Chapter 25
Future of Machine Translation: Musings on Weaver's memo

Alan K. Melby

Abstract

Machine translation (MT) has seen fluctuating periods of growth and attention since the 1950s, with the primary leading paradigms shifting from rule-based (RBMT) to statistical (SMT), and now neural (NMT). This chapter aims to analyse the seeds of these paradigms through a 1949 memorandum by Warren Weaver that presents five hypothetical approaches to MT. They are: word-for-word translation; disambiguation using micro context (i.e. co-text); an approach based on formal logic; a cryptography approach to translation as decoding; and translation based on invariants (universal symbolic representation of meaning). This chapter shows how some developments external to the translation industry have enabled the realization of Weaver's vision and discusses various connections among the seeds, the enabling developments and the three machine-translation paradigms. The chapter concludes with a discussion of possible directions for the future of MT and its connection with human translation.

Keywords: machine translation (MT), rule-based MT (RBMT), statistical MT (SMT), neural MT (NMT), neural networks, deep learning, symbolic representations, linguistic levels, understanding

Introduction

For the purposes of this chapter, machine translation (MT) is defined as a fully automatic process that starts with a text in one language and produces a corresponding text in another language, using a machine of some kind. The MT output can be used 'as is', repaired by a post-editor, or made available to a human who translates from the source text. The earliest known description of MT dates back to the 1930s, when Troyanskii obtained a patent in this area; however, the system described in his patent was never implemented (Hutchins 2000). Operational MT systems would have to wait for the availability of general-purpose electronic computers in the late 1940s. This chapter is about the three major operational machine-translation paradigms that have been successively dominant since the 1950s, and their relationship to an influential 1949 memorandum (Weaver 1949b). Those three paradigms are commonly referred to as follows, based on their respective computational architectures:

- RBMT (rule-based machine translation);
- SMT (statistical machine translation); and
- NMT (neural machine translation).

RBMT systems operate by applying rules hand-crafted by humans. Typically, there are rules for dividing a sentence into words; rules for looking up the words in a dictionary (usually involving base-form reduction); rules for syntactic analysis (based mostly on data retrieved from the dictionary); rules for syntactic transfer, sometimes with a limited semantic component to deal

with ambiguity and always to adjust the intermediate syntactic representation to comply with the word order and other target-language requirements; and rules for generating text in the target language from the intermediate representation. SMT and NMT are very different from RBMT in that they operate by applying the results of automatically analysing training data, which consist of many source texts and their human translations. They are described as data-driven rather than rule-based, even though all MT systems depend on some kind of data.

One could claim that all MT systems are rule based, in that they are computer-software applications implementing an algorithm, which is a set of rules to be followed; however, rules in data-driven systems, especially NMT, are not inspectable by humans, because they are not symbolic. In contrast, algebraic formulas (e.g. x = 3*y + 4) are inspectable and symbolic, consisting mainly of numeric symbols, called variables and constants, and symbols for arithmetic operations, e.g. addition and multiplication. One way to distinguish between inspectable rules and the opaque rules in an NMT system is to call them symbolic vs. sub-symbolic (e.g., Lieberman 2016). This distinction between inspectable (symbolic) rules in RBMT and non-inspectable (sub-symbolic) rules in NMT is fundamental to the discussion at the end of this chapter regarding possible limitations of NMT.

RBMT was dominant from the 1950s through the 1980s. The 1990s were a transition period between RBMT and SMT, and, starting around the year 2000, SMT gained momentum and began to replace RBMT as the main focus of MT research and development. As of 2018, NMT rapidly is becoming the dominant data-driven architecture. However, sufficient training data for a viable NMT system are available only for a handful (perhaps twenty) of the over-four-thousand languages in the world. For the rest (over 99%), either RBMT or SMT or, in most cases, human translation are the only options. Thus, older paradigms are still relevant.

Rather than summarizing the already well-covered histories of RBMT and SMT and trying to capture the current evolving state of NMT, this chapter takes a different approach. It attempts to demonstrate that the seeds of all three MT paradigms, i.e. RBMT, SMT and NMT, already existed in 1949, with the latter two awaiting key enabling developments outside the translation industry so that those seeds could take root. The bulk of the chapter consists of a description of (1) five seeds in the 1949 memo written by Warren Weaver; (2) five enabling developments external to the field of translation since 1949 and (3) connections between the seeds, the enabling developments and the three paradigms. The chapter ends with a description of some remaining challenges for MT and a proposal for answering the central unanswered question about Weaver's memo: does it foreshadow yet another paradigm, beyond NMT, involving the symbolic representation of meaning and machine understanding?

Discussion of the Weaver memorandum

Five approaches to MT in the Weaver memorandum

This section will examine the approaches described in the Weaver memo: (1) word-for-word translation (which he dismissed as a dead end); (2) using surrounding words (called micro context or co-text) to resolve ambiguities; (3) logic, as applied to treating the brain as a machine; (4) cryptography, as in describing translation as a decoding process; and (5) invariants (suggesting the need for a universal symbolic representation of meaning).

Weaver's memo was a major factor in the early interest in MT. John Hutchins, who is generally recognized as the main historian of MT from its early days up to the SMT era, has noted: 'In July 1949, ... Weaver wrote the memorandum which was to launch MT as a serious subject of research in the United States, and subsequently throughout the world' (Hutchins 1997: 203). The first approach, word-for-word translation, was already being implemented when Weaver wrote his memo. An implementation plan for the other four approaches was not included in the memo, because enabling technology and language resources were not yet available. Given the huge technological differences between the world in the mid-1950s and the world at the beginning of the 21st century, Weaver's ideas were truly remarkable for his time.

Word-for-word translation

Weaver reports (Weaver 1949b: 6) that on May 25, 1948, he visited Andrew Booth's computer laboratory in London. Booth and his colleague, Richard H. Richens, were interested in using computers for translation. At the time, they were only interested in mechanizing a dictionary. Weaver's comment on Booth and Richens' work explains why he dismisses the word-for-word approach: 'They had, at least at that time, not been concerned with the problems of multiple meaning, word order, idiom, etc.' The problems identified by Weaver in this one sentence, namely, ambiguity (multiple translations for the same word, depending on the context), differing word orders across languages, and multiword expressions that cannot be translated word-for-word (e.g., idiomatic expressions), have been and continue to be challenges for MT. Booth and Richens did recognize the problem of looking up inflected forms in a dictionary whose entries are keyed only by the base form of a word, but they did not have any sophisticated approach to computing base forms through morphological analysis. They simply removed the last letter of a word, successively, in hopes of finding a shorter form in the dictionary.

