Knowledge and choices in machine translation

SERGEI NIRENBURG

The field of machine translation (MT) has a long and turbulent history. Indeed, it may have been the first non-numerical application suggested in the 1940s for the then nascent field of computer science, and it seemed a very attractive application to try to develop. This opinion was bolstered by the following considerations. First, in the era of information explosion translation becomes a very critical business. As in every other business, automation is developed to enhance efficiency. Second, because translation is a common task regularly performed by humans, the specification of the task is relatively straightforward: the conceptual design of a potential MT system can be modeled after the organization of the translation process performed by humans. Third, the dictionary look-up, which may account for a very significant part of the time spent on translation, can be reduced to an insignificant level when on-line dictionaries are used. Feasibility considerations tended to be influenced by the belief that since translation is such a common everyday task, performed with relative ease by humans, it must be easy to automate. This was the rationale behind the exciting development of MT research from the late 1940s till the early 1960s.

It is common knowledge that the early MT projects failed to reach their stated goal of building systems of good-quality fully automated translation in broad domains. A number of good surveys of the history of machine translation (notably, Hutchins, 1978, 1986 and Zarechnak, 1979) discuss the reasons for the failure of MT research to achieve this goal. One must not forget, however, that the early MT efforts provided insight into the study of language and its processing by computer. They yielded many positive results by contributing to the development of modern linguistics, computational linguistics and, eventually, artificial intelligence (AI).

The principal mistake of the early MT workers was that of judgment: the complexity of the conceptual problem of natural language understanding was underestimated. It soon became clear that both the variety and the sheer amount of knowledge that must be used in any solution to this problem are enormous. In Section 1.1 we will discuss the types of knowledge needed to understand natural language text. As soon as it became obvious that the good-quality fully automated translation in broad domains was not immediately feasible, MT workers started to develop alternative methodologies, with the aim of rendering their systems more readily useful. A number of such methodologies coexist in the field now, and we will discuss
their strategies and achievements in Section 1.2. Section 1.3 contains brief overviews of the rest of the chapters in this book.

1.1. Knowledge in MT

The task of MT can be defined very simply: the computer must be able to obtain as input a text in one language (SL, for source language) and produce as output a text in another language (TL, for target language), so that the meaning of the TL text is the same as that of the SL text. It is clear that finding a way of maintaining invariance of meaning is the crucial problem in MT research. Indeed, the differences among the existing machine translation efforts can be summarized in terms of the solutions that they propose for the problem of finding means of expression in TL for the various facets of meaning of the input text units.

A number of important questions are raised at this point.

1. What is the meaning of the text?
2. Does it have any component structure?
3. How does one represent the meaning of a text?
4. How does one set out to extract the meaning of a text?
5. Is it absolutely necessary to extract meaning (or at least all of the meaning) in order to translate?

Question 1 is a basic problem in linguistics and philosophy of language. We cannot even circumscribe all of its facets here. This book is devoted to the theory and methodology of one of the application areas of linguistics and AI. Therefore, we will take a more operational approach to discussing the problem of meaning, as made manifest in the rest of the above questions. Components of meaning that have to be taken into account include morphological, syntactic, lexical-semantic and contextual (inferential-semantic and pragmatic) facets of meaning. Question 3 relates to the problem of knowledge representation, either in the style it is done in AI or in a more traditional sense, as in lexicography. Question 4 highlights the computational problems of such an enterprise as MT. Question 5 strongly implies the negative answer which reflects the hopes of many MT researchers and the practical limitations of the state of the art in MT.

All of the above problems are difficult. No definitive solutions have been suggested for them at this moment in the development of the field of computational linguistics and AI. It is in this light that one must interpret Question 4. Is there a possibility that success in a particular application area, such as MT, is not contingent on producing workable solutions for the above problems? The rest of this section is devoted to a discussion of the depth of meaning analysis necessary for determining the translation of a text.

Human translators use dictionaries as sources of information about SL and TL. The type of dictionary that is most used by humans is the bilingual
dictionary that connects units of SL and TL. By design, this SL-TL mapping seeks to preserve meaning. And since meaning is the invariant between the SL and TL texts in translation, such dictionaries must serve the purpose adequately. An important point to remember, however, is that bilingual dictionaries, as we know them, are designed for human use. People possess a great ability to ‘make sense’ of language units. This makes the task of the lexicographer simpler in that not all aspects of meaning have to be absolutely laid out; people will able to understand even a flawed explanation. The situation is quite different when the dictionary is used by a computer program. Let us illustrate the types of dictionaries and processing modules that will become necessary in this case.

Suppose the system obtains the German text (1) as input.

(1) Das Buch liegt auf dem Tisch.

The English translation of (1) is (2).

(2) The book is on the table.

The dictionary necessary to perform this translation is as follows:

<table>
<thead>
<tr>
<th>German</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>auf</td>
<td>on</td>
</tr>
<tr>
<td>Buch</td>
<td>book</td>
</tr>
<tr>
<td>das</td>
<td>the</td>
</tr>
<tr>
<td>dem</td>
<td>the</td>
</tr>
<tr>
<td>liegt</td>
<td>is</td>
</tr>
<tr>
<td>Tisch</td>
<td>table</td>
</tr>
</tbody>
</table>

The translation program is supplied that substitutes the English words for their German counterparts, one by one. Do we have an MT system? Yes, a system of machine translation of the German sentence (1) into the English sentence (2). Can we use the same system to go from English into German? No, because we have a one-to-many relationship from the to das and dem. Our knowledge, as recorded in the dictionary, is insufficient to resolve this ambiguity, and we have no additional knowledge to help us make the proper choice.

