Machine translation

In brief

CAT *Automàtica (traducció)*

origins

The *translatio/traductio* part of the term is clearly analogical here; the differences are clearly evoked by modifying it in English by prepending it with the term *machine*, coming from Latin *machina,* itself an early borrowing from a form of Greek *μηχανή* (*mekhané*), meaning 'device, gear, or contrivance'. The variants of *automatic* used in romance languages to refer to machine translation are modern reflexes of the Greek term *αὐτόματον (autómaton),* a neutral adjective meaning 'self-moving, spontaneous', used to refer to the way in which machines act without any human intervention. Interestingly enough, Russian originally used *автоматический перевод* (*avtomaticheskiy perevod*), but *машинный перевод* (*machinniy perevod*), parallel to *machine translation*, *is* much more common now.

other names

Automatic translation (used by Yehoshua Bar-Hillel and others in the 1960s); *mechanical translation* (used in the 1950s and 1960s and then mainly in Japan in the 1980s, but with a different meaning in the field of Torah translation); *automated translation* (sometimes in European Commission parlance).

abstract

Machine translation is the process by which a computer system produces, from a source-language computer-readable text, a target-language computer-readable text which is intended to be an approximate translation of the former.

Machine translation, a mature technology today, has radically changed the way in which people perceive multilingual communications, as nowadays anyone having access to the Internet can use it, for instance, to make sense of web content written in a different language. Of course, it has also impacted translation as a profession (and the way it is perceived by the general public). After defining machine translation and distinguishing it clearly from other computer-aided translation technologies and giving a brief historical review, from the early rule-based systems of the 1950s to the statistical systems of the 1990s and the early 2000's to the advent of the "deep-learned" neural approaches in the twenty teens, this article describes how machine translation is used by ordinary people and in professional computer-aided translation workflows, and how it can be evaluated, both when considering adoption or during its development. It also describes the main technological approaches: on the one hand, rule-based machine translation and, on the other hand, corpus-based machine translation in its two flavours: statistical and neural, both to allow professional translators to make informed decisions about the technology and to raise the awareness of the general public about what to expect from this technology and how to use it where applicable. To close, some active research lines in the field of machine translation are outlined.

record

Mikel L. Forcada Zubizarreta

2022

Forcada Zubizarreta, Mikel L. 2022."Machine translation" @ *ENTI (Encyclopedia of translation & interpreting).* AIETI.

https://doi.org/10.5281/zenodo.6369130

https://www.aieti.eu/enti/machine_ENG/

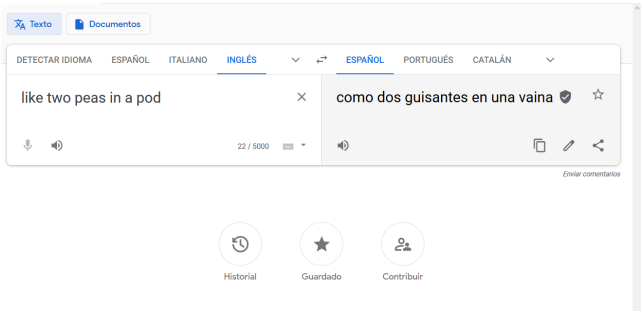
Entry

CAT *Automàtica (traducció)*

contents

Introduction

Machine translation (MT) is the process by which a computer program produces, from a source-language computer-readable text, a target-language computer-readable text which is intended to be an approximate translation of the former, and does so without any human intervention.



Raw MT is not always ready to use: The English idiom Like two peas in a pod, *meaning 'very similar', is literally translated into Spanish.*

MT should be clearly distinguished from other translation technologies used by translation professionals, such as computer-aided translation, based on translation memories, where professionals use previously existing translations of related text segments to translate new segments, or even from other natural-language processing technologies applying computational linguistics such as automatic speech-to-speech translation, that is, automatic interpreting. MT is a text-to-text technology and is fully automated.

But, is raw translation really a translation? Usually, it cannot be used as a professional translation would; as Sager (1994) put it in his pioneering book, "there is no single situation in which [they] would be equally suitable". This means that, for example, MT output can rarely be published as it is; this does not mean, however, that MT is useless. Indeed, customers of professional translators are starting to be aware that there is a technology out there that may help them do their job and

therefore demand better prices, while the general public is gradually getting used to read raw machine translation output and even make purchasing decisions based on machine-translated reviews or descriptions. As a result, professional translation is perceived a dispensable luxury in many applications. Professional awareness of the usefulness and the limitations of MT is therefore crucial in contemporary translation practice.

back to top

¶ Two main uses of machine translation, *assimilation* and *dissemination*

The wide availability of MT, a mature technology today, has radically changed the way in which people perceive multilingual communications, as nowadays anyone having access to the Internet can use it, in many cases for free, make sense of text (usually web content) written in a language that they cannot read. When this happens, MT is said to be used for *assimilation* or *gisting* purposes. Assimilation is by far the most common use of MT. Franz Josef Och, the scientist then in charge of Google's MT, said already in 2012: “In a given day we translate roughly as much text as you’d find in 1 million books. To put it in another way: what all the professional human translators in the world translate in a year, our system translates in roughly a single day.” (Och 2012).

