D.O.G. Übersetzung Dokumentation Software
Dokumentation ohne Grenzen

Machine Translation

White paper of the D.O.G. GmbH / Author: Dr. François Massion

This white paper is a machine translated and post-edited version of the German original

Machine Translation

An alternative that pays off

In these difficult times, companies are more than ever looking for ways to cut costs. At the same time, they want to generate additional revenue as quickly as possible and believe that customers abroad can contribute to this. Translate more, translate faster and pay less? It sounds like an attempt to square the circle.

Almost everyone has experimented with automatic translation programs like Google Translate or DeepL. The results are impressive. Although not everything is correct, and these systems also make gross errors, the following question is justified: Can't at least a portion of the translation requirements be translated automatically?

Anyone who is seriously considering the use of machine translation (MT) is faced with many unknowns.

- **How does MT work?**
- **Are there different programs for this?**
- **How good is the quality?**
- **How much can be saved as a result?**

The aim of this white paper is to provide some answers and recommendations.

Content

Brief historical review

MT actually has a very long history. At the beginning of the Cold War, it was important for the great powers and especially for the United States and the former Soviet Union to quickly understand what the other side was writing about.

The first breakthrough was achieved in January 1954, when an IBM 701 mainframe computer translated 60 Russian sentences into English at the "legendary" speed of 2.5 lines per second.

More than 65 years have passed since this event. In the intervening period, there have been many ups and downs. The initial enthusiasm was followed by a phase of disillusionment. A succession of different technologies was used: rule-based translation then example-based approaches in the 1960s. In the 1990s, SMT (Statistical Machine Translation) marked a turning point. SMT determines statistically from numerous parallel texts which words or expressions in the source and target language occur together. The translation results are relatively good despite the clumsy style. SMT is still used successfully to this day.

Finally, in 2016, Google Translate ushered in the era of Neural Machine Translation (NMT). NMT systems use methods of artificial intelligence (AI) and are characterized by their very fluid style. The results are surprisingly good in some cases, so good in fact that today they sometime achieve human parity, i.e. some of the translations are as good as human translations.

How machine translation works today

To assess what you can expect from MT and what errors may occur, you should first understand how neural machine translation works.

Machine translation systems use models that are trained using neural networks. Machine learning specialists train a model for each language pair.

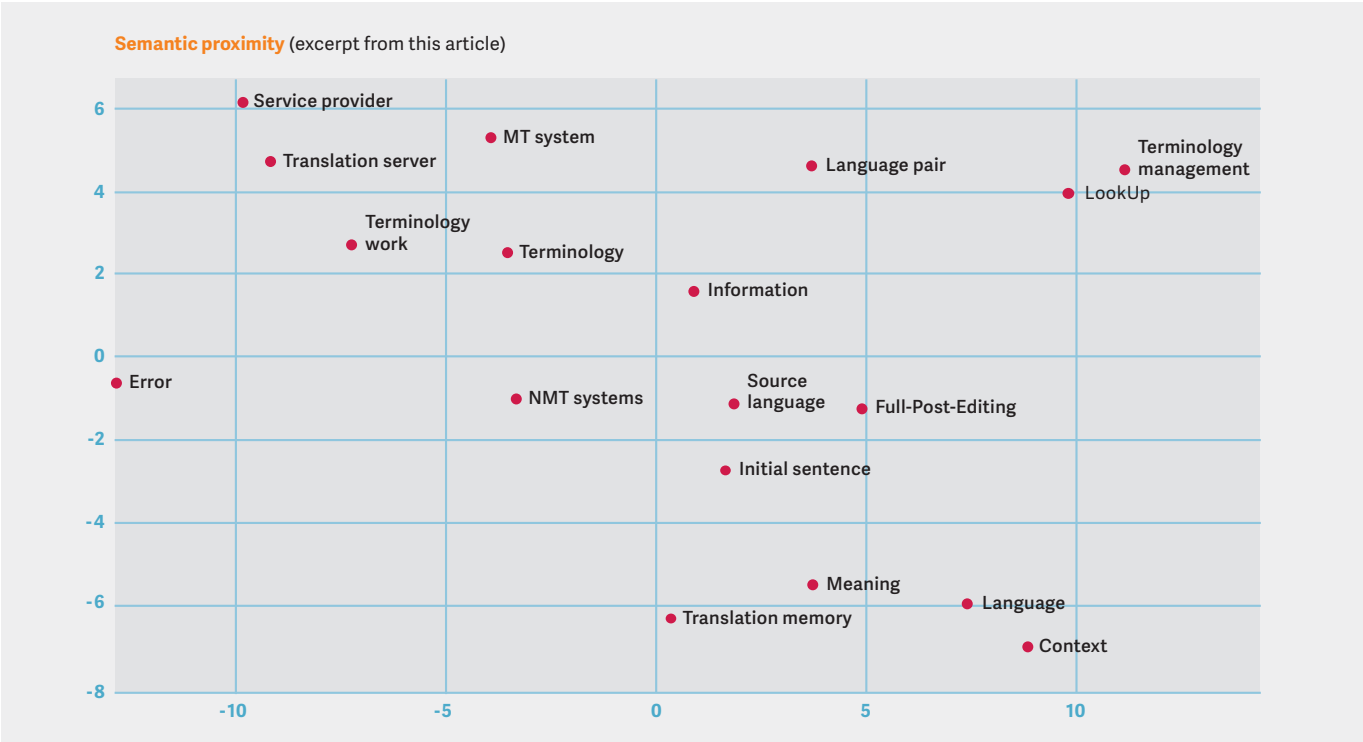


Figure 1: Learned semantics

A translation program uses a learned model to translate new unknown texts. The language combination German-English uses a different model than the combination English-German.

