The METAL System. Status 1995

T.Schneider
Munich

1. MT Economics

The development of an operational machine translation system is not just a difficult but an exceedingly expensive task. The step from laboratory prototype to useful product is huge; actually, the existence of a miniature experimental system for a minimal segment of controlled language input is of no relevance. Problems of ambiguity, of over-generation or of combinatorial explosion do not arise until the heterogeneity of linguistic expressions in large samples of a language requires large grammars and large lexicons. Moreover, real-life texts do not always conform to the rules of school grammars - sometimes due to the sloppiness of the authors, sometimes because the sublanguage is characterized not only by a separate lexicon but also by a different syntax. Nevertheless, a productive MT system has to be able to handle a wide range of different types of text to be commercially viable.

And this requires huge investments.

Whenever in industry a decision is made about the development of a new product, a market analysis is required. In spite of its long history, machine translation is still a new technology, and there are very few objective and reliable figures available as to its potential market.
So on the one hand there is no certainty about economic success, on the other hand it is a well-known fact that innovation involves error and no certainty about technical viability. Thus business controllers in industry are generally very hesitant to approve machine translation projects which are likely to require many years of R&D effort by highly qualified and expensive specialists. One other aspect is also generally underestimated: marketing, user support, quality control and maintenance of system versions are at least as costly as the development itself.

Experiences with the development of METAL point out that the phase from zero to prototype constituted only 5% of the total product development cost. In the light of these figures, it is questionable if the present policy of giving public funding only up to the prototype level can entice competent industrialists to enter the field of machine translation.

However, the absence of machine translation systems for a multitude of languages may prove to be more costly than the granting of higher levels of funding for legitimate developers. For to cope with the problems of multilingual communication, with ever increasing volumes of textual information to be transferred across linguistic boundaries, there is no alternative to machine translation. Conventional means of translation would not be able to sustain information flow and knowledge transfer.

2.0 System Structure

The METAL system structure has been described in detail elsewhere (Schneider 1987,1989,1992; Thurmair 1990), so a summary of recent changes will be more suitable here.
While former versions of METAL were implemented on LISP machines with a multi-user UNIX front end, the system has now been ported to UNIX work-stations as servers, with smaller UNIX machines or PCs as clients. This permits a far better integration into standard office environments as well as a better distribution of tasks for the user. In implementing the client/server architecture, incrementally the individual components, originally written in LISP, were redesigned and re-written in C++, with a marked increase in performance, for some functions by a factor of 10.

2.1 Converter

An operative machine translation cannot content itself with the translation of individual sentences entered from the keyboard. One of the major application of machine translation is in the area of technical documentation, and here most of the documents are heavily formatted, strewn with graphics or tables. To avoid having to manually extract text portions and re-inputting them after translation, an automated deformatting and reformatting process is required.
The necessary programs are not simply filters. Filters will map the layout and formats of one DTP system onto another, losing some information in the process which cannot be represented in the target DTP system.
A converter has to preserve all the layout information of the original because the translated target texts ought to look exactly like the original.
It is not a trivial task to design such a program. The heterogeneity of desk top publishing systems and editors and the lack of standards, not to mention the internal changes from one version of a given DTP system to the next, do not permit the completely automatic treatment of random formats. Besides the (solvable) problems of words being deleted in the translation process which serve as anchors for graphics, there are also the excessively creative use of multiple layer graphics on the part of authors and other surprises. So the design of a converter is not a trivial task, and even though the program is not part of machine translation per se, no system would be operative without it.
During the last two years, additional converters were developed for METAL so that by now the most widely used DTP systems are covered. In some cases, individual customers with "exotic" office environments have developed their own solutions so that METAL could be integrated in the internal documentation processes.

2.2 Network Access

There are two distinct applications of machine translation, one for the dissemination of information in a foreign language, the other for the acquisition of information from foreign-language sources. The dissemination of information usually requires high quality which presupposes human intervention, in tuning the system lexicon as well as in post-editing the machine output. Target groups for this application are normally users with high volumes of documents within specialized subject fields,

as in technical documentation. This market is rather limited in relation to the effort necessary to produce a high-quality MT system.
In the past, the design of systems for the acquisition of information had not received the same degree of attention (at least in the civilian sector). By now it has become apparent that there is a large potential for MT use by non-translators, by casual users who come across a foreign-language text and need to know if the content is relevant for them.
To translate a twenty-page article by conventional means just to decide if its content might be of use is prohibitively expensive. By contrast, a quick and rough machine translation of the same text can be much more economical and still fulfill the same purpose, i.e. to get a glimpse of the content. Obviously, for the occasional translation requirement of this type it would not make sense to invest in a personal MT system. Therefore, we have received more and more requests for on-line access to METAL, on a pay-as-you-translate basis.
Consequently, network access to METAL was developed, with a simplified user interface, the automation of parameter setting and counting routines. After an in-house prototype had operated reasonably successfully, the integration into a large-scale service network was implemented. Within the LINGO project, the use of METAL for non-translators is made available via public network. Commercial operation is scheduled to begin in 1996.