This is the first time we observe that MT research can be viewed as a process of accumulating knowledge that facilitates making correct choices of output.\(^1\)

\(^1\) It is interesting to note, in passing, that the problem of disambiguation can be reduced to the general problem of search, as it is known in AI. The most important and pertinent property of this problem is that the average time of search is in inverse proportion to the amount of knowledge that the system is able to use to establish priorities at every choice point.
Now, to translate (2) into Russian, we will need the dictionary as follows:

<table>
<thead>
<tr>
<th>English</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>book</td>
<td>kniga</td>
</tr>
<tr>
<td>is</td>
<td>#</td>
</tr>
<tr>
<td>on</td>
<td>na</td>
</tr>
<tr>
<td>table</td>
<td>stole</td>
</tr>
<tr>
<td>the</td>
<td>#</td>
</tr>
</tbody>
</table>

Sentence (2) will be, therefore, translated into Russian as (3).

(3)  Kniga na stole.

Interestingly enough, (3) is also the translation of (4):

(4)  The book is on a table.

The above suggests, somewhat unexpectedly, that articles do not have meaning in English. This, of course, is not true. We will return to the question of how the articles influence the translation later. Note also that the Russian word stole is in fact one of 10 different words (corresponding to different case and number values) that will each correspond to the English table meaning a piece of furniture. We will not, however, discuss morphological analysis here. Morphological analyzers have been built for many languages, including such morphologically rich ones as Hebrew (Nirenburg and Ben Asher, 1984) and Finnish (Koskenniemi, 1984).

Once again, we can see that it is impossible to use the same dictionary for a back translation of (3) into English. There is no indication where to put (if at all) the words that correspond to empty strings (marked ‘#’) in Russian. This example shows that more knowledge has to be introduced into the system. For instance, class nouns, when used in their singular forms in English sentences, must be preceded by an article (e.g., a, the), a demonstrative (e.g., this, that), a possessive (e.g., my, their), a question-word (e.g., what, which, whose), or the quantifier one. This is a part of the knowledge about the syntax of English which an automatic translation system must possess.

Another type of syntactic knowledge used in the disambiguation process relates to lexical categories of words.

(5)  The coach lost a set.

Without the knowledge of the syntactic structure of (5) it is impossible to decide whether coach is a noun or a verb; lost, a verb or an adjective; set, a noun, a verb or an adjective. The knowledge of English syntax is sufficient to eliminate this 12-way ambiguity and choose the correct reading. This type of knowledge is recorded in a grammar of English (or any other SL in an MT system). A special processing unit (a syntactic parser) applies this knowledge to the input text and produces its syntactic structure. The
dictionary can now have a separate entry for every distinct syntactic reading of an SL word (that is, set will appear as noun, adjective and verb).

Unfortunately, in some cases, syntactic knowledge is not sufficient for complete disambiguation. Thus, on purely syntactic grounds it is impossible to determine whether in (6a) the conjunction connects two nouns or a noun and a noun modified by an adjective (in other words, whether the chairs are also white). Note that one cannot neglect to extract this type of knowledge, because the form of a potential translation may depend on the intended meaning. Thus, (6a) will be translated into Hebrew as either (6b) or (6c).

(6) (a) White tables and chairs  
(b) shulhanot levanim vekisaot (white tables) and (chairs)  
(c) shulhanot vekisaot levanim: white (tables and chairs)

Further types of ambiguities that syntactic knowledge fails to take care of include prepositional phrase attachment and decomposition of noun compounds in English.

Even more profound evidence of the insufficiency of syntactic analysis for MT is presented by the commonplace lexical-semantic ambiguity of natural language. Thus, in a standard English-Russian dictionary (Halperin, 1972) the words coach, lose and set, in their correct syntactic meanings, detected by the syntactic analysis, have six, ten and thirty-four readings, respectively. This is a 2040-way ambiguity. Incidentally, even though syntactic disambiguation leaves us with this multiple ambiguity, syntactic analysis is still very useful. After all, if the correct parts of speech are not detected, the number of readings for the three words is eleven, fifteen and ninety-six, respectively, producing a 15840-way ambiguity.²

<table>
<thead>
<tr>
<th>coach</th>
<th>lost</th>
<th>set</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ways ambiguous, syntactically and semantically</td>
<td>11</td>
<td>15</td>
<td>96</td>
</tr>
<tr>
<td>ways ambiguous, syntactic ambiguity eliminated</td>
<td>6</td>
<td>10</td>
<td>34</td>
</tr>
</tbody>
</table>

As another example, consider the Russian sentence (7).

(7) Novaja partija byla luchshe vo vseh otnoshenijax

Taken out of context, this sentence does not contain any clue as to the meaning of partija. The texts in (8) illustrate the correct translations when the context is provided.

