Multilingual Translation Model Based on Object-Oriented Design

Eric Wehrli
LATL-Department of Linguistics
University of Geneva
Eric.Wehrli@lettres.unige.ch

February 28, 2006

Abstract

The increase of cross-cultural communication triggered notably by the Internet intensifies the needs for multilingual linguistic tools, in particular translation systems for several languages. The LATL has developed an efficient multilingual parsing technology based on an abstract and generic linguistic model and on object-oriented software design. The proposed project intends to apply a similar approach to the problem of multilingual translation (German, French, Italian and English).

1 Research plan summary

2 Research plan

2.1 State of research in the field

Machine translation (MT) is arguably one the oldest natural language applications. Over the last 50 years, a large number of systems have been developed and large variety of architectures/designs have been proposed, based on linguistic approaches or, more recently, on statistical methods (See Hutchins, 1986, 2006, King, 1987, Hutchins & Somers, 1992, among many others, for a general overview of MT. See Ney (2005) for a review of statistical MT, Wahlster (2000) for several aspects related to speech-to-speech translation, Hutchins (2006) for commercial aspects of MT. For a discussion of some of the problems faced by MT, see Arnold (2000), and Volk (1998) for the problem of idiom translation.).

When it comes to general-purpose high-quality machine translation, a majority of researchers consider that some variant of the (linguistic) rule-based transfer approach is probably the most appropriate design. There will be, however, wide disagreements with respect to issues such as the precise nature of the linguistic descriptions required for translation, and especially the nature of the representations on which the mapping from one language to another will take place, the so-called transfer level (Eberle, 2001).

One of the problems faced by MT, that we would like to address, is the need for systems capable of handling not just one pair of languages but several pairs, ultimately a very large number of language pairs. The globalization of commercial and cultural exchanges, for instance over the Internet, creates an ever increasing demand for multilingual linguistic tools and in particular multilingual translation systems. To take another example, in the West-European context, where no less than 15 languages are used for everyday life as well as for administrative, scientific and commercial purposes, the explosive need for translation tools is by no means satisfied by the existing commercial systems. To quote Boitet (2001) "Despite considerable investment over the past 50 years, only a small number of language pairs is covered by MT (...), and even fewer are capable of quality translation or speech translation."

The problem of the quadratic growth of the number of translation pairs ($n \cdot (n-1)$, for n languages) has often been taken as an argument against transfer-based translation models as opposed to systems based on interlingua (see Boitet, 2001, who argues in favor of UNL (Unified Networking Language)). In the proposed project, we would like to argue against such a view, and show that the above-mentioned problem can be very effectively restricted. We intend to show that this goal can be achieved (i) by the use of an abstract level of representation – which abstracts away from several surface-structure cross-linguistic differences (ie. word order, morphological cases, etc.), and (ii) an object-oriented design, which makes possible the use of generic code while allowing for language-pair specific properties and processes through type extension and redefinition of methods.

This approach is very similar to the one we have applied quite successfully to the development of the multilingual Fips parser (Wehrli, 2004, 2006). In the latter case, the modularity of the resulting system is such that the addition of a new language module can be achieved without any modification to the overall system, and no recompilation of the other modules.

2.2 Applicant's previous work and research in the field

The principal investigator, and more generally the LATL lab, have been involved in NLP research for almost 20 years, mostly in the area of parsing (Wehrli, 1997, 2004), but also in lexical database design, machine translation (cf. Etchegoyhen & Wehrli, 1998, L'haire et al., 2000, Ramluckun & Wehrli, 1993, Wehrli, 1998), translation tools (Wehrli, 2003), speech-to-speech translation (Gaudinat et al. 1999), multi-word expressions (Wehrli, 1998, 2000, Seretan et al. 2004).

2.3 Detailed research plan

2.3.1 Objectives and goals

This project pursues several objectives:

- Design of a MT model using an abstract syntactic level of representation as transfer level and an object-oriented design for its software realization. The conjunction of those two choices is expected to yield a MT system, with generic and well-motivated linguistic structures, and lean and easy to maintain software.
- This MT model will be applied to the four main languages used in Switzerland (German, French, Italian and English).
- As a by-product, a word-in-context translation tool for the 12 language-pairs will be developed (TWiC)

To a large extend, this proposal can be viewed as an extension of our current Multilingual Fips project (SNF 101412-103999). For one thing, the availability of the Fips parser for the targeted languages is a crucial element of the project and, second, the software design of the proposed MT model closely matches the one developed for the multilingual parser. In both cases, the goal is to set up a generic system which can be redefined (through type extension and method redefinition) to suit the specific needs of, respectively, a particular language or a particular language-pair.

2.3.2 Methodology

The translation algorithm follows the traditional pattern of a transfer system. First the input sentence is parsed by the Fips parser producing an information-rich phrase-structure repre-

sentation with associated predicate-argument representations. The transfer module maps this source-language abstract representation to a target-language representation. Given the abstract nature of this level of representation, the mapping operation is relatively simple and can be sketched as follows: recursively traverse the source-language phrase structure in the order: head, right subconstituent, left subconstituent. Lexical transfer (the mapping of a source-language lexical item with an equivalent target-language item) occurs at the head-transfer level (provided the head is not empty) and yields a target-language equivalent term often, but by no means always, of the same category. Following the projection principle used in Fips, the target-language structure is projected on the basis of the lexical item which is its head. In other words, we assume that the lexical head determines a syntactic projection (or meta-projection).

