Machine Translation

Past, Present and Future

Daniel Stein
Universität Hamburg, Hamburger Zentrum für Sprachkorpora
ds@daniel-stein.com

The attempt to translate meaning from one language to another by formal means traces back to the philosophical schools of secret and universal languages as they were originated by Ramon Llull or Johann Joachim Becher. Until today, machine translation (MT) is known as the crowning discipline of natural language processing. Due to current MT approaches, the time needed to develop new systems with similar power to the older ones, has decreased enormously. In this article, the history of MT, the difference to computer aided translation, the current approaches and future perspectives are discussed.

1 History of Machine Translation

Although the first systems of MT were built on the first computers in the years right after World War II, the history of MT does not begin, as often stated, in the 1940s, but some hundred years ago. In order to judge current developments in MT properly, it is important to understand the historical development.

1.1 Universal and secret languages

Most likely the first thoughts on MT emerged out of two philosophical schools that dealt with the nature of language and resulted in similar insights, although stemming from different directions. The first was directed at creating secret languages and codes in order to communicate in secrecy. The second evolved from the ideal of a universal language which would allow communication without borders in the times after Babylonian language confusion.

Noteworthy proponents of the movement of universal languages were the Catalan philosopher Ramon Llull (1243 to ca. 1316, often remarked by the latinized version of his name Raimundus Lullus) and the German philosopher and mathematician Gottfried Wilhelm Leibnitz (1646-1716). Llull developed a theory of logic that allowed objectifying the reasoning on God and the world by means of a formal language. His ideas were later used by Leibnitz is his theory of monades (1696), in which he tries to develop a set of the smallest units of meaning ("termini primi") to compose all thinkable thoughts. Other attempts were started by a precise determination of the inventory of the world in form of a taxonomy in order to find all sayable things (cf. Gardt, 1999).

In the long history of secret languages and hidden codes, the German physician and alchemist Johann Joachim Becher developed a system in 1661 that is especially interesting in the frame of MT as it appears to be very similar to the first technical approaches in the late 1940s. It is called „Character pro notitia linguarum universal" and offers „Eine geheimschriftliche Erfindung, bisher unerhört, womit jeder beim Lesen in seiner eigenen Sprache verschiedene, ja sogar alle Sprachen, durch eintägiges Einarbeiten erklären und verstehen kann" (Becher, 1962) („A secret and currently unknown language invention that enables everyone to explain and

Translation: Computation, Corpora, Cognition. Special Issue on Language Technologies for a Multilingual Europe, ed. by Georg Rehm, Felix Sasaki, Daniel Stein, and Andreas Witt. Volume 3, Number 1. June 2013.
ISSN 2193-6986

understand different and even all languages after a one day orientation by reading in their own language."). The approach is based on dictionaries that are related to each other by number codes, which is more or less identical to what was then called "mechanical translation". But despite the obvious relationship to Becher, the influence of the school of universal languages to MT was small. In contrast, with the uprising of the science of secret languages, cryptology continuously gained in importance.

In World War II, the decipherment of the German ENIGMA code is regarded as a crucial point. The British team around Alan Turing, located in Bletchley Park, was responsible for this urgent project and achieved the breaking of the code by means of statistical methods that were processed on computing machines. Without their knowledge, the scientists laid the foundations for practical MT.

Considering the experiences of Bletchley Park, the exchange of letters by Warren Weaver and Andrew Booth is regarded as the date of birth of MT. Weaver wrote

> „[...] it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the ‚Chinese Code'. If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation? " (Weaver, 1955).

1.2 Evolution of MT

Although the mathematical methods prove useful for cryptology, they turned out to be inadequate for more challenging and complex translation tasks. Accordingly, the subsequent systems that were developed were based on dictionaries and selectively used syntactic operations (this was the time, when J.J. Bechers article on the universal language was republished with the subtitle "A programming approach from the year 1661"). From today's point of view, these approaches were remarkably naïve.

