Book Reviews
SPEECH COMMUNICATION JOURNAL
Issue 32(2), October 2000

Translation Engines: techniques for Machine Translation

Arturo Trujillo, Springer-Verlag, Applied Computing, Heidelberg
ISBN 1-85233-057-0, 1999

Reviewer : Malek Boualem
France Télécom R&D
2, avenue Pierre Marzin - 22307 Lannion - France

Except for W.J.Hutchins books, no other book gives such a wide overview of the major concepts of machine translation (MT) and machine-aided translation (MAT). Indeed this volume of 300 pages presents (almost all) the major translation approaches argued with linguistic and technical implications for each kind of application. Moreover neighbouring concepts that are spontaneously involved in MT/MAT are introduced such as multilingual text processing, formal language theory and artificial intelligence. The audience is well defined by the author in the preface. This book is useful for anyone who is interested in translation. Linguists, computer specialists and professional translators can find useful resources in this book for their research, development, studies and other activities. Moreover the author's writing style is quite fluent making the key concepts accessible to non-specialist readers. Note that the author does not present a new approach to MT/MAT in this book, his own approach is mainly descriptive and less suggestive. Four parts containing ten chapters are intelligently arranged allowing the reader to go through them progressively. A conclusion and a list of complementary references end each one of the ten chapters.

Part 1, entitled "Background", contains two chapters. Chapter 1, entitled "Introduction", includes a very brief history presenting the main phases of MT and MAT: initial efforts (1940s and 1950s), ALPAC report (1966), first operational systems (1970s), first commercial systems and translation on personal computers (1980s and 1990s), and translation on the internet. The main strategies for machine translation are also listed: direct, transfer and interlingua. A short section about artificial intelligence ends this chapter. Chapter 2, entitled "Basic terminology and background", starts with a section that introduces some linguistic notions that are relevant to MT/MAT: phonetics and phonology, lexicon, morphology, syntax, semantics, pragmatics and stylistics. A second section of this chapter gives some formal theories in mathematics and in computer science that are relevant to computational linguistics: formal language theory, automata theory, dynamic programming, probability and statistics, and Hidden Markov models. This chapter ends with a review of Prolog, on which much of the notation used in the book is based.

Part 2, entitled "Machine-Aided translation", also contains two chapters. Chapter 3, entitled "Text processing", actually has a strong relevance to MAT and to multilingual translation. It introduces important topics in text processing that are relevant in translation of languages that are not based on the Latin script. Text format preservation in translation is discussed in this section, especially for HTML documents. Chapter 4, entitled "Translator's Workbench and Translation Aids", includes a section on Translator's Workbench (TWB). Here the author lists a variety of features and tools that are suitable for TWB however he did not discuss their availability and realistic use, especially for non Latin-based languages. Moreover the author did not give enough information about workflow in translation. A section on Translation Memory (TM) includes an interesting discussion (also useful for information extraction) about similarity measures between sentences to be translated and sentences stored in the TM. Finally bilingual and subsentential alignment is discussed and substantial methods for sentence, word and terminology alignment are presented.