Weaver's statements about word-for-word translation using dictionary lookup are still valid. This approach is considered inappropriate for anything but a simple word-by-word gloss of a source text. In addition, it is generally agreed that morphological analysis of some kind, in conjunction with automatic dictionary lookup, is needed for effective dictionary lookup of inflected forms in all but the morphologically simplest languages.

Weaver then describes four additional possibilities (co-text, logic, cryptography and invariants), each planting one or more 'seeds' that go beyond word-for-word translation, resulting in a total of five approaches.

Using micro context (co-text)

The first of the four possibilities is to examine the words in a text through a 'slit in an opaque masque' (or a computer-based equivalent) that is lengthened until one can see N words on either side. This approach can be described as a 'sliding window' into the text. Weaver writes that, although it is difficult even for a human to translate a word without seeing any surrounding words, 'if N is large enough, one can unambiguously decide the meaning [translation] of the central word'. Thus, Weaver has clearly described the n-gram approach that is frequently used in computational linguistics. If N is one, you have a tri-gram (the word in question plus one word on each side, for a total of three). If N is two, you consider a sequence of five words, and so on.

Take, for example, the word ‘fire’. In a word-for-word approach, multiple translations are possible for this word in most, if not all, languages. To begin with, is it a noun or a verb? Without some context, the grammatical category cannot be determined. Suppose that, by some method, the machine determines that it is being used as a verb in the source text. Then, it is still ambiguous. For example, in Portuguese, three major translation options for the verb ‘fire’ are: fogo, disparar and despedir. If ‘fire’ appears in the tri-gram ‘set fire to’ then the translation is almost certainly ‘fogo’. However, the tri-gram ‘to fire the’ does not resolve the ambiguity. Moving from an N of 1 (a tri-gram) to an N of 2 (a 5-gram) we find real-life examples, e.g. ‘want to fire the gun’ and ‘need to fire the intern’. These 5-grams apparently resolve the ambiguity between ‘disparar’ and ‘despedir’ in Portuguese, but only based on a human-like understanding that a gun is an object that shoots a bullet when fired and that an intern is a person that loses his or her position when fired (i.e. dismissed).

True ambiguity may be possible, even for a human, even if N is fairly large, but the number of possible translations is drastically reduced when N is even one or two. At this point, Weaver introduces the notion of statistics. He asks what value of N would be needed, within technical writing limited to some domain, for the target-language word choice to be correct 98% of the time. He notes that computers at that time (the late 1940s) had insufficient storage to handle all possible phrases 2N+1 words long, for any substantial value of N, but he is hopeful that ‘a reasonable way could be found of using the micro-context to settle the difficult cases of ambiguity’. Implicit in Weaver's hope is that a method can be found to use micro context to resolve most ambiguities, and that such a method can be programmed into a computer.

We will later see how developments independent of the translation industry have made feasible the use of micro context, and we will find that Weaver's suggestion that statistical methods are relevant was also clearly prescient. In this chapter, we will call micro context ‘co-text’ (see Melby and Foster 2010). Thus, the micro-context section of Weaver’s memo planted two seeds: co-text and statistics.

Language and logic

For his second possibility beyond word-for-word translation, Weaver cites McCulloch and Pitts (1943)[i] describing a computational model of the brain. That paper turned out to be a seminal work in the development of neural networks, as applied to artificial intelligence (AI) (Piccinini 2004).

Weaver may or may not have been fully aware of the paper’s future significance but hoped that it could be the basis for solving ‘a useful part of the problem of translation’ in reference to a Turing machine. First proposed as an abstract idea by Alan Turing, a contemporary of McCulloch, in 1936, a Turing machine is an idealized mathematical model that performs its functions in a sequence of discrete steps. In effect, this describes how computers work. However, what might that ‘useful part’ be in Weaver’s memo? When this section of the Weaver memo was presented to a professor of philosophy specializing in logic (Ryan Christensen), the following response was received (by personal communication [email] on 31 January 2018):

> The relevant part of the [McCulloch and Pitt] paper is the last paragraph of section 3. It states (but does not prove, contra Weaver) that a neural net, supplied with various aids, is Turing complete—i.e., can compute any function that a Turing machine can. I'm not sure

> exactly how that relates to machine translation. One guess is ... [that] M&P are talking about human brains, and trying to describe them formally. If human brains—which can translate—turn out to be equivalent to computers, then computers should be able to translate, too. Of course, this equivalence is only in computing power, so it can deal only with the logical, as opposed to the 'literary,' part of language.
>
> *(Christensen 2018)*

Along with Professor Christensen, we can only speculate about exactly which part of the translation problem Weaver was thinking of when he cited McCulloch and Pitts, but it is clear the Language and Logic section of his memo planted another seed: neural nets (i.e. treating the human brain as a machine consisting of a network of nodes roughly equivalent to neurons). Thus, we have identified three seeds, so far, in the first two approaches beyond word-for-word translation.

Translation and cryptography

For his third possibility, Weaver cites the then-recent work of Claude Shannon on the mathematical theory of communication; again, bringing up the relevance of statistics, which is already on our list of seeds. Weaver clearly had more than a passing acquaintance with Shannon's ideas, having published an article on Shannon's work in *Scientific American* (Weaver 1949b), the same year he wrote the memo. Weaver points out in his memo that Shannon's theory of communication was influential in the field of cryptography, which had been important during World War II.

Applying cryptography to translation, Weaver states that it is 'very tempting' to say that a book written in a foreign language, say Russian or Chinese, is 'simply a book written in English which was coded into the [foreign language]'. Thus, to translate from a language other than English, you would only need to break the code and decode (that is, decrypt) the foreign language text, and thus obtain the original English text. The notion of treating a source text that was authored in a foreign language, say Russian, as an encrypted version of the English target text seemed bizarre and completely backwards to many people in the 1950s, especially when a Russian source text is written by a native Russian speaker and there is no English involved. Hutchins (1998) wrote: 'As it happens, it was not long before researchers in machine translation recognised the fallacy of the argument'. Nevertheless, although initially rejected by MT researchers, the idea of using cryptography in this way, sometimes called the Noisy Channel approach[ii], was taken very seriously much later (starting in the 1990s), and Weaver's idea turned out to be the fourth seed: cryptography.

