[*MT Summit III*, Washington, DC, USA, July 1991]

The METAL System. Status 1991

Thomas Schneider
Siemens Nixdorf Information Systems
Otto-Hahn-Ring 6
D-8000 Munich 83, Germany

Abstract

The METAL system which originally evolved from a cooperation between the University of Texas and Siemens became a product in 1988. METAL is implemented on multi-user workstations with a LISP server in the background. It is integrated into the office environment and permits automatic deformatting and reformatting of documents. METAL is characterized by recursive grammars, best paths parsing and a modular lexicon structure. Recent changes in system design have focussed both on internal structure and on user interface. Experiences with productive use have proven METAL's cost-effectiveness but have also shown the need for increased cooperation between developers and end-users.

1 Introduction

The METAL machine translation system, originally evolved from a cooperation of the University of Texas at Austin and Siemens, existed as a prototype for German to English in 1978. After another ten years of development it finally reached the level of a viable product in 1988, As any machine translation system of today's state of the art, METAL is not intended for the impeccable translation of literary texts but for the rapid dissemination of information of a factual nature, i.e. the types of text in which style is of less significance than the explicit information content, which have to be translated fast and are of sufficient volume. METAL is used for two distinct applications: the translation of documents for publication, and the translation of documents for purposes of information gathering. In both cases, METAL has been used successfully.

2 General System Description

Since the METAL system structure has been described in more detail previously ([Schneider 87], [Schneider 89]) the following outline will just summarize the main features.

2.1 Hardware

METAL is implemented on a hardware package consisting of a number of translator work stations (Siemens MX-300or MX-500) and a translation server in the background (Symbolics 36xx or XL). Translators work at a menu-driven terminal which supports various tasks: import of texts from other DP systems, automatic deformatting, administration of translation jobs, postediting, automatic reformatting etc. The actual translation process runs on the server in the background and does not interfere with the translator's work.

2.2 Grammar

METAL uses an augmented phrase structure grammar. The application of the individual rules is restricted by tests on features and values of the constituents and on their interaction. As they are recursively applied a limited number of these rules is able to analyze surface structures not explicitly stated in the grammar. This has proven to be a viable alternative to the (mistaken) assumption that a large number of single rules, each describing a specific surface structure, would eventually cover all legal utterances of a given natural language. Natural languages are infinite systems, and grammars have to allow for an infinite number of surface structures.

The prioritized chart parser operates on an agenda. Unlikely paths are eliminated via preferential weightings calculated from lexical and grammatical data. Still, in a source language with an almost free word order like German, the analysis of a complex sentence may require the comparison of several thousand interpretations. (This of course reduces any attempt to implement a sophisticated MT system on a small PC to absurdity). Transfer to the target language is attempted only after a sufficiently probable interpretation of the source language sentence has been reached. If no plausible interpretation of the complete sentence is found, a fail-soft routine is invoked which outputs a translation of the individual phrases that could be analyzed. Quite often, this output is still a grammatically correct translation which can serve as a basis for postediting.

2.3 Lexicon

METAL operates on monolingual lexicons and one transfer lexicon for each language pair. The monolingual lexicons contain morphological, syntactic and semantic information needed for the analysis and/or the generation of a language. The transfer lexicon provides a link from the source language to the target language, indicating

in which contextual environment and in which subject field the source language entry should point to a specific target language entry. The advantages of such a structure are obvious: the extensive grammatical information contained in the monolingual lexicon needs to be carried only once even if one lexeme in one of the languages corresponds to many different entries in the other language. This saves not only storage space but also a lot of coding effort. Moreover, the monolingual lexicons can be re-used in various language combinations,

Often, the "size" of the lexicon is considered to be a major criterion for the assessment of MT systems. However, the number of entries alone is not conclusive, as a system lexicon may contain full forms or stems, or derive all forms from lexemes (as METAL does). Another aspect is the way in which compound words and discontiguous phrases are handled. Are they generated by the grammar (as in METAL) or do they have to be carried as lexical entries? Moreover, the main application of machine translation is in subject-specific texts, not in general language or literary texts. In most European languages, the set of the most frequent 5000 words makes up approximately 90% given text (on the average). Beyond this limited set, the point of diminishing returns is soon reached. Increasing an undifferentiated general lexicon to more than 100000 words, for example, would not increase text coverage decisively. On the contrary, many unpleasant ambiguities would he introduced which can be avoided in a modular structure.

