Translating with human and machine power

Artificial Intelligence has allowed us to boost productivity in machine translation. So where to from here? Which way to go with regard to pre-editing and training with one's own data? What we need is the expertise of technical communicators.

Text by Rachel Herwartz

In Neural Machine Translation (NMT), a number vector is generated for each word in the input sentence ("input word"). During the machine translation process, an encoder enriches each input word with contextual information from the entire input sentence. A decoder generates the output words and builds the output sentence word by word. The probability of each following word results from the words of the input sentence and all previously generated output words (Figure 1).

The success of this translation method is based on the use of artificial neural networks. In the brain, nerve cells (neurons) are connected to each other by synapses, thus creating "neural networks".

Image: © baona/istockphoto.com

Artificial neural networks (ANNs) are programs that learn what the expected outputs are for certain input values by linking data during training. This is how AI systems can autonomously recognize traffic signs on a road and distinguish pedestrians from cars. But if a truck tips over and blocks the road crossways, the system will fail. The situation does not correspond to the "learned" scenario. Neural machine translation follows the same pattern. So translations have "something to do with AI" these days.

Training of a neural system

To train the NMT engine, i.e., the AI system, a number vector is randomly assigned for each word (Figure 2, left). The training material consists of bilingual sentence pairs. This is how generic systems such as those from Google, Microsoft, or DeepL were trained with millions of data. With each new sentence pair added to the training material, the engine "learns" which words appear repeatedly in similar contexts. These words converge during training and thus also acquire similar coordinates (Figure 2, right). Therefore, the prediction of the system is not based on grammatical rules, but on frequencies of distribution of words in the training material.
In the example on the right in Figure 2, the four words "mouse", "cat", "hungry" and "house" have already moved closer together through training.

Dealing with homonyms

In Figure 3, the word "bat" appears near the verbs "fly" and "throw" after training. This is because "bat" is a homonym in English, i.e., a word with several meanings. However, "bat" meaning "Fledermaus" (animal) seems to be more frequent in the training material than the less frequently occurring meaning "Schläger" (sports equipment), which refers to "a piece of wood with a handle, made in various shapes and sizes, and used for hitting the ball in games such as baseball, cricket and table tennis" (Oxford dictionary).