Language and invariants

Weaver considered his fourth possibility to be the most promising and most sophisticated. It involves digging 'so deeply into the structure of languages as to [go] down to the level where they exhibit common traits'. Weaver touches on this search for deep levels of representation at several points in his memo. The deeper the representation, the more similar languages become, despite dramatic differences at the surface level. He writes, 'Perhaps the way [to solve the translation problem] is to descend, from each language, down to the common base of human communication—the real but as yet undiscovered universal language—and then re-emerge by

whatever particular route is convenient'. Thus, Weaver assumed the existence of a universal basis underlying all human communication, independent of any particular human language.

This idea of going deeper and deeper below the surface of languages until they all more and more look alike was a common objective within the RBMT community, starting in the 1960s, and was expressed by Bernard Vauquois, a prominent early researcher in MT, in a diagram known as the 'Vauquois triangle' or 'Vauquois pyramid'. The point on the diagram where the representations of the source and target texts become identical uses a language-neutral intermediate representation, often called an interlingua. In discussing the Vauquois triangle and the interlingual approach, Saers (2011) assumes, along with others in the translation field, that translation between human languages can be viewed as a sequence of transductions, that is, conversions, between formal representations; however, that transduction hypothesis is separate from the notion of multiple linguistic-representation levels, with the deepest level being a language-neutral interlingua. The idea of using an interlingua as an intermediate representation in MT is the fifth and final seed planted by Weaver. It was nurtured by Vauquois and others.

The interlingua approach is distinct from the approach of using another natural language, sometimes called a pivot language, as an intermediate representation. For example, one could translate from Russian to Japanese by first translating from Russian to English as a pivot language, and then from English to Japanese. However, this multi-stage approach introduces two opportunities for ambiguity and loss of detail. An interlingua is presumed to be unambiguous and capable of representing nuances of meaning in a language-independent fashion. The term 'pivot language' is itself ambiguous in the MT literature, sometimes referring to an interlingua (see Boitet 2000) and sometimes to a natural language.

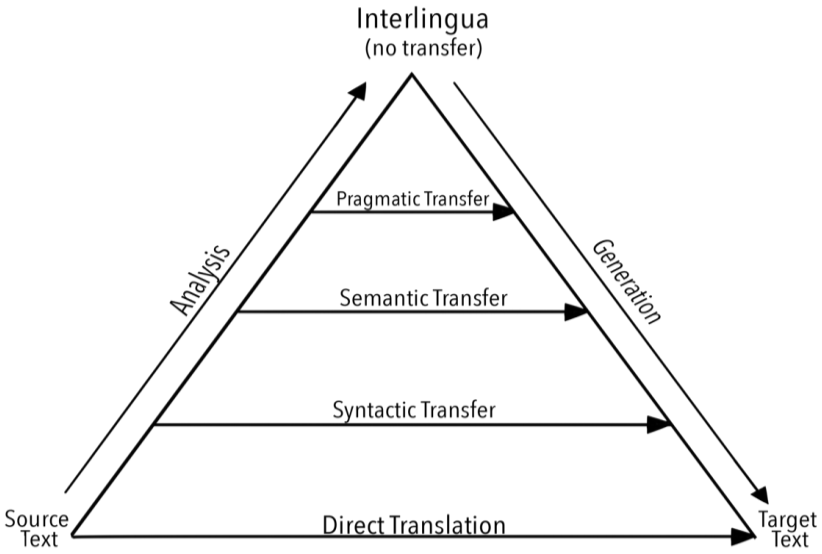

Figure 1. Diagram of the Vauquois triangle

As can be seen in Figure 1, the layers (sometimes called strata) correspond to the traditional (and still accepted) linguistic levels: morphology, syntax, semantics and pragmatics.

It is important to note that, although RBMT systems have explored all four linguistic levels using symbolic representations that can be discussed by humans, SMT and NMT systems are focused on the bottom of the triangle, the word (morphological) level. Something relevant to syntax and semantics might be going on in NMT, but it is not accessible to humans through a symbolic representation. An interlingua is also assumed to be symbolic; thus, there is no straightforward connection between NMT systems and an interlingua, as envisioned by Weaver and others.

Five enabling developments

Weaver was so far ahead of his time that it took many years before some of the seeds he planted could be acted on. Five enabling developments that took place independent of the needs of the MT community would turn out to be key in MT's subsequent history. They are the emergence of: large quantities of machine-processable text, affordable computational power, machine learning, a substantial demand for translation and deep structure in linguistics. They are all essential to SMT and NMT, except linguistic deep structure, as further discussed below.

Machine-processable text

One of the most important developments outside of translation was word processing. It resulted in the storage and transmission of machine-processable text; i.e. text where each character is represented as a distinct numeric code point, not as a digital image. It was not until 1969, two decades after the Weaver memo, that IBM introduced an electric typewriter with word-processing features and the capability of storing text in machine-processable form on magnetic cards (Flores 1983). This inevitably led to the storage of text on computers and transmission from one computer to another. As the desire for a 'paperless office' grew, huge volumes of text were stored on computer hard drives; initially attached to desktop computers, then attached to local servers and eventually stored in remote servers referred to as 'the cloud'.

Another crucial development in machine-processable text, independent of the translation industry, was the development of Unicode, which unified hundreds of different character-encoding systems into one standard ('What is Unicode?' 2017, Graham 1999). Unicode allowed for a robust representation of multiple languages in a single textual database[iii] (see also Wright in this volume).

Together, the widespread use of word processing (which was utilized by approximately 10% of translators at the 1991 ATA conference and nearly 100% at the same conference in 2000, with an incremental increase each year, based on personally conducted surveys) and the acceptance of Unicode (which has risen from less than 3% in 2001 to 93%, as of the time of this chapter's writing [Davis 2018]) have facilitated the creation of huge corpora in a number of languages. The only necessary translation-specific development was algorithms for automatic segmentation and alignment, resulting in bitext[iv] (parallel) corpora (see Simard in this volume). However, even these developments were not initially aimed at MT, but rather at translation memory lookup in an interactive tool controlled by a human translator.

Affordable computing power

A second enabling development outside the field of translation was the rapid increase in processing speed and storage capacity in computer hardware, without a proportional increase in cost. At the time of the Weaver memo, computers were physically huge, expensive, power-

hungry machines based on vacuum tubes, but they did not have the processing power of a hand-held graphing calculator, let alone a modern smartphone. It took the invention of the transistor at Bell Labs in 1948 and later the development of integrated circuits at Texas Instruments by Jack Kilby[v], along with major advances in hard-drive technology, to make available affordable computers with the processing power needed for the more recent paradigms in MT. None of these developments was driven by the need for MT, but such developments, rather importantly, enabled the translation process.

