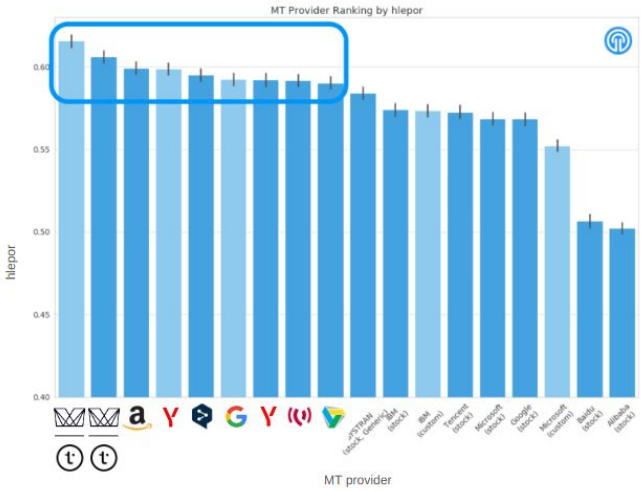
Independent tests confirm ModernMT as a quality leader in 6 languages tested with COVID data

INTENTO
Independent evaluation of commercial machine translation engines - *july 2020*

translated.

Average hLEPOR Scores in Russian

Custom NMT Evaluation with COVID data (July 2020)



INTENTO
Independent evaluation of commercial machine translation engines - *july 2020*

We have also found that the most reliable quality assessments are those that are done by humans. **While automated scoring can provide a rough idea of relative quality it is difficult to understand what small differences in score mean**. Human preferences are much clearer and emphasize the factors that matter most to human readers or editors.

translated.

ModernMT quality with COVID data set on ZH but Selecting MT systems based on scores is problematic

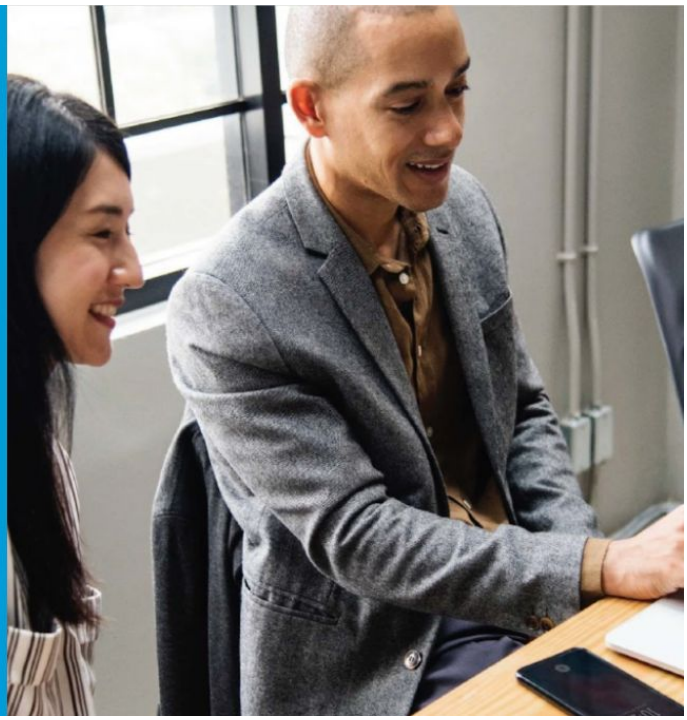
Source: INTENTO
Independent evaluation of commercial machine translation engines trained with COVID training data - *july 2020*

Small score differences may be meaningless

provider	hlepor	sacrebleu
Baidu Translate API (stock)	0.74	44.69
Tencent Cloud TMT API (stock)	0.74	44.82
ModernMT Enterprise Edition (custom)	0.73	45.87
Google Cloud Advanced Translation API (custom)	0.73	42.18
Amazon Translate (stock)	0.73	44.12
Yandex Cloud Translate API v2 (stock)	0.73	44.37
Alibaba-General (stock)	0.73	43.00
Alibaba-E-commerce Edition (stock)	0.73	42.89
DeepL API (stock)	0.73	44.21
GTCOM (stock)	0.73	44.18
ModernMT Enterprise Edition (stock)	0.72	43.79
PROMT Cloud API (stock)	0.72	43.33
Google Cloud Advanced Translation API (stock)	0.72	39.87
SYSTRAN PNMT API (stock, Generic)	0.72	42.33
Microsoft Translator API v3.0 (custom)	0.72	41.27
Microsoft Translator API v3.0 (stock)	0.72	42.50
IBM Watson Language Translator Service v3 (custom)	0.71	40.01
IBM Watson Language Translator Service v3 (stock)	0.70	39.40