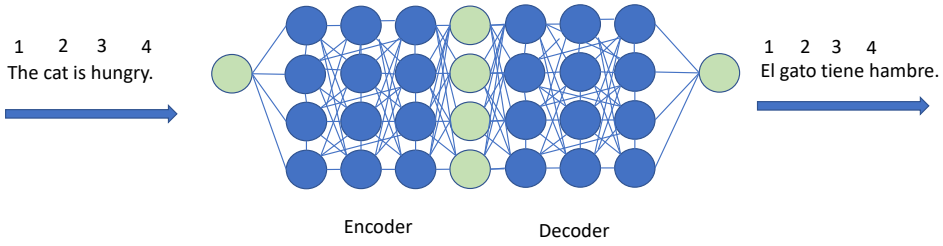

Figure 1: Encoder decoder architecture

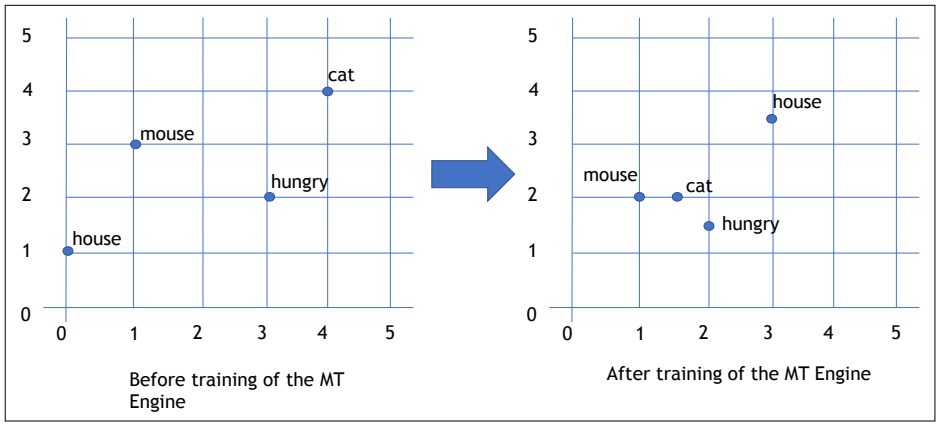

Figure 2: Example of number vectors before and after training

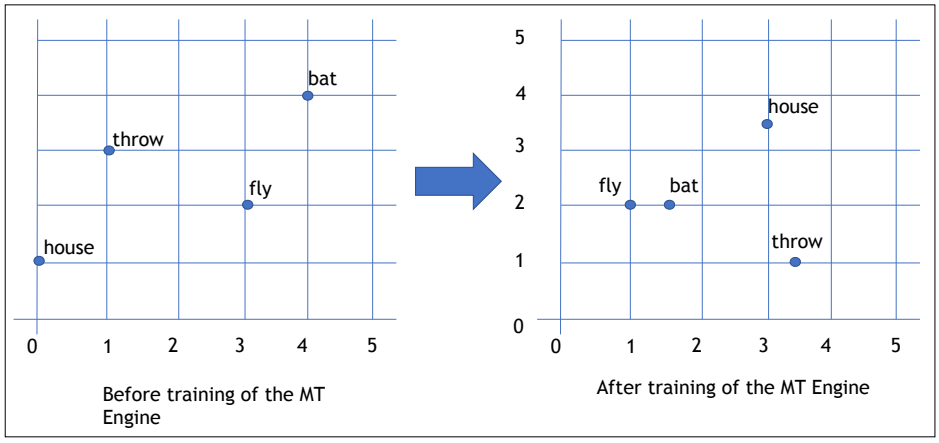

Figure 3: Dealing with homonyms such as "bat"

If one meaning prevails in the training material, the NMT engine will translate the homonym based on this meaning, thus again taking into account the frequency distribution. As already shown in Figure 1, the encoder-decoder architecture takes into account the context during translation. For this purpose, each word of the source sentence and all already translated words of the target sentence are considered in the generation of the target sentence. Thus, sentences such as "This bat was thrown" and "This bat was flying next to our house" were corrected accordingly during the course of translation (see Figure 4 and 5).
Still, this is mainly down to mathematics and statistics rather than grammar and semantics. Adding an unfamiliar word can change the result, as shown in Figure 6.

By adding a word in another language ("Schiedsrichter" instead of "referee") "bat" was translated as "Fledermaus" despite the existing context of "thrown".
This highlights the fact that terminology control is essential, especially for NMT. [2]
Besides, there is no guarantee that the decoder has considered all encoder outputs when it reaches the end of the sentence. The reason for this is that the decoder – during translation – can already include input words that it finds useful. This can lead to unwanted additions ("over-translation") or omissions ("under-translation") in the target texts. So-called "hallucinations", in which words are repeated several times, can also occur. [3]

Ambiguous personal pronouns

The NMT also fails with regard to personal pronouns whose assignment only becomes clear with context. Thus, the following German example cannot be resolved by a machine: „Immer wenn Marie ihre Katze füttert, beißt sie sie" (see Figure 7). Common sense tells us that the cat bites Marie and not vice versa, so a human translator would automatically choose the correct translation: "Whenever Marie feeds her cat, it bites her". The machine, however, cannot rely on world knowledge and misinterprets the ambigious pronoun. Systems like Google Translate will sometimes deliver the correct sentence in target languages like English or Spanish. More often, they will produce translations implying that it is Marie who bites the cat.