On the other hand, MT is used by professionals as a source of help when producing translations that will be published more or less widely. When this happens, MT is said to be used for *dissemination* purposes. MT has therefore impacted translation as a profession (and the way it is perceived by the general public), so much that the *retronyms* "human translation" and (unfortunately less) "professional translation", are gaining usage in order to refer to *translations* that have not been produced by an MT system.

back to top

¶ Post-editing, pre-editing, controlled languages

One common way in which professional translators take advantage of machine translation for dissemination purposes is by editing or, as it is usually termed, *post-editing* its output to turn it into a fit-for-purpose translation, when this is economically possible. When text is translated from a single source language to several target languages (for instance, in multilingual document management workflows), proactively *pre-editing* the source text (expectedly more expensive than post-editing as it requires understanding what kind of input makes the system produce errors) may be a way to partially avoid *post-editing* of content in MT output for more than one language. Pre-editing itself may be avoided if authors may be constrained in the way they produce source texts. This is done by defining and enforcing (using assisted editing) a *controlled language* that uses rules to restrict the lexicon and the structures of the source language to avoid machine translation problems.

back to top

¶ A brief history of MT

The idea of mechanical translation had been around for a while. Among the precedents of modern MT, Warren Weaver’s (1949) memorandum is often cited. Weaver’s two main contributions were: (a)

the idea that a text in the source language is like a “scrambled” version of the text in the target language, and that translation would be similar to a “descrambling” process that would intelligently use the statistics and probability theories used in communication theory; and (b) the idea that instead of directly translating from one language to another it may be more useful to look for representations of text that are deeper and try to leverage what languages have in common. These two ideas underlie much of the technology that ensued.

MT was one of the first applications of the pioneering electronic computing systems. An MT system (developed by IBM and Georgetown University) was demonstrated in public for the first time in 1954. It translated into English a set of 49 sentences in Russian using a *direct* word-for-word approach with a dictionary of 250 words with some adjustments performed using six grammar rules. Despite the limited results, the public and the industry were led to believe that in a few years quality translations of scientific and technical documents could be achieved. Research flourished with generous public funding but progress towards the goal of fully automated high-quality machine translation was too slow. In 1966, the Automatic Language Processing Advisory Committee (ALPAC) published a report that recommended that the resources devoted to MT research be used for better-defined and less ambitious natural-language processing tasks and for the development of translation aids such as machine dictionaries. Research in the United States almost halted —but it did not completely die out—, while it continued in Europe and Japan. Indeed, in the 1970s, a commercial system, Systran (still available, but now a completely new system), was adopted both by the United States Air Force and by the European Commission.


LANGUAGE AND MACHINES

COMPUTERS IN TRANSLATION AND LINGUISTICS

A Report by the
Automatic Language Processing Advisory Committee
Division of Behavioral Sciences
National Academy of Sciences
National Research Council

Publication 1416
National Academy of Sciences National Research Council
Washington, D. C. 1966

The ALPAC report (1966).

The 1980s saw how efforts were directed towards *indirect* translation using different levels of source-language analysis to build intermediate representations in systems such as GETA-Ariane (in Grenoble; Hutchins and Somers, 1992, ch. 13), SUSY (in Saarbrücken; Hutchins and Somers 1992, ch. 11), Mu (in Kyoto; Nagao, Nishida & Tsujii 1984), DLT (in Utrecht; Hutchins & Somers 1992, ch. 17), Rosetta (in Eindhoven; Hutchins & Somers, 1992, ch. 16), the systems developed at the Carnegie-Mellon University (in Pittsburgh; Hutchins & Somers 1992, sec. 18.1) and those developed by two international projects: Eurotra, funded by the European Community (Hutchins & Somers 1992, ch. 14), and the Japanese CICC project with participants in China, Indonesia and Thailand (Tanaka, Ishizaki, Uehara *et al.* 1989). While Eurotra did not succeed in delivering a usable MT system and was abandoned in 1992, it stimulated research on language technologies throughout Europe, and led to the conception of commercial systems such as Siemens’ *Metal* (Hutchins & Somers 1992, ch. 15), which were based on linguistic principles such as syntactical parsing and transformations. All of these systems relied on experts writing dictionaries and rules, and computer programs that would apply them to texts (rule-based machine translation, see below).

But toward the end of the eighties a new approach emerged at IBM (Brown, Cocke, Della Pietra *et al*. 1988; Hutchins & Somers 1992, sec. 18.3), even if it only started to be known in the early 1990s. A new system called Candide could extract detailed statistical information from a sentence-aligned

version of the Hansards, the bilingual English–French proceedings of the Parliament of Canada, *learn* probabilistic models of machine translation and efficiently apply them to new text, almost without any linguistic expertise involved in its development. The resulting system was not too distant from Weaver's (1949) statistical approach of *descrambling* the source text to translate. *Statistical MT* (see below) had come into the scene to compete with the *rule-based* MT approaches, and began displacing them. Until around 2015, many machine translation systems were based on *phrase-based* statistical machine translation, an evolution of the original IBM approach and the most used variety. This was due to the availability of free software to train and implement machine translation systems, such as Moses. While software was free, good training data was now the key; companies appeared which would ask customers to mix their data with the companies' own general data to build specialized systems, all through a regular Internet browser.

