Motivations, aims and architecture of the LIDIA project

Ch. Boitet
GETA, IMAG-campus
(UJF & CNRS)
BP 53X, 38041 Grenoble Cedex
France

Introduction

At the first Machine Translation Summit in Hakone, 2 years ago, I had been asked to present the research directions envisaged at GETA (Groupe d'Etude pour la Traduction Automatique). At that time, we were just emerging from a 3-year effort of technological transfer (CALLIOPE), and considering many directions for future work. Very soon afterwards came the time to choose between all open possibilities.

Besides 3 main research themes ("static" grammars, lexical data bases and software problem linked with multilinguality), we have recently embarked on the LIDIA project to crystallize the efforts of the team. It may be interesting here to explain briefly the motivations, the aims, and the oven architecture of this project.

I. Motivations

1. Not another project with the same approach

Since 1970, GETA has worked on the 2nd generation "multilevel" transfer approach, with a combination of declarative and heuristic linguistic programming. It has developed a large generator of MT systems (Ariane-G5) for the VM/CMS environment, supporting now 5 specialized (symbolic, rule-based) languages for linguistic programming. This generator, and the lingware engineering methodology developed by B.Vauquois and many others over the years, have given rise to a, variety of experiments, ranging from small mockups to a full-fledged operational system (B'VITAL's French-English for aviation manuals) through laboratory prototypes (Russian-French at Grenoble, English-Malay at Penang).

We feel that there is no need for further development of that type of system in a research lab. 0f course, the research themes mentioned above remain interesting *per se,* but the scientific program has been completed. The best quality we can obtain is somewhat higher than in 1970, but the average quality is rather better, with a comparable amount of effort. This seems to be chiefly due to the introduction of "traces" in the interface structures (beside the "pivot" information) and to progresses in lingware engineering. Moreover, all other current systems are now based on comparable principles and show comparable results, again under similar conditions of development.

More precisely, such systems can only offer "heavy" MT : cost-effectiveness is only attained if the translation load is in the order of 10000 pages of homogeneous domain and typology during at least 2 years. Moreover, the average quality seems to be inversely proportional to the variety of texts handable by the system.

2. A system to answer new needs.

With the advent of powerful micros used for redaction and translation, the need for some really automatic help is steadily growing. At the same time, there is a trend to use one's national language. But heavy MT is not a solution in the majority of cases. We had recently an uproar when the Pasteur Institute decided to use only English in its most famous publications, in order to reach more readers. But there is no system to help a researcher write in his/her native language and translate into another.

Also, many French companies are now looking forward towards integrated Europe and facing the problem of translating their technical documentation in all other languages. Typically, some of them have already begun to produce it on CD-ROMs under a hypertext system such as HyperCard (Renault). All would accept a measure of control in the redaction of the texts to get them automatically translated.

This motivates us to turn to translation from French, while we always worked from other languages in the past, and to look for some kind of "light" MT.

3. Improve drastically the average quality and coverage of MT

This is necessary to meet the above needs, but impossible in the context of 2nd generation systems, which rely only on the *linguistic* knowledge, which includes:

- core knowledge about the language;
- specific knowledge about the corpus (domain, typology);
- intrinsic semantics (a term coined by J.P. Desclés to cover all information formally marked in a natural language, but referring to its interpretation, such as semantic features of concreteness, location,...);

but not:

- extrinsic semantics (static knowledge describing the domain(s) of the text, e.g. in terms of facts and rules);
- situational semantics (describing the dynamic situations and their actors);
- pragmatics (overt or covert intentions in the communicative context).

Considering our desire to "democratize" MT one day, we did not want to go the way of CMT (KBMT project), which consists in coding a huge knowledge base of a domain and of integrating it with the linguistic knowledge base.

But there is a far better knowledge source, namely the author of the considered text. S/he is the only one, in particular, to solve many ambiguities with certainty. Heuristics can never reach 100%. Hence, our last motivation is to produce far better translations than in the past, by asking the author to write better (lexically, grammatically and stylistically) and to clarify the final version in order for the system to reduce ambiguities of all types.

II. Aims

1. To realize a prototype

For the prototype, we have stated the following constraints :

- translate from French into at least Russian and German (inversion of previous systems);
- plan for written and oral outputs ;
- organize the dialog with the author in two steps, one for *standardization,* the other for *clarification;*
- create, modify and possibly revise the texts on a Macintosh under HyperCard;
- distribute processing between the workstation (all functions realizable in real-time) and Ariane-G5 on an IBM (mini or PS2/370) linked through a network.

