In: TAMA '94 Proceedings. 3rd TermNet Symposium. Terminology in Advanced Microcomputer Applications. 24.-25.11.1994. Wien: 1995, 207-218

Harald H. Zimmermann

Value-added Coding of Electronic Dictionaries for the LOGOS Machine Translation System

0 ABSTRACT

Machine translation requires dictionaries with special codings of morphologic, syntactic and semantic information. This relates to the format, content and depth of the coding scheme.

The author describes methods of extraction of terminology and dictionary data from bilingual text files (text and vocabulary alignment). In addition, semi-automatic coding processes are discussed which are based on internal data and their ability to integrate with the LOGOS MT software.

1 REQUIREMENTS OF A MODERN LANGUAGE TECHNOLOGY

In general. the technical prerequisites are there, now, in order to market machine translation technology:

- Processing speed is sufficiently high for mass market PCs and MT applications.
- In addition, telecommunication services are available providing transmitter and user interfaces.
- There is sufficient storage capacity available locally and under favourable conditions.

No longer can we see any obviously fundamental or additional problems for the usage of MT despite some weaknesses which are to be attributed to the MT world itself (lacking standards or insufficient application of existing quasi standards), despite some practical adaptations and conversions to be implemented, for example to specific platforms, client-server architecture, word processors.

In the near future, development and enhancement of MT systems will rather be oriented towards and subject to market conditions and the industry (and not only the software industry). Among those are:

- the availability of tools suitable for production and development which help to reduce software development costs,
- the improvement in the transparency of the systems: the user should be aware of the performance of an MT system in quality and speed. The developer and the customer should be in a position to make further developments and adaptations inexpensively in order to meet their specific needs,
- the possibility to make reliable statements and provide reliable examples as far as quality is concerned.

It is not possible to achieve a total transparency and to completely compare different systems. However, in my opinion, it is now time that MT systems suppliers (in their own interest) and qualified users get together to define their own standards and requirements or quasi standards.

2 FUNDAMENTAL TECHNIQUES OF EXPANDING LEXICAL INVENTORIES

In the following it is assumed that any kind of development of an MT system provides the basis for a certain lexical core inventory. It has to be made sure that the common basic vocabulary in particular is available in the system for each language pair.

This is rather the rule than the exception. In this connection, we do not deal with the coding of an auxiliary (at least not its basic meaning), at the same time we also assume that - de facto - all function words (co-ordinates, prepositions, articles, pronouns...), all irregular verbs (relating non-compounds), and all common adverbs are coded and available in the system. This is true of the leading MT systems available on the market, or at least to be expected, whereas it is not important whether rare or obsolete words such as "erkiesen", an irregular verb in German, is really coded in a system.

With all MT systems, the user has the possibility to add lexical entries himself/herself, although there are different modes of entry. Here we can say, however, that the first non-trivial problems occur. I would like to state some of the decisive factors.

2.1 Which priority should a user entry have?

There is no question, of course, that the situation is different for words unknown to the system than for overlaps with entries already in the system. In general this is a question of transparency of decisions in the system and finally of the deeper insight of the user into the system functionality and way of processing. This competence can hardly be assumed for a user of mainly a mass market low cost MT software. If, however, problems arise less frequently later on, because the basic system provides enhanced dictionaries and generates quality translations, then the user may do without a first-hand transparency of system decisions or hotline help.

Indeed the qualified and professional user has to be offered more explanations for system decisions, if they so require. One example for such a support is a visualization of analyses and results in a tree structure. By means of such a tool users could optionally see why a lexical or lexical and context based translation does not "match" in one case or the other.

2.2 Which subject-matter or text type oriented possibilities should be provided for mark-up and how should they be included in a strategy?

MT systems offer more or less enhanced types of categorisation for this purpose. LOGOS knows a so-called subject-matter code combined with a generic code which links deeper subject matter areas according to priority. The users also have analogous possibilities to provide "their" entries either with these categories or to define their own.

Despite these tools and options for differentiation (also similarly offered by other systems) there is still quite a lot left to be done, for example to make the existing differentiation more flexible

and to make the inventory more transparent to the user in its architecture and "effect" in the actual application.

2.3 Which vocabulary has actually been in the system and how do I - as a user - adapt it lexically to my data as rapidly as possible?

Here we have two ways, roughly distinguished, which could also be combined:

Adaptation of the system "on the job", i.e. within a specific application. This may well be a reasonable approach wherever there is sufficient lexical "subject matter inventory ".

LOGOS as well as other systems offer a tool producing a not-found-words list, in German it also includes words found due to decomposition analysis plus their proposed transfer.