Around 2013, a new form of translation based on so-called *deep learning* started to dispute the hegemony of statistical MT. This new *neural MT* (see below for details) uses methods from a mature field of artificial intelligence called *artificial neural networks*; its results are shown in laboratory tests to be comparable or better than the best statistical MT results available. As statistical MT, *neural* MT learns from bilingual texts, and is behind popular contemporary commercial MT systems such as Google Translate, Microsoft Translator and DeepL.

back to top

¶ General-purpose and special-purpose MT

As regards its sensitivity to the actual genre or subject matter of texts, two types of MT can be distinguished: general-purpose MT and special-purpose or task-tuned MT. General purpose MT (that is, systems such as Google, Microsoft Translator, DeepL, etc.) tries to satisfy the needs of everyone and every type of text: it is usually free or almost free, but it cannot satisfy the needs of a specific translation task, because it tries to address all at once. MT is getting better by using the *co-text* (adjacent text) but is still far from taking *context* (the whole communication circumstances that also contribute to the actual interpretation of the text) into account. Special-purpose or task-tuned MT may be therefore better at translating texts in a particular genre or subject but one usually has to pay for it; there is indeed business in adapting MT for specific tasks. Task-tuned MT may save translators from some of the boring (mechanical) part of their work: it provides quick and affordable raw translations, makes almost no typographical errors or misspellings, and tends to be terminologically consistent.

back to top

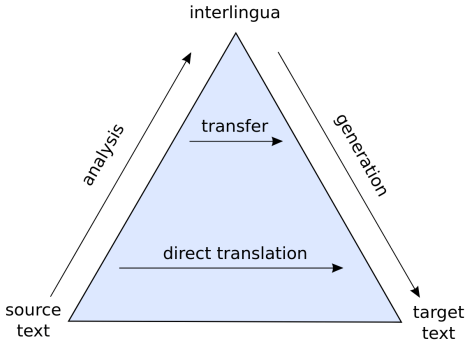
¶ MT approaches

Machine translation technologies can roughly be divided in two groups: *rule-based* MT and *corpus-based* MT, the latter with two main varieties *statistical* MT and *neural* MT. It is important to take into account that most current machine translation systems simplify the problem of translating text to translating its sentences one by one (this may be changing in neural MT). This "short-sighted" view of the text is however not only found in machine translation: in computer-aided translation, translation memories usually operate upon sentence-sized translation units.

Rule-based machine translation

Rule-based machine translation (RBMT), the dominant approach to machine translation from the first attempts in the 1950s until the 1990s, can still be found in systems such as Apertium. Rule-based machine translation progresses from word-for-word translation, adding rules that may or may not span the whole sentence.

To develop a RBMT system, on the one hand, translation experts compile dictionaries in electronic form and write rules that analyze the source text and transform structures of the source language into equivalent structures of the target language. Experts have to turn the intuitive and incompletely formalized knowledge of translators about the translation task into rules that have to be coded in an efficiently computable way; this may lead to rather radical simplifications, which, however, if chosen well, can be useful in most cases. On the other hand, computer experts write programs (called MT *engines*) that look up dictionaries and apply (in the expected order) the rules to the source text to analyze and translate it.



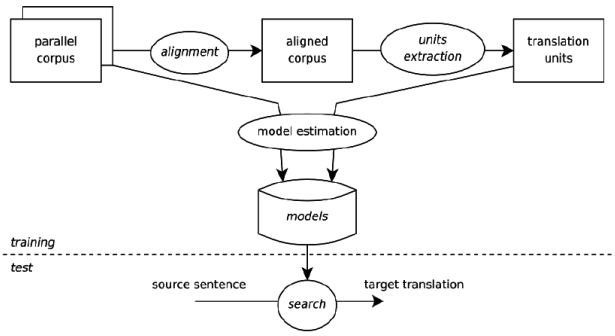
Rule-based machine translation is usually indirect and operates in three stages: analysis, transfer and generation. When analysis is so deep that transfer is not necessary, we have interlingua systems.

Historical approaches to machine translation such as the so-called *direct* or *transformer* approaches started with roughly word-for-word translation followed by *finishing* rules that tried to turn it into a grammatical target-language text. Most current rule-based machine translation systems may be described as following a three-stage *transfer* approach: in the first stage, *analysis*, source text is analysed (morphologically, syntactically, semantically) into a source-side abstract representation; the second stage, *transfer,* replaces source-language lexical items with target-language lexical items (*lexical* transfer) and transforms source-language structures into target-language structures (*structural* transfer) to generate a target-side abstract representation; finally, the third stage, *generation,* produces actual source-language text from it. In the extreme case in which analysis and generation are so deep that the source-side and target-side abstract representations are the same and there is no need for transfer, we talk about *interlingua* systems (the term *interlingua* actually designates the language-neutral abstract representation).

The output of RBMT systems is usually consistent but mechanical, lacking fluency. RBMT is famous for having trouble solving ambiguity both at the lexical level *(“replace” → “put back”/ “substitute”)* and at the syntactical/structural level *(“I saw the girl with the telescope” → “I saw the girl by looking through the telescope”/ “I saw the girl who had a telescope”).* The customization of RBMT to produce a special-purpose or task-tuned MT system is quite costly as it requires experts to edit dictionaries and rules.

Statistical Machine Translation

Statistical machine translation was the first approach to corpus-based machine translation (CBMT). Pierre Isabelle (Isabelle, Dymetman, Foster *et al*. 1993: 205) is often quoted as saying that "existing translations contain more solutions to more translation problems than any other available resource". This is the basic idea behind corpus-based machine translation, an approach in steady growth since the mid 1990s. Corpus-based machine translation programs *learn* to translate from huge corpora of bilingual texts where hundreds of thousands or millions of sentences in one language have been paired or *aligned* with their translation in the other language (that is, huge *translation memories*); untranslated target-language monolingual corpora may be used in some approaches too. In the case of corpus-based machine translation, the role of translation experts might seem less important unless one considers the fact that training corpora contain the work of (ideally, but unfortunately not always professional) translators. Such large corpora are seldom available for less-translated languages or domains, and this restricts the applicability of corpus-based MT.



Statistical machine translation (Karan Singla 2015).

Statistical MT, invented in the late 1980s, applied commercially since about the turn of the millennium, *learns* and uses probabilistic models that are estimated by counting certain events in the bilingual training corpus (for example, how many times a given word is next to another given word in the target sentence, or how many times a given word appears in the source sentence when another given word appears in the target sentence).