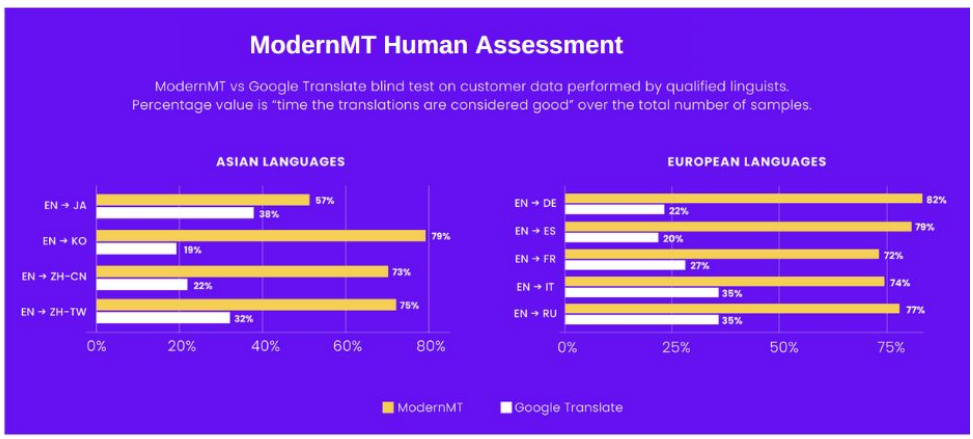
Human evaluations are generally more reliable than any automated score, and should be used whenever possible, to better understand quality differences

Human assessments also identify differences that matter to downstream human post-processing (editing or comprehension)



Human evaluation by client

Notice that human evaluations have a much clearer picture of the differences than hLepor or other metrics-based evaluations



Human assessments reflect editor preferences and the perception of the ease of modification and review of MT output

So while these scores that provide some rough indication of MT system output quality are useful, the evaluation study originally done by Polyglot Technology shows clearly that a continuously learning adaptive MT system like ModernMT provides many benefits and continuously improving output.

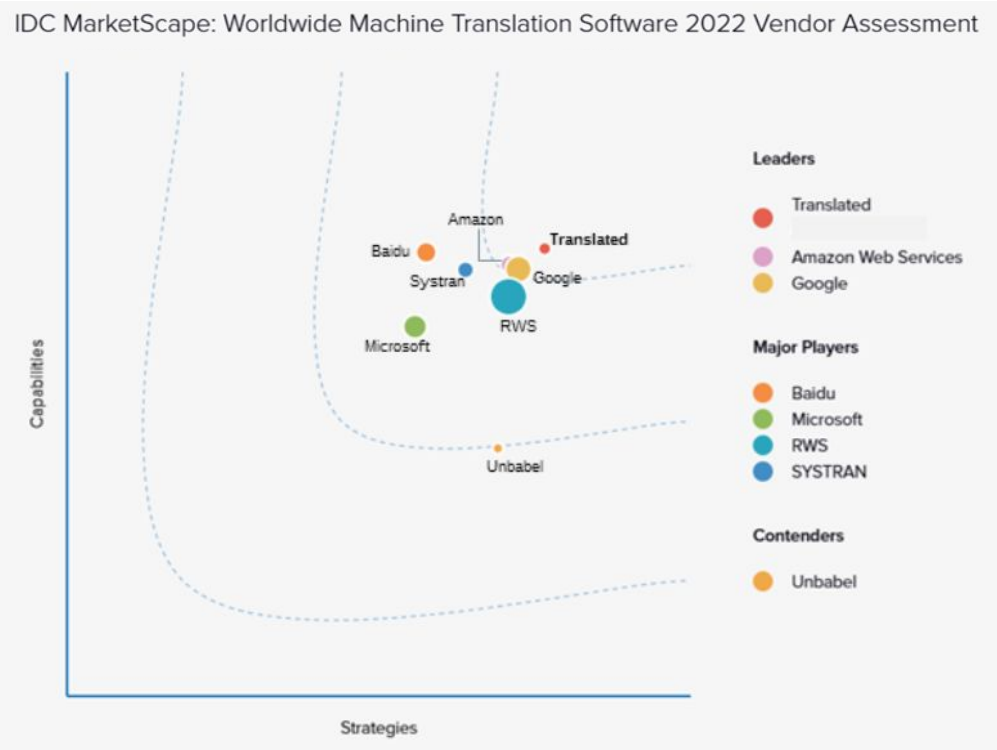
The following chart highlights what matters in the MT selection perspective to produce the most successful outcomes.

