Brief history of MT[edit]

The beginnings (1940s)[edit]

First computers[edit]

The obvious prerequisite for MT are computers. They started to appear in 1940s though it depends on what one considers as a computer.

Generation zero computers: Z1–3, Colossus, ABC, Mark I, Mark II. Then came first generation computers: ENIAC, MANIAC.

It is worth realizing that in 1947, RAM could store only 100 numbers and simple operation as summing took a fraction of a second.

Information boom[edit]

Roughly at the same time, the world started to produce and proadcast much more information than ever before. In 1922, regular BBC radio started to broadcast, in 1936 BBC TV followed.

Early believes[edit]

The view of translation was quite naive at that time. Some researchers[*citation needed*] saw translation as a repeated activity, ideal to be executed by computers. Why not: computers were successfully exploited in breaking war ciphers, they seemed suitable to crack language too.

The early boom (1950s)[edit]

In 1950 Warren Weaver sent a memorandum to 200 addressees in which he outlined some problems of MT:

•polysemy (ambiguity) is a common phenomenon,
•intersection of logic and language,
•connection with cryptography and
•universal properties of languages.
His view can be seen in the famous quote[*citation needed*]:

When I look at an article in Russian, I say: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.
　　—Warren Weaver, source??

The early interest in MT was pursued at several institutions: University of London (Andrew D. Booth), MIT, University of Washington, University of California, Harvard etc.

In 1952, the first public conference was held at MIT, two years after the first showcase of a working MT system followed.

Among the first topics were

•morphologic and syntactic analysis,
•meaning and knowledge representation and
•creating and working with electronic dictionaries.

At that time, Alan Turing was focusing on artificial intelligence, but he was not involved in MT research. [*citation needed*]

Georgetown experiment[edit]

The first working prototype of MT was publicly demonstrated in IBM, New York on January 1st, 1954. It was an example of using a computer for a non-numerical task.

The experiment showed translation of 60 sentences (probably carefully chosen) from Russian to English. The system contained a dictionary with 250 words and a rudimentary grammar with 6 rules.

Since the resulting translations were accurate, the demonstation provoked strong enthusiasm among researchers and gave rise to many projects in USA and SSSR.

Theoretical linguistics (Noam Chomsky) and artificial intelligence (Alan Turing) thrived.

To do:
connect paragraphs

But it became clear very soon that with increasing coverage of MT systems, their output quality suffered.

In 1950s, computers were used for generating art for the first time, love poems (1952)[*citation needed*].

The first PhD thesis on MT was defended (1954), the Journal of Machine Translation started to be published (1954), the First International MT conference was held in Londong (1956), Noam Chomsky wrote his famous Syntactic Structures (1957), the first book about MT (an introduction) was published in Paris (1959).

Besides USA, MT was on radar in USSR, Japan.

The disappointment (1960s)[edit]

In 1959, he famous critic of MT Yehoshua Bar-Hillel wrote about unsatisfactory status of MT. He claimed that computers are not capable of resolving one important phenomenon in language: lexical ambiguity. He coined term *fully automated high-quality translation* (FAHQT) and claimed it to be unreachable.

His famous example where computers are supposed to struggle with is *Little John was looking for his toy box. Finally, he found it. The boxwas in the pen. John was very happy.* where *pen* is clearly used in a meaning different from the common meaning of the word. But for computers to disambiguate it, they would need to have knowledge of the world and that boxes are not usually inside a writing tools.

Probably connected to his and others' criticism, money supply for MT projects began to shrink.

MT in USSR was focused on translation of English scientific papers (abstracts) at that time.

In 1962, Association for MT was founded in USA. Around that time, Peter Toma, leaves Georgetown[*citation needed*] to start developing AUTOTRAN to became later Systran, one of the most successful MT software in the following decades.

ALPAC report[edit]

What was lethal for MT research was so called ALPAC report (Automatic Language Processing Advisory Commitee) prepared in 1966 on behalf of U.S. National Academy of Science.

The commitee prepared analyses and evaluation of MT quality and usability and recommended to reduce expenditures for MT research to U.S. government[*citation needed*]. It claimed that researched underestimaged the complexity of natural language understanding and subsequently, the report had profound negitive impact on MT field.

Despite, MT research was unaffected in Europe, USSR and Japan, but USA took 15 years after the drop in financial support to catch up with the rest of the world in the pursuit of MT.