How the model performs can perhaps be best illustrated using the example of language acquisition by children. When a small child learns a language, it remembers words and word sequences that are repeatedly used together over a period of years. The more often words appear together, the better the child remembers them: ***I am tired. I want to sleep*** is normal for the child and forms part of its learned "model", while ***I am tired. I would like an airplane***, may be familiar words, but they do not fit its "model".

Encoder and Decoder

A neural machine translation system uses two linked networks: the encoder and the decoder.

Huge collections of translated sentences in which each initial sentence is matched with its translation serve as the input. The translation is, so to speak, the target that the system must learn without errors.

The encoder reads the source sentences and learns which words have a common context. Since a machine understands only numbers, it converts the meaning of words and sentences into number vectors. At the end of the learning process, the encoder delivers a sentence vector for each output sentence, which represents the meaning of the sentence in numbers.

In a second step, the machine learns how to translate this vector into a foreign language. This step is controlled by the decoder with the source sentence as input and the translation as output. In several intermediate steps, the decoder learns features that are relevant for the translation of words of the source language.

The learned model can now translate new, previously unknown sentences. For each sentence, the algorithm calculates the translation that has the highest probability. Since language is not always precise and often ambiguous, the proposed translation may sometimes be incorrect or may be stylistically less adequate than other correct translations that have scored less points with the algorithm.

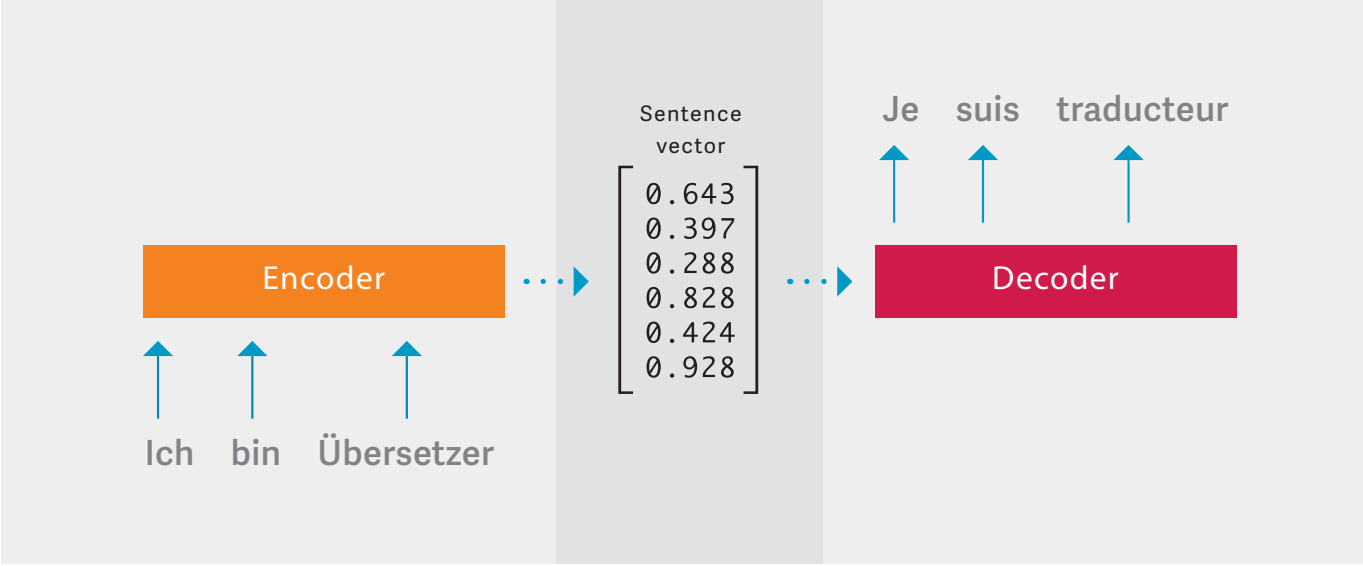


Figure 2: Encoder and Decoder

System selection: generic or individual?

If you want to introduce machine translation, there are several options available. The choice is not easy because not all types of machine translation are the same. The more you want to use machine translation, the more you need a customized solution. You will need systems that have been trained with your corporate language and are familiar with your specialized terminology.

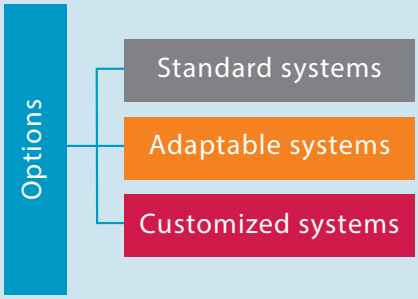
The heart of a machine translation system is a neural network. All systems have this in common. A significant proportion of the systems in use today use open source libraries such as OpenNMT (https://opennmt.net/), which they configure and enhance with their own scripts and programs.

The closer you move towards a customized MT system, the more the generated translations correspond to the language usage of your company. This results in less correction work.

This is the case, for example, if a company prefers certain terms (***battery cover*** instead of ***battery lid***) or expressions (***open battery cover*** instead of ***please open battery cover***). In addition, since many words can have different meanings depending on the context, publicly available translation systems often provide the most common meaning. They do not necessarily provide the specific translation used by your company (the so-called "long tail" phenomenon). Words such as ***performance***, ***device***, ***system*** or ***disc*** have more than one "correct" translation. Furthermore, a customized solution offers greater security regarding data protection.

Three options

Depending on how generic or company-specific the translation results are to be, companies can choose between three approaches:



1. You can use translation tools such as Google Translate, DeepL or Microsoft Translator that are publicly available directly or indirectly via API. These systems are ready to use and already configured. However, they do not support company-specific terminology and subjects.

