1. Overview of machine translation problems

Definition: Machine Translation (MT)

Machine translation (MT) is a sub-field of computational linguistics (CL) or natural language processing (NLP) that investigates the use of software to translate text or speech from one natural language to another. The core of MT itself is the automation of the full translation process, which is different with the related terms such as machine-aided human translation (MAHT), human-aided machine translation (HAMT) and computer-aided translation (CAT). [1]

A history of machine translation

1. Genesis (1933–1945)

The first automated translation systems were independently created in 1933, by George Artsrouni in France and Petr Troyanskii in the USSR. Unfortunately, neither really took hold in engineering or research circles, for different reasons. Artsrouni's system, which was a mechanically automated retrieval system that could function as a dictionary, generated a lot of interest in the French administration but could not come to fruition before the start of the Second World War. Troyanskii's system, which also started as an automated dictionary but grew to incorporate a memory as well as electronic components (those were at the time still mechanical computers !) was ignored by the Soviet scientific establishment.

Elsewhere in Europe, events that would prove (only slightly) more impactful were unfolding at the same time. From 1932 to 1933, the Polish Cipher Bureau — most notably Marian Rejewski, who the Marian NMT system is named after — broke the code of early German Enigma machines. During the Second World War itself, cryptography became a key topic and mobilized significant intellectual and financial resources. After the war ended, with the cold war rising, machine translation became a topic of interest to both superpowers' intelligence communities. A key problem, for example, was automatically translating scientific articles from the other side, as scientific output out-scaled the number of competent translators.

In this context, the 1949 Weaver memorandum on translation[4] was a landmark in the US, advocating that automated translation was becoming possible thanks to the newly created computer. It proposed several approaches, like storing the rules of language in the machine or learning statistical similarities between sentences, with even a mention of early efforts on perceptrons.

2. Rule-based MT (1949–1984)

Early rule-based MT (1949–1967)

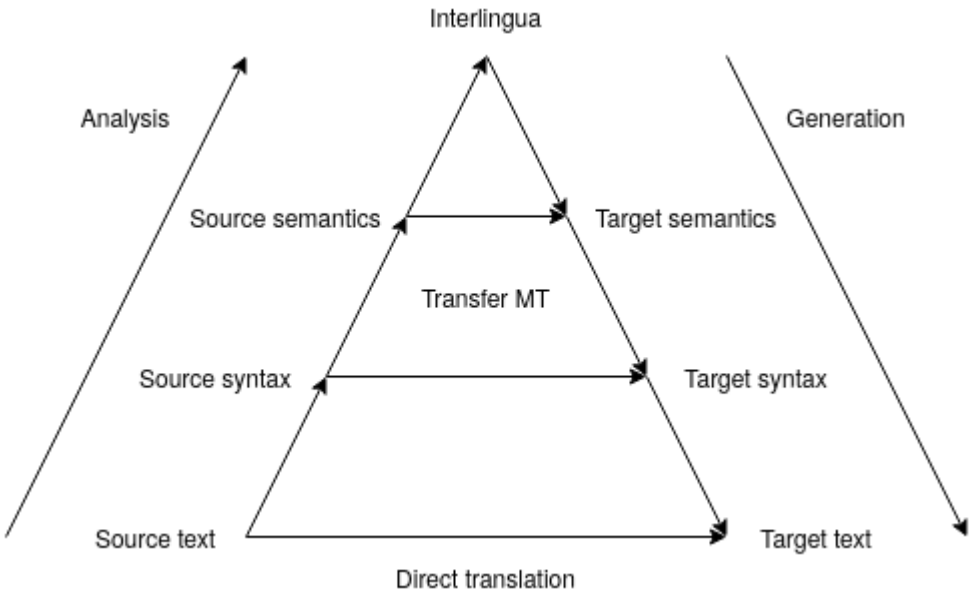
After the publication of the Weaver memorandum, research in machine translation began in earnest in the United States, mostly focused on translating Russian scientific articles into English. Translation systems of the time can generally be placed on a scale between empiric and linguistically-grounded approaches. For example, on the empiric end of the scale, research at the RAND corporation proceeded in cycles of translation and editing. First, start from a few basic rules, observing the result on a predetermined corpus of Russian texts. Then, revise the glossary and grammatical rules of the system, and repeat the cycle, in a sort of expectation-maximization algorithm performed by humans (perhaps an ancestor of graduate student descent ?) On the other hand, academic research, especially at MIT, focused on finding intermediate representations between the source and target sentences. Sufficiently expressive representations, it was hoped, could allow for general-purpose translation. Another goal was building an interlingua, e.g. a representation of semantic meaning independent of language; Noam Chomsky was introducing universal grammar at the same time.