Projections which have been analyzed as arguments of a predicate undergo a slightly different transfer process, since their precise target-language properties may be in part determined by the subcategorization features of the target-language predicate. To take a simple example, the direct object of the French verb *regarder* in (1a) will be transferred in English as a prepositional phrase headed by the preposition *at*, as illustrated in (2a). This information comes from the lexical database. More specifically, the French-English bilingual lexicon specifies a correspondence between the French lexeme [NP regarder NP] and the English lexeme [NP look [PP at NP]]. For both sentences, we also illustrate the syntactic structures as built by Fips:

(1)a. Paul regardait la voiture.

b. [$_{\texttt{TP}}$ [$_{\texttt{DP}}$ Paul]regardait$_i$ [$_{\texttt{VP}}$ e$_i$ [$_{\texttt{DP}}$ la [$_{\texttt{NP}}$ voiture]]]]

(2)a. Paul was looking at the car.

b. [$_{\texttt{TP}}$ [$_{\texttt{DP}}$ Paul] was [$_{\texttt{VP}}$ looking [$_{\texttt{PP}}$ at [$_{\texttt{DP}}$ the [$_{\texttt{NP}}$ car]]]]]

Adding a language to the system Given the general model as sketched above, the addition of a language to the system requires (i) a parser and (ii) a generator. Then for each language pair for which that language is concerned, the system needs (iii) a bilingual database and (iv) a (potentially empty) language-pair specific transfer module. These four elements will now be discussed in turn.

parser The Fips multilingual parser is assumed. Adding a new language requires the following tasks: (i) grammar description in the Fips pseudo-formalism, (ii) redefinition of the language-specific parsing methods to suit particular properties of the language, and (iii) creation of an appropriate lexical database for the language.

generator In our translation model, target-language generation is done in a largely generic fashion (as described above with the projection mechanism). What remains specific in the generation phase is the selection of the proper morphological form of a lexical item. The generator is thus largely a matter of morphological generation. For that purpose, we rely either on the (fully inflected) lexical database or the FipsInflection morphological engine.

bilingual database The multilingual lexical database contains the information for the lexical transfer from one language to another. For storage purpose, we use a relational database management system. For each language pair, a relational table is used and contains the associations between lexical items of language A to lexical items of language B. In addition to these links, the table contains transfer information such as translation context, preferences between one to many translations, argument matching for predicates (mostly for verbs). The table structures are identical for all pairs of languages.

What is challenging in this project is that it necessitates as many bilingual tables as the number of language pairs considered, i.e. n(n-1)/2 tables. For instance, for 4 languages, it requires 6 bilingual tables, for 5 languages 10 tables. We consider that an appropriate bilingual coverage (for general purpose translation) requires at least 50'000 correspondences per language pair. In order to achieve the building-up of the bilingual database during the duration of the project, we plan to use semi-automatic generation for part of it. For this purpose we will derive by transitivity a bilingual lexicon using two existing ones. For instance, if we have bilingual correspondences for language pair A→B and B→C, we can obtain A→C. The generated correspondences will then be validated using some semantic information. The Wordnet database will be used for this purpose. However, the correspondences that could not be checked this way necessitate to be checked manually.

language-pair specific transfer The transfer from language A to language B requires no language-pair specification if the language structures of A and B are isomorphic. Simplifying a little bit, this happens among closely related languages, such as Spanish and Italian for instance. For languages which are typologically different, the transfer-specific module must indicate how the precise mapping is to be done.

In this project, we will restrict ourselves to languages for which we already have both a fully functional parser and generator, that is English, French, German and Italian[1]. We, thus, intend to develop the 12 transfer-specific modules, although we will not have the possibility to develop the necessary bilingual databases to have complete translation systems for the 12 pairs. Given our previous work in the domain of translation (and the resources we own), we will primarily focus our work on the following six pairs.

- English-French
- French-English
- German-French
- French-German
- Italian-French
- French-Italian

For the other six pairs, only a demo version will be provided, to show the validity of the design and the potential of the approach.

[1]Spanish and Greek might also be considered.

L-to-L translation A major development tool for this project will be the translation of a language into itself, for instance French $\longrightarrow$ French. Given the general translation model, translating, French into French should only require (i) a French parser, (ii) a French generator and (iii) a French-to-French bilingual database. The fourth component can be discarded since we obviously have an isomorphic language pair. In other words, to validate the parser, the generator and the generic transfer components, the translation of documents from one language to itself will be extremely valuable. Furthermore, the evaluation of the quality of that "translation" will be easy to achieve and does not require any translation knowledge. This methodological tool will be particularly useful to test the quality of the analysis, the quality of the morphological generation, and the quality of the lexical information (subcategorization and other selectional information). It will also test the ability of the system to properly generate multi-word expressions.