Machine learning

The term 'machine learning' was coined by Arthur Samuel, ten years after the Weaver memo. Samuel (1959), before focusing on the question of how a computer could be programmed to learn to play checkers, states: 'At the outset, it might be well to distinguish sharply between two general approaches to the problem of machine learning'. He called the first method the Neural Net Approach, clearly demonstrating his awareness of neural nets. The second method, which he states is 'much more efficient' is to create a network 'designed to learn only certain specific things'. Samuel correctly selected the second approach because it was 'capable of realization' at that time. The use of general-purpose neural networks in machine learning became feasible much later, with affordable computing power (the previous enabling development). However, there is a more fundamental issue in machine learning than computer power; namely, whether the knowledge acquired by a machine-learning system is represented symbolically or sub-symbolically (i.e. non-symbolically).

From an essay by the *Machine Learning* journal's founding editor (Langley 2011), we learn that in the 1980s, the emphasis in the field was on symbolic representations of 'learned knowledge, such as production rules, decision trees, and logical formulae' that were part of systems that carried out reasoning and involved language understanding. Indeed, the original call for papers discouraged submissions on neural networks and other non-symbolic approaches. Nevertheless, by the mid-1990s, papers on symbolic representations had nearly disappeared from the literature, having been displaced by work on 'less audacious tasks', e.g., classification and regression.

The approach to machine learning called 'deep learning' simply refers to a neural network with more than the basic two layers (input and output). According to Sze *et al.* (2017), the idea of deep learning was suggested in the 1960s, while RBMT was dominant, but it was not until 2012 that it really took off, thanks to the use of relatively inexpensive computers with multiple GPUs (graphical processing units) in neural network systems.

Demand for translation

Contrary to the expectation of some proponents of artificial languages, such as Esperanto, local languages, tied to national culture, have become more important rather than less important than they were in 1949, as demonstrated by the current 24 official languages in the EU. In addition, although English has become the second language for an increasing fraction of the world, it is not the first language for larger and larger areas of the world. Despite globalization in trade and a decrease in the differences between products available in different parts of the world, the demand for translation has increased, partially driven by the localization of software apps and websites, which involves both the translation of text and adaptation of non-textual elements. For example: The United States Department of Labor reported recently that the employment of translators was

expected to increase approximately 18 percent from 2016 to 2026, and a 2014 survey indicated that the success of international businesses with localized websites was greater than those that did not offer a localized 'global customer experience' as it relates to language (United States Department of Labor 2016).

Another aspect of the demand for translation is the acceptance of various grades of translation, depending on the translation's audience and purpose (see Bowker in this volume). For example, a low-grade translation, e.g. a raw (unedited) MT that contains various types of errors, is often preferable to end-users in an instant-messaging or rapidly evolving technical-support environment than no translation at all, especially if fall-back procedures are available, such as immediately asking for clarification of a garbled translation, or requesting interaction with a human when a machine-translated technical-support article is unusable. Wendt (2010) has been promoting the usefulness of low-grade translation for over a decade, although he does not use the term 'grade', in a context where translation errors are not fatal. This increase in the demand for translation of various grades is a major motivation to develop new and better approaches to MT.

Deep structure in linguistics

The notion of deep structure is based on the assumption that human language as written and spoken is based on an underlying representation in which languages are more similar than at the surface. The best-known proponent of deep structure in linguistics is Noam Chomsky, the driving force behind Generative Grammar. Generative Grammar's early versions were called Transformational Grammar, which was kicked off by a short but highly influential 1957 book by Chomsky. Even though Chomsky eventually backed off from the idea of going deeper and deeper until the structures of all languages became identical, some of his students declined to follow his shift in direction and, in 1967, developed what became known as Generative Semantics (Newmeyer 1986). This approach starts with a representation of meaning, and derives the syntactic structure of a sentence from its semantic representation (reversing the Generative Grammar approach, which derived the semantics from the syntax).

Presumably, in Generative Semantics, the deep representations of equivalent sentences in various languages would be the same or very similar, even if they looked very different on the surface. Deep structure in Generative Semantics was clearly headed toward an interlingua, even though it was not focused on translation. It was intended to provide insight into the relationship between various levels in language. Another linguist who has pursued the notion of language levels is Sydney Lamb, whose approach was originally called Stratificational Grammar in the late 1960s[vi].

Since 1949, many linguists have explored various aspects of universals in language; however, the search for universals was not part of mainstream linguistics when Weaver wrote his memo.

In 1949, the notion of structural transformations was already present in linguistics, as elaborated by Chomsky's mentor, Zellig Harris (1946/1951). However, until the 1960s, American Structuralism, which was purely descriptive and emphasized differences between languages rather than similarities, was the dominant linguistic paradigm in the United States, and semantics was considered off limits. Therefore, Weaver's idea of using an interlingua to facilitate translation had to wait for later developments within the fields of linguistics. However, unlike the first four enabling developments, the search for deep-structure models in linguistics has not

flourished[vii]. One might wonder about a connection between deep structure in linguistics and deep learning in AI. There is no direct connection. In linguistics, a deep structure is symbolic, while in AI, the results of deep learning are sub-symbolic and un-inspectable by humans.

Connections between paradigms and seeds

Now, in light of enabling developments external to MT (machine-processable text, affordable computing power, machine learning, demand for translation and deep structure), it is straightforward to make connections between the three successively dominant paradigms in MT (RBMT, SMT and NMT) and the seeds planted by Weaver (co-text, statistics, neural nets, cryptography and an interlingua), even though those connections have not been explicitly acknowledged in the MT field.

RBMT

RBMT systems can be seen as an attempt to take co-text into account and move through linguistic levels toward an interlingua. The focus has been on the syntactic level rather than the semantic or pragmatic levels, and co-text has been treated mostly at the syntactic level. However, many RBMT systems have relied too heavily on word-level lookups in dictionaries and have not considered co-text sufficiently. RBMT systems might have been able to benefit further from deep structure in linguistics, if the notion of deep structure had been pursued more successfully by linguists, and if robust, automatic monolingual converters between surface and deep structure had been developed by computational linguists for purposes other than machine translation.