2. To attack some scientific points

Among interesting scientific questions which the LIDIA project might tackle, there are :

- the organization and production of *dialogues* (in the source language) from partial or ambiguous analyses;
- the refinement of the notion of type of text to that of *type of fragment* (typically, the textual content of a HyperCard field or button) and its use to help analysis ;
- the coherence of the lexical databases necessarily present on the two processors ;

In a second phase, we would like to :

- test the real potential of using interface m-structures (Vauquois' "multilevel structures") as pivot structures (e.g., to connect Russian-French to French-German and study the quality of the resulting Russian-German) ;
- see whether the availability of rich structures really helps in producing more natural oral output (by computing the prosody) than can be done in conventional text-to-speech systems with no deep analysis, and at what added cost.

III. Architecture

Of course, this is only a very preliminary sketch. As will be evident, we have been inspired by the CRITIQUE system of IBM, by the English-Japanese Alvey project (UMIST), by the English-Malay workstation SISKEP (USM, Penang) and of course by our experience with Ariane.

1.. *On the documentation station*

The source text is made of fragments contained in the fields, buttons and menus of a HyperCard slack. By observation or by design, we associate a type_of_fragment to each such element. One such type is the *label* (typically, a menu item such as **Save as...).** Another one might be the *imperative_infinitive* (in French : "Eteindre le disque dur" — shut down the hard disk), etc.

To each card, we associate a "shadow" MT-card, which contains the images of the fragments, grouped in translation units, which may sometime not coincide with the fragments. For example, the subject of a sentence is often in one field and the rest in another. One translation unit may have as many as 3 associated fields, the first containing the input to MT, the second its surface structure, in linear form, and the third its m-structure.

The input to MT is not identical to the text seen and manipulated on the "real" card. We put it in a suitable transcription. It is also enriched by lots of marks, the first being the type of fragment, and others resulting from the dialog. For example, "diplôme" might be rendered as "diplome-1" for its first meaning (degree) and "diplôme-2" for the second (diploma). Of course, the set of "meanings" should be established and maintained with regard to the target languages.

All that implies the presence on the workstation of several resources :

- a multilingual label dictionary, containing the conventional translations **(Enregistrer sous...** and not **Sauver comme...).** It should be modifiable by the user, and appropriately reflected in the MT dictionaries on the MT station.

- a morphological analyzer (lemmatizer) used for spellchecking as well as for accessing on line thesauruses or classical dictionaries ;

- a thesaurus used for standardizing the terminology (each term should either point to the recommended equivalent term or be recommended). Again, the user should be free to modify it in any way (e.g., each of "plane", "aircraft", "ship" or "airplane" is recommended in some firm and not in others);

- a morphological generator to produce the correct forms of the recommended terms when the user accepts the suggestion of the system (the dictionaries of the generator may be the same as that of the analyzer);

- a dialog generator working from the structures sent by the MT station, coupled with a transformer modifying the structures according to the answers);

- mechanisms for ensuring the coherency between the structures and the text, which the author may want to modify while processing of the previous version is under way.

Note that the text on the user card should be generated from the text on the MT card and not be directly in order to keep track of previous clarifications.

2. *On the MT station*

There are at least the following components :

- an ambiguity detector, producing a surface tree (the text is directly produceable through a left-right traversal of the leaves) with special attributes signalling all kinds of ambiguities (attachment, referents, semantic relations, syntactic functions, traces...);

- a deep analyzer producing deep m-structures ;

- transfers and generators into the target languages.

This calls for an organization by servers, quite easy to realize thanks to the total modularity of Ariane-G5 and to the flexibility of the VM system, which simulates a network of "virtual" machines.

3. Communication

It seems necessary to communicate as simply as possible, by exchanging textual files buffered in waiting lists. We have yet to determine the best physical solution (Mac-mainframe or other). In an case, both operating systems allow background processes.

Final remarks

With the LIDIA project, we are trying to capitalize on many old ideas and experiments to introduce HAMT on the desks of the writers of technical documentation and scientific papers. We don't aim at professional translators. That poses interesting problems (all questions should be asked: without reference to the target languages) and opens new possibilities.

Among them, the most exciting is perhaps that intermediate road between free text and controlled! language. With free text, one must study corpuses and try to approximate them in grammars. Quality can be very good in restricted conditions (METEO), but certainly not for the general use envisaged! here. On the other hand, the controlled language approach leads to very high quality (TITUS), but! seems to necessitate a very constraining environment.

What we propose here is to define (by observation) a variety of "fragment types", and, knowing from the document itself what type to expect at a particular point, to gently nudge the writer to write accordingly.

Finally, note that this approach is by no means in itself restricted to hypertext documents. In some recent text processing systems such as WinText™ (used to write this paper), it is possible to define "styles". A style contains the information on the font, the size, the stress, the color, *and* the natural language. It would be easy to add another attribute for the type of fragment, and perhaps to do even more by allowing expert users to include directly some marks (like [(and]) for special parentheses), reflected in the MT text but normally hidden in the normal text, in order to avoid questions by the systems.