A hybrid system to translate Russian technical documents was demonstrated in 1954 at Georgetown University. Deemed very impressive at the time, it spurred investment in the United States and seeded interest elsewhere, mostly in the Soviet Union and in Europe, where research concentrated on the theoretical approach. Systems at the time relied on the work of extensive teams of linguists: they translated human instructions into code rather than learning word correspondences on their own.

Knowledge-based MT (1967–1984)

The 1967 ALPAC report is generally held to have been the end of that first phase of machine translation hype, after it made the case that American research funding should be directed to machine-aided human translation rather than fully automated machine translation. After its publication, research funding dried up in the United States, leaving machine translation research efforts in Canada and Europe, and the Soviet Union.

One influential concept to understand the evolution of rule-based MT during that time is the Vauquois pyramid, reproduced here.



First, the system attempts to understand the source text (analysis) and to represent this understanding. Then, it produces text in the target language (generation) from this representation. This was christened knowledge-based MT: the goal was to have ever more complete and general representations, moving up the pyramid, as opposed to earlier rule-based systems' direct translation or only syntactically-informed translation. However, transfer machine translation, operating at a lower level, remained more effective and powered the systems of the time. Those were mostly domain-limited technical use cases like Canada's Météo system.

3. Data-driven MT (1984-present)

Example-based MT (1984–1993)

By the 1980s, computers had gotten a lot more powerful, especially in storage capacities. This allowed for larger databases of text, which remained yet to be systematically exploited. One early idea to do so was example-based MT, which was first proposed in 1984 in Japan. Example-based systems made the observation that beginner-level foreign language speakers rely on sentences they already know to produce new ones: an example between French and English is shown in the figure. Similarly, they relied on databases of known examples to produce new translations, querying the closest one. Although those ideas would eventually be subsumed in the broader framework of statistical MT, they were the first example of data-driven translation.

Statistical MT (1993–2013)

Underneath all this, a revolution was brewing, as a few outside developments came to fruition in the 90s. Statistical speech recognition started showing strong results on the back of advances in automata theory and hidden Markov models; computers became more powerful and accessible still; and high-quality, abundant datasets appeared, such as the Hansards accounts of the Canadian parliament.

Nevertheless, the statistical approach quickly proved fruitful, as IBM's models 1–5 became references in machine translation. Those were powered by the expectation maximization algorithm to learn both alignments between languages — which and how many words in the source and target correspond to each other — and a dictionary to translate after computing alignments. In a sense, they were direct descendants of the early RAND empirical approach: instead of being fed instructions by teams of linguists, the computer could learn all of the relationships from data on its own. By the 2010s, statistical methods had asserted their hegemony, as they powered virtually all of the internet-based translation services that comprise the bulk of translation use.

Neural MT (2013-present)

In his 1949 memorandum, Warren Weaver briefly touches upon early perceptron research as a promising avenue for machine translation. 60 years later, neural networks had made significant progress in other tasks, but had yet to be convincingly applied to translation. The first functional neural language models appeared in 2011, powered by recurrent neural networks. Translation could then be reformulated as a conditional language modeling task: instead of predicting the most likely next word, predicting the most likely next word conditioned on the source text. The first modern machine translation paper appeared within a few years, in 2013. It consisted of an encoder model that produced a representation of the input with a convolutional neural network and of a decoder model that generated text from that representation with a vanilla recurrent neural network (RNN).