Considering the traditional linguistic levels of morphology, syntax, semantics and pragmatics as ranging from being close to the surface to rather deep, RBMT systems have generally done a good job with morphology and syntax. Some have explored semantics and pragmatics, e.g., Lytinen and Schank (1982) and Uchida (1989). However, the approaches to an interlingua in these systems have not been pursued. This is firstly because genuine interlinguas are extremely difficult to create and also probably because of the shift toward sub-symbolic AI since the 1990s and the shift toward language-specific investigations in linguistics.

Needless to say, RBMT systems have been able to take advantage of the widespread availability of machine-processable source texts. In addition to very narrow domains, RBMT performs well between very similar languages that could be considered dialects of each other (Cherivirala *et al.* 2018). Here, a syntactic representation serves as an interlingua, when sufficiently close dialects have no significant differences at deeper linguistic levels.

The lack of currently operational systems using an interlingual approach for substantially different language pairs can be attributed to the fact that mainstream linguistics is not focused on finding universals, which raises the aforementioned philosophical question of whether there really is a universal basis underlying all human communication, as posited by Weaver and many others. The opposite might turn out to be true: Outside of well-defined domains, human languages may differ at all three levels of traditional linguistics: syntax (sentence structure), semantics (meanings of words at the sentence level) and pragmatics (meaning as it relates to interpretations of the real world and imagined worlds) (Melby 1995).

SMT

SMT can be seen as following up on three of Weaver's seeds: co-text, statistics and cryptography, thanks to the enabling factors of affordable computing power and machine learning applied to huge bitext (parallel) corpora (see Simard in this volume). The main reason that SMT became the dominant paradigm is economic, as it can produce an MT system automatically from a translation memory database. One example is the Moses open-source SMT system, which accepts a translation memory file and then generates an SMT system trained on the translation units (i.e. aligned pairs of source and target segments) contained within the translation memory. In the late 1990s, the Logos RBMT system took three years to add a new source language and a year to add a new target language (Mike Dillinger, past president of IAMT, personal communication 25 September 2018). Only a decade later, given a large, clean translation-memory corpus, a new language pair can be developed in weeks rather than years.

NMT

Weaver's reference to McCulloch and Pitts was a seed in the garden that grew modern neural networks and thus NMT, which is fuelled by the same external developments as SMT.[viii] Thus, NMT mainly benefits from two of the five seeds that Weaver planted (co-text and neural nets), and it depends on four of the five external developments listed above (machine-processable text, affordable computing power, machine learning and demand for translation).

SMT requires more computing power (speed and storage) than RBMT (at least the limited systems that have been developed), and NMT, at least in 2018, requires significantly more power and larger bitext training corpora than SMT.

This is why they both had to wait for affordable computing and the accumulation of massive amounts of machine-readable text, to enable Weaver’s seeds to grow, as summarized in Table 1.

Table 1. Summary of the relationship between Weaver’s seeds, the enabling developments, and the three paradigms

Weaver's seeds	RBMT	SMT	NMT
Co-text	✓	✓	✓
Statistics		✓	✓
Neural nets			✓
Cryptography		✓	
Interlingua	✓		
Enabling developments			
Machine-processable text	✓	✓	✓
Affordable computing power		✓	✓
Machine learning		✓	✓
Demand for translation	✓	✓	✓
Deep structure	✓		

Critical discussion: what comes after NMT?

Is NMT the final paradigm for MT? Will deep learning using the existing type of neural nets be continuously improved until it produces truly intelligent machines that can perform every intellectual task, including translation, as well as humans?

Some machine-learning experts believe that neural networks are, indeed, the ultimate answer. For example, Andrew Ng, a prominent figure in the AI community, suggested in 2016 that deep learning would 'now or in the near future' be able to do 'any mental task' a person could do 'with less than one second of thought' (Marcus 2018). If Ng is right, NMT is the final paradigm for MT. However, others have serious concerns about deep learning. For example, some have reservations about neural nets because they are opaque at two levels: architecture and correcting mistakes.

Deep-learning architecture

Multiple deep-learning architectures are available for any particular task. It is unclear exactly which deep-learning architecture to use in a given situation and how each works, as illustrated by the following critical observation made in December 2017 by Ali Rahimi, an AI researcher at Google in San Francisco, California, as reported by Hutson (2018):

> Rahimi charged that machine-learning algorithms, in which computers learn through trial and error, have become a form of 'alchemy'. Researchers, he said, do not know why some algorithms work and others don't, nor do they have rigorous criteria for choosing one AI architecture over another.

It is problematic that a prominent deep-learning insider such as Rahimi views the process of selecting a deep-learning architecture as simply unprincipled guessing. When will the alchemy of current AI become like chemistry, and what will it then look like? Would this significant evolution in machine learning usher in a new paradigm for MT?

Correcting mistakes

Once a deep-learning architecture has been selected and trained on a set of data points, it is put into use. NMT systems are typically presented with an input sentence from a source text and they produce an output sentence, presumably a translation of the source sentence. As with human translators, NMT systems sometimes make mistakes; however, the crucial difference is that NMT provides no direct method for finding out why the system made a particular mistake and to correct the mistake, based on why it happened. Paul Voosen (2017) describes the 'black box' in AI and how AI developers are tackling it. For example, 'counterfactual probes' are used to understand what is inside the black box by varying the inputs to the AI system to see which changes affect the output, and how.

The article reports that such probes can identify specific words that influence a decision by a neural network; however, this 'says little about the network's overall insight.' The article further reports that Mark Riedl, director of the Entertainment Intelligence Lab at the Georgia Institute of Technology in Atlanta (USA), states, 'If we can't ask a question about why [AI systems] do something and get a reasonable response back, people will just put [them] back on the shelf.' Riedl was talking about an AI system that plays a video game. He wants to allow a human to ask

why the system is making a particular move during the game. Applying this to a machine-translation system, we want to be able to ask the system why it translated a word or phrase the way it did.

Something remarkable about the Voosen (2017) article is that none of the proposed solutions to the black-box problem with pure deep learning suggest the use of symbolic representations. Why not? There seem to be two camps in the AI community. One thinks that deep learning is sufficient and that all use of symbols should be ignored. Thus, they would not suggest a symbolic representation as a solution to the black-box problem. The other camp thinks that deep learning is only part of the answer. Gary Marcus, a Professor of Psychology at New York University is from the second camp. Marcus (2012) provides an accessible history of the question, and his recent updated article (Marcus 2018) confirms that the currently hard-line deep-learning camp is still dominant and not interested in a dialog with the pro-symbolic camp.