Another component which is available as a prototype is a nominal translation memory which uses post-edited noun phrases in the target text and links them to the LOGOS core. In this case, the problem of "real" lexicalization is being "covered up". The results, however, are available later on for the enhancement of the lexicon. (I do not want to elaborate on this system development project which is pursued by LOGOS in the US).

Integration of lexical or textual customer data bases within the framework of a lexical adaptation phase.

2.4 Which coding methods should be offered to the user?

The user and also the system developer or service centre is offered a fine-tuned dialogue-oriented (interactive) coding system. LOGOS offers two tools: ALEX (the automatic lexicographer) for basic lexical coding including simple nominal word groups and SEMANTHA for context-based entries such as verb valency coding including the subcategorization of valencies with transfer rules, which are defined in so-called SEMTAB rules.

There is no doubt that MT systems need a deeper coding adapted to the appropriate processing strategy. If I consider the basic requirements, there are no fundamental differences among the high-end MT systems such as METAL, SYSTRAN or LOGOS.

This entails for the user a certain linguistic or strategic competence, either to be there as a prerequisite or to be taught by the supplier. One goal of the techniques which I will describe in the following is to replace this competence step-by-step by means of help tools or at least to considerably assist the user.

3 MOST RECENT DEVELOPMENTS AT LOGOS IN THE CONTEXT OF LEXICAL ADAPTATION TO CUSTOMER DATA

In the following I will exclude the area of interactive coding by means of the LOGOS tools ALEX and SEMANTHA which are part of the standard LOGOS software license.

The questions and topics resulting from the discussion above can be summarized and what follows is a description of what LOGOS has developed as front-end tools and beta test products:

- Is it possible to speed up the process of making an inventory of customer terminology (on a relevant level) and to make it cost-efficient? The answer is the front-end tool "Semi-Cod" offered by LOGOS.

- How can existing and machine-readable data be made available for the coding of LOGOS MT applications in a cost-efficient manner? The answer is the front-end tool for specific alignment and text vocabulary offered by LOGOS.

3.1 SemiCod

The LOGOS technique for semi-automatic coding (SemiCod) supports qualified coding of lexical entries. Currently, nouns and nominal groups of the following language pairs can be coded: German into English, German into French, English into German, English into French.

The coding is done via an interface which is being generated analogously to the process of human coding with ALEX. Both ways of coding generate the same file which is then processed with the usual automatic update routines by LOGOS (f. ex. STEMGEN and MERGE) in order to use these coding results for transfer and translation later on. This means that in certain areas SemiCod virtually "replaces" human tasks in the process of ALEX coding.

The technique has the following features:

(1) The software is available on PC-DOS and may run as a stand-alone product. Later on an MS-Windows solution will be provided which is not relevant for the performance of the tool as such, neither are the changes in basic data structures in the front-end dictionaries.

(2) PC based front-end dictionaries are used for coding. This mainly includes:

(a) the existing material of data in the LOGOS MT system. They were extracted from the UNIX-based LOGOS system and integrated into a PC solution. All data have been split into the following parts:

- a bilingual LOGOS PC translation dictionary which contains source and target information, SMC codes, basic word classes (for all of the above mentioned language pairs). Included are all words (also verbs and adjectives): we are currently working on an enhancement of the inventory by extracting the so-called SEMTAB files for translations according to context-based criteria. By means of this technology, the "main dictionary" of LOGOS is becoming transparent on a PC basis. This dictionary is used - among other things - to externally match with new inventories (customer data) and compare them with the existing Logos inventory,

- a monolingual LOGOS (word) semantic dictionary for the source languages German and English. It mainly contains single words (head words) and links to possible semantic codes (only nouns). In addition, sample words have been extracted from the so-called PROMPT files of the ALEX coding tool, they were marked and added to the inventory. In the first case we deal with entries which are "real" in

LOGOS transfers, in the second case we have entries from the ALEX coding tool which also may contain meanings, i.e. semantic interpretations of an entry which have not (yet) occurred in a real transfer entry,

- a monolingual morphologic-paradigmatic dictionary which contains the head word and a link to the systematic inflection morphology coding.

(b) As we use the IDX indexing method developed by Softex for the semi-automatic coding process, the following monolingual dictionaries - relevant for IDX - are being used:

- a so-called identification dictionary (for German and English each) including word stems and information about inflection, decomposition and derivation,
- a so-called relations dictionary which includes a number of relations, for example ranging from the Ablaut and Umlaut stems to the head word stem, among other things (for German) also including partial word links.