(8) (a) [The old Liberal Center was too doctrinaire for his taste.]  
The new party was better in all respects.

² It certainly says something about the disambiguating powers of humans that in actual discourse we can assign a single meaning to this sentence effortlessly.
The choice of the appropriate TL correlate was facilitated for humans by the context. A Russian-English dictionary that would be able to distinguish the alternatives will have to have special context identification markers, as in (9). These context markers are semantic in nature, and they suffice for human translators. MT systems should, however, possess special means of identifying the semantic context. This task has been a central concern of the subfield of natural language processing within AI.

(9) Partija, noun, feminine. 1. (political) party; 2. (commerce) batch; 3. (chess) game.

It has been soon recognized that in order to automate semantic analysis one has to devise a principled way of representing the meaning of the input text with the help of a complete system of semantic markers and then provide rules of using such a representation to extract the necessary knowledge about the context. The large number of semantic markers necessary to describe a reasonably rich subworld, the fact that they stand in well-defined relationships to other markers, and the total absence of the material of a natural language in the representation are characteristics of AI-oriented approaches to meaning extraction. The notations used in AI systems are rich enough to be called semantic interpretation languages. In AI they are called knowledge representation languages; in MT they are traditionally called *interlinguale*.

One of the first attempts to use semantic analysis, that is, to use a knowledge representation language in MT, was made within the Yale AI school. The Conceptual Dependency knowledge representation language (e.g., Schank, 1975) was used to represent the meaning of the input sentence. The experimental MT systems that were built in this school used background knowledge about the world to infer information not explicitly mentioned in the input sentence, in order to be able to disambiguate it. This background knowledge usually described typical event sequences, called *scripts*, that are common in certain subworlds. Thus, the knowledge that the event described in (5) belongs to the *tennis* script helps disambiguate this sentence completely by suggesting the appropriate readings for the words *coach* and *set*. The bilingual dictionaries in such systems take the form of discrimination nets with choice points marked by particular units of semantic knowledge. These nets were used at the generation stage of the system and connected *meanings* and TL units. Of course, the meanings to be used in a particular generation instance were first obtained through semantic analysis of the input text.
Schematically, the process of translation in such systems can be illustrated as follows (we use an example from Carbonell et al., 1981). Suppose, for instance, that the sentence (10) has been supplied as the input to the translation program. (10) will be analyzed (translated into the conceptual dependency language) as (11). The dictionary that the analyzer will use connects the English verb *hit* with a frame in which there will be a slot for *action*, occupied by the marker PROPEL (which is *not* an English word but rather an interlingua concept!) and a slot for *force*, which will be filled by the marker ABOVE-AVERAGE. The slots for *agent, object* and *instrument* will be listed in the dictionary without fillers. It is the responsibility of the analyzer to create the representation of a new event with the representation of *hit*, taken from the dictionary, as its nucleus and to find fillers for the slots that are unoccupied. It is predominantly for the purpose of identifying these slot fillers that the system uses scripts and other background knowledge.

(10) Mary hit John.

(11) (event EV001
(action PROPEL)
(agent MARY)
(object JOHN)
(instrument *UNKNOWN*)
(force *ABOVE-AVERAGE*)
(intentionality *POSITIVE*))

In order to translate (11) into Spanish, a discrimination net such as that in Figure 1 has to be used. The aim is to choose the appropriate Spanish verb to render the English *hit*.

In this particular case *pegar* will be chosen.

The type of analysis performed by this type of system is, however, far from sufficient. A number of choices still remain unresolved even at this level of semantic processing. It appears that the script information is insufficient to resolve all the text ambiguities. Additional types of choices remain unaccounted for. Let us briefly illustrate them.

(12) Will you please start working on the project?
(13) Bud’te doby, nachnite rabotu nad proektom.
(14) Ne mogli by vy nachat’ rabotu nad proektom?

If (12) is uttered by a boss in a conversation with a subordinate, it should be translated into Russian as (13); if it is uttered in a conversation between a homeowner and a reluctant housepainter, it should rather be translated as (14). This example highlights the influence of the *speech act* character of the utterance (an order in the first case; a plea in the second) on the representation of its meaning and, therefore, its eventual translation.
Figure 1. An example of a generation dictionary structured as a discrimination net.

(15) (a) Chelovek voshel v komnatu.
(b) V komnatu voshel chelovek.
(c) Voshelel voshel v komnatu...
(d) Voshelel v komnatu chelovek.
(e) V komnatu chelovek voshel.
(f) Chelovek v komnatu voshel.

(16) (a) The man came into a/the room.
(b) Into the room came a man.
(c) Came the man into the room...
(d) A man came into the room.
(e) It was a man that came into the room.
(f) It was into a/the room that the man came.