2. You can use platforms like Systran, Kantan or Globalese. They allow the training and configuration of a MT system to a limited extent.

3. You can develop your own system or have a specialized service provider develop a customized system that is trained with texts and information material from your company.

What mistakes do MT systems make?

The errors made by NMT systems have some things in common:

Error type	Initial sentence	Translation	Comment
Inconsistent or non-compliant terminology	• ***Anzeige und Bedienelemente*** • ***Bedienelemente – Leiste*** • ***Auf der rechten Seite sind die Bedienelemente angeordnet.***	• ***Display and operating elements*** • ***Control elements – Bar*** • ***The controls are located on the right side.***	NMT systems calculate every translation separately. As a result, they cannot consistently adhere to specified terminology.
Misunderstood homonyms	***NRW plant neue Corona-Auflagen***	***NRW plans new corona coatings***	This happens more often with generic terms.
Additions	***Qualifikation Personal***	***Required qualification of personnel***	Not easy to detect in an otherwise correct translation.
Context/reference error	***It must be saved.***	***Es muss gerettet werden.***	It (the file) must be saved.

As mentioned at the beginning, NMT systems sometimes produce startingly good translations. Anyone who translates a text using MT will encounter a combination of these good or seemingly good translations and incorrect or erroneous translations.

The exact type of error depends on the basic MT technology (SMT or NMT) used and on the selected model (generic, adaptable or customized).

Why do NMT systems have so much trouble with terminology?

One of the greatest challenges of neural machine translation systems (NMT systems) is to take terminology specifications into account. Unlike its technological predecessor (SMT, statistical machine translation systems), it is virtually impossible to give a state-of-the-art system a predefined terminology list.

This is because NMT systems are complex neuronal networks that learn in countless operations how words are translated. They store this information as a series of numbers in large vectors that reflect the complex relationships between words of both languages. The algorithms are incapable of using pure translation tables (like: ***Scheibe*** = ***disk***).

see p. 6, fig. 3: 'Meaning' of a word for a NMT system →


```
In [22]:  1  print(model['übersetzen'])

[-0.3972198   0.15762967  0.46194917  0.17740306 -0.36603695  0.40951872
 -0.18695748 -0.4747253  -0.37309977 -0.51228845  0.24443704 -0.15607826
 -0.27267137 -0.25051364 -0.03269976 -0.23351745  0.41749892 -0.23172306
  0.41134125  0.33964443 -0.20515428  0.28367534  0.0628183   0.21196295
 -0.03653887 -0.3595316   0.05570595  0.9420088  -0.01160865 -0.0248226
  0.08398762 -0.27247164  0.2707084   0.32289538 -0.11952855  0.05280286
 -0.02814898  0.2553159  -0.7383995   0.4192747   0.30214256  0.23328382
 -0.2331925   0.07106207 -0.38610226 -0.12182928 -0.06937908  0.20980522
  0.0566839  -0.24953169 -0.68693984  0.7137946   0.6530615  -0.40889168
  0.0336898   0.57632774  0.36043662  0.02917454  0.9916566  -0.16568148
 -0.00262833  0.5754901   0.20918594 -0.10612661  0.01525148  0.02797612
  0.23608086  0.3928159   0.68412286  0.10793792  0.02760796 -0.34926915
  0.30073845  0.12777413 -0.16146977 -0.441489   -0.48704648 -0.3150954
 -0.08529658 -0.3245209  -0.47360054  0.7129007  -0.02352894  0.04178019
  0.41401875 -0.30696347 -0.03826707 -0.41467956  0.48962304 -0.16006011
  0.28656143  0.39558902 -0.17324285  0.31181908  0.45250472 -0.50092775
  0.25884745 -0.24865492  0.35699815 -0.208062  ]
```

Figure 3: 'Meaning' of an word for a NMT system

Data security

Much of the information a company produces is confidential: specifications, quotation, test reports, minutes of meeting, etc. If employees have this information automatically translated via public or unsecured translation servers, this poses a high risk. For German or European companies, the risk is even greater if the servers concerned are located outside the European Union or Germany.

For this reason, the location of the translation server and the security measures taken are very important when deciding on the introduction of a system. There are various solutions that you can evaluate here in terms of cost and security:

- ✔ For example, you can host the translation server on your own premises, which gives you a high degree of flexibility and security, but on the other hand, involves time, resources and costs for the deployment and maintenance of the infrastructure.
- ✔ Alternatively, an external service provider can host the server on its own premises or with one of the large established public cloud providers such as Amazon AWS, Google Cloud Platform (GCP) or Microsoft Azure.

This allows you to achieve a high level of security for your data and your intellectual property.

Post-editing

Machine translation alone is only useful in certain situations, such as translating live chats or when large amounts of information need to be available rapidly. Here, too, systems trained with the texts of your company are more suitable.

The most common working model is MT in combination with post-editing. The output of the machine translation system is corrected by a human, the post-editor. Post-editors are professionals who have the appropriate training because (1) machines make different mistakes than humans and (2) depending on the quality goals for the

translation, not everything that is not perfect needs be corrected.

There is now a standard for post-editing (ISO 18587), which provides for two levels of correction: light post-editing, in which mainly content and meaning errors are corrected, and full post-editing, in which the translation must be of similar quality to that of a human translator.

At the end of the day, full-post-editing produces results that are just as good as a conventional translation. If you are interested in gaining an impression of the output, you can visit the English version of our website, which we created using this procedure (MT + post-editing) (www.dog-gmbh.de/en). This enabled us to reduce production costs and, most importantly, we were able to publish the English version of our website much faster.