Part 3, entitled "Machine Translation", is the major part and contains four chapters. The first one presents standard computational linguistics techniques for the analysis and the generation of natural language sentences, with an outline for their implementation. The three other chapters present the main MT techniques. Chapter 5, entitled "Computational linguistics techniques", includes a first section on computational morphology and the two-level model. A second section on syntactic analysis gives pointers to some approaches and presents short descriptions of syntax aspects: sign structure, agreement, complement structures, unbounded dependency constructions, relative clauses, etc. A section on parsing presents a chart-parsing algorithm to derive syntactic and semantic analyses according to a grammar. And a last section on generation presents general aspects of natural language generation (NLG) and a short description of the Semantic Head-Driven generation approach used in MT. Chapter 6, entitled "Transfer machine translation", presents classic translation problems and the way they are handled in transfer-based MT systems. This chapter describes three different types of transfer systems. A first section describes the syntactic transfer MT and includes a fine classification of translation divergence argued with numerous examples from English and Spanish. A second section describes the semantic transfer MT based on the QLF semantic representation (Quasi-Logical Form), which is a representation of language meaning based on predicate logic. A third section describes the lexicalist MT by highlighting its non-recursive character compared to the syntactic and semantic transfer methodologies. This chapter ends with a comparative discussion about these three major MT transfer strategies. Chapter 7, entitled "Interlingua machine translation", describes two approaches to multilingual MT that are based on a language-neutral representation of sentence meaning: interlingua-based MT (also known as pivot language-based MT) and knowledge-based MT (KBMT). The interlingua MT approach is described through the Unitran English-German-Spanish MT system (B.J.Dorr). A common representation language, called Lexical Conceptual Structure (LCS, R.Jackendoff) is described in this section. I found that the analysis process into LCS is described in detail while the generation process from LCS is underrepresented. The Knowledge-Based MT approach presented here follows the ideas proposed and tested by Sergei Nirenburg. A knowledge representation framework is presented in which the domain model is described as an ontology. Analysis and generation processes are also described in this section. Finally, translation divergences are briefly discussed for both approaches and a comparative discussion ends this section. What I think is missing in this chapter about the interlingua MT approach, is a reference to the ARIANE system and to the important R&D activities in interlingua MT at the GETA research center (Groupe d'Etudes en Traduction Automatique, France). Moreover, although NLG has been described in section 5.5, it would have been rather suitable to discuss NLG applicability to the interlingua MT in this place. Chapter 8, entitled "Other Approaches to MT", describes four other MT approaches that illustrate alternative solutions to various problems in translation. Two of these approaches are classified as corpus-based approaches: Example-Based MT (EBMT) and Statistical MT. They rely on large amounts of bilingual corpora to achieve translation. The two other approaches described here are classified as rule-based approaches: Minimal Recursion Semantics (MRS) and constraint-based approach. I think that although Mel'čuk's works are cited in the book, a wider discussion of paraphrases and lexical semantics would have been desirable in the previous two chapters.

Part 4, entitled "Common issues", is the last part and it discusses two common issues in MT: disambiguation and evaluation. Chapter 9, entitled "Disambiguation", discusses disambiguation in the analysis and transfer stages of MT. Categorial (part-of-speech), structural, lexical and transfer ambiguities are considered. A first section presents the POS tagging of words together with some tagging models such as the stochastic and the rule-based tagging (E.Brill). A brief discussion about designing taggers is also introduced. A second section discusses the disambiguation of syntactic analysis. Methods for processing structural ambiguities are presented such as psycholinguistic (heuristic) preferences and probabilistic context-free grammars (PCFGs). A third section underlines some techniques for word sense disambiguation that are particularly useful for interlingua MT: selectional (or sortal) restrictions, frames and semantic distance and corpus-based approaches. A last section presents some techniques for transfer disambiguation. Chapter 10, entitled "Evaluation", is the last chapter of the book. It discusses strategies for assessing the translation quality of a system as well as its cost-effectiveness. The author presents a variety of concepts and a number of strategies for evaluating translation software. The first part lists the different groups interested in translation evaluation and who are involved in the creation, deployment, use and maintenance of translation software. The second part presents a variety of common evaluation issues.

The author ends the book with an interesting conclusion giving some trends in machine and machine-aided translation. An appendix of useful and relevant resources on the Internet is also given. These resources concern a variety of aspects related to machine and machine-aided translation: text processing, internationalization and localization, character sets and fonts, input methods, computational linguistics, machine and machine-aided translation, disambiguation and evaluation. This book has a very rich list of references covering most of the works, during the last years, that are related to computational linguistics and to machine and machine-aided translation.

Finally I would like to point out some minor editorial mistakes:

- Page 24: F(d) is given by : min(F(b)+12, F(c)+9), instead of min(F(b)+12, F(c)+10)
- Page 24: F(x) = min(F(y1)+d(y1,x) , F(y2)+d(y2,x) …, instead of F(y2)+d(y1,x) …
- Page 47: The ASCII standard defines 128 codes, not 127 codes as written.
- Page 53: For Arabic glyphs, the term "independent" is more commonly used than "on its own".
- Page 72: Subsection 4.3.1 exists but not subsections 4.3.2, 4.3.3, …
- Finally, the author frequently uses initials and abbreviations of concepts instead of literal textual forms; thus the reader has to frequently access the index.

I end my review by saying that this book is excellent and I highly recommend it to anyone who is interested in machine and machine-aided translation.

Dr Malek Boualem
France Télécom R&D