milengo

Machine Translation FAQ

Machine Translation (MT) has taken the localization industry by storm over the last few years. This is hardly surprising, since this technology gives rise to faster and significantly more cost-effective translation processes. But many companies face a series of challenges in implementing MT. How can machine translation be used without risking low quality? And which MT platform produces the best results? In this brochure, Milengo will answer your most pressing questions about machine translation, so that you learn the truth behind the hype surrounding MT and gain a clear overview of this revolutionary technology.

What do the terms “machine translation” and “post-editing” mean?

MT is a subfield of computational linguistics that refers to the use of software to translate text or speech from one language to another. MT systems and platforms – which can be adapted individually to the text type and content to improve the quality of translated material – form its technological foundation. In some cases, its output is so good that it can be used directly. Having said that, human translators are usually brought on board to work on these raw translations in a process referred to in the industry as post-editing (PE). Here, the raw machine translation is revised until the pre-defined quality level is achieved.

How exactly does machine translation work?

Modern MT systems are based on synthetic neural networks. These consist of programmed learning algorithms that were designed to mimic the structure of the human brain. These neural networks are fed texts in a source language and the corresponding translations in the target language so that they learn the relationship between the two languages. These texts can be reference translations of contracts or technical documentation, to name two examples. The performance capability of neural networks depends on the type, quality, and quantity of training data that they are fed. These linguistic algorithms determine the most statistically likely translation to produce a machine-translated target text.

Neural machine translation (NMT) is a relatively new phenomenon and celebrated its commercial breakthrough in 2016. Compared to earlier MT technology, NMT covers a significantly wider range of language combinations and produces substantially more fluent and higher quality texts. This makes the post-editor’s work much easier, and makes tangible productivity gains a possibility for companies.

What kind of productivity gains can be expected from using machine translation?

First things first – potential productivity gains depend on the quality of the pre-translated machine output. However, the localization workflow used also plays a role – for example, more time has to be factored in if the client’s style guide has to be observed on top of Milengo’s post-editing guidelines.

The following throughputs can be expected for the corresponding workflows:

- Machine translation: approx. 2,000 to 2,500 words per day
- Full post-editing: approx. 3,000 to 4,500 words per day

In short, the potential productivity gains are considerable, and in our experience, can be up to 60 percent compared to a conventional translation workflow.

When does it make sense to use machine translation?

We recommend using MT in the following application scenarios:

- Your company has to produce a higher translation volume while sticking to its existing localization budget
- The content wouldn’t be translated otherwise (too expensive, high text volume, localization by a human translator would take too long)
- Global customer communication (e.g., self-service offers within the framework of customer support)
- Extensive product knowledge is to be made available globally (e.g., knowledge bases)
- Real-time or short-lived content (e.g., chats, e-mails)

What kind of content is most suitable for a workflow made up of MT and post-editing (MT-PE)?

Neural MT produces translations that read very well, even among native speakers. That makes its areas of application highly diverse. However, the post-editor has to ensure that company terminology is consistent and that no information is lost – the latter of which can very well occur when NMT is used.

The greatest productivity gains are achieved with large text volumes that feature clearly defined terminology

and aren't made up of long-winded sentence structures. In an ideal scenario these texts have been written in controlled language. Excellent quality can still be achieved using neural MT even if this prerequisite hasn't been met.

We recommend that clients think about using MT-PE for the following areas:

- E-commerce (product specifications/descriptions)
- E-learning
- Business communication (like e-mails, internal documentation)
- Technical manuals
- Assembly instructions
- Component and product catalogs
- Software UI
- Online support/documentation
- Knowledge bases
- User-generated content
- Customer support

Which languages are most suitable for MT?

How good a machine translation ends up being depends primarily on the source text and the respective text type. Satisfactory results can be achieved in almost all target languages when English is used as the source language, so long as the content is "MT-compatible."

In our experience, the best translation quality is achieved with Romance languages like Portuguese, Spanish, French, and Italian as well as Germanic languages like German and Dutch. A further advantage of neural MT is that languages like Japanese, Chinese, and Korean – which proved problematic for earlier MT technology – have since come on leaps and bounds. This allows companies to bolster their localization efforts on these markets and thus make significant cost savings. But NMT tends to also produce solid results for source languages other than English.

Which factors influence the quality of machine translation?

Despite its many advantages, machine translation technology can't turn water into wine. The quality of a machine translation all depends on the quality of the source text. If a set of technical instructions is riddled with spelling and formatting errors or even inconsistencies in terms of its content, then these errors will be replicated in the MT output. Post-editors will then need more time to bring the text up to scratch. But excessively long, convoluted sentences can sometimes make things difficult for the MT system to correctly understand the sentence structure and replicate the meaning of the original text.

Providing additional reference material like terminology lists can also serve as a helpful aid to post-editors. Illustrations and diagrams in technical instructions make it easier for the post-editor to understand the context within the text, while reference translations allude to the client's corporate language, and can be used to further train the MT system.

To what extent are MT systems "capable of learning" ?

MT systems aren't static entities. Rather, machine performance can be increased exponentially by means of continuous training. We recommend post-edited content in particular for these training cycles. It's also important that machine-translated texts are regularly subject to examination by human experts. This is the only way of identifying frequently occurring categories of errors that can be subsequently prevented by using pre- and post-processing scripts.