In a standard conceptual analysis system all six Russian sentences in (15) will be assigned the same meaning. The word order permutations are, however, significant in that they contribute to the meaning of the sentence. (16) lists the English translations of the nonemphatic readings of the sentences in (15). The sentences in (16) differ in what is considered (by the speaker) already known and what is considered new information in them. Establishing these distinctions is known as thematic analysis of text or the functional sentence perspective. Note that, while in Russian these distinctions are
marked by word order, in English word order is accompanied by the choice of indefinite and definite articles (indefinite articles typically introduce noun phrases that are new). Thus, we find that English articles do, after all, carry a meaning. Note that in (15a) the new information can be either the prepositional phrase \( v \ komnatau \) or the entire verb phrase \( voshel v \ komnatau \). In the former case the indefinite article must be used; in the latter, the definite.\(^3\)

Sometimes it is difficult or impossible, while processing a text, to evoke a standard script or even a more general memory organization packet (MOP; Schank, 1982) that relates texts to typical abstract settings and events remembered from past experience. But it may help enormously if just the subworld to which the text belongs can be determined. Thus, (17) (from Anderson et al., 1977) will be translated in two clearly distinct manners depending on whether the text belongs to the subworld of jail or prizefighting. Of course, it is not an easy task to detect the subworld automatically.

(17) Rocky slowly got up from the mat, planning his escape. He hesitated a moment and thought. Things were not going well. What bothered him most was being held, especially since the charge against him had been weak. He considered his present situation. The lock that held him was strong, but he thought he could break it.

One must also have means of understanding elliptic constructions. Ellipsis is an ordinary and necessary feature of all input texts, not an aberration. Thus, one has to be able to understand (18) if it appears in a text after (19). Special types of knowledge and processing arrangements are necessary for this task.

(18) Six, to be precise.
(19) There are several flights from Atlanta to Pittsburgh on Tuesday.

The problem of anaphoric reference also involves a number of knowledge-based choices. Knowledge is needed for the computer to be able to find the referents for \( there \) and \( they \) in (20), as well as the beginning of the list of problems referred to there. Note that one cannot in the general case translate (20) without understanding, that is, simply using corresponding pronouns in TL. This is because in some languages additional choices have to be explicitly made. For example, translating (20) into Hebrew, one will have to make a choice between the masculine and the feminine gender form of \( they \). In order to do this, one has to determine the anaphoric referent for the pronoun. Anaphoric phenomena cover not only pronouns but also definite noun phrases. In order to translate (21) (the example is from Brown and Yule, 1983, p. 56) one has to understand that all the italicized noun phrases in this text refer to one concept token. Indeed, a

\(^{3}\) The remaining ambiguity in (15f) is of a different origin: the new information in it is that the man came into a/the room and not into some other place.
straightforward translation of this passage into, say, Russian would be difficult. Thus, a human translator would not use *ispamec* (Russian for *Spaniard*) to translate the noun phrase in the last sentence. It will be similarly quite difficult to render in Russian the indefinite noun phrase *a man...* if complete understanding of coreference relation in this text is not achieved.

(20) There they found many additional problems.

(21) *Priest is charged with Pope attack (Lisbon, May 14)*

A *dissident Spanish priest* was charged here today with attempting to murder the Pope.

*Juan Fernandez Krohn, aged 32, was arrested after a man armed with a bayonet approached the Pope while he was saying prayers at Fatima on Wednesday night. According to the police, Fernandez told the investigating magistrates today he trained for the past six months for the assault. He was alleged to have claimed the Pope ‘looked furious’ on hearing the priest’s criticism of his handling of the church’s affairs.

If found guilty, the *Spaniard* faces a prison sentence of 15-20 years.

*The Times, May 15, 1982.*

It is approximately at this level of understanding of the input text meaning that the number of legal readings becomes comparable with that recognized by human translators. The state of the art in the MT field is such, however, that no actual or even experimental MT system at present can detect all or even much of semantic, contextual and rhetorical meaning.

Potentially, however, there are additional levels of sophistication that can be added to an MT system. For instance, style considerations will often be important during the generation stage of the operation of an MT system. The study of these factors (one can call it computational stylistics) has not yet fully gained the attention of the research community. Natural language generation as such has been extensively studied in the context of dialog, question-answering systems (most often, to generate computer responses in the natural language front ends of database systems). The general problem of generation of texts has been studied to a much lesser degree.

Another example of a possible avenue of research which is still *terra incognita* is the study of the craft of the human translator as an expert task. Translators are experts. In this age of expert systems it is surprising that no thought has been given to studying the process of translation from this point of view. Some difficulties arise, of course. One important question is, whether the expertise of the translator transcends the knowledge of the two languages and the knowledge of the subject area of the translation. If the answer is yes, then the expert system approach may prove interesting, because it will provide a way of extracting this additional knowledge and learning to use it in a computer system. Naturally, it may prove impossible to extract that knowledge efficiently. It seems, however, that such knowledge exists: we intuitively believe that experienced translators know
how to do translations better than novices. This may have to do with their
ability to express their understanding of the SL text better. Indeed, it is the
case that the translation agencies consider it more important for a (human)
translator to know TL better than SL.

1.2. The choices

In recent years the MT activity in the U.S., Europe, and Japan has
markedly intensified. Many new projects appear on the map, and a number
of existing projects both amplify their activities and attempt to change to
newer models of MT research. In each such case there are a number of
strategic and tactical decisions to be made with respect to the nature of the
MT system(s) that will be the result of proposed research and development.
In this section we will briefly discuss the spectrum of possibilities in MT
research.