Statistical machine translation systems must be *trained*, on a large corpus of bilingual text aligned sentence to sentence and, optionally, an even larger target-language monolingual text corpus. The larger the corpora, the better each possible translation item (word, structure, transformation) is statistically represented; said otherwise, the larger the corpora, the more likely it is that words and structures in future texts are covered.

The training generates (a) probabilistic dictionaries containing words and segments of more than one word, where probabilities are associated with translation units such as *pursuant to/con arreglo a,* which is very likely, or *pursuant to/excepto si,* which is rather unlikely; (b) models that assign probabilities to each possible sequence of words in the target language (so that *two out of three are* is more likely than *two three out of are*), and, finally, probabilistic word reordering models (for instance, to obtain *el auto nuevo de Peter* from Peter's new *car*).

Using these correspondences between short stretches of source and target words learned from the training corpus (usually called *phrase pairs* even if they do not have to be phrases in a linguistic sense), statistical MT covers the source with the source *phrases* in every possible way, concatenates the corresponding target *phrases* in almost every possible way, and chooses the most likely way of doing so.

For instance, the Basque sentence

Hilaren 21ean irekiko dute Ipar eta Hego Euskal Herriaren arteko muga

could be translated into English as

> ***The border between the North and South Basque Country will be opened on the 21st of this month***

using some of the millions of phrase pairs automatically extracted from a Basque–English corpus such as

1. Hilaren→of this month
2. 21ean→on the 21st
3. arteko muga→the border between
4. Euskal Herriaren→Basque Country
5. Ipar eta Hego→North and South
6. irekiko dut→will be opened

by first slicing the source as

> *Hilaren / 21ean / irekiko dute / Ipar eta Hego / Euskal Herriaren / arteko muga*

and then using the target language model to score possible reorderings of the elements

1. *of this month*
2. *on the 21st*
3. *will be opened*
4. *North and South*
5. *Basque Countr*
6. *the border between*

and come up with a reasonable reordering. Another *legal* reordering would be

> ***On the 21st of this month the border between the North and South Basque Country will be opened*.**

But, how are translation and target probabilities obtained from parallel corpora used to translate? They are used as partial scores that are combined for each possible candidate translation of the sentence and weighted using a kind of “scale” to assign an overall score to each possible translation of a given sentence: the best translation will be the one that gets the best overall score. Obviously, not all possible translations are scored (but many are; approximations are used to search only among those that are a priori most likely). This scoring is a computationally very intensive process; this is why we have only had feasible statistical MT for about twenty years. Before, computers were not fast enough and could not store the parameter tables of the huge probabilistic models they use.

The scale has weights for each partial score, but what are the weights? The *tuning* of these weights is usually done on a small part of the bilingual corpus which has not been used for training (the *development* corpus); each sentence there is translated into the other language using different values for the weights, and the similarity of the machine-translated output is automatically and quantitatively compared to the translation in the target-side of the development corpus. Then, weights are chosen so that this similarity is as close as possible for the whole corpus. The assessment of this similarity is usually quite crude; the most popular measure of this kind, called

BLEU (Papineni, Roukos, Ward *et al*. 2002) counts the coincidences of groups of one word, two words, three words, etc. between the machine-translated and the reference sentence and combines them into a single measure that ranges between 0 (completely unrelated) and 100% (exact match). As a result, one important property of statistical translations is that they resemble those found in training corpora; this provides a clear opportunity for customization by selecting the adequate training material.

There is virtually no need for translation experts in statistical machine translation to build the system: the experts were the ones who produced the translations used to train and *tune* it. Statistical machine translation also has important limitations. An important one is that translations can look very fluent (due to the weight of the target-language probability model in the scale) but they can also be unfaithful, for example, because they have unnecessary additional words or because they miss some necessary words.

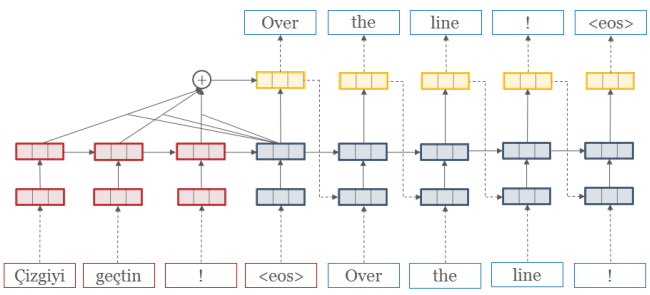
Neural machine translation

The new neural MT has been commercially exploited since 2016. It is based on *artificial neural networks* which are inspired (vaguely) in the way the brain learns and generalizes. In this case, they learn and generalize from the observation of bilingual corpora (Forcada 2017, Casacuberta & Peris 2017). In fact, the major publicly available online systems from Google, Microsoft, and so on, have turned neural and there are in addition, new “born neural” systems such as DeepL.

Neural machine translation has been reported as needing more data than statistical MT (Koehn & Knowles 2017), but these claims have recently been challenged (Sennrich & Zhang 2019). It is also commonly found to be more sensitive than statistical machine translation to noisy data (that is, when data contains sentence pairs which cannot be considered mutual translations) (Belinkov & Bisk 2018). Neural MT is considered competitive with statistical MT in many applications (Koehn & Knowles 2017, Sennrich & Zhang 2019), but comprehensive comparative evaluations in real-world applications are still scarce (see, e.g., Jia, Carl & Wang 2019; Shterionov, Superbo, Nagle *et al.* 2018; Klubička, Toral & Sánchez 2018).