The constant threat of the Cold War caused euphoria in government and military circles regarding the anticipated possibilities of MT. Until 1966, great amounts of money were spent in order to develop MT systems, mostly for the English-Russian language constellation. But with the publication of the famous Automatic Language Processing Advisory Committee (ALPAC)-report, on behalf of the US administration, the CIA and the National Science Foundation, the funding decreased immediately, due to the prognosis that MT is neither useful nor does it seem to provide any considerable advance or meaningful progress (cf. Hutchings, 1996). With the exception of some practically oriented teams in Europe and the USA, research and development of MT expired.

In order to react to the results of the ALPAC report and the reduction of resources, the discourse became more classically scientific and tried to integrate linguistic knowledge on a broader basis, above all, semantic analysis. The results that have been achieved by these approaches were promising and so, in the middle of the 1970s and in the course of the rapid development of technology and the introduction of the first personal computers, MT research was revitalized and headed to a continuously increasing popularity from the beginning of the 1980s.

Ten years later, however, in the middle of a syntax- and semantic-based MT system era, an IBM research group led by Peter F. Brown published an article (Brown, 1988), which suggested the return to statistical methods for a new MT system. Technological advances and the increased availability of language resources as machine readable parallel corpora had changed the underlying conditions significantly. Thus, the results seemed very promising, especially regarding the extremely condensed time that would be necessary in order to create a state of the art MT system. Due of this, the majority of MT research switched to statistics-based MT in the following years, as it was possible to create comparable MT systems without ten years of work and the expertise of a team of linguists. A few days of time and a very good bilingual corpus ("bitext") was enough for a prototype.

Since then there has been a lot of development in statistical MT (SMT). While the first systems were only trained to compare the probabilities of co-occurring words, later approaches tried to use groups of words instead, n-grams of different sizes. But pure SMT seems to hit its frontiers as there are several shortcomings and problems confusingly similar to those of rule-based MT systems and it seems to be impossible to solve them by just using bigger corpora. Hence, the focus in MT research adapts again. Actually, various trends are discussed simultaneously, e.g. SMT for lesser resourced languages or example-based methods. Since the middle of the 2000s hybrid approaches that combine SMT with linguistic knowledge were often seen ("context-based" or "knowledge-based MT") and a new trend of the last years is to use corpora that are not parallel but at least comparable. One of the most recent interesting developments links interestingly back to the beginning of MT, ie. as well to the famous memorandum by Warren Weaver as to the creators of secret languages mentioned above: After the success of Kevin Knight, Beáta Megyesi and Christiane Schaeferin in deciphering the Copiale codex (cf. Knight, 2012), a German 18th century text with freemasonry background, the use of decipherment strategies in MT undergoes a renaissance (cf. Dou, 2012).

2 Machine Translation vs. Computer Aided Translation

An important distinction exists between MT and computer aided translation (CAT). While the (today not that often announced) goal of MT is a so called FAHQT (fully automatic high quality translation), in CAT, tools and methods that assist human translators in the translation process are to be researched and developed. A well-known and widely used example of CAT is the use of translation-memory systems (TMS). A TMS combines a user friendly translator front end with a database that saves all translations that have been done in a certain project (the translation memory), as well as a component that analyzes the units that are to be translated for similarities with the ones in the translation memory. If a similarity beyond a threshold is found, the system enables the translator to modify the translation or, in cases of 100% similarity, just pastes it. Certainly, this kind of tool turned out to be impressively useful for translators in the domains technical documentation or software localization. But of course CAT is not designed for the translation of literary texts – the localization of video games seems to be situated in between these poles, as the texts are often combinations of technical and literary writing. Further components of a TMS may involve MT for units with lower similarities, the automatic transliteration of numbers, dates and other placeable elements, or the implementation of user made dictionaries for terminology management (cf. Seewald-Heeg, 2002).