MT research in Canada[edit]

At that time in Canada, at Université de Montréal, MT research hit a few successes. Researchers developed a few working prototypes of MT systems, namely TAUM-73, TAUM-METEO. These were the first systems to incorporate proper analysis of the source language and synthesis of the target language.

English-French (and French-English) pairs were the topic. One project TAUM Aviation focusing on translation of technical manuals were cancelled.[*citation needed*]

Later, METEO system was used for weather forecast translation between 1981 and 2001. It was developed by John Chandiou.

Systran[edit]

At the end of 1960s, Systran, on of the oldest companies developing MT systems was founded. The software of the same name became very popular and was a basis for Yahoo Babelfish later. It was used also by Google until 2007.

It started as rule-based system but since 2010, Systran is a hybrid system incorporating also statistical methods.

The renaissance (1970s and 1980s)[edit]

The first Soviet MT: AMPAR for translation from English to Russian. Since 1976, Systran has been used as an official MT system at European Economic Community. Xerox started using Systran. A project proposes using Esperanto as interlingua but was turned down.

Rule-based systems using interlingua started to appear. In 1980, project Rosetta has started using logical formulae as interlingua.

First data-driven (example-based MT) has appeared. MT systems were good enough to generate a revenue and were commercionalized. Trados, the first companty to develop CAT tool was founded in Stuttgard in 1984.

EU project EUROTRA has started.

To appreciate the context: in 1983, IBM introduces its 8-bit ASCII code and in 1987, Unicode project set out. Word Wide Web proposal saw the light of the day in 1989.

The rise of SMT (1990s)[edit]

IBM has contributed the world with another gem: stastical MT was born in early 1990s. SDL (current CAT market leader) was founded in UK in 1992 to later acquire Trados. Verbmobil project which gave born to some MT methods has been running between 1992 and 1999.

AltaVista's Babelfish logged 500,000 requests per day in 1997[*citation needed*]. The first online commercial MT service iTranslator arose.

During this decade, rule-based systems still dominated the field.

The new millenium[edit]

Statistical methods took over the field and first hybrid systems started to appear. New translation pairs (languages) are added to repertoire of MT systems as new data are gathered and digitized.

NIST launches first round of MT system benchmarking in 2001.

EuroMatrix, a large scale EC funded project has started in2006 and Moses, highly successful (for being open source) statistical MT engine is born in a year later.

To do:
add a prognosis graph from that time

The computing power is steadily growing and Google is one of the leaders. For instance, using a new big-data technique MapReduce, researchers managed to sort trillion of 100 byte records on 4,000 computers and on 48,000 harddrives over just 6 hours. The computing power allows processing billions of words in blink of an eye and thanks to projects like Moses, MT became available to everyone.

To do:
new data structures, suffix arrays and trees

To do:
Google N-grams

At the same time, new parallel data are developed. There are special events (LREC) where new resources for languages from all over the world are regularly presented. Also under-resourced languages are treated and in general, MT quality improves slowly but steadily.

2010 and beyond[edit]

There are various interests in source and target languages. Under-resourced language are often neglected. In EU, the focus is on all official languages (English, Bulgarian, Czech, Croatian, Danish, Estonian, Finnish, French, Irish, Italian, Lithuanian, Latvian, Hungarian, Maltese, German, Dutch, Polish, Portugese, Romanian, Greek, Slovak, Slovene, Spanish a Swedish). English speaking countries consider it as the main target language. Global markets and companies on the contrary want to bring their products to people around the world so the target languages are those of developed countries and the source language is usually English.

In some sense, some languages (and language pairs) are larger (usually better covered in digital media) than others. They also achieve better translation quality (English-to-Spanish, English-to-French).

Statistical methods are enriched with both linguistically-based techniques (syntax, semantics) and neural language models to achieve state-of-the-art results.

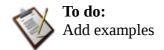Google Translate is considered a gold standard.

Morphologically-rich language are usually harder to translate to.

English-to-XXX and XXX-to-English language pairs prevails heavily.

Template:TODO:citations and more fluent text

Since 2015, statistical methods are slowly replaced by neural network techniques in on top of the leaderboard.

Machine translation is now available everywhere thanks to smartphones. It is used for *gisting*, for instant translation of web pages (probably the most common use of MT[*citation needed*]), to speed up human translation within CAT tools, cross-language information retrieval (CLIR), for instant messaging and other e-communication on mobile deviced, for translating speech-to-speech and even image-to-image.

To do:
Add examples

Further reading on history of MT[edit]

Online resources[edit]

- Statistical MT
- MT Archive