What Really Matters?

The speed with which systems adapt to unique customer content.	Integration with production management platform (TranslationOS & MateCat).
The ease with which systems ingest corrective feedback and learn.	Automation of the MT model improvement process.
Overall system management and maintenance	Trained Human resources available to drive improvement process.

The ease with which ModernMT is able to adapt to changing enterprise use cases is a key factor in driving this rating.

Published January 2023



The Most Responsive MT Solution

"ModernMT is arguably the most advanced implementation of responsive MT to date."

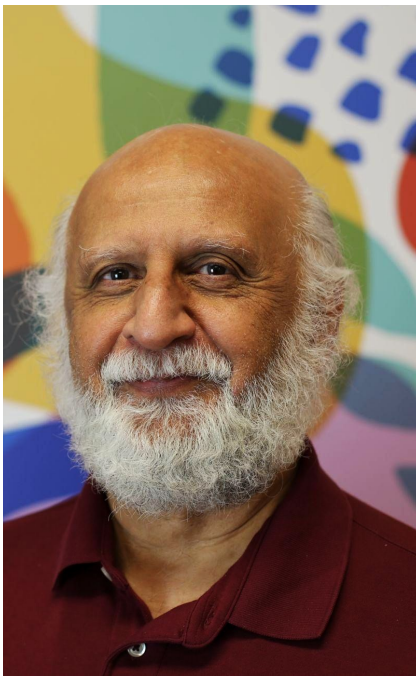
"Our analysis is shows that ModernMT offers compelling advantages over previous-generation neural MT and has a strong position as a market leader in this sector. It also has opportunities for continuing improvement that will allow it to deliver compelling offerings in the future."

CSA Research, January 2023

Authors Of the Research

After graduating in computer science Achim Ruopp worked as a translator, opening his eyes to the myriad of possibilities of using computers in language translation.
He has been involved in enabling computers to process different languages and the translation business ever since. He participated in a wide range of projects in machine translation research and industry adoption of machine translation and natural language processing. Achim's goal in sharing his knowledge, experience and the latest developments is to break down barriers in cross-language communication.
Achim performed his original comparative analysis in April 2023 as a principal at Polyglot Technology and selected the public domain Autodesk data to provide an open and transparent basis for analysis and comparison.

Kirti Vashee is a Language Technology Evangelist at Translated Srl, and was previously an Independent Consultant focusing on MT and Translation Technology deployment in the enterprise.
His MT journey started with SMT pioneer Language Weaver in 2005 where as VP of Sales & Marketing he drove revenues from $1M to over $11M in three years. He subsequently worked with other MT technology developers, including RWS/SDL, Systran, and Asia Online.
He is the moderator of the Automated Language Translation (MT) group with over 13,000 members on LinkedIn, considered an elite and top engagement group, and is also a former board member of AMTA (American Machine Translation Association).
Kirti is active on Twitter (@kvashee) and is the Editor and Chief Contributor to a respected blog that focuses on MT, AI, and Translation Automation, and Industry-related issues.



Thank you.

translated.