3 Typology

As described above, in the course of the years several approaches to the task of MT have evolved. Today, the most important ones are rule-based MT (RBMT) and SMT. Although they sometimes may still be understood to be concurring approaches, the general view seems to be that both statistical, as well as linguistic approaches may serve as tools in the machine translation toolkit that may be freely combined in order to improve results. In the next paragraphs the two main representatives and the most common alternative approaches will be discussed (cf. Jekat, 2010).

3.1 Rule-based

RBMT today is often considered the "classical approach" and is still regularly used in commercial solutions, although, with the withdrawal of Systrans "Babelfish", the most popular representative of this approach disappeared. The results of RBMT systems range from useful to hilarious, depending on the concrete text and its complexity with regard to common problems as resolution of anaphors or lexical ambiguities, as well as the language pair, even the translation direction, and if the text is of a certain domain or contains special terminology (which is, given a prepared system, generally easier to process than general language).

A loose distinction between three levels of complexity of MT is common and the results, as well as the expenses, differ significantly: Direct, transfer and interlingual translation.

The majority of RBMT systems is of the transfer level and process text in three successive steps:

1. Analysis
2. Transfer
3. Generation/Synthesis

3.1.1 Direct Translation

MT systems that are based on direct translation simply replace words on a word by word basis and only rely on a parallel dictionary – so they neither do analysis nor transfer or generation. Often, positional changes are also included in order to follow the word order of the target language. This approach is only of interest for a few possible application scenarios, but in general it may rather be considered a theoretical measure to illustrate revenues and expenses. Historically, however, this is how the first systems were designed.

3.1.2 Transfer

Transfer translations define a set of rules ranging from morphology and syntax to semantics and context. Regarding the complexity of these rules there are no limits and tens of thousands of rules, combinations and exceptions may be coded. But in practice there seems to exist a point where higher complexity does not indicate better results anymore. Instead, internal conflicts and contradicting rules produce arbitrary new errors. The majority of the existing RBMT systems can be considered a part of the transfer level.

3.1.3 Interlingua

The third level of complexity, Interlingua, is based on the utopia of a neutral language that would be able to represent all meaningful information of every utterance in every language. On the scale presented above for Interlingua systems there is no need to transfer from one language to another as they use a common meta language that is able to express the meaning of both in an unambiguous way. This universal language ("Interlingua") would be the target language for every translation in the first place and in the next step it would be the source for the composition of meaning in the target language. Unfortunately, such a language has not been found until today, although several attempts have been made, beginning with the thoughts of Lull and Leibnitz, over to "semantic primitive" as in the work of Anna Wierzbicka (Wierzbicka, 1996) and later on in experiments using constructed languages as Esperanto or Lojban. Although this approach is considered optimal, it should be noted that even a perfect interlingua could make things potentially even more complicated due to its abstraction. (Nicholas, 1996).

3.2 Statistics-based

As mentioned above, the new rise of SMT began in 1988 when IBM researcher Peter Brown presented a new approach of MT that was solely based on statistic measures (Brown, 1988) on the second TMI conference of the Carnegie Mellon University. The basic principle is that every translation decision is made due to conditional probabilities, i.e. the probability that an event will occur, when another event is known to occur (or already occurred). As a resource, instead of complex rule sets, large parallel corpora are needed.

3.2.1 Functioning

From a formal point of view, SMT works like this: In order to translate the arbitrary French sentence f to English, one can consider all possible and impossible English sentences e as potential translations of f. But some are more probable translations than others. p(e | f) is the probability that e is a valid translation of f. Philosophically spoken, we assume that the speaker of f in the initially thought e and then internally translated e to f before he uttered it. This construction is used to define the goal of SMT: Find the original sentence e which is the most probable translation. Please note, that this assumption is similar to Weavers remark about understanding Chinese as English that is encrypted with the Chinese code.

This ideal situation is confronted with the impossibility to access all sentences of a language. Therefore, SMT works with approximations, so-called models. A bilingual aligned corpus defines the translation model that represents all possible translations between two languages, i.e. the larger the translation model, the better the expected results. Generally, every word is considered a potential translation of all the others, but the probability is the highest for those that they are aligned to.