Work of the post-editor doing **light post-editing**

✔ **Goal:** Create an understandable text and change as little as possible, but:

- ✔ Correct errors of meaning
- ✔ Re-word ambiguous translations
- ✔ Correct wrong numbers
- ✔ Correct misleading spelling mistakes

✖ **NOT:**

- ✔ Improve the style (e.g. literal translation)
- ✔ Correct minor grammar errors

Work of the post-editor doing **full post-editing**

✔ **Goal:** The translation should be comparable to a human translation. This means:

- ✔ Correct errors of meaning
- ✔ Correct wrong numbers
- ✔ Correct spelling mistakes
- ✔ Optimize style and grammar
- ✔ Standardize terminology

Correction technology - ErrorSpy

How do you efficiently and reliably check large quantities of machine-translated texts? Since costs and time are important factors for the introduction of MT solutions, hu-

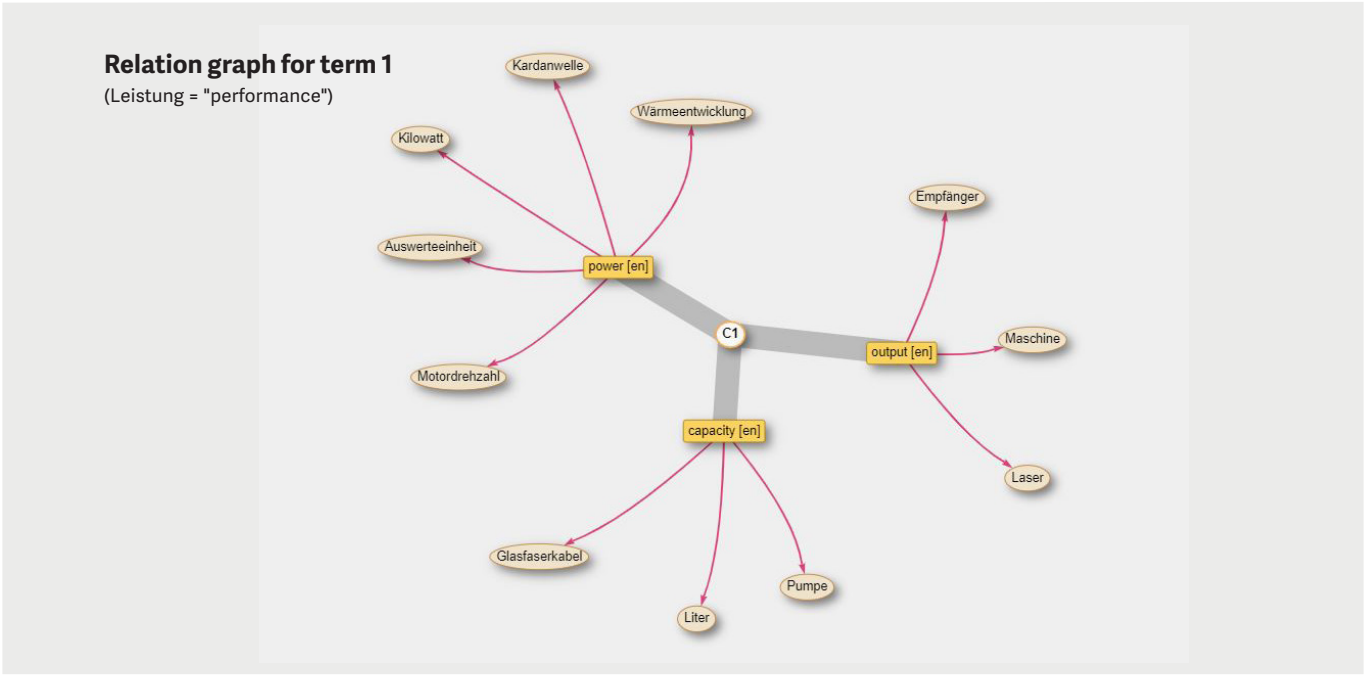


Figure 4: Contextual relations in LookUp

man post-editors are in a quandary. They must not work too slowly, otherwise the advantages of MT + post-editing will be lost. But they must also not work too quickly, otherwise they will overlook important errors. This applies even more so if the machine translation sounds good at first.

Therefore, what could be more obvious than to use the help of automatic checking systems to support the post-editor? After all, it is already common practice in industry: optical sensors automatically detect defective products on the assembly line, spellchecking systems discover spelling mistakes in documents, etc.

For almost 20 years, D.O.G. has been developing and using ErrorSpy, a proven quality assurance software solution that checks translations for a whole range of errors (numbers, terminology, completeness, standardization). ErrorSpy notifies post editors of potential errors.

The powerful integrated editor is a great help for error correction. About 5 years ago, we added new checking functions that are especially aimed at machine translation. Context checking is a key feature in this respect.

MT systems find it particularly difficult to use the correct translation in context. Words like the German "***Leistung***", for example, have very different English translations depending on the device in question. Is it a motor that generates a power (EN: ***power***)? Is it a pump that produces a flow rate in litres (EN: ***capacity***)? Or is it a machine that produces a certain number of screws per hour (EN: ***output***)? Thanks to the contextual relations stored in the terminology, ErrorSpy can detect incorrect MT translations in context.