(c) In addition and as an option, further SOFTEX dictionaries, i.e. translation dictionaries may be processed.

(3) If a (new) word is to be coded, at first the software automatically checks whether the translation/transfer is already available in LOGOS. For the remaining ones, the software distinguishes between an entirely new entry or an entry which is new to a subject matter. In order to describe it in a simple way, we will look at the case of the entirely new entry.

(4) If there is no basic word class information for the new entries (in the case of nouns and nominal groups this includes information about gender and number, among others), the software will at first (this is optional) try to determine the basic word class according to different comparison techniques. In doing this, a PC-based dictionary is created which corresponds with the coding dimensions of the standard SOFTEX dictionaries and with the coding level of the LOGOS PC transfer/translation dictionary.

(5) On the basis of the data resulting from (3) or (4) an input file is now created for the actual SemiCod processing.

(6) While taking into account all the different front-end dictionaries, the automatic enhancement of data is induced which means adding paradigmatic-morphological and word-semantic features. These are now stored in a largely neutral output format. Besides these features, the software also determines the syntactical structure of more complex nominal phrases. In the case of compounds the software will do the coding on the basis of the head element if the total entry is not identified. I would like to distinguish between complete, ambiguous-complete and incomplete results - these are further differentiated internally.

(a) "Complete" is when all features have been identified and there is only one solution.
(b) "Ambiguous-complete" is an entry for which we have several possibilities as a solution. Generally, these are several word semantic categories.

(c) "Incomplete" is an entry, when a necessary feature could not be determined (source or target related, e.g. an inflection-paradigmatic mark-up, a semantic category).

In the case of (a) there is the risk, obviously, that there was a wrong assignment. This may also be caused by to the fact that a reduction rule (for example in the word-semantic part) did not match in this particular case. In the case of (b) we would have to either decide by human intervention (as with the ALEX tool) or to eliminate the problem by enhancing reduction rules. In the case of (c) the front-end dictionaries are to be enhanced, however, the cause could also be a typo in the lexical inventory. Technically speaking, we are able to achieve the coding of up to 800 to 1.000 words and their transfers per hour on a 386 or 486 PC in case of "complete" entries.

(7) As a final step the still largely neutral codings are automatically converted into the relevant internal LOGOS format (the so-called RDICT-Format). What I mean by largely neutral is that they are indeed neutral except for the word-semantic categorization which is LOGOS system-based.

(8) Besides, the resulting data are used in order to enhance relevant special dictionaries. This way the system is "learning". This is also true of the LOGOS PC dictionaries.

Advantages

- The technique can easily be adapted to new LOGOS developments. It will not be a problem to adapt the system output interface - now still being quite bulky - to the new lexicon solution which will be based on a relational database.
- Despite its complexity the technique is more transparent and more flexible. Systematic errors may be easily detected and corrected.
- Due to the learning effect, major cost reductions are expected during the enhancement of a language pair and the development of new pairs. One example is the possibility to invert entries (German-English - English-German) when using the complementary MT system.
- The recording of customer data (besides the coding of verbs and SEMANTHA rules) can be accelerated considerably. As - according to experience - most entries are nouns or noun phrases (usually more than 90 %), a practical application makes sense even as early as now in its beta phase.

Availability

Right now, SemiCod is not yet ready for the customer and their production purposes. It is, however, being used as a service tool already in its beta testing phase. The main field of application is the enhancement of the LOGOS MT dictionaries internally.

3.2 Developments of text and word alignment at LOGOS

I will briefly present some front-end tools which have been used or are under development within the cooperation of SOFTEX and LOGOS.

3.21 SoftAlign

This language independent tool is designed to "parse" user text corpora with text parallel translations into the smallest possible parallel segments. These could be sentences, whereas a source sentence could correspond to one or more target sentences and vice versa. A prerequisite is the same and paralleled number of paragraphs in the source and target text. The technique is mainly statistically based, formal segment separators are used and periods for abbreviations are distinguished from end of sentence periods.

3.2 WordExtract

The starting point is a machine readable text corpus from the customer and its "text parallel" translation (language pairs: German/English/French). This tool allows to work on the following subjects while taking into account existing electronic lexicons:

(1) Identification of known translations (this may contribute to set a subject matter or customer code mark-up);
(2) To determine potential word equivalents / equations (new terminology).

The quality of results produced by WordExtract is determined by the existing data volume and the extent of "translation equivalency" in the sense of a most similar source-oriented structure and choice of words.

If larger quantities of data are available, the results of this terminological enhancement process are considerable.