Connection with the fifth Weaver seed

Circling back to Weaver's seeds, I emphasize that Weaver's fifth seed (levels of symbolic representation, with the deepest level being an interlingua) is missing from NMT, because NMT is based on sub-symbolic deep learning. The fundamental question regarding the Weaver memo is whether this fifth seed foreshadows a new paradigm in MT or whether deep learning with sub-symbolic results from training is the ultimate paradigm. As a theoretical linguist and a translator, it is obvious to me that Marcus takes a more viable position than deep-learning hard liners. Language experts, including translators, can switch between using language in real time and talking about language. Real-time reading, listening and interacting with other humans may well involve sub-symbolic activity in the brain, but talking about language involves symbols.

Written language is a prime example of a symbolic system that is central to human mental activity. Words are probably the most common symbols used by humans, yet the result of training a multi-layer neural net, as opposed to the output produced by applying it, is not represented in human language or any other symbolic representation.

What does Weaver have to say about the question of symbolic representations? He does not treat the issue explicitly, since at the time he wrote the memo, sub-symbolic representations of knowledge were not even under consideration. Weaver's closing remark was that 'it' (a computer program that could convert between human languages thanks to levels of representation below the surface level of words) 'could not fail to shed much useful light on the general problem of communication'. It is clear that Weaver, in mentioning communication, was talking about machines understanding language. At some point in the development of MT systems, we will have to confront the possibility that machines do not truly understand what they are translating and that understanding, in a fully human sense, is essential to further progress.

Perhaps what is missing in NMT is some kind of integration that ties symbols to real neural networks, not the simplified networks currently in vogue. What would that integration look like? Perhaps, the levels of symbolic representation that are key to such an integration have been around for a long time, right under our noses, in the field of theoretical linguistics, but are not even considered in current AI work. Those levels of symbolic representation, morphology, syntax, semantics and pragmatics, are relevant to MT, whether or not there is a level beyond pragmatics at which all languages are identical.

Each successive level brings in more aspects of context. NMT deals explicitly with co-text, which consists of the words immediately surrounding the word in question, and with bitext, thanks to its training data, which consists of text and its context in another language; i.e. its human translation. However, NMT is not equipped to relate a text to its context in the real world. This third aspect of context is sometimes called 'non-text'. See Melby and Foster (2010) for a taxonomy of aspects of context, focused on translation.

Sydney Lamb (1999, 2016) has written extensively on how linguistic levels in what he calls relational networks, could be tied to networks of realistic neurons. Wilks (1979) provides a detailed view of the connection between MT and AI from an era during which symbolic AI was still in vogue. Symbolic representation of meaning is not dead, outside the current approach to machine learning.

The elephant in the room is the question of understanding. Does an MT system need to understand what it is translating?

A related question is whether a human must understand what s/he is translating.

A highly experienced financial translator, Robin Bonthrone, responded in this fashion:

> …Yes, translators still need to have an understanding of the text. In fact, specialized translators (and that, surely, is the future of human translation) need to have an in-depth understanding at several levels [linguistic, subject matter and context]. And this is one of the things that sets them apart from translators who don't understand the text. So actually, I despair a bit when I hear that experienced colleagues claim that translators don't have to understand the source text. That's the wrong message to send, especially for younger colleagues hoping to still work as translators in 10 or 20 years' time.
>
> (*Personal communication, 2 November 2018*)

The bottom line for translators is that, with the currently dominant MT paradigm, whose proponents refuse to even attempt to integrate symbolic representations with the sub-symbolic results of training a deep-learning system, it is not possible to ask an NMT system why a particular translation did not meet the specifications for a project. This serious failing might indicate that NMT systems do not understand what they are translating and that this is an inherent limitation to the paradigm.

Despite, or perhaps because of, the black-box nature of NMT systems and the lack of desire to integrate symbolic representations into deep learning, the debate about how to tell whether a computer understands language is usually not interesting to MT developers. They focus on how well their systems can translate. They avoid the philosophical questions. Therefore, I propose that translators and MT developers focus on operational approaches to evaluating translation service providers, whether they involve humans (once the bitext training material has been gathered) or not. Melby (2012) describes such an operational approach, called TTT (an intentionally ambiguous acronym standing for Translation Turing Test and various Translation Technology Tests). Instead of debating the nature of meaning and understanding, the spirit of TTT is to expect a translation service provider (human, machine, or hybrid) to produce translations that meet detailed specifications and respond appropriately to errors detected through a third-party translation-quality evaluation and described in human language. If a machine-

translation system behaves like a competent human provider, it is assumed to understand what it is doing.

As of this writing (2018), NMT-based systems are far from being able to demonstrate understanding according to the TTT approach. Weaver, if he were around today, would probably be pleased that four of the five seeds he planted have sprouted and would agree with the position that the fifth (a multi-level, if not universal base for language) is the most promising approach, and might even support TTT as an approach to testing understanding. It is up to the reader to evaluate my claims about Weaver's memo, which remains as an invaluable legacy.

Future of MT

In closing, I predict that if NMT remains opaque (i.e. un-inspectable), its translation quality will eventually plateau, especially for demanding specifications, and it will remain unable to support human-language interaction with the requester concerning specifications and errors. People will then start acknowledging the elephant in the room in their search for a new paradigm. Interest in whether machines actually understand language will then be rekindled, and Weaver's fifth approach, which implies the search for a symbolic representation of at least syntactic, semantic and pragmatic levels of meaning, will again be pursued.

Related Chapters

Chapter 5: Building and using parallel text for translation
Chapter 7: Multinational language service provider as user
Chapter 8: Application of technology in Patent Cooperation Treaty (PCT) Translation Division of the World Intellectual-Property Organization
Chapter 9: Small and medium-size enterprise (SME) translation service provider as a technology user: Translation in New Zealand
Chapter 10: Freelance translators' perspectives
Chapter 19: Post-editing of machine translation
Chapter 22: Translation technology research and human–computer interaction
Chapter 28: Copyright and the reuse of translation as data
Chapter 30: Technology and translator training

Further Reading

Hutchins, J. (1986) *Machine Translation: Past, Present, Future*, Chichester: Ellis Horwood.

> Hutchins (1986) is one of the most comprehensive histories of machine translation during the RBMT period. It ends with a note about the emergence of statistical methods and hints at SMT.