As mentioned , neural MT is called *neural* because it is performed by software that simulates large networks of *artificial neurons*, which are in turn highly simplified versions of biological neurons. The activation or excitation of an artificial neuron depends on the activation of other artificial neurons and the strength of the connections through which their signals (and external input signals) are received by them. Signals coming from excited neurons through positive connections will tend to excite the neurons receiving them; signals coming from inhibited or depressed neurons through positive connections will tend to inhibit them. With negative connections, the behaviour is the opposite.

For a particular neural network to have a specific behaviour — that is, specific patterns of activation of neurons in the network — when processing a series of learning examples, it must be trained by changing the strengths of the connections. In artificial neural networks, neurons usually form *layers*, that is, groups of



neurons that receive signals only from neurons in the previous layer and send signals only to neurons in the next layer. *Deep learning* is said to occur when there are many of these layers, that is, when *learning* is performed by a *deep* neural net.

The structure of a NMT system: the sentence Çisgiyi geçtin! *is translated to* Over the line *(<eos> is an end of sentence marker).*

An important concept in neural networks is that of *representation*. The activation values of neurons in a layer are said to form a representation of the information they are processing. For example, in a trained neural network, the *vector* or fixed-length list (0.35, 0.28, –0.15, ... 0.88), where each of the (possibly hundreds of) values represents the activation of neurons in a layer, could form the representation of the word *study*, while (–0.35, 0.90, –0.12, ... 0.73) could be that of the word *snake*. The learned representations have interesting properties. Similar concepts turn out to have mathematically similar representations, which may be seen as if the neural network were learning their semantics during training.

Most neural machine translation architectures work by sequentially reading one by one the words of the source sentence to progressively build a representation of the whole sentence *(encoding),* and, once built, *extract* from it one by one the words of the translated sentence *(decoding),* taking into account the words already written. More precisely, at every step, each unit in the decoder calculates, for all of the words in the vocabulary, the likelihood of each possible target word and usually the most probable word is selected, in a way resembling the next word prediction feature in the keyboard of our smartphones.

There is a wide variety of neural MT designs: *recurrent encoder–decoder architectures* (Sutskever, Vinyals & Le 2013), augmented with *attention* mechanisms where each step in the process of building the source sentence representations is examined each time an output word is produced (Bahdanau, Cho & Bengio 2015) or even *transformer* architectures (Vaswani, Shazeer, Parmar *et al.*, 2017) where no explicit representation of the whole sentence is built and where attention is paid to past outputs too.

Neural MT is a completely new technology and this has important implications. On the one hand, it requires specialized, powerful hardware, as the calculations needed to simulate artificial neural networks with many thousands of units and millions of connections are mathematically very intensive. In particular, an evolution of computer graphics cards called GPUs (graphics processing units) are usually added to specialized computers. As mentioned above, neural MTs is also reported as requiring large amounts of data for bilingual learning. Neither such specialized hardware nor the large amounts of training data are usually available to most professional translators, who therefore have to resort to third parties to train neural MT systems and perform MT for them. This is actually a business model in the translation industry, as was the case with statistical MT; neural MT companies can add the translation memories to their general bilingual corpora to create a task-tuned MT setup system for them.

But, on the other hand, neural MT also produces quite a different type of output. First, as decoding starts from representations of the complete sentence, it is difficult to know the source of each target word (in statistical MT, one can easily trace the source *phrases* corresponding to each target *phrase*). Second, as happened with statistical MT, it can occasionally produce grammatically fluent texts that are not, however, a translation of the source sentence, and in fact, neural translation does

this more often. Errors are generally of a semantic nature: words not seen during training can be replaced by words with a similar meaning (*appliance* → *device*), even with dangerous results (*Tunisia* → *Norway*); paraphrases may even occur (*Michael Jordan* → *the Chicago Bulls shooting guard*). To mitigate the problem of translating words that are not seen during training, training material is usually automatically segmented into *sub-lexical* units, and this can lead to invented words such as *engineerage* ('engineering') or *recruitation* ('recruiting') made up of sub-lexical *translations* of sublexical units. All this means that professionals post-editing MT output have to pay even more attention as errors are likely to be quite subtle.

back to top

¶ Corpora for machine translation

Putting together and managing the corpora needed to train corpus-based (statistical or neural) machine translation faces important challenges:

- One has to acquire and assemble very large collections of parallel texts, that is, texts in one language with an equivalent text in the other language. This may not be available for some languages or for some text genres.
- Ensuring whether a text is a proper translation of another text may not be a trivial matter in some cases (as re-purposing of texts may have led to parts not being strictly translation). When harvesting text from multilingual websites, aligning each document with its translation has to be done automatically and this is prone to errors.
- Then, one has to segment the bilingual text in sentences and align them sentence by sentence (with the exception of translation memories produced in computer-aided translation environments). For such large collections, one has to rely on *automatic* segmentation (using simple rules based on punctuation and format) and sentence alignment (using statistical methods), which may introduce errors.
- Machine translation training requires putting together a development set with a few thousands of sentence pairs —which are as representative as possible of those expected when implementing the system— to guide training and obtain performance figures.

Each sentence in either side of the parallel corpus has to be divided in smaller units called *tokens*. In Western languages it is not too difficult to divide most of a text in word tokens using punctuation, whitespace (blanks, tabulators, end-of-line markers), and some rules regarding contractions, etc. But there are languages which are written in *scriptio continua* and require the use of separate linguistic processors (for instance based on dictionaries) to divide them. More recently, sentences are automatically divided in *sub-lexical* tokens using statistical or neural methods, regardless of language. For instance, using algorithms such as byte-pair encoding (statistical; Sennrich, Haddow & Birch, 2016) or the newer *SentencePiece* approach (neural; Kudo & Richardson 2018).

back to top

¶ Machine translation evaluation

In order to assess a machine translation, it is necessary to take its purpose into account. Let us consider first an *adoption* scenario, where we want to decide whether we are going to use machine translation or which of several machine translation systems we are going to choose. Ideally, one

would try to devise an experiment which is representative of the actual task where we are expecting MT to help, and then measure the usefulness of machine translation output in that task. As said above, machine translation may be applied with assimilation or dissemination purposes. Machine translation evaluation has historically been a controversial issue and as a result, there are many approaches.