The deeper the desired level of analysis, the more difficult it is to achieve
and therefore (given the state of the art in the field) the less feasible it is at
present to build an MT system that would benefit from that depth of analysis.
Indeed, even syntactic parsers and grammars of sufficient generality cannot
be taken off the shelf and used without major modifications. Semantic
analyzers are scarce and provide at best a partial coverage of the world of
concepts necessary for translation. Modules capable of analyzing the rhetor-
ical content of an utterance or, say, its discourse structure are even more
remote. A significant amount of research is currently underway on the
problems of semantic and pragmatic analysis of natural language. But this
research is predominantly theoretical at the present time.

Is it possible to simplify, or avoid having to produce, systems of complete
automatic analysis and still achieve tangible results in MT? The answer is a
qualified yes. One can do it, at the expense of decrease in both the linguist-
ic and conceptual coverage and the degree of automation of the translation
process.

There are two major avenues of circumventing the problem of completely
automatic disambiguation. First, one can restrict the grammar and the
vocabulary of the input text in such a way that most of the ambiguity is
eliminated. This is the sublanguage, or subworld, approach to MT. Second,
one can drop the requirement of complete automation and allow humans to
get involved in the translation process. As we will see, there are a number
of ways in which this process of machine-aided translation can be organ-
ized. The difference between these approaches is not only in the tactics of
interspersing automated and manual steps in the process of translation, but
also in the nature of the subtasks for which humans are responsible.

Those who contemplate building an MT system must weigh the particular
requirements in terms of quality, allowed development time and breadth of
coverage of their projects before deciding what level of automation is the most appropriate for them. A simplified rule of thumb is: the less time allowed for research and development, the shallower the analysis module and, therefore, the deeper the involvement of humans in the translation process.

1.2.1. Restricting the ambiguity of SL text

The star example of the sublanguage approach is the operational MT system TAUM-METEO, developed at the University of Montreal and delivered to the Canadian Weather Service for everyday routine translations of weather reports from English into French. The system operates very successfully, practically without human intervention. Its vocabulary consists of about 1,500 entries, about half of which are place names. The syntactic constructions that occur in the variant of English used as SL in TAUM-METEO are a very proper subset indeed of the set of English syntactic constructions. There is very little semantic ambiguity in the system, since it is always expected that ambiguous words are used in that of their meanings which belongs to the subworld of weather phenomena. This project is further described in Chapter 2 of this book. Finding such well-delineated, self-sufficient and useful sublanguages is a very difficult task in general.

1.2.2. Partial automation of translation

The demand for MT is high and growing. Most of it occurs in subject areas whose corresponding sublanguages are much richer, and, consequently, less feasible, than that of weather forecasts (as the Montreal group quickly learned when they tried, with quite limited success, to extend their system to the subject area of aircraft maintenance manuals). To deal with the demand for automation of translation today one has to think of ways of using the knowledge and know-how already available in the field to provide a certain degree of automation of the translation process. Practically all operational and experimental MT systems require some human involvement, and it is safe to say that, in the immediate future, this involvement will decrease but not disappear completely.

There are two major classification dimensions for MT systems featuring partial automation. They may be classified according to

- the actual share of work performed by the computer (the degree of automation);
- the strategy of human involvement: whether the humans work on the text before, during, or after the computer deals with it (there is, of course the possibility of combining some or all of these three strategies);
In accordance with the degree of automation involved, MT systems range from relatively simple interactive editing and dictionary look-up tools for the use of human translators (this type of activity is known as *machine-aided human translation*, MAHT) to quite sophisticated systems, most of them still experimental, that involve syntactic and sometimes even semantic analysis and provide a much higher level of automation of the translation process (these systems perform *human-aided machine translation*, HAMT). The achievement of fully automated good-quality translation remains a distant but, for some projects, persistent goal.

Examples of MAHT systems are various products from such companies as LOGOS, Weidner and ALPS. The philosophy of this approach is best expounded by Alan Melby (cf. Chapter 9 of this book). He recognizes three levels of interaction between man and machine in developing what he calls a translator workstation. Level One Workstation, the least sophisticated one, essentially presupposes a complete and convenient word processing environment with convenient means of accessing on-line dictionaries and encyclopaedias. Level Two Workstation adds the spelling checks, concordances and text dictionaries and presupposes that the text to be translated is in machine-readable form. Level Three Workstation involves a degree of automatic processing, possibly including some analysis. Melby is not specific about what particular means are available at this level, but it is clear that such a workstation is somewhere in between MAHT and HAMT.

With respect to the strategy of human involvement, the three basic possibilities have come to be known as *pre-editing*, *post-editing* and *interactive editing*. A human pre-editor reads the input text and modifies it in such a way that the MT system is able to process it automatically. Difficult and overly ambiguous words and phrases are replaced with those that the editor knows the program will handle. A human post-editor, conversely, obtains the output from an MT system and eliminates all inaccuracies and errors in it. An interactive editor engages in a dialog with the MT system, in which the human resolves ambiguities that the machine is not capable of resolving itself. It is, of course, necessary to build a special interface to maintain this kind of dialog. In it, the types of questions asked can vary: the computer may ask the human to provide a TL correlate for an ambiguous SL unit or may ask to be provided with the meaning of an SL unit, in the language in which the meanings are represented in the system.