An additional monolingual corpus of the target language is defined as the language model. It represents all valid sentences (or better words or word sequences which is a more operable abstraction) of a language. A search algorithm then determines the sentence, by finding the highest product of the values sentence validity (language model), word translation and word order (translation model). The result is the most probable translation.

The concrete probabilities used by the computer are estimated with Bayes' Theorem.

$$1)\quad Pr(e \mid f) = Pr(e) * Pr(f \mid e) \,/\, Pr(f)$$

This sentence can be reduced to the search of the maximum value of the terms Pr(e) ("Probability that e has been said by someone") and Pr(f | e) ("Probability that someone would translate e to f")

$$2)\quad \hat{e} = \underset{e}{argmax}\ Pr(e) * Pr(f \mid e)$$

Brown used the English-French parallel "Hansard" corpus that consists of protocols from the Canadian parliament. Hence, this is where the example languages e and f derive from.

In the beginning SMT was mainly based on Browns original model, i.e. the target language utterances were derived according to Shannon's Information Theorem out of a noisy channel translation model. But since 2002, when Och and Ney proposed a system in which the noisy channel was replaced by a discriminative log linear model (cf. Och and Ney, 2002), this approach has established itself as de facto standard as it allows to add additional features next to the language and translation model (cf. Chiang, 2012).

3.2.2 SMT types

The analysis of whole sentences makes little sense: How often is it possible to translate the exact same sentence that is already present in the translation model? As long as an SMT system does not have a corpus that indeed contains all (or at least almost all) possible sentences of a language, it is useful to reduce the considered unit. Therefore, there is the differentiation between word-based and phrase-based SMT.

3.2.2.1 Word-based

Word-based is the original approach and analyzes data on the level of simple lexical units. This means that a word in the source language has to correspond to one word in the target language. But unfortunately, it is quite often the case that a word has to be translated with more than one simple lexical unit, e.g. the English verb "slap" has to be translated to Spanish "dar una bofetada". This is a construction that is possible to model with word-based SMT but the opposite direction, i.e. to come from "dar una bofetada" to "slap" is impossible. And as a matter of fact, so called multi-word expressions (MWE) are by far the biggest part of the lexicon of any natural language - but that does not answer the question as to which concepts are expressed through MWE in which language.

A related problem is that words may belong together although there are other words between them (e.g. so called separable verbs in German). It is impossible to translate them correctly when the relation is not considered, as the word "ab" which is derived as the construction "reiste … ab" from the verb "abreisen" in the German sentence

"Ich	*reiste*	*schon*	*nach*	*vierzehn*	*Tagen*	*wieder*	*ab"*
"I	*checked*	*yet*	*after*	*fourteen*	*days*	*again*	*out"*

("I left after only fourteen days").

This is especially problematic for languages with a strongly deviating syntax, e.g. in regard to the position of the finite verb.

3.2.2.2 Phrase-based

Phrase-based SMT is an approach that tries to solve the problems mentioned above and is common for actual SMT systems. But the term phrase does not indicate that the systems are able to identify, analyze or separate linguistically motivated phrases, e.g. noun phrases that may be composed of (complex) determiners and (compound) nouns. The word phrase rather refers to sequences of succeeding words (n-grams) that were derived from data.

The use of n-gram-based phrases in SMT allows to address some of the shortcomings of word-based SMT, i.e. it is possible to translate one word with many and vice versa. Additionally, the broadened context enables better disambiguation algorithms. E.g. it is impossible to decide whether „pretty" is to be translated as "schön" or as "ziemlich" without knowing if the next word is "flower" or "much" and thus cannot be translated properly by word-based SMT but by phrase-based. Depending on the size of the word sequences (i.e. the n-gram window) it might also be possible to address problems regarding different word order or other syntactical phenomena. Hierarchical phrase-based SMT, also known as syntax-based SMT is an advanced approach that allows the use of tree-based syntax data in the phrase-model (cf. Koehn, 2010).