A detailed description of ErrorSpy can be found here:
www.dog-gmbh.de/en/products/errorspy

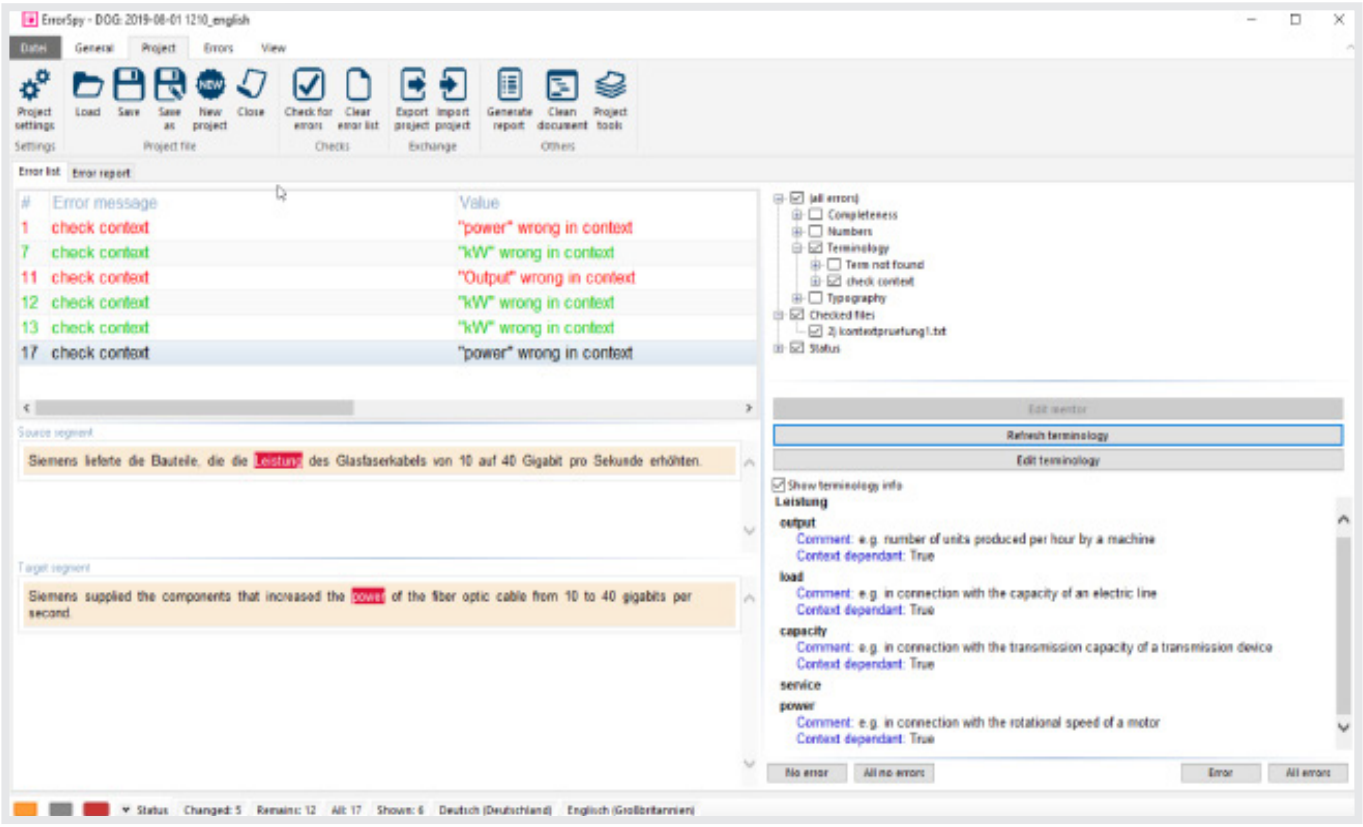


Figure 5: Unique context check in ErrorSpy

What can you do to optimize MT results?

Machine translation systems struggle with texts that have long sentences, contain spelling mistakes, are written very inconsistently or are formulated inaccurately.

Companies can certainly influence the results of machine translation by writing "better" and "machine-compatible" texts. This is especially true for companies that produce a lot of information material and where different employees write texts over the years. Then it is well worth writing in a controlled language. A style guide summarizes the rules for optimized writing (sentence length, syntax, use of verbs, etc.). You can specify your company's terminology and define which synonyms are preferred and which are prohibited (***cruise control***, not ***speed limiter***). The terminology management system LookUp from D.O.G. GmbH is an excellent tool for this purpose (www.dog-gmbh.de/en/products/lookup).

Some rules for the optimization of your texts:

1. Check the spelling: MT systems have difficulty handling spelling mistakes.

2. Pay attention to correct punctuation in the sentence: commas, semicolons, dashes, single quotes, etc. They help MT systems to understand the sentence correctly.

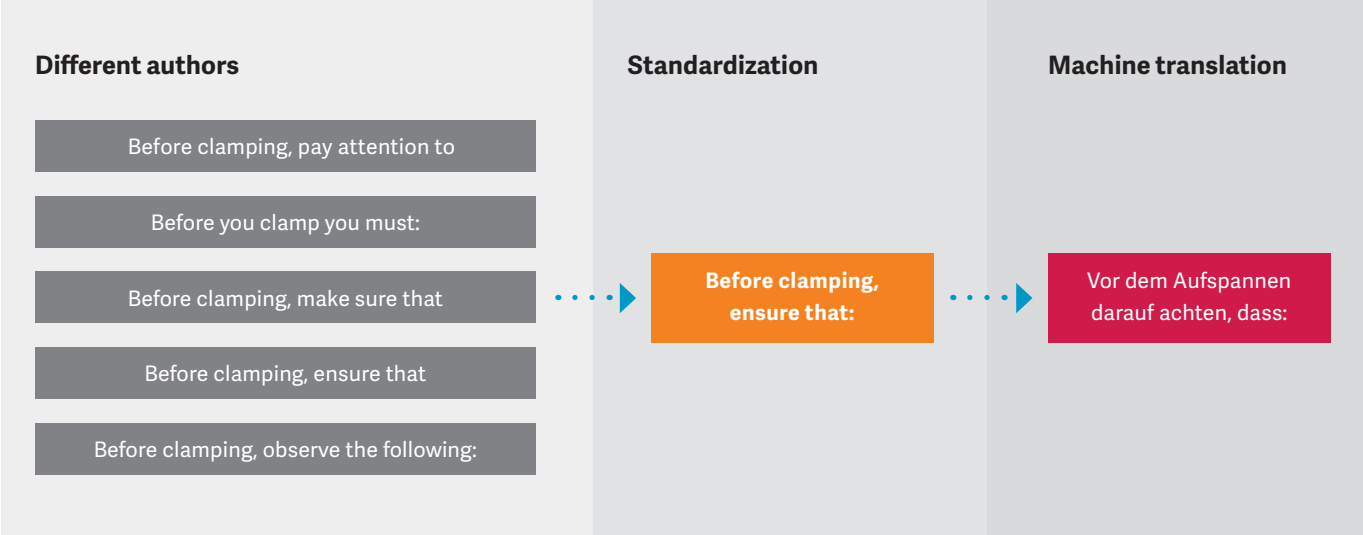
3. Write simply: short sentences, few subordinate clauses.