For those new to the MT field, the MT Archive is a reliable source of articles, journals, conference series and books through 2017 (http://www.mt-archive.info).

Kay, M. (2017) *Translation: Linguistic and Philosophical Perspectives* (Studies in Computational Linguistics, Book 221). Stanford, CA: Center for the Study of Language and Information.

Kay discusses another take on the Weaver memo.

Koehn, P. (2010) *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Among the extensive literature available, Koehn (2010) is one of the most comprehensive introductions to SMT, and features an extensive bibliography that can be used for further reading about SMT.

Forcada, M. (2017) 'Making Sense of Machine Translation', *Translation Spaces* 6 (2), 291–309. doi: 10.1075/ts.6.2.06

NMT is evolving at a breakneck pace as of this writing (2018), so a comprehensive and definitive description is infeasible at this time; however, Forcada (2017) presents a highly accessible introduction.

For current MT research, see the conferences organized by the International Association for Machine Translation (http://www.eamt.org/iamt.php) and its three regional organizations in Asia, Europe, and the Americas, as well as the WMT (Workshop in Machine Translation) series (e.g. see http://www.statmt.org/wmt18/ for the 18th workshop).

Notes

[i] According to Google Scholar, the McCulloch and Pitt paper had been cited over 15,000 times by late 2018.

[ii] A recent commentary on Shannon's work, including the notion of a Noisy Channel, is available in an MIT news article (http://news.mit.edu/2010/explained-shannon-0115).

[iii] Unicode allows multiple languages to be stored in the same file without the fragile representations of non-English text used previously (e.g., ISO 2022). ISO 2022 was fragile because it was 'stateful'. The interpretation had to begin at an escape character, which could be far away from the desired text. Thus, a single-bit error in an escape character could cause hundreds or even thousands of subsequent bytes to be misinterpreted. Early computers and word processors used single-byte approaches without escape characters, mostly EBCDIC and ASCII, which allowed only English characters and punctuation. They included a few accents (e.g. acute, grave, circumflex and tilde) separate from the character they modified, to be represented directly, but no accented characters could be represented as a single code point.

[iv] A bitext is a source text with a corresponding translation; segmented and aligned, usually at the sentence or paragraph level. For more information, see Harris (1988) and Melby (2015).

[v] For this work, Kilby shared the year 2000 Nobel Prize in physics: see https://www.nobelprize.org/prizes/physics/2000/summary/

[vi] For major updates to Lamb's approach see Lamb (1999, 2016).

[vii] The program of the January 2019 conference of the Linguistic Society of America (https://www.linguisticsociety.org/node/9647/schedule) indicates that pursuing a universal deep structure underlying all languages is not part of the current research agenda for theoretical linguists.

[viii] For more on the history of neural networks and the influence of McCulloch and Pitt's article, see Rojas (2013). For in-depth information about McCulloch, see a recent biography by Abraham (2016) and an analysis of McCulloch's article (Piccinini 2004).

References

Abraham, T. H. (2016) *Rebel Genius: Warren S. McCulloch's Transdisciplinary Life in Science*, Cambridge, MA: MIT Press.

Bar-Hillel, Y. (1964) *Language and Information*. Jerusalem: The Jerusalem Academic Press.

Boitet, C. (2000) 'Bernard Vauquois' contribution to the theory and practice of building MT systems: A historical perspective', in W. J. Hutchins (ed.), *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*, Amsterdam & Philadelphia: John Benjamins Publishing Company, 331–348.

Cherivirala, S., S. Chiplunkar, J. N. Washington and G. K. B. Unhammer (2018) 'Apertium's Web Toolchain for Low-Resource Language Technology', *Proceedings of Technologies for MT of Low Resource Languages at the 13th Conference of The Association for Machine Translation in the Americas (AMTA) (LoResMT 2018) Boston*, MA: AMTA, 53–62.

Chomsky, N. (1957) *Syntactic Structures*, The Hague: Mouton & Co.

Davis, M. (2008) 'Moving to Unicode 5.1.', *Google Blog*, 5 May 2008. Available online: https://googleblog.blogspot.com/2008/05/moving-to-unicode-51.html [last accessed 12 Jan. 2019].

Eco, U. (1997) *The Search for the Perfect Language*, New York: Fontana Press.

Flores, I. (1983) *Word Processing Handbook*, New York: Van Nostrand Reinhold Company.

Forcada, M. (2017) 'Making Sense of Machine Translation', *Translation Spaces* 6(2), 291–309. doi: 10.1075/ts.6.2.06

Graham, T. (1999) 'Unicode: What Is It and How Do I Use It?', *Markup Languages: Theory & Practice* 1(4): 75–102. doi:10.1162/109966299760283210.

Harris, B. (1988) 'Bi-text, a new concept in translation theory', *Language Monthly* 54(March): 8–10. Available online: http://mt-archive.info/LangMonthly-54-1988-Harris.pdf [last accessed 12 Jan. 2019].

Harris, Z. (1946/1951) *Methods in Structural Linguistics*, Chicago: University of Chicago Press.

Hutchins, J. (1986) *Machine Translation: Past, Present, Future*, Chichester: Ellis Horwood.

Hutchins, J. (1997) 'From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947–1954. A Chronology', *Machine Translation* 12(3), 195-252.

Hutchins, J. (1998) 'Milestones in Machine Translation, no. 2: Warren Weaver's memorandum, 1949', *Language Today* no. 6 (March 1998), 22–23.

Hutchins, J. (2000b) *Warren Weaver and the Launching of MT. Early Years in Machine Translation*. Available online: http://www.hutchinsweb.me.uk/Weaver-2000.pdf [last accessed 12 Jan. 2019].

Hutchins, J. and E. Lovtskii (2000a) 'Petr Petrovich Troyanskii (1894-1950): A Forgotten Pioneer of Mechanical Translation', *Machine Translation* 15(3): 187–221. Available online: http://www.jstor.org/stable/40009018 [last accessed 12 Jan. 2019].

Hutton, M. (2018) 'AI researchers allege that machine learning is alchemy', *Science.* Available online: http://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy [last accessed 29 December 2018].

Kay, M. (2017) *Translation: Linguistic and Philosophical Perspectives* (Studies in Computational Linguistics, Book 221). Stanford, CA: Center for the Study of Language and Information.

Koehn, P. (2010) *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Lamb, S. (1999) *Pathways of the Brain: The Neurocognitive Basis of Language,* Amsterdam & Philadelphia: John Benjamins Publishing Company.