What quantity of training data is needed to set up an MT system?

Our rule of thumb is "the more, the better." However, the quality of the training data is always crucial. A fairly decent system can be created using just 10,000 high-quality sentence pairs, but we recommend at least 20,000 units for an optimal performance. The bigger the corpus, the more reliable the results will be.

How risky is it to use raw machine translation?

On the basis of our years of experience with MT-PE

workflows, we would strongly advise you not to simply take on machine translated texts without a further quality assurance step – no matter what the content is. That's because neural MT can be deceptive – at first glance, the machine translation will often appear flawless due to the readability of its content and style. It's only when you take a closer look that you'll notice inconsistences and omissions in the target text.

NMT is particularly "notorious" for its haphazard approach to terminology. The same term can be translated in different ways throughout the document or even omitted for no apparent reason. In some cases, NMT can completely distort the meaning of a sentence. Only a post-editor, who compares the source text with the raw machine translation, can reliably eliminate errors of this nature.

But post-editing isn't the only service that Milengo uses to improve the quality of machine translations. Client-specific and adaptable pre- and post-processing scripts eradicate typical issues found in raw machine translations such as additional and missing spaces as well as punctuation and formatting errors.

Which services does Milengo offer in the area of MT?

Milengo's MT services are oriented primarily towards companies who would like to use machine translation in combination with post-editing (MT-PE).

We rely on carefully trained MT systems and can select the most suitable system for your language pair and industry. We can also set up further company-specific MT systems for our clients on request. To this end, we require access to relevant reference translations and terminology lists (if available). We then test and evaluate the MT system to optimize the raw machine translation's quality and integrate it into your localization workflow. Milengo also attends to the continuous maintenance and optimization of your MT systems. What's more, we offer companies post-editing services for machine pre-translations that we have produced using our own MT solution.

You can find further details on Milengo's MT services in our Machine Translation Data Sheet, which you can download from our website. On top of this, we offer free pilot projects for companies who want to determine the extent to which our MT service can fulfil their requirements in terms of quality and cost savings.

Which MT technology is used at Milengo?

Milengo almost exclusively uses NMT systems due to neural MT's considerably better quality compared to statistical MT. We pursue a technology-neutral approach in doing so, which means that we are not tied to any specific system, but instead maintain close contact with multiple MT providers. This is a good strategy for us, as the market for MT technology is currently undergoing rapid change.

We usually gain an overall picture of a client's requirements and carry out extensive tests with potentially suitable MT systems, before selecting the best possible approach.

Which projects offer the greatest potential savings?

The greatest cost savings can be expected in the localization of content that doesn't already exist in our clients' translation memories. To put it in simpler terms, these should be totally "new" texts, and not just an updated version of an existing product catalog, for example.

The per-word rates for post-editing at Milengo reflect the productivity gains measured during an MT-PE pilot project. Our clients can expect cost savings of between 25 and 40 percent on average compared to conventional translation workflows.

How does Milengo evaluate the quality of machine translations?

On the one hand we use automated evaluation standards that are widespread in the localization industry like BLEU and TER. BLEU scores (Bilingual Evaluation Understudy) measure the difference between human and machine translations. Translated segments (usually sentences) are used as the evaluation benchmark. An average value for the entire text is calculated based on a comparison of these segments, which then serves as an indication of the quality and readability of the MT output. The closer the machine translation comes to the human translation, the better the score.

TER scores (Translation Edit Rate) on the other hand reveal how many steps are required in the post-editing process to achieve the right (human standard) translation from the machine translation. In simpler terms, it relates to the number of changes to the text that a post-editor has to make to bring the translation up to the quality standard required. The lower the TER score, the higher the quality of the MT output.

However, you're best off consulting a human post-editor, who will be able to tell you what's good or bad about the output. That's why alongside the aforementioned automated measuring processes, Milengo has developed its own model for manually evaluating MT output, which we also use for MT pilot projects. The post-editor examines a section of the MT output in a real project context, categorizing the most frequently occurring errors, identifying systemic translation problems, and determining any potential for improvement.

What's included in a free MT pilot project from Milengo?

- Customer support
- Agreement on the project scope
- Evaluation of the most suitable MT system for the client's requirements
- Training of the MT systems (taking client data/ terminology into account)
- Machine pre-translation of a selected client text
- Classification of the machine translation quality
- Post-editing of the client text, including documentation of the productivity gains achieved
- Preparation of a client report detailing the project workflow, evaluated MT systems, and optimization potential identified
- The report, the post-edited translation, and a cost proposal for a potential future partnership are then sent back to the client

Will my data be kept safe at Milengo?

Milengo always signs a non-disclosure agreement with its clients, which obligates us to protect their company data before kicking off our collaboration. We never use publicly accessible translation services. The MT systems we use are hosted on secure cloud servers so that client content is kept confidential at all times.

→ We will be happy to personally consult you on our MT solutions – simply send us an e-mail at sales@milengo.com

Milengo has been helping companies globalize their products for more than 25 years. Our clear focus on the IT, hardware, industrial engineering, and e-commerce industries ensures we pay special attention to the requirements of these markets: industry-specific expertise, speed, and cost-efficiency.

think global Milengo GmbH
sales@milengo.com +49 (0)30 47 37 59 96
www.milengo.com Wilhelmine-Gemberg-Weg 5-7, 10179 Berlin, Germany

milengo