The METAL lexicon is organized as follows: There are modules for function words (FW), for general vocabulary (GV) and for common technical vocabulary (CTV) organized in a tiered hierarchy. From the next level down, users can define and structure their own modules and tailor them to their specific application. In other words, the classification of lexicon modules is flexible so that e.g. someone working in Pharmacy need not carry subject areas like Turning Machines which would be quite expendable in this case. For in-house application at Siemens Nixdorf, there are for example modules like Data Processing (DP) with submodules Software (SW), Hardware (HW) etc. Furthermore, it is possible to define transfers on the basis of a specific customer, a specific product or project or a specific target country. Thus a text translated for a recipient in Great Britain will show "boot" instead of the American "trunk", and a text intended for Spain will automatically have "ordenador" instead of the Colombian "computadora".

Before a translation run is started, the modules appropriate to the subject area of the text are indicated. If the syntactic and semantic criteria for the selection of a lexical entry are met and there are several candidates for transfer, then the one tagged for the subject area of the text is chosen. This assures that subject-specific transfers receive priority over more general transfers. Since no MT system can possibly contain all of the estimated 30-50 million (!) existing terms of the sum of all subject fields, it is imperative that an MT user is given an efficient lexicon interface. In METAL, users update their own lexicons with the aid of the INTERCODER, an integrated expert system. It predicts the morphological and syntactic behavior of new lexicon entries and generates the necessary coding; the missing pieces of information are inferred from a set of rules and from partial information already contained in the lexicon. The INTERCODER has proven its usefulness by reducing coding time by a factor of 10.

METAL users may code all word classes (outside of function words). Even though grammar rules are not accessible for the user, the lexicon permits significant syntactic transformations through the call of predefined macros, e.g. changing the roles of arguments, adding or deleting constituents etc. The design of this user interface required a lot of effort since great care had to he taken not to overburden translators with linguistic detail.

2.4 Office Environment

An operative productive MT system needs to do more than simply translate individual sentences entered from the keyboard. Most of the texts which have to be translated quickly and are of great volume such as technical documentation are heavily formatted. In some texts, more than half of the characters on a page may be non-translatable material, notably flow charts, diagrams, tables and various control characters for format and layout. It would be highly uneconomical to manually extract the text portions to be translated and afterwards manually re-input them. Therefore, METAL has been integrated into a chain of processes from text acquisition via automatic deformatting and translation to automatic reformatting procedures. A translation run usually goes through the following steps: Once the text has been received in machine-readable form, a set of programs on the SINIX system checks the pages for tables, graphs etc and generates a layout file of the page. The individual translation units, usually sentences, are automatically recognized, numbered consecutively and extracted from the layout file. They are written into a text file and transferred to the LISP machine for translation. After the translation run, the file containing the target language text is returned to the SINIX work stations for postediting. Afterwards, the text is automatically merged with the layout information contained in the page mask so that the target language text is available with all the formatting information and with the same layout as the original.

Before a text is translated, it is advisable to run a comparison of text and system lexicon. Analysis and transfer rely on grammatical information contained in the lexical entries, and if too many words of the text are missing translation quality will suffer. In METAL, such a lexicon check produces several files. One contains the words not found in the system lexicon (it also doubles as a spelling checker), one lists all the compounds not contained in the lexicon for which a translation was generated by the system. The third file is a text glossary providing source language term and target language term with the relevant subject field. These three files permit a quick decision as to how much effort is required to ready the system for a specific task.

3 Recent Changes in METAL

The addition of other language pairs as well as the experiences with productive use in various environments have led to several changes in system design and user interface. The original intention had been to build a language-independent system with a modular structure so as to utilize existing components in other language combinations. In designing new language pairs, it has become evident that while the intent was largely followed, a certain bias towards the languages under development was retained. Since language-specific adaptations of the core software get expensive (and difficult to maintain) a so-called METAL Interface Representation (MIR, described in [Thurmair 90]) has been developed. It will permit the direct linking of various source language analysis modules to various generation modules without modification, thus greatly reducing development cost.