Let us first consider a *gisting* or *assimilation* situation where directions to perform a task (a cooking recipe, instructions to install and configure a mobile-phone application) are the *raw* MT output. A measure of success in the associated task is clearly an indicator of the usefulness of machine translation. Note, however, that, on the one hand, many uses of machine translation do not have a well-defined task associated to it (browsing the catalogue of an online store, reading sports news from another country, skimming through a forum on menopause) and that, on the other hand, setting up a representative experiment with enough text, subjects and situations may actually be quite costly. One may approximate this by using the typical methods used in second-language instruction: after reading a machine-translated text, subjects may complete reading comprehension questionnaires (Jones, Shen & Herzog 2009; Scarton & Specia 2016) with questions in the target language (also quite costly to prepare), or closure (*cloze*) tests where they are asked to complete professionally-translated target sentences where some words have been deleted (Forcada, Scarton, Specia *et al*. 2018; a bit cheaper if one already has professional translations of portions of the source text available).

In *dissemination* applications, machine translation will be used by professional translators, either as raw material that they will post-edit into a translation that is adequate for the purpose, or perhaps as a source of inspiration. An experiment is quite straightforward: after selecting a set of representative texts, a group of translators would be asked to translate them from scratch, or with the help of one or more machine translation systems, and then the *effort* it took them to produce the translation in each scenario would be measured. For example, one could measure *how long* it took them to translate a thousand words, or, conversely, measure their productivity, that is, *how many words* they translated per hour. If a machine translation system is helpful, translators will translate faster than without it; if machine translation system *A* is better than machine translation system *B*, the productivity with *A* will be larger than with *B*. Getting significant results may involve commissioning expensive translation tasks and several translators, which makes this evaluation also quite expensive.

Many approaches to evaluation try to mitigate the cost of task-based measurements. One way to do this is by collecting *subjective judgements*, usually regardless of task. A recent popular way of doing this is called *direct assessment* (Graham, Baldwin, Moffat *et al*. 2017): a crowd of target-language monolingual users are shown a professional translation of a sentence (say, in grey) and a machine translation thereof (say, in black) and they are asked to what extent they agree with the statement that “the black text adequately expresses the meaning of the grey text”, and they are shown a sliding button that they can place anywhere between 0% and 100%. Statistical processing (and sometimes filtering) of many such judgements leads to indicators that have been found to show reasonable correlation with actual measurements of usefulness.

But what if one wants to evaluate machine translation, not for adoption, but rather during *development*, repeatedly for different versions of a system? Then all of the above methods, which involve expensive setups using humans as subjects, are inapplicable. In particular, imagine a

statistical or neural machine translation being trained: one needs to measure its performance periodically during training, for instance, to decide when to stop training (so that the system does not *memorize* too deeply the training set). This calls for *automatic evaluation metrics*, which usually work as follows: a development set of, say, a couple thousand source sentences paired with *reference* professional translations are machine-translated with the system being developed, and the *similarity* between the machine-translated output and the reference (or references, if one can afford more than one professional translation) is automatically determined using a simple indicator (for example, computing how many edits would be necessary, as in *word-error rate* to produce the (closest) reference, or how many stretches of one, two, three and four words are present both in the MT output and the reference(s), as in the popular indicator called BLEU discussed above). The correlation of these indicators with actual usefulness has been proven to be limited, but they are anyway used massively in view of their convenience and even sometimes presented as actual indicators of translation quality when computed on independent *test sets*. Note also that they are not particularly *cheap* as they require the prior existence of a sizable set of reference translations.

In fact, there is a field of research called *machine translation quality estimation* that studies ways to *predict* the usefulness of machine translation when no reference professional translations are available, by simply examining the source and the machine-translated output (Specia & Shah 2018).

back to top

Research potential

Since the advent of corpus-based machine translation, the historical role of linguists and translation experts is seen as being accessory, as most of the research about how machine translation systems work is performed by researchers with a scientific or technological background (computer engineers, statisticians, data scientists, etc.).

But for those language pairs (and text genres) that cannot afford parallel corpora of the size needed to train corpus-based systems, rule-based machine translation may still be an alternative. Here, translators and linguists are still needed to create and manage the dictionaries and rule sets. As an example, Apertium (Forcada, Ginestí-Rosell, Nordfalk *et al*. 2011), an open, collaborative machine translation platform, focuses on languages with less resources and makes it possible to perform machine translation research.

Most corpus-based machine translation research is nowadays neural machine translation research and has a strong technological component. Here are some examples:

- new *architectures*, that is, neural network designs or training strategies;
- training machine translation systems with little or no parallel (bilingual) content (*monolingual* or *unsupervised* machine translation); generation of additional *synthetic* data: for instance, generation of additional source-language synthetic data from *natural* target-language data by training a reverse (target→source) machine translation system (*backtranslated* data; Sennrich, Haddow & Birch 2016);
- more efficient learning algorithms (important if one considers that current neural machine translation systems need to go several times through the whole set of training data before they actually start learning something);

- using more co-text (document context) and context (images, diagrams, etc.) as input to the processing of each sentence;
- constrained translation, so that machine-translated text satisfies the restrictions of a certain medium such as subtitles or menu items.