Interactive editing is among the topics discussed in Chapter 5 of this book. It is suggested there that the dialog can include interactions as follows:

The word 'pen' means:
1) a writing pen
2) a play pen
NUMBER >>
To resolve referential ambiguity, a system can ask in the following manner:

The word 'she' refers to
1) 'Cathy'
2) 'my mother'
3) 'the sailboat'

The interactive word sense disambiguation can indeed be accomplished relatively easily. The referential ambiguity, though, will present a problem, because the program will have to be able to find the candidate (pronominal and other) referents, which is a non-trivial task in itself. The design of the interactive component to perform syntactic disambiguation may be difficult and the component itself cost-ineffective, simply because it is not an everyday task for a human to compare syntactic structure trees. Semantic analysis, however, has a stronger disambiguating power and, therefore, syntactic disambiguation can be rendered unnecessary in an interactive editing system that relies on human intervention to choose the appropriate word senses.

Until very recently human intervention came almost exclusively in the form of post-editing, whereby the product of an MT system is submitted for editing by a human editor, exactly as human translations are in better translation agencies. The important feature of this approach is that the posteditor is not required to know SL. In practice, however, many of the outputs of such systems are so garbled that it becomes a major problem to edit them without dipping into the SL text.

In the systems that espouse the HAMT strategy with post-editing (and a majority of current experimental systems belongs to this group) feasibility and cost-effectiveness become the major criteria for success. The post-editing approach is based on the premise that MT can (and should) be performed without a complete understanding of SL texts by the computer. This belief is justified in terms of feasibility. What this approach means is that an MT system is essentially an aid in human-controlled translation.

The quality and depth of the disambiguation process, as determined by the quality of underlying conceptual models of language and world used in an MT system, is an alternative criterion for judging MT systems. If one accepts the position that the nature of human involvement should be not in correcting the (erroneous) texts produced by the system but rather in providing the system with additional disambiguation knowledge that was not recorded in its knowledge base, one becomes able to judge an MT system in terms of this latter criterion. The methodological basis for this approach is interactive editing, with the dialog aimed at gaining disambiguation knowledge, not the actual TL correlates that will be eventually obtained based on this knowledge. This approach is compatible with the research strategy of gradual movement toward fully automated translation and,
therefore, its success depends on significant advances in basic research. The former approach is more of the engineering variety in that it aims at partial automation within the realm of what is feasible today. The tension between proponents of these two approaches enlivens the MT research scene of the 1980s.

1.3. About this book

This book can be naturally divided into six parts:

Part I (Chapters 1 and 2): 'The State of the Art in Machine Translation'
Part II (Chapters 3 and 4): 'Machine Translation and Linguistic Theory'
Part III (Chapters 5 through 9): 'Methodologies for Machine Translation'
Part IV (Chapters 10 through 12): 'Machine Translation and Artificial Intelligence'
Part V (Chapters 13 and 14): 'Research Tools for Machine Translation'
Part VI (Chapters 15 through 17): 'Case Studies of Machine Translation Projects'

A brief outline of these parts and their constituent chapters follows.

Part I sets the stage for all the subsequent discussion. This chapter introduces the problems of MT and presents the variety of research and development avenues that the workers in the field can opt for.

Chapter 2, Current strategies in machine translation research and development, by Allen B. Tucker, summarizes the situation in the field today. It starts with a description of the methodologies that have been used in MT over the years. The direct translation strategy that sought to juxtapose the elements of SL and TL with very little analysis is discussed first. The transfer approach, which involves more analysis steps than the direct approach, makes the connection between SL and TL more indirectly, by comparing, typically, both lexical units and syntactic structures of the two languages. Finally, the interlingua strategy presupposes the connection between the languages of translation via a language-independent representation. Chapter 2 goes on to discuss the types of analysis modules employed by current MT systems, the problem of the choice of translation domains and the important question of evaluating performance of MT systems.

The rest of Chapter 2 is devoted to brief surveys of a number of representative MT systems, both operational and experimental. The seminal early work at Georgetown University is evaluated, together with the systems, like SYSTRAN, based on that effort and built later. The TAUM-METEO operational system is then described, and the METAL system developed at the University of Texas. The section on MT projects currently in the research stage includes entries devoted to the EUROTRO project of the European Economic
Community, the Japanese government MT project, the SUSY project at the University of the Saarland in Germany, the DLT (Distributed Language Translation) project in Utrecht, The Netherlands, and the TRANSLATOR project at Colgate University, U.S.A.

Part II of this book discusses the influence that the study of linguistic theory must have on the research in such an applied area as machine translation. Theoretical results in linguistics are known not to be directly transportable into applications. Many workers in MT (as well as other computational linguists) made repeated attempts to import the theoretical constructs and the way of reasoning about language developed within various linguistic theories, notably the transformational grammar paradigm and the systemic grammar paradigm. Such attempts have been only partially successful. As a result, computational linguists set out to develop their own views of language and language use (which, naturally, stressed the discovery, parsing, procedures and, more recently, the reconstitution, generation, procedures for meanings of utterances in natural languages).