At the time, neural MT was still underperforming compared to statistical MT, and it required two main developments from 2014 to eventually come out on top. First, vanilla RNNs were replaced with long short-term memory RNNs (LSTMs). Then, learnable attention mechanisms were re-purposed from their computer vision roots and added to LSTMs. By 2016, Google Translate had switched to neural MT. Transformer-based models, which do away with the recurrent network part and only use iterated attention modules, have become the norm in recent years as they scale better than LSTMs with compute time and available data. The Helsinki models we're releasing today all rely on this architecture. The power of transformers was quickly noticed outside of machine translation and, combined with pre-training, they now form the backbone of most modern NLP applications. [2]

Classification of machine translation errors

In order to find the errors in a translation, it is useful to have one or more reference translations in order to contrast the output of the MT system with a correct text. However, as it is well known in the machine translation community, there are several correct translations for a given source sentence, which poses a difficult problem for automatic evaluation and comparison of machine translation systems.

The classification of the errors of a machine translation system is by no means unambiguous. In work [3] is proposed the classification scheme, that has a hierarchical structure as shown in Figure 1. In the first level the errors are splitted in five big classes: "Missing Words", "Word Order", "Incorrect Words", "Unknown Words" and "Punctuation" errors.

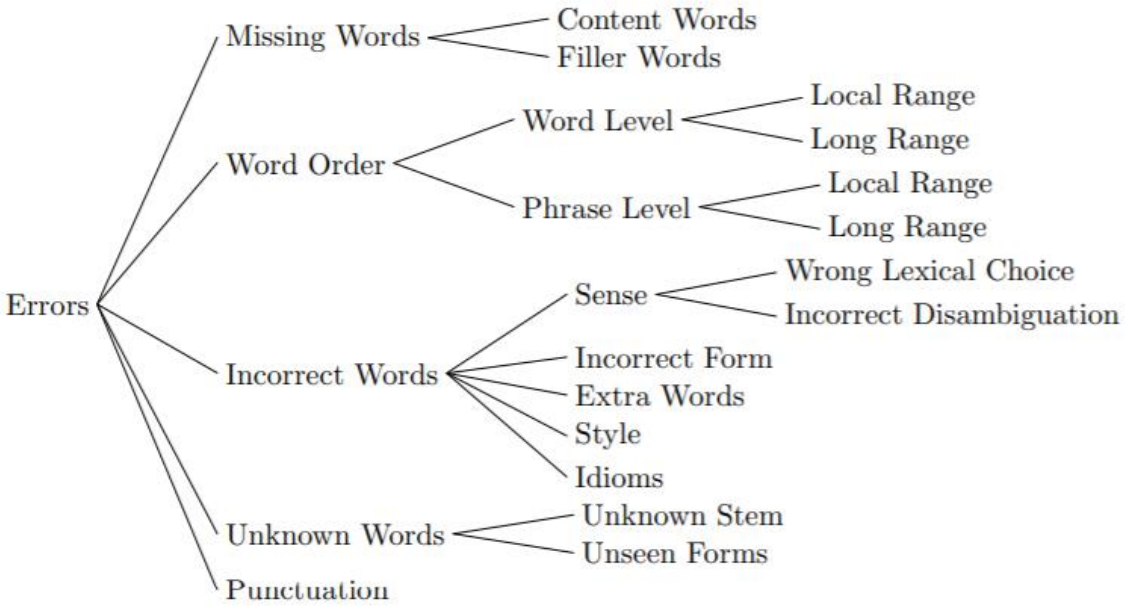


Figure 1. Classification of translation errors [3]

A "Missing Word" error is produced when some word in the generated sentence is missing. It is possible to distinguish two types of errors, when the missing words is essential for expressing the meaning of the sentence, and when the missing word is only necessary in order to form a grammatically correct sentence, but the meaning is preserved. Normally the first type of errors are caused by missing "main words" like nouns or verbs, but this not always the case, as for example a missing preposition can alter the meaning of the sentence significantly. This first type of errors is of course more important and should be addressed first. For each of these divisions one could further distinguish which lexical category ("Part of Speech") is missing, as different word types may have different treatments. For simplicity these subclasses are not included in Figure 1.