2.3 Lexicon

In the practical use of machine translation, one of the main tasks for the user is to keep the system lexicon updated. Entries need to be enhanced with morphological, syntactic and semantic information, and even with the aid of automated coding tools like the METAL INTERCODER, extending lexicon coverage to a new subject field can be time-consuming. In order to facilitate the introduction of machine translation for potential customers, the METAL dictionaries were considerably enlarged. The German to English system, for example, by now contains some 230,000 entries (plus all the possible compounds which are generated automatically).
In addition, many subject-specific sets of terminology were compiled and coded. There are components for e.g. banking, telecommunications, textile industry, electrical engineering, data processing etc.
As before, users will update the system lexicon according to their own needs but the existence of preceded dictionaries reduces the time needed to become operational.

2.4 Grammar

Since real-life texts tend to show grammatical structures which cannot always be predicted, it is not surprising that a fair amount of grammar development or revision takes place after the installation of a "completed" system. Based on feedback from customers, the METAL grammars have been improved incrementally. Unfortunately, in machine translation it is true that, beyond a certain threshold, even small gains in quality require enormous investments, and at some point, diminishing returns may lead to a decision to cease further development. This is not yet the case in METAL. Recent modifications of the English analysis grammar for example led to a 8 % increase in correct parses for a large set of benchmark texts.
Additional language pairs such as German-Danish, French-English and English-Spanish have been released, and development has begun on

Catalan, Italian and Portuguese. After a study phase, which tested the applicability of the linguistic approach of METAL to non-European languages, an operative Russian to German system was implemented, and an extensive feasibility study produced a detailed implementation plan for Arabic.

3. Tools

It is understood that not all documents ought to be translated by machine, be it for reasons of style or special requirements of the target group. Still, other tools below the level of machine translation can be used to advantage. METAL has been enhanced by the addition of a terminology data base which can be queried directly from the editor. Authoring tools check the text for illegal terminology and suggest preferred terms. They also generate indices and produce document-based glossaries.
An integrated Translation Memory stores sections that had been translated before. It is usually invoked before a translation run to save duplicate work. The automated document version comparison is used not only in translation but also in monolingual applications. Especially in technical documentation, the manual describing the new version of a given product may differ from the first manual in only some of the chapters. It would make sense to deal with only the new material, but identifying the changes “by hand” can be very time-consuming. Automating this process has proven to lead to higher productivity.
Integrating machine translation and add-on tools can create gains higher than the sum of the parts. Streamlining processes as in the automatic Error Message Response System at SAP can lead to significant advantages on the World Market.

4.0 Other Applications

From a linguistic base technology designed in a highly modular way, applications outside of pure MT can be derived.
Large corporations, working in different countries, often have the problem that communication within the organization is flawed. Even if eg. English is adopted as the one and only corporate language, it will be used by many employees whose native language is not English and whose linguistic skills are limited. Laws on product liability have raised the awareness that throughout a technical documentation, terminology has to be standardized to avoid misunderstandings. Similarly, follow-up costs can be prevented if the syntax of intra-corporate communication is unambiguous and easy to interpret. Not only does this avoid mis-conceptions but it also speeds up the reading and comprehension process.
In the European SECC project, the English METAL analysis module is used to check a source text for syntactic correctness and appropriateness of style. Illegal constructions are marked for revision. And instead of translating from English into French, it is possible to “translate” from faulty English into correct English.
The METAL technology will also be used by General Motors to define, implement and verify a controlled language for the corporation.
Some other projects are still in the planning stages: the re-use of

existing METAL components for multilingual information retrieval, based on semantic content rather than just key words, and for message routing. If documents arrive within a network, a decision has to be made as to who should receive the information. A manual selection process may take too long, especially if large volumes of data are involved and if the information is critical. Analyzing the content of an incoming message and comparing it automatically to a predefined profile will ensure that the document reaches the appropriate recipient immediately.

Development costs for high quality MT systems are extremely high, and so far the translator market has not been receptive enough to cover these costs. Therefore, new directions had to be pursued: addressing the needs of casual users by simplifying interfaces and operation, adding tools for translation tasks unsuitable for MT, deriving monolingual applications for the office environment, and re-using existing modules for information retrieval and message routing. It remains to be seen if the implementation of the European Information Highway will provide a new boost to the development of natural language processing systems. From both a political and an economic view they are indispensable.

Literature:

Schneider 1987. "The METAL System. Status 1987", MT Summit, Hakone
Schneider 1989. "The METAL System. Status 1989", MT Summit II, Munich
Schneider 1992. "User Driven Development: METAL as an Integrated Multilingual System". META vol 37, no.4, pp.583-494
Thurmair 1990. "METAL. Computer Integrated Translation" in J.McNaught, (Ed). Proceedings of the SALT Workshop 1990. Manchester