3.2.3 Pro and Contra

The great advantage of SMT is the possibility to create a working MT system without any knowledge of the source or target language and their special features. As a matter of fact, the translation quality of an unadapted, i.e. pure SMT system, is generally weak (mainly depending on the used corpora). However, they are still comparable to RBMT systems and – in the view of decades of language rule modeling – a ground breakingly fast approach to proportionate robust MT systems, both in terms of time and money. So MT becomes tangible for languages that do not dispose sufficient man power to create a work intensive RBMT system, but for which sufficient resources (i.e. bitexts) exist (which for instance is the case for most of the official languages of the European Union).

In terms of translation quality it can be stated that RBMT and SMT are similarly error-prone, but have some principal differences regarding the error type. Thus, one can easily observe that RBMT systems produce better sentences in terms of word order, syntax and coherence, but SMT systems produce better translations in terms of word choice, disambiguation etc. Multi-word expressions or proverbs may also be translated without the effort of enumerating them beforehand (but only if they are present in sufficient number in the corpora to be identified statistically). Hence, one can state the basic philosophy of SMT as "bigger corpora means better results".

However, the disadvantages of SMT are closely related to the advantages. Due to the fact that every translation is produced on opaque calculation processes on gigantic text quantities, it is nearly impossible to identify the potential causes of failures. Therefore, manual correction efforts for systematic errors are laborious and may often result in just adding better examples manually in order to change the statistical measure of a misinterpretation. Additionally, it is necessary to mention that for certain language pairs immense problems may arise, especially if they consist of a fundamentally different structure in terms of flexion, word order, use of pronouns, number and kind of temporal forms, etc. For instance, the translation of German separable verbs often results in a missing finite verb which is essential to the sentences' meaning. According to this, it becomes evident that the best translations are obtained when the SMT is created, trained and used for a special domain. The simple philosophy of SMT mentioned above also includes a disadvantage: If bigger corpora mean better results it means that a corpus can be too small but never big enough.

3.2.4 Parallel, comparable and low resource corpora

Another access point to improve SMT are the requirements of language data for training and translation purposes. As described above the first approaches obligated the use of large sized parallel corpora, i.e. corpora in which every sentence is aligned to a translated version of itself – for every language pair. Nevertheless, large parallel corpora exist for many language pairs, the corpora generally consist of parliament protocols and translations or something similar, e.g. from the European Parliament (EuroParl) or the already mentioned Canadian Hansard Corpus. Therefore, the use of political and economic terminology is highly overrepresented compared to corpora with standard language. But the creation of parallel corpora for other language domains constitutes a complex and laborious task even for languages with many speakers, but it is, as a third shortcoming, very hard to manage for lesser resourced languages where the corpus not only needs to be compiled or translated, but simply written in first place. Due to this, a new approach is working with so-called comparable corpora, i.e. corpora that are not parallel but related to each other as e.g.

Wikipedia articles. Changes in the processing of the translation model in another approach resulted in the use of larger monolingual corpora and smaller parallel ones. Bridging through similar, but higher resourced languages, e.g. in the case of using Spanish as a bridge to translate English to Catalan, is also a way to deal with this.

3.3 Hybrid Systems

Hybrid approaches try to combine the advantages of several systems. This is especially the case for SMT: There are numerous articles describing the combination of SMT with syntactic preprocessing, semantic disambiguation or similar applications. Often the combination of approaches broadens the scope of research possibilities for unfavorable language pairs, may it be due to strong divergence in terms of flection and word order, or due to the fact that one or both of the languages in question are lower resourced ones. But although there was quite a lot of effort in this research direction and most of the approaches have indeed improved the translation quality (at least a bit) there does not seem to be a breakthrough in sight.

3.4 Perspectives

MT research has experienced some highs and lows in its history. Although a FAHQT is no longer the often proclaimed single goal of MT, the last years have been characterized by increasing MT research funding and diversification of the topics of interest. This may be due to the fact that freely available state of the art MT systems, e.g. by Google or Microsoft, have shown the high usability of MT, even though the systems are not perfect.