4. Use standardized terminology, if possible without generic terms (***throttle valve*** instead of ***valve***): MT systems need precise terms.

5. Do not use filler words (***namely, however...***)

6. If possible, do not use references to words outside the sentence (***the telephone*** instead of ***it***, etc.): MT systems are lost without context information.

7. When formatting your documents, make sure that they do not interfere with the work of the MT system: e.g. hard returns in the middle of a sentence.



Terminology for machines

Well-maintained corporate terminology is a decisive advantage when using machine translation systems. Whereas terminology work used to focus on communication between or with people, MT opens a new dimension: terminology for machines. Because MT systems sometimes make completely different errors than humans, terminology work must take this into account.

Using extracted terminology, quality assurance systems such as ErrorSpy can detect incorrect terms that a human would never use. These terms, which are typically incorrectly translated by machines, can be included in the terminology and given a usage label such as "prohibited". This may include words of the general language as well as proper names or product names.

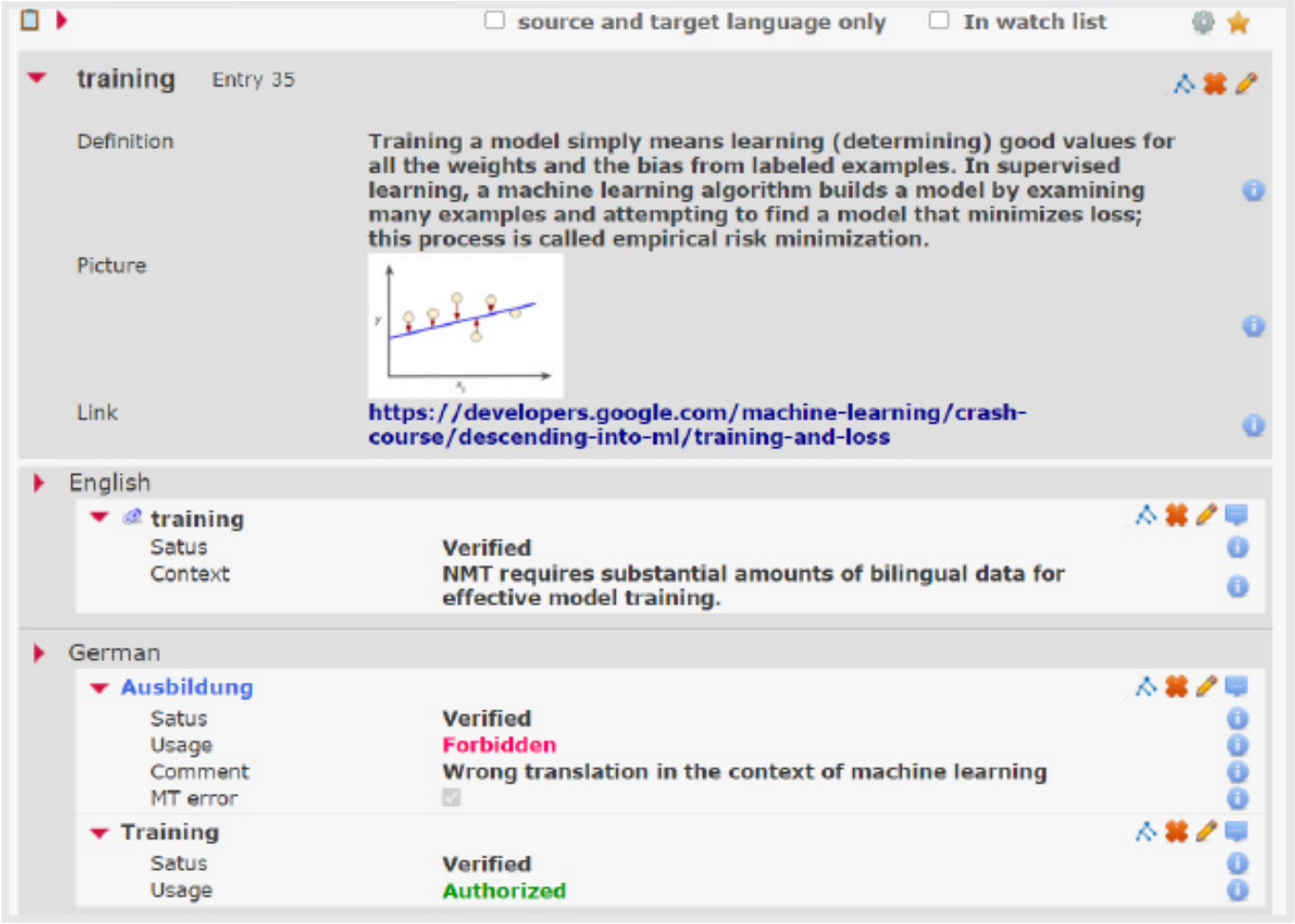


Figure 6: Terminology entry used to recognize MT errors

Better training material

It is not only the optimized writing of texts that improves the results of MT, but also the quality of the material used to train an MT model. In keeping with the motto "Garbage in, garbage out", the systems are only as good as the training material used.