Figure 7: Example from Google Translate (6/7/2022) DE – EN/ES

Advantages of NMT

Texts translated with NMT systems are more fluent and of much higher quality than those of rule-based (RBMT) or statistical translation (SMT) systems that have prevailed in the past. Translation quality can be measured not only by a human translator but also by automatically calculated scores such as BLEU (Bilingual Evaluation Understudy).
In BLEU, a machine-generated translation is compared with the reference translation of a human translator. The bigger the gap between machine translation and original translation, the lower the respective BLEU score. However, we should keep in mind that even human translations never match this first reference translation 100%. Even a translator who re-translates his own text months later might make slight deviations. Using the BLEU score, various studies show a significant increase in the quality of neural MT over other types of machine translation. [4] This applies, for example, to the language combinations German/English and English/German as well as translations from Chinese into English.
In an assessment of NMT in the language combination English/German, human translators found 50% fewer word order errors, 17% fewer lexical errors, 19% fewer morphological errors, and a total of 26% less post-editing effort than for RBMT. [5]
Language service providers who have been using machine translation for some time also report a huge increase in quality when switching from SMT to NMT. For example, NMT elements can be adopted in full more often, and there is less rework overall.
For some experts, translations – in particular from English into German and vice versa – have now reached such a high quality that "it might make sense to replace a large part of pure translation exercises with post-editing courses" during translator training. [6]
Nevertheless, the quality of MT can usually be improved even further. This can be achieved by training it with your own material and by prior pre-editing.

Figure 4: Example from Google Translate EN – DE

Figure 5: Homonyms such as "bat" are recognized only by adding more context such as a verb.

Figure 6: Unknown words can affect the result badly.

Check and adjust

In parallel with "post-editing" (PE), defined in DIN ISO 18578 as the post-editing of a

text that has been translated exclusively by machine, "pre-editing" refers to a revision of the source text to achieve a better result in machine translation. According to the standard, this is the responsibility of the translator or post-editor, although he or she often does not feel called upon to revise the source text. If, on the other hand, technical communicators compose their texts in compliance with the applicable spelling rules, with correct formatting, and according to the rules of translation-oriented writing, these work steps are not far removed from pre-editing.

The effects of controlled language on the quality of machine translation have been researched for a long time. [7] Rule-based MT achieved great successes as early as 1996: "By integrating pre-editing based on Siemens Documentation German [SDD - Siemens Dokumentationsdeutsch, R.H.] into the translation process with the in-house MT system, Siemens was able to largely dispense with post-editing." [8]

In a 2017 study of pre-editing in statistical MT, over 90 percent of the source text was edited so that the text could be translated in a sufficiently high quality using a statistical machine translation system. [9] Unfortunately, these successes have not yet been clearly demonstrated for NMT. Based on nine selected rules from the tekom guideline 2013 (Rule-Based Writing) [10] and a corpus of ten German user manuals, Marzouk [11] found that the BLEU score for NMT is 83 percent, which is twice as high as that of SMT and RBMT. However, this good result applies to the pre-edited as well as the non-pre-edited texts.

The following rules of the tekom guideline were selected for the analysis:

- Mark up interface texts with quotation marks
- Avoid paraphrasal verb structures
- Introduce conditional sentences with "if"
- Make clear pronominal references
- Avoid participial constructions and passive voice
- Avoid constructions with "to be" and "to"
- Delete superfluous prefixes
- Do not omit parts of words

Evaluating these translations, human translators often made the criticism that the translated text, which had been optimized by controlled language in the source language, "doesn't sound as nice" in the translation. [12] However, this evaluation applies not only to the translated text but also to the controlled language in the source text.