The next category concerns the word order of the generated sentence. Here it is possible to distinguish between word or phrase based reorderings, and within each of these categories between local or long range reorderings. In the case of word based reorderings, it is possible to generate a correct sentence by moving individual words, independently of each other, whereas when a phrase based reordering is needed, blocks of consecutive words should be moved together to form a right translation out of the generated hypothesis. The distinction between local or long range is difficult to define in absolute terms, but it tries to express the difference between having to reorder the words only in a local context (within the same syntactic chunk) or having to move the words into another chunk.

The widest category of error are the "Incorrect Words" errors. These are found when the system is unable to find the correct translation of a given word. Here it is distinguished five subcategories. In the first one, the incorrect word disrupts the meaning of the sentence. Here it is possible further to distinguish two additional subclasses, when the system chooses an incorrect translation and when the system was not able to disambiguate the correct meaning of a source word in a given context, although the distinction between them is certainly fuzzy.

The next subcategory within the "Incorrect Words" errors is caused when the system was not able to produce the correct form of a word, although the translation of the base form was correct. This is specially important for inflected languages, where the big variability of the open word classes poses a difficult problem for machine translation. How to further analyze the errors that fall into this category is very much dependent of the language pair we are considering. For example, for the Spanish language, being a highly inflected language, it is useful to distinguish between bad verb tenses and concordance problems between nouns and adjectives or articles.

Another class of errors is produced by extra words in the generated sentence. This kind of error was introduced mainly when investigating the translation of speech input, as artifacts of spoken language may produce additional words in the generated sentence.

The last two classes are less important. The first one ("Style Errors") concerns a bad choice of words when translating a sentence, but the meaning is preserved, although it can not be considered completely correct. A typical example is the repetition of a word in a near context. In this case a human translator would choose a synonym and avoid word repetition. The second one concerns idiomatic expressions that the system does not know and tries to translate as normal text. Normally these expressions can not be translated in this way, which causes some additional errors in the translation.

Unknown words are also a source of errors. Here it is possible further to distinguish between truly unknown words (or stems) and unseen forms of known stems.

A variation of this category has a special importance for the Chinese-English language pair. For the majority of European languages, or even languages that share the same alphabet, unknown proper names can be "translated" simply by copying the input word to the generated sentence, without further processing. Chinese characters, however, can not be translated into English by itself, and a conversion, sometimes guided by the pronunciation, is required.

Lastly there can also be punctuation errors, but, for the current machine translation output quality, these represent only minor disturbances for languages without fixed punctuation rules.

The error types so defined are not mutually exclusive. In fact it is not infrequent that one kind of error causes also another one to occur. So for example, a bad word translation can also cause a bad ordering of the words in the generated sentence. [3]

References

1. Chan Sin-wai. Routledge Encyclopedia of Translation Technology / 1st Edition. London: Routledge, Taylor & Francis Group, 2015. – 105 p.

2. Teven Le Scao. A brief history of machine translation paradigms. URL: https://medium.com/huggingface/a-brief-history-of-machine-translation-paradigms-d5c09d8a5b7e.

3. Vilar, D., Xu, J., d'Haro, L. F., & Ney, H. (2006). Error analysis of statistical machine translation output. Proceedings of the 5th international Conference of\n Language resources and Evaluation (LREC). Genoe, pp. 697-702.