Many companies that translate regularly may have access to translation memories (TM). TMs are databases of previously translated sentences that their translators usually create. Once their contents have been cleaned up and optimized, these translation memories are an excellent source for training MT systems.

Cost and time advantages

The most frequently mentioned reasons for machine translation are time and cost savings. In fact, for some people the time factor alone is reason enough to use machine translation.

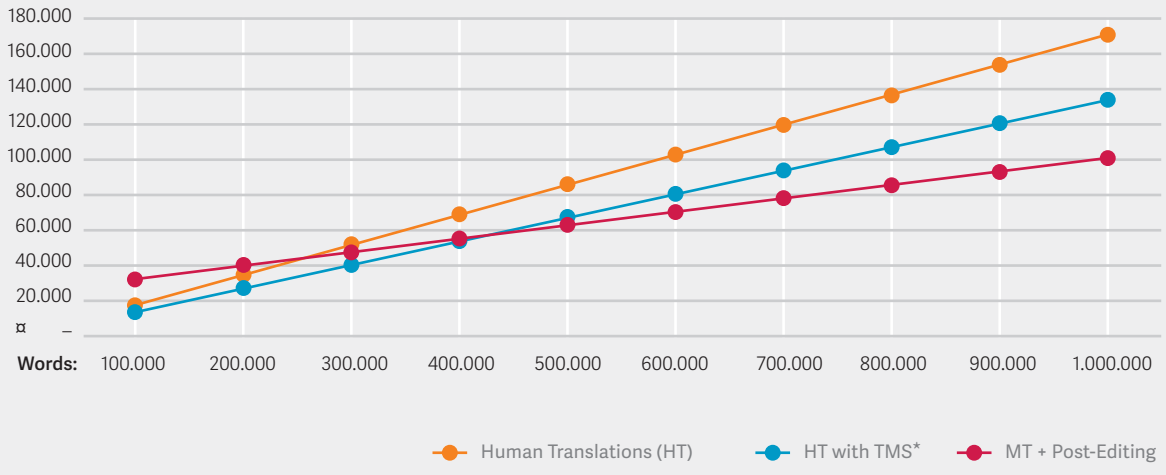
If a web shop is translated 4 weeks earlier in German, French or Spanish, sales can start 4 weeks earlier. And that can be worth a lot.

All kinds of cost information circulate on the Internet, so it is not always easy for the layperson to distinguish between serious and dubious statements. Costs are mainly incurred:

- ✔ for training a model and subsequently for the regular maintenance of the trained model including the technical infrastructure
- ✔ for post-editing machine translations.

Many providers charge a low base rate per translated word for the first cost component and a word rate for post-editing, depending on the amount of editing required. Costs can be 20-50% lower than for traditional translations. The final savings depend on how demanding the texts are and whether the use of translation memories has already resulted in significant savings.

Example: Cost comparison
One language with 30% repetitions



** TMS = Translation Memory System*

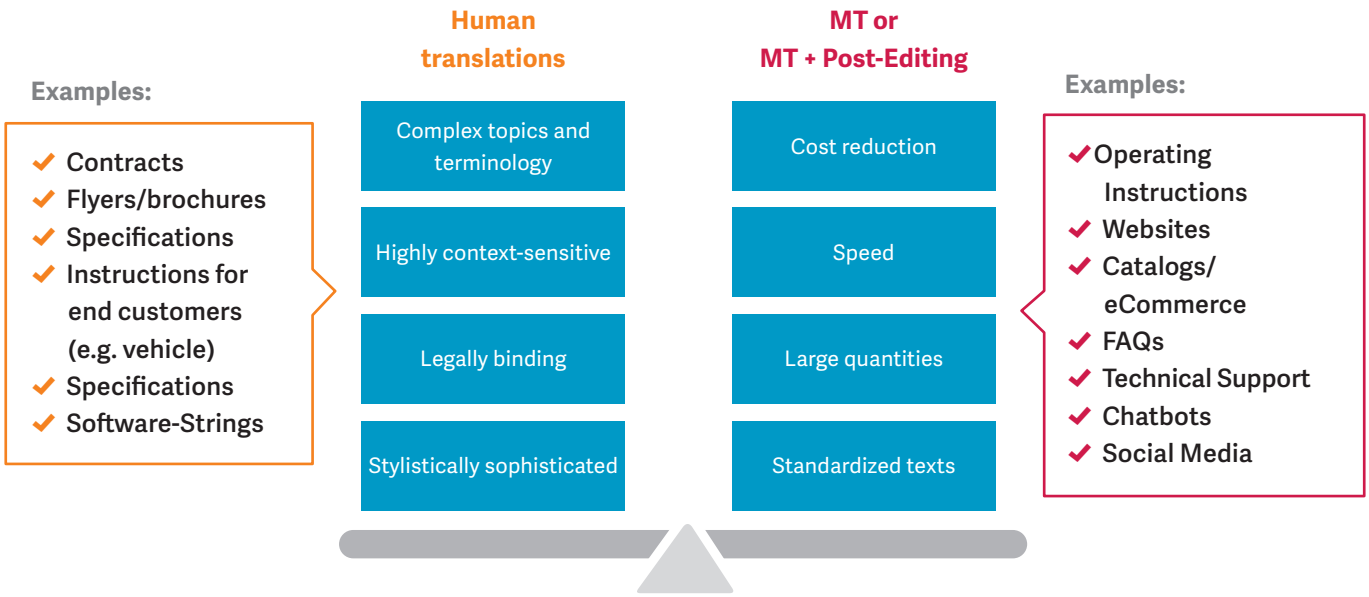
When to use MT?