Translation service provider Dr. François Massion emphasizes that certain characteristics help identify difficulties that might arise when using MT. These characteristics include: word count per sentence, medium word length per sentence, punctuation, number of verbs per sentence, complexity of sentence structure, and ambiguity. [13] This leads translators to not pre-edit the source text but divide segments into these three categories:

- Not suitable for MT
- Suitable for MT (and post-editing)
- Does not need MT

In 2021, the experts Miyata and Fujita repeated their 2017 pre-editing study, but this time for Neural Machine Translation. Whereas previously it was recommended that sentences are kept as short and simple as possible for machine translation, they now stated regarding NMT: "Rather, it is more important to make the content, syntactic relations, and word senses clearer and more explicit, even if the ST becomes longer." [14]

Domain-specific or generic training

More specific training material can further improve MT results. Despite the "black box" that is NMT, we can already influence the connections between words during training in such a way that certain connection paths are preferred over others.

For example, you can teach the NMT to prefer information "from subject area X", "from text area Y" or "for target group Z". This can be remedied by "domain-specific" trained engines, such as the Matecat translation system, where one can currently choose between 36 different "subjects" for each translation job.

Self-trained data

Other vendors leave it up to their customers to use the generic engine or to use a system created or enriched with their own data. Google AutoML shows for each translation the result of the generic engine as well as that of the self-trained engine. A study, updated annually, provides an up-to-date overview of current systems. [15]

However, a distinction must be made between training an empty system and "retraining" or "customizing" a generic system. As the former would require at least one million sentence pairs per language direction and per subject area, the only realistic option – except for governmental agencies and large corporations – is to enrich an existing system with one's own data. [16] In this way, good results can be achieved with as few as 5,000 to 15,000 sentence pairs per language direction.

More proprietary data can sometimes degrade the MT quality in the systems.

It is also possible to prioritize training pairs, for example by preselecting "good" segments, which are then weighted higher than the others.

Good training material in terms of bilingual sentence pairs must be free of errors, especially in spelling and punctuation, and must not contain detached sentence fragments or superfluous tags or formatting. This is

Figure 8: Workflow of (re-)training and customization

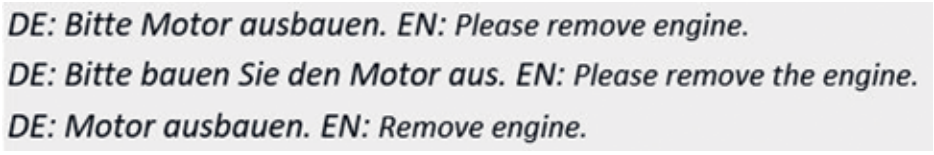

Figure 9: Call to action

References

[1] Burchard, Aljoscha/Porsiel, Jörg (2017): „Was kann maschinelle Übersetzung und was nicht." In: Porsiel, Jörg.

[2] Herwartz, Rachel (2021): „Neuerung mit Folgen." In: technische kommunikation.

[3] Van Genabith, Josef: (2020): "Neural Machine Translation". In: Porsiel, Jörg: *Maschinelle Übersetzung für Übersetzungsprofis*.

[4] Vashee; Kirti (2017): "Neural Machine Translation; A Practitioner's Viewpoint." In: Porsiel, Jörg: *Maschinelle Übersetzung. Grundlagen für den professionellen Einsatz*.

[5] Bentivogli, Luisa et al. (2016): "Neural versus PhraseBased Machine Translation Quality: A Case Study." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

[6] Hansen-Schirra, Silvia/Schaeffer, Moritz/Nitzke, Jean (2017): „Post-Editing: Strategien, Qualität, Effizienz." In: Porsiel, Jörg.

[7] Lehrndorfer, Anne (1996): „Kontrollierte Sprache für Technische Dokumentation – Ein Ansatz für das Deutsch." In: *Krings, Hans-Peter: Wissenschaftliche Grundlagen der technischen Kommunikation*; and Lehrndorfer, Anne (2001): „Kontrolliertes Deutsch. Linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der Technischen Dokumentation."

[8] Marzouk, Shaimaa/Hansen-Schirra, Silvia (2020): „Kontrollierte Sprache im Zeitalter der neuronalen Maschinellen Übersetzung." In: Porsiel, Jörg.