Chapter 3, *Linguistics and natural language processing*, by Victor Raskin, makes a case for the indispensability of using linguistic theory and especially the expertise of a well-trained linguist for any application that involves analysis of natural language. It describes the spectrum of phenomena in morphology, syntax and semantics that the linguists have knowledge about and explains why it is rarely possible to apply this knowledge without modifications in systems for automatic language processing. A special discussion is devoted to the necessity of developing an explicit theory and methodology of linguistic applications. MT is taken as one example of a linguistic application, and specific problems in translation are discussed. Another linguistic application illustrated in this chapter is that of the study of sublanguages. This last topic is the bridge to a more dedicated discussion in Chapter 4.

Chapter 4, *The significance of sublanguage for automatic translation*, by Richard I. Kittredge, is devoted to the following three questions:

What is sublanguage?
Why is sublanguage analysis important for automatic translation?
How can a translation system take advantage of sublanguage properties?

The discussion of the first question is necessarily concise, because it is only recently the study of sublanguages has become a focus of attention in the field, and no definitive theory of sublanguages has been proposed as yet. This chapter is a step in that direction. The chapter describes three examples of sublanguages: the language of weather reports, the language of stock market reports and that of aircraft maintenance manuals. It proceeds to discuss the influence of the size and type of sublanguages on the nature and
computational tractability of various processes of automatic analysis and dictionary augmentation and maintenance.

Part III of this book is devoted to an in-depth discussion of the methodologies for MT research. The emphasis here is not on the well-established approaches and techniques but rather on the new opinions and proposals that have not yet been fully implemented but represent the way MT will be done in the years to come. It is safe to say that the concerns and suggestions highlighted in this part will remain prominent in the field in the year 2000 and beyond.

Chapter 5, Knowledge-based machine translation, the CMU approach, by Jaime G. Carbonell and Masaru Tomita, starts by surveying the variety of MT paradigms in existence and goes on to suggest that the most appropriate directions for practical MT research are the knowledge-based and the interactive strategies. Each of these is discussed in some detail, and the particular proposal for an MT system at Carnegie-Mellon University is presented.

Chapter 6, The structure of interlingua in TRANSLATOR, by Sergei Nirenburg, Victor Raskin, and Allen B. Tucker, describes the design of the knowledge representation medium used for representing concepts and assertions in the subworld of the knowledge-based MT project TRANSLATOR. The chapter’s main methodological thrust is the opinion that it is unrealistic to expect to achieve good-quality translation in a system with shallow analysis. It is claimed that decision-making during translation must be based on a large number of meaning components. Thus, the knowledge to be extracted from the input text must include not only the syntactic and compositional-semantic data, but also information about the discourse structure of the input text, its speech-act character, thematic relations, and more.

Chapter 7, Basic theory and methodology of EUROTRA, by Doug Arnold and Louis des Tombe, describes the methodological position of the EUROTRA MT project. The methodological approach described in this chapter is quite different from that presented in Chapter 6, in that it seeks to eliminate the need for an involved analysis of the input text. EUROTRA is a multinational effort, in which a number of research groups in different EEC member states are responsible for producing the grammars of their (usually native) languages and the bilingual dictionaries from these languages to the rest of the EEC languages. Logistically, it is a very ambitious enterprise and, therefore, its results are expected to have a strong impact on the state of the field of MT. This chapter presents results of the work in the ‘kernel’ group of EUROTRA and concentrates on the overall methodology (the four levels of specificity in MT research) and the particular proposal for the grammar formalism to be used by all the various teams within the project.
Chapter 8, *Machine translation as an expert task*, by Roderick L. Johnson and Peter Whitelock, contains further methodological observations concerning the current non-feasibility of the fully-automated high-quality automatic translation. A case is made here for studying and eventually simulating the behavior of a human translator in an automatic translation system. Specifically, the concept of *contrastive knowledge* that the translators have about the two languages they work with is highlighted. It is claimed that there is more to this concept than the information one obtains in a typical bilingual dictionary. The chapter goes on to discuss the distribution of work between people and machines in MT and concludes by proposing improvements to existing interaction strategies.

Chapter 9, *On human-machine interaction in translation*, by Alan Melby, discusses the methodology for translation systems at the lower end of the automation scale. He considers the ways of enhancing the technology and methodology of human-machine interaction and ways of their application to the specific problem of MT. Some of the possibilities here were mentioned in 1.2 above.

Part IV of this book deals with the new horizons for automation in MT. A sampling of conceptual problems in the field of artificial intelligence (processing ill-formed natural-language input; analyzing the discourse structure of a text; generating a natural language text from speakers' intentions and conceptual representations) is presented here, together with some current ideas about the ways of solving them. If MT research is not facing these problems in their entirety now, it will have to tackle them before long. Therefore, it is very important for a person contemplating an MT effort or thinking about augmenting an existing one to be aware of the work done in AI.