As the number of METAL users grows, the need for the exchange of lexicon modules has also increased. The METAL Lexicon Interchange Format (MLIF) defines how conflicts between entries from different sources with different owners are resolved. This assures that lexicon modifications introduced by one user arc not overwritten by someone else's entries. In this context, lexicon entries have been compacted so as to require less storage space.

One of the most difficult areas in the development of a productive system is the integration into the office environment. We are faced with a multitude of incompatible word processing systems and DTP editors. Each of them insists on defining layout information and graphics differently. So while METAL in the past had been able to deal with a number of different formats it could by no means cover all of them. In consequence, potential users tied to existing hardware environments were hesitant about an installation. Fortunately, the Office Document Interchange Format (ODIF) has become a European standard. An interface between METAL and ODIF allows for the transport of all documents with all formatting information. Vendors of word processing systems and DTP editors will have to observe the ODIF standard and implement an interface. While filters to and from certain editors will continue to exist (e.g. Xerox Viewpoint, Interleaf etc) the main text import and export pivot will be ODIF.

4 Experiences

Besides the German to English system which has been operational for quite some time, German to Spanish, Dutch to French and English to German have been released. Other language pairs are still under development. There are at present approximately 25 installations, sometimes involving more than one language pair. Most of the users work in industry, trade and banking. The primary application is in the translation of documents for publication (which of course requires postediting of the machine output). Recently, however, there has been a marked increase in interest in the use of METAL for purposes of information gathering. Should this trend continue, additional features such as automatic recognition of subject area etc may need to be implemented. In most cases, the hardware configuration required to run METAL is not available at a customer's site. So the investment for a METAL installation has to include hardware cost as well. Cost/benefit analyses have shown that a METAL installation is cost-effective at a volume of 2000 pages per year or higher. But even below that threshold, other criteria may be significant: faster throughput, faster availability of information or consistency of terminology throughout all documents. In the beginning, many translators were sceptical about the use of an MT system (which may be due to the exaggerated claims and unfulfilled promises of the past). Our experiences however show that translators are quite willing to use METAL as a tool which reduces routine work but does not challenge their position as responsible intellectual head of the translation process. However, it is imperative to clearly state the limitations of machine translation so as not to create false expectations (for a user report, [Little 90]).

Machine translation today is not a technology that one can buy off the shelf, install and use effectively, manual in hand. Legitimate MT systems are highly complex and not self-explanatory. Our experiences with METAL installations show that one week of initial training is usually sufficient to enable a translator to operate the system. However, a lot of the questions arise during practical application. So it has become obvious that users often need additional help even beyond a second training course. Many translators do not have the linguistic background to immediately understand the interaction between grammar and lexicon coding. So several users have requested additional courses in this area, as well as in system administration.

For specific questions which may occur during productive use, a hotline has been installed. Even if it may not be possible to diagnose a possible software or grammar error immediately, at least an intermediate response may point out ways to circumvent the error. In the meantime, the METAL users have formed a user group (MAUS) which meets twice per year. Such an institution has proved most helpful for both users and developers. Change requests can be discussed with all users concerned and assigned a priority, and the danger that developers invent features which are not really needed is diminished. MAUS also provides a forum for users to exchange information about available lexicon modules, about tricks of the trade and ways to improve the operation. Certainly, the METAL system can and wilt be further improved and extended. One aspect however always needs to be kept in mind: Is it really worth it to achieve a marginal increase in translation quality if it means a sizeable investment of linguistic work? Is it desirable to issue a new release with perhaps noticeable improvements—if it means that all users have to recode their lexical entries? Such decisions can only be made jointly by the METAL team and MAUS. No MT system can evolve effectively without user input.

References

[Little 90] P. Little, “METAL - machine translation in practice”, in C. Picken, ed., *Translating and the Computer II.* London: Aslib, 1990.

[Schneider 87] T. Schneider, “The METAL System. Status 1987”, in M. Nagao, ed., *Proceedings of the Machine Translation Summit,* Tokyo: Omsha 1989.

[Schneider 89] T. Schneider. “The METAL System. Status 1989”, in Ch. Rohrer, ed., *Proceedings of the Second MT Summit,* Frankfurt: DGD 1989.

[Thurmair 90] G. Thurmair, “Complex Lexical Transfer in METAL”, in *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language,* University of Texas at Austin, 1990.