[9] Miyata, Rei/Fujita, Atsushi (2017): "Dissecting human pre-editing toward better use of off-the-shelf machine translation systems". In: *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*.

[10] tekom-Leitlinie (2013): *tekom-Leitlinie* (2014): "Rule-Based Writing – English for Non-Native Writers."

[11] Marzouk, Shaimaa/Hansen-Schirra, Silvia (2020): „Kontrollierte Sprache im Zeitalter der neuronalen Maschinellen Übersetzung." In: Porsiel, Jörg.

[12] Marzouk/Hansen-Schirra (2020).

[13] Massion, Francois (2020): "NMT im Einsatz bei einem Dienstleister: von der Systemauswahl bis zum fertigen MÜ-Workflow." In: Porsiel, Jörg.

[14] Miyata, Rei/Fujita, Atsushi (2021): Understanding Pre-Editing for Black-Box Neural Machine Translation. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

[15] Nimdzi (2021): Nimdzi Language Technology Atlas: the Definitive Guide to the Language Technology Landscape www.nimdzi.com/languagetechnology-atlas

[16] Winter, Tom/Zielinski, Daniel (2020): Terminologie in der neuronalen maschinellen Übersetzung. In: Porsiel, Jörg.

[17] Lieske, Christian (2020): Metadata and Machine Translation. In: Porsiel, Jörg.

[18] Burchard/Porsiel (2020).

not a given for many companies' translation memories, which must therefore be cleaned up before training (Figure 8).

Training material standardization

Translation memories often lack metadata attributes such as "user guide" and "marketing material".
And sometimes they have accumulated as unstructured MT over the years: "A linguistic data management is still missing at both client and vendor side." [17]
However, when it comes to requirements, one can go one step further. Even if a text type such as "technical documentation" is identified with the help of metadata and an engine is trained with it, it must be ensured that the sentences have consistent wording patterns, for example in the call to action (Figure 9). In addition, one should not use translation memories that mix British and American English. Otherwise, the engine "picks up" these variants during training and then randomly outputs one or the other variant in the translation.

Enriching the training material

If you don't have enough bilingual material of your own, it is possible to enrich the training material by a machine back-translation or the generation of further sentence pairs by placeholders and sentence duplications, for example with numbers or dates. However, you can also purchase domain-specific language pairs or use free open-source collections for training.
For rare language combinations, it might be advisable to translate first into English and then from English into the target language via a pivot translation or relay language.
For languages with a similar structure, you can also work with multilingual sentence pairs. Initially, you only train one language pair, for example, German/Spanish. Other Romance languages such as Italian, French, Portuguese, and Romanian can then be "attached" to this pair.

Integrate terminology

Terminology must be integrated into all process steps of a machine translation: in the training material of the MT engine, in the pre-editing of the source text, during the ongoing MT process, in the post-editing by the human translator, and the final quality assurance.

New tasks

When selecting a machine translation tool, there is still no "one-size-fits-all" application. Rather, there are solutions available that fit particular language pairs, text types, subject areas (domains), and possibly, target groups. [18] Technical writing is responsible for editing the source language with meaningful metadata, standardized wording patterns, and controlling the terminology.

Technical writers are also well versed in preparing translation-ready source texts. As all source texts and their translations can serve as training material for machine translation systems, the quality of the source texts also affects the quality of the machine translations as well as the post-editing effort.

When translation memory systems were introduced, coordinating technical writers and translators was a challenge. Today, we are facing new challenges coordinating the machine translation process. We will only master them if technical writers, post-editors and translators work together.

ABOUT THE AUTHOR

Prof. Dr. Rachel Herwartz is a consultant for the design of translation and terminology processes.

She is the managing director of TermSolutions GmbH, a provider of terminology software and consulting. She heads the "Translation Management" course and is responsible for the online certificate courses "PostEditing" and "Machine Translation" at the International University of Applied Sciences SDI Munich.

@ rachel.herwartz@termsolutions.de
www.termsolutions.com