The combination of approaches to hybrid systems, e.g. the use of linguistic information and statistical data has become one of the most researched fields in MT over the last decade. The integration of syntax into phrase based SMT systems reanimates the search for the right kind of linguistic data (e.g. multi-word expressions, linguistic motivated phrases, etc.) to be integrated as well as the kind of preprocessing that is needed for it (syntax trees, support of variables, etc.). This way, type and state of the resources are rated more appropriately than in the beginning of SMT research. This is also relevant in the context of domain adaption, i.e. the identification of data that are necessary to represent a closed domain and the expansion to new fields as it turns out that the automatic translation of specialized domains is more reliable.

Recently there has been a shift from the "traditional" language pairs in MT, namely English, Russian, German, French, Spanish and in the last years also Chinese and Japanese to the lesser resourced ones. Especially the expansion of the European Union has been a starting point for growing research in this area as there are speakers of 23 languages that demand participation at eye level and in their mother tongue for a daily growing amount of texts and offers as e.g. ecommerce. And the automatic translation between language pairs that do not involve English also reinforces attempts to deal with complex problems of morphology.

Another topic of still growing interest is the automatic evaluation of translations – whether it is with the focus on metrics that go behind the currently standard metric BLEU (e.g. by using syntax information) or it is with the focus on reusing good translations as additional training data.

4 References

Becher, Johann Joachim. 1962. Zur mechanischen Sprachübersetzung. Ein Programmierversuch aus dem Jahre 1661. Allgemeine Verschlüsselung der Sprachen. Stuttgart: Kohlhammer.

Brown, Peter F. et al. 1988. A statistical approach to french/english translation. In *Recherche d'Information Assistee par Ordinateur*, 810-829.

Chiang, David. 2012. Hope and Fear for Discriminative Training of Statistical Machine Translation. In *Journal of Machine Learning Research,* 13, 1159-1187. Washington: Sparc.

Dou, Qing and Knight, Kevin. 2012. Large Scale Decipherment for Out-of-Domain Machine Translation. In *Proceedings of the EMNLP*.

Gardt, Andreas. 1999. Geschichte der Sprachwissenschaft in Deutschland. Vom Mittelalter bis ins 20. Jahrhundert. Berlin: de Gruyter.

Hutchings, John. 1996. ALPAC: The (in)famous report. In *MT News International. Newsletter of the International Association for Machine Translation*, 14. International Association for Machine Translation.

Knight, Kevin et al. 2012. The Secrets of the Copiale Cipher. In *Journal for Research into Freemasonry and Fraternalism*, ed. Cécile Révauger, Jeffrey Tyssens and Andrew Pink. Sheffield: Equinox.

Koehn, Philipp. 2010. Statistical Machine Translation. New York: Cambridge University Press.

Nicholas, Nick. 1996. Lojban as a Machine Translation Interlingua in the Pacific. In *Fourth Pacific Rim International Conference on Artificial Intelligence: Workshop on 'Future Issues for Multilingual Text Processing'*, 31-39. Cairns.

Och, Franz-Josef and Ney, Hermann. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *ACL 2002: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 295-302. Philadelphia: Association for Computational Linguistics Stroudsburg.

Seewald-Heeg, Uta. 2002. CAT - Intelligente Werkzeuge anstelle unzulänglicher Automaten. In *Computerlinguistik: Was geht, was kommt? Computational Lingustics: Achievements and Perspectives*, ed. Gerd Willée, Bernhard Schröder and Hans-Christian Schmitz, 263-267. St. Augustin: Gardez! Verlag.

Weaver, Warren. 1995. Translation. In *Machine Translation of Languages. Fourteen Essays*. Ed William N. Locke and Andrew D. Booth. New York: Technology Press of MIT.

Wierzbicka, Anna. 1996. Semantics: Primes and universals. Oxford: Oxford University Press.