But even if one does not get into such technical aspects of machine translation systems, there is a wide variety of aspects about how machine translation is deployed in the real world which open avenues to research by translators and linguists.

In *assimilation* scenarios, that is, the most common ones in which ordinary people consume it raw:

- Evaluation methodologies based on judgements and or measurements and their ability to predict the actual usefulness or acceptability of machine-translated text in a variety of scenarios and tasks.
- Understanding how people make use of raw machine translation, either observing their behaviour (eye tracking, keyboard and mouse logging) or asking them to report (think-aloud protocols, post-task questionnaires); assessing the triggering of meta-cognitive strategies to deal with machine-translated content and its relation to the processing of text produced by non-native speakers, etc.; studying its acceptability and attitudes towards it.

In *dissemination* scenarios, that is, when professional translators use machine translation as help:

- Improving methodologies to measure possible reductions in translation effort, particularly during post-editing.
- Translation studies of the nature of text produced by machine translation systems (for instance, the ability of neural systems to perform operations such as reductions and expansions as professional translators do).
- Studying the ability of evaluation indicators (automatic and manual, obtained from judgements and from measurements, etc.) to predict post-editing effort and improvement of these indicators.
- Integration of one or more machine translation systems and other translation technologies in the computer-aided translation environments of freelancers and agencies; automatic selection of the most convenient technology (“technology brokering”) for each professional, for each translation job, or even for each segment; ergonomics, changes in translation workflows, acceptability and its effect on actual productivity, etc.

In both scenarios, linguistic and translational studies are needed of the types of errors produced by each kind of machine translation technology and their effect on the usefulness of their output.

Finally, the use of machine translation in *computer-aided language learning* or, particularly, in the actual training of translators has become very important in view of the fact that machine translation systems are currently readily available almost to any learner with an Internet connection.

back to top

References

Bahdanau, Dzmitry; Kyunghyun Cho & Yoshua Bengio. 2014. “Neural machine translation by jointly learning to align and translate.” @ arXiv preprint arXiv:1409.0473. [+info] [quod vide]

Belinkov, Yonatan & Yonatan Bisk. 2017. “Synthetic and natural noise both break neural machine translation.” @ *Proceedings of the 6th international conference on learning representations ICLR 2018*. [+info] [quod vide]

Brown, Peter F.; John Cocke; Stephen Andrew Della Pietra; Vincent Joseph Della Pietra; Frederick Jelinek; Robert Leroy Mercer & Paul S. Roossin. 1988. “A statistical approach to language translation”. @ *Proceedings of the 12th international conference on computational linguistics COLING-88,* Budapest, 1988, 71-76. https://doi.org/10.3115/991635.991651 [+info] [quod vide]

*Casacuberta Nolla, Francisco & Álvaro Peris Abril. 2017. “Traducció automàtica neuronal”. @ *Tradumàtica* 15, 66-74. https://doi.org/10.5565/rev/tradumatica.203 [+info] [quod vide]

Forcada, Mikel Lorenzo; Mireia Ginestí-Rosell; Jakob Nordfalk; Jim O’Regan; Sergio Ortiz Rojas; Juan Antonio Pérez Ortiz; Felipe Sánchez Martínez; Gema Ramírez Sánchez & Francis M. Tyers. 2011. “Apertium: A free/open-source platform for rule-based machine translation”. @ *Machine translation* 25/2, 127-144. https://doi.org/10.1007/s10590-011-9090-0 [+info]

*Forcada Zubizarreta, Mikel Lorenzo. 2017. “Making sense of neural machine translation”. @ *Translation spaces* 6/2, 291-309. https://doi.org/10.1075/ts.6.2.06for [+info] [quod vide]

Forcada Zubizarreta, Mikel Lorenzo; Carolina Scarton; Lucia Specia; Barry Haddow & Alexandra Birch. 2018. “Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting”. @ *Proceedings of the 3rd conference on machine translation WMT18,* Brussels, 2018, 192-203. https://doi.org/10.18653/v1/W18-6320 [+info] [quod vide]

Graham, Yvette; Tim Baldwin; Alistair Moffat & Justin Zobel. 2017. “Can machine translation systems be evaluated by the crowd alone”. @ *Natural language engineering* 23/1, 3-30. https://doi.org/10.1017/S1351324915000339 [+info] [quod vide]

*Hutchins, William John & Harry L. Somers. 1992. *An introduction to machine translation*. London: Academic Press. [+info]

Isabelle, Pierre; Marc Dymetman, George Foster; Jean-Marc Jutras; Elliott Macklovitch; François Perrault; Xiaobo Ren & Michel Simard. 1993. “Translation analysis and translation automation”. @ *Proceedings of the fifth international conference on theoretical and methodological issues in machine translation TMI’93: MT in the Next Generation* (Kyoto, 1993), 15-22. [+info] [quod vide]

Jia, Yanfang; Michael Carl & Xiangling Wang. 2019. “Post-editing neural machine translation versus phrase-based machine translation for English-Chinese”. @ *Machine translation* 33/1-2, 9-29. https://doi.org/10.1007/s10590-019-09229-6 [+info]

*Jones, Douglas; Wade Shen & Martha Herzog. 2009. “Machine translation for government applications”. @ *Lincoln laboratory journal* 18/1, 41-53. [+info] [quod vide]

Klubička, Filip; Antonio Toral Ruiz & Víctor Sánchez Cartagena. 2018. “Quantitative fine-grained human evaluation of machine translation systems: A case study on English to Croatian”. @ *Machine translation* 32/3, 195-215. https://doi.org/10.1007/s10590-018-9214-x [+info]