Chapter 10, *Reflections on the knowledge needed to process ill-formed language*, by Ralph M. Weischedel and Lance A. Ramshaw, makes the case that processing ill-formed language requires contributions from morphological, syntactic, semantic and pragmatic knowledge. The particular kinds of knowledge required from each area are discussed, as well as the nature of the problem of combining those multiple knowledge sources. Various recent systems have made good starts at handling ill-formed input using knowledge from some of the above areas, but a substantial amount of fundamental work must still be done if the AI systems are to understand language as robustly as humans do. This chapter concentrates on the pragmatic knowledge necessary for understanding information-seeking dialogs. Based on this discussion, the chapter offers important perspectives on the knowledge and architecture needed in any task that involves understanding natural language.
Chapter 11, *An integrated theory of discourse analysis*, by James Pustejovsky, is devoted to an increasingly intensive area of research in computational linguistics that deals with the laws of the rhetorical structure and meaning of texts. A number of difficult questions related to various specific topics in discourse analysis are discussed in this chapter. Critical surveys of existing approaches are followed by suggestions with respect to partial solutions of certain discourse analysis problems. Thus, the chapter argues that, in order to attain a sufficient depth of discourse analysis, one must distinguish between the syntactic and semantic components of the rhetorical meaning. The chapter also discusses CICERO, a knowledge-based system for discourse analysis that incorporates some of these theoretical ideas.

Chapter 12, *Natural language generation: complexities and techniques*, by David D. McDonald, describes the progress in this area over the last fifteen years. It discusses the conceptual differences between the tasks of natural language comprehension and generation. Next, it analyzes the popular misconception (of which MT workers are also typically guilty) that generation is a simple process. It then goes on to discuss the succession of generation techniques, from the early direct replacement paradigm to the modern grammar- and knowledge-oriented approaches. It concentrates on the multi-level description-directed generation approach used in the Mumble generation system. The chapter concludes with some thoughts about the relevance of the current generation research in AI to the task of machine translation.

Part V is devoted to more practical problems than the other parts of this book — the ways to establish and maintain an optimum research environment for an MT project and the use of information resource tools that can be made available for an MT project.

Chapter 13, *The research environment in the METAL project*, by John S. White, describes the working environment that has been developed over the years at the Linguistics Research Center at the University of Texas. It discusses the difference of the software maintenance problems in small experimental projects and large, production-oriented MT efforts, such as METAL. The importance of software maintenance in the latter type of projects becomes infinitely higher, with the accumulation of language and world knowledge. The chapter describes in some detail the software development tools in METAL, starting with those facilitating development of grammars and continuing to those used for augmentation and maintenance of dictionaries. It also describes the translation tools, such as the sophisticated trace and statistics packages that help to test the system ‘in action.’

Chapter 14, *Knowledge resource tools for accessing large text files*, by Donald E. Walker, provides an overview of a research program under development at Bell Communications Research. The objective of this
program is to develop facilities for working with large document collections in order to provide more refined access to the information contained in these 'source' materials than is possible through current information retrieval procedures. The tools being used for this purpose include machine-readable dictionaries, encyclopaedias and other sources of knowledge. A distinction is made in the chapter between the texts (sources) and tools (resources), and the relationship between them is discussed. Two specific systems (both under development) are described as illustrations of this overall approach: one aiming at extracting the overall subject from the text; the other, at concept elaboration while reading text. The chapter concludes with a discussion of implications of this type of research for MT.

The final part of the book, Part VI, contains Chapter 15, The role of structural transformation in a machine translation system, by Makoto Nagao; Chapter 16, An experimental in lexicon-driven machine translation, by Richard E. Cullingford and Boyan A. Onyshkevych; and Chapter 17, Integrating syntax and semantics, by Steven L. Lytinen. This part will allow the workers in MT to make use of the experience accumulated in these three projects with respect to the implementation of various theoretical proposals as well as practical 'everyday' problems that inevitably pop up in developing large software systems.

Chapter 15 describes the state of the MT project at Kyoto University in Japan, which is one of a significant number of projects supported by the Japanese government. It describes the experience of building an actual MT system that is based on the transfer methodology and is augmented by semantic processing. It goes on to show how the basic transfer approach has to be augmented by the pre-transfer and post-transfer processing 'loops,' and how a phrase structure transformation component has to be added to the process of generation. The chapter presents a number of examples of Japanese-to-English translation.

Chapter 16 describes a system of automatic translation from Ukrainian into English in which the focus of processing is at the level of the lexicon, rather than the grammar. The approach expects the analyzer to map SL language units into an interlingual form, which then is mapped into the surface structures of the TL (optionally, after a certain amount of inferencing activity takes place to annotate the interlingual form). The particular representation language used for describing the world knowledge in the system (ERKS, for Eclectic Representations for Knowledge Structures) is presented and illustrated.

Chapter 17 argues that one has to use both syntax and semantics in the analysis stage of an MT system and suggests an approach to this type of integration. It provides a good operational example of how one can reconcile positions that can seem to some to be quite opposite. Indeed, the early
conceptual analyzers attempted to disregard as much of the syntactic knowledge as was possible without hindering the extraction of meaning. At the same time, some of the natural language processing (including MT) paradigms disregarded semantics altogether. This paper argues for the integration of syntactic and semantic processing in such systems, while maintaining a separate body of syntactic (a grammar) and semantic (a world knowledge base) knowledge. An experimental MT system is described that performs this type of analysis, including even the construction of a separate syntactic representation during the parsing process.