The decision to use MT or not is not an all-or-nothing decision. Companies that choose MT still produce part of their translations in the traditional way, using human translators. Which documents or information are suitable for MT depends on a number of factors and priorities. We have summarized in the following diagram some of these decision factors.

For many texts and publications, MT is usually deployed in combination with post-editing. In situations where the texts are comprehensive and the sentences are written in a uniform style, the results of MT are relatively good, especially from MT systems trained with your own texts. Operating instructions, catalogs and web stores, websites and training materials are just a few examples.

The best solution is to rely on a translation service provider that offers both traditional and machine translation with a trained translation system. D.O.G. GmbH has specialized in this area.

If you want to learn more, you can visit our website and read how we train and use MT systems for our customers. Of course, you can also contact us directly and request a quote or arrange a consultation.



Brief overview of services offered by D.O.G.

✔ We advise you on the pros and cons of different alternatives.

✔ We work with you to develop a specification that defines systems, workflows, integration requirements and quality guidelines.

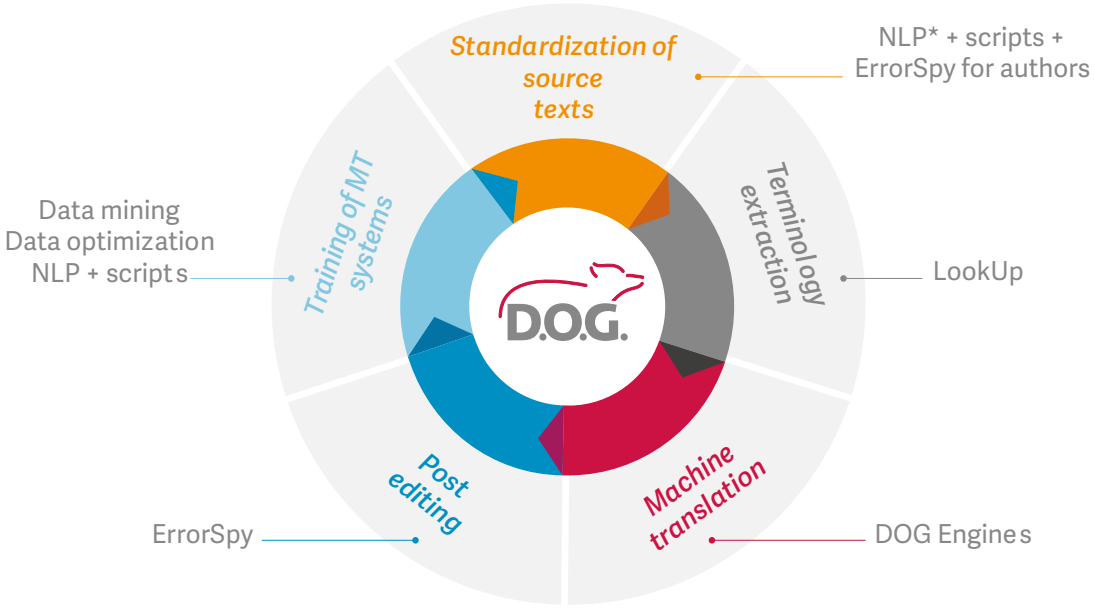
✔ We use the most suitable MT system for you, e.g. a neural machine translation system (NMT), which we train with your data.

✔ We set up quality management for your MT texts and deploy a team of post-editors. Our post-editors correct the machine-generated texts according to agreed criteria. We use our quality assurance software and terminology agreed with you, which we can maintain together with you in our terminology management system LookUp.

✔ We continuously maintain the language resources such as translation memories and terminology, which are important for optimal training of the translation engine.

✔ You have a permanent D.O.G. contact person who coordinates a team of developers, post-editors and translators for your projects.

D.O.G. services and ecosystem:



** NLP = Natural Language Processing*

Further reading:

✔ DIN ISO 18587:2018-02 Translation services – Post-editing of machine translation output – Requirements (ISO 18587:2017). *Beuth Verlag GmbH, Berlin.*

✔ Koehn, Philipp (2020). Neural Machine Translation. *Cambridge University Press.* doi. org/10.1017/9781108608480

✔ Massion, François. 2020. DeepL und Terminologie. *Edition - Deutscher Terminologie-Tag e.V. (DTT) – P. 18-25.*

✔ Porsiel, Jörg (Ed.). 2020. Maschinelle Übersetzung für Übersetzungsprofis *BDÜ Weiterbildungs- und Fachverlags GmbH, Berlin.*

Dr. François Massion

01.10.2020

Your contact person:

Dr. François Massion
Managing director
e-mail: francois.massion@dog-gmbh.de
Tel. +49 (0)7152 35411-0

Dokumentation ohne Grenzen GmbH
Neue Ramtelstr. 12
D-71229 Leonberg
Tel. +49 (0)7152 35411-0
Fax +49 (0)7152 35411-50
www.dog-gmbh.de

e-mail: info@dog-gmbh.de
Tel. +49 (0)7152 35411-11 or -0

Cover photo: Adobe Stock - Design: www.simonehaasis.de