Koehn, Philipp & Rebecca Knowles. 2017. “Six challenges for neural machine translation”. @ “Proceedings of the first workshop on neural machine translation”, 28-39, Vancouver. https://doi.org/10.18653/v1/W17-3204 [+info] [quod vide]

Kudo, Taku & John Richardson. 2018. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing”. @ *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations*, 66-71. https://doi.org/10.18653/v1/D18-2012 [+info] [quod vide]

Nagao, Makoto; Toyoaki Nishida & Jun’Ichi Tsujii. 1984. “Dealing with incompleteness of linguistic knowledge in language translation–transfer and generation stage of Mu machine translation project”. @ *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, 420-427. https://doi.org/10.3115/980491.980577 [+info] [quod vide]

Och, Franz Josef. 2012. “Breaking down the language barrier—six years in”. Google official blog. [quod vide]

Kishore Papineni; Salim Roukos; Todd Ward & Wei-Jing Zhu. 2002. “BLEU: a method for automatic evaluation of machine translation”. @ *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311-318. https://doi.org/10.3115/1073083.1073135 [+info] [quod vide]

*Sager, Juan Carlos. 1994. *Language engineering and translation: Consequences of automation*. Amsterdam: John Benjamins. [+info]

Scarton, Carolina & Lucia Specia. 2016. “A reading comprehension corpus for machine translation evaluation”. @ Calzolari, Nicoletta; Khalid Choukri, Thierry Declerck, Sara Goggi, Mario Grobelnik, Bente Maegarrd, Joseph Mariani, Helène Mazo, Asunción Moreno Bilbao, Jan Odijk & Stelios Piperidis, eds. 2016. *Proceedings of the tenth international conference on language resources and evaluation LREC'16 (*Portorož), 3652-3658. [+info] [quod vide]

Sennrich, Rico; Barry Haddow & Alexandra Birch. 2016a. “Improving neural machine translation models with monolingual data”. @ *Proceedings of the 54th annual meeting of the Association for Computational Linguistics* 1: Long Papers, 86-96. https://doi.org/10.18653/v1/P16-1009 [+info] [quod vide]

Sennrich, Rico; Barry Haddow & Alexandra Birch. 2016b. “Neural machine translation of rare words with subword units”. @ *Proceedings of the 54th annual meeting of the Association for*

Computational Linguistics, 1715-1725. https://doi.org/10.18653/v1/P16-1162 [+info] [quod vide]

Sennrich, Rico & Biao Zhang. 2019. “Revisiting low-resource neural machine translation: A case study”. @ Korhonen, Anna; David Traum & Lluís Màrquez (eds.) 2019. *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019)*, 211-221. Firenze: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1021. [+info] [quod vide]

Shterionov, Dimitar; Riccardo Superbo; Pat Nagle; Laura Casanellas; Tony O’Dowd & Andy Way. 2018. “Human versus automatic quality evaluation of NMT and PBSMT”. @ *Machine Translation* 32/3, 217-235. https://doi.org/10.1007/s10590-018-9220-z [+info]

Specia, Lucia & Kashif Shah. 2018. “Machine translation quality estimation: Applications and future perspectives”. @ Moorkens, Joss; Sheila Castilho, Federico Gaspari & Stephen Doherty, eds. 2018. *Translation quality assessment. Machine translation: Technologies and applications* 1, 201-235. Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_10 [+info]

Sutskever, Ilya; Oriol Vinyals & Quoc Viet Le. 2014. “Sequence to sequence learning with neural networks.” @ Gharahmani, Zoubin; M. Welling & C. Cortes, eds. 2014. *Proceedings of advances in neural information processing systems 27: Annual conference on neural information processing systems,* Montreal, 3104-3112. [+info] [quod vide]

Tanaka, Hozumi; Shun Ishizaki; Akira Uehara & Hiroshi Uchida. 1989. “Research and development of cooperation project on a machine translation system for Japan and its neighboring countries”. @ *Proceedings of the MT Summit II* (Munich),16-18. [+info] [quod vide]

Vaswani, Ashish; Noam Shazeer; Niki Parmar; Jakob Uszkoreit; Llion Jones; Aidan N. Gomez; Łukasz Kaiser & Illia Polosukhin. 2017. “Attention is all you need.” @ *Proceedings of the 31st conference on neural information processing systems NIPS 2017,* Long Beach, 5998-6008. [+info] [quod vide]

Weaver, Warren. 1949. “Translation”. @ Locke, William N. & Andrew Donald Booth (eds.) 1955. *Machine translation of languages: Fourteen essays,* 15-23. Cambridge: MIT. [+info]

Credits

Mikel L. Forcada Zubizarreta

Prof. Mikel L. Forcada (1963) graduated in Science in 1986 and got his Ph.D. in Chemistry in 1991. Since 2002 he is full professor of Computer Languages and Systems at the Universitat d'Alacant. Prof. Forcada has been president of the European Association for Machine Translation (EAMT) since 2015. From the turn of the millennium on, Prof. Forcada's research interests have mainly focused on the field of translation technologies. He is the author of more than 70 articles and book chapters, of which about 40 are about translation technologies. In 2004, after heading several publicly- and privately-funded projects on machine translation he started the free/open-source machine translation platform Apertium (with more than 40 language pairs) and the free/open-source software project Bitextor (which crawls Internet sites to harvest parallel corpora). He is also co-founder of Prompsit Language Engineering (2006).

Licensed under the Creative Commons Attribution Non-commercial License 4.0

Asociación Ibérica de Estudios de Traducción e Interpretación (AIETI)