Lamb, S. (2016) 'Linguistic structure: A plausible theory', *Language Under Discussion* 4(1). Available online: http://www.ludjournal.org/index.php?journal=LUD&page=article&op=view&path%5B%5D=30&path%5B%5D=21 [last accessed 12 Jan. 2019].

Langley, P. (2011) 'The changing science of machine learning', *Machine Learning* 82: 275–279.

Lieberman, H. (2016) 'Symbolic vs. Subsymbolic,' (a presentation), MIT. Available online: http://futureai.media.mit.edu/wp-content/uploads/sites/40/2016/02/Symbolic-vs.-Subsymbolic.pptx_.pdf [last accessed 1 Oct. 2018].

Locke, W. N. and A. D. Booth (eds.) (1955) *Machine Translation of Languages: Fourteen Essays*. Cambridge, Mass.: MIT Press.

Lytinen, S. L. and R. C. Schank (1982) 'Representation and translation', *Interdisciplinary Journal for the Study of Discourse* 2(1–3): 83–112 (1982 article reprinted in 2009). *Cambridge Dictionary Online*. Available online: https://dictionary.cambridge.org/us/dictionary/english/machine-translation [last accessed 15 December 2018]

Marcus, G. (2001) *The Algebraic Mind: Integrating Connectionism and Cognitive Science.* Cambridge, Mass.: MIT Press.

Marcus, G. (2012) 'Is "Deep Learning" a Revolution in Artificial Intelligence?', *New Yorker*, 25 November 2012. Available online: https://www.newyorker.com/news/news-desk/is-deep-learning-a-revolution-in-artificial-intelligence [last accessed 29 December 2018].

Marcus, G. (2018) 'The deepest problem with deep learning', in *Medium.* Available online: https://medium.com/@GaryMarcus/the-deepest-problem-with-deep-learning-91c5991f5695 [last accessed 29 December 2018].

McCulloch, W. and W. Pitts (1943) 'A logical calculus of the ideas immanent in nervous activity', *Bulletin of Mathematical Biophysics*, 115–133.

Melby, A. K. (2012) Human and Translation Quality: Definable? Achievable? Desirable? Available online: http://www.ttt.org/melbyak [last accessed 12 Jan. 2019]

Melby, A. K. and C. Foster (2010) 'Context in Translation: definition, access, and teamwork', *The International Journal of Translation and Interpreting Research* 2(2): 1. Available online: http://trans-int.org/index.php/transint/article/view/87 [last accessed 12 Jan. 2019].

Melby, A. K., A. Lommel, and L. Morado Vásquez (2015) 'Bitext', in S-W, Chan (ed.), *Routledge Encyclopedia of Translation Technology*, London & New York: Routledge, 409–424.

Melby, A. K. and T. Warner (1995) *The possibility of language: A discussion of the nature of language, with implications for human and machine translation*. Vol. 14. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Newmeyer, F. J. (1986) *Linguistic Theory in America*, 2nd ed., Cambridge, Mass.: Academic Press.

Piccinini, G. (2004) 'The first computational theory of mind and brain: a close look at McCulloch and Pitts's "logical calculus of ideas immanent in nervous activity"', *Synthese* 141(2): 175–215.

Qun, L. and X. Zhang (2015) 'Machine Translation – General', in S-W, Chan (ed.), *Routledge Encyclopedia of Translation Technology*, London & New York: Routledge, 224–249.

Rojas, R. (2013) *Neural Networks: A Systematic Introduction*. New York: Springer Science & Business Media.

Saers, M. (2011) *Translation as Linear Transduction: Models and Algorithms for Efficient Learning in Statistical Machine Translation*, Doctoral thesis, Uppsala University, Sweden. Available online: http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-135704 [last accessed 12 Jan. 2019].

Samuel, A. L. (1959) 'Some studies in machine learning using the game of checkers', *IBM Journal of Research and Development* 3(3): 210–229.

Sze, V., Y.H. Chen, T.J. Yang, and J. S. Emer (2017) 'Efficient processing of deep neural networks: A tutorial and survey', *Proceedings of the IEEE*, 105 (12): 2295–2329.

Turing Machine. (n.d.). In the Editors of Encyclopaedia Britannica (Ed.), *Encyclopædia Britannica*.

Uchida, H. (1989) 'ATLAS II: A machine translation system using conceptual structure as an interlingua', in M. Nagao (ed.) *Machine Translation Summit*, Tokyo, 93–100. Available online: https://pdfs.semanticscholar.org/0504/48982b6e976d9cbe41c9199137dc9dd08e18.pdf [last accessed 12 January 2019].

United States Department of Labor (2016) 'Interpreters and translators', *Occupational Outlook Handbook*. Available online: https://www.bls.gov/ooh/media-and-communication/interpreters-and-translators.htm#tab-6 [last accessed 12 Jan. 2019].

Vauquois, B., G. Veillon and J. Veyrunes (1967) 'Un métalangage de grammaires transformationnelles', *Conférence Internationale Sur Le Traitement Automatique Des Langues*, 1. Available online: https://aclanthology.coli.uni-saarland.de/papers/C67-1019/c67-1019 [last accessed 12 Jan. 2019].

Voosen, P. (2017) 'The AI detectives', *Science Magazine* 357(6346). Available online: http://science.sciencemag.org/content/357/6346/22/tab-pdf [last accessed 7 Jul. 2017].

Weaver, W. (1949a) *Translation*. Reprinted in: W. N. Locke and A. D. Booth (eds.), *Machine Translation of Languages*, New York: The Technology Press of MIT, 15–23.

Weaver, W. (1949b) 'The Mathematics of Communication', *Scientific American* 181(1): 11–15.

Wendt, C. (2010) 'Better translations with user collaboration—integrated MT at Microsoft', presented at *The Ninth Biennial Conference of the Association for Machine Translation in the Americas (http://mt-archive.info/AMTA-2010-Wendt.pdf)?* [last accessed 12 Jan. 2019].

What is Unicode? Available online: http://unicode.org/standard/WhatIsUnicode.html [last accessed 24 July 2017].

Wilks, Y. (1979) 'Machine translation and artificial intelligence', in B. M. Snell (ed.), *Translating and the Computer*, Amsterdam: North Holland Publishing Company. Available online: https://www.researchgate.net/publication/241488724 [last accessed 12 Jan. 2019].