Word translation As a by-product of the proposed system, we will get a word in context translation system (TWiC, translation of words in context, see Wehrli, 2003) for all the language-pairs. Twic is designed as a translation tool for readers of on-line documents in foreign languages, who have a fairly good knowledge of the language but have some terminological deficiencies. When faced with an unknown word (or phrase), the user clicks on it, the whole sentence is then retrieved, parsed by Fips, so that the lexical lookup in the bilingual database is restricted to the lexeme(s) identified by the parser. The result is a less noisy system, with proposed translations which are restricted to readings which are compatible with the (syntactic) context. Furthermore, since the parser is capable of identifying multi-word expressions (collocations, idioms, etc.), TWiC also provides phrase translation when appropriate.

2.4 Timetable and milestones

The projet will start on October 1, 2006 and lasts 24 months. It will be lead by the principal investigator and conducted by a team composed of 2 Ph.D. students paid by the project (2 candoc positions) along with several members of the LATL (linguists and computational linguists, along with a few graduate students). We will rely, in part, on C. Laenzlinger's work and experience in both syntax and computational linguistic projects, as well as on L. Nerima (for whom 1/6 position is requested), who will be in charge of the design, development and the maintenance of lexical databases (monolingual and bilingual). Furthermore, we will also benefit from Dr. Wolgang Weck's experience in software design – who will act as software consultant, as he has already done for the multilingual Fips project. We will also collaborate with Oberon microsystems, AG, in Zurich for specific software issues.

Here is the proposed timetable, divided into 4 semesters:

semester 1 Architecture of the overall system (OO-design); Elaboration of the "reflexive" translation modules (L$\longleftrightarrow$L), for the 4 languages (French, English, German, Italian).

semester 2 Elaboration of the language-specific modules for the six basic language-pairs; Word translation system for the six language-pairs (en$\longleftrightarrow$fr, ge$\longleftrightarrow$fr, it$\longleftrightarrow$fr)

semester 3 Elaboration of the language-specific modules for the other six language-pairs;

semester 4 Evaluation and reports

Work on the bilingual databases as well as improvements of the Fips parser for the four languages will occur throughout the project.

2.5 Significance of the planned research

The significance of the project lies in (i) the use of an object-oriented design to drastically simplify the addition of language-pairs in an MT system (as well as simplify its maintenance), and (ii) the development of a translation model/system which can handle the four "national" languages (German, French, Italian and English). This system will be made available on the Internet, minimally as part of the TWiC (Translation of words in context) words and sentence translation services on our web pages, but could also be more wildly distributed/commercialized depending on both the quality of the translations and the state of the market.

Arnold, D. 2000. "Why translation is difficult for computers" in H.L. Somers (ed.) *Computers and Translation : a handbook for translators*, John Benjamin.

Boitet, C. 2001. "Four technical and organizational keys to handle more languages and improve quality (on demand) in MT" in *Proceedings of MT-Summit VIII*, Santiago de Compostela, 18-22.

Eberle, K. 2001. "FUDR-based MT, head switching and the lexicon" in *Proceedings of MT-Summit VIII*, Santiago de Compostela, 93-98.

Etchegoyhen, T. and E. Wehrli, 1998. "Traduction automatique et structures d'interface" in *Proceedings of TALN 1998*, 2-11.

Fellbaum, Ch. (ed.), 1998. *Wordnet: An Electronic Lexical Database*, Cambridge (MA), The MIT Press.

Gaudinat, A., J.-Ph. Goldman, H. Kabre and E. Wehrli, 1999. "Syntax-Based French-English Speech-to-Speech Translation on Restricted Domains", *Proceedings of EACL-99*.

Hutchins, J. 1986. *Machine Translation : past, present, future*, Chichester, Ellis Horwood.

Hutchins, J. 2003. "Has machine translation improved?" in *Proceedings of MT-Summit IX*, New Orleans, 23-27.

Hutchins, J. 2006. "Current commercial machine translation systems and computer-based translation tools: system types and their use", http://ourwordl.compuserve.com/homepage/WJHutc

Hutchins, J. and H. Somers, 1992. *An Introduction to Machine Translation*, Cambridge, Academic Press.

King, M. (ed.) 1987. *Machine Translation Today : the state of the art*, Edinburgh University Press.

L'haire, S., J. Mengon and C. Laenzlinger 2000. "Outils génériques et transfert hybride pour la traduction automatique sur Internet" in *Proceedings of TALN 2000*, 253-262.

Ney, H. 2005. "One Decade of Statistical Machine Translation" in *Proceedings of MT-Summit X*, Pukhet, Thailand.

Ramluckun, M. et E. Wehrli, 1993. "Its-2: An interactive personal translation system" in *Proceedings of EACL-1993*, Utrecht.

Seretan, V., L. Nerima and E. Wehrli, 2004. "Multi-word collocation extraction by syntactic composition of collocation bigrams" in Nicolas Nicolov et al. (eds.) *Recent Advantces in Natural Language Processing III*, Amsterdam, John Benjamins, 91-100.

Volk, M. 1998. "The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems" in N. Weber (ed.) *Machine Translation . Theory, Applications, and Evaluation. An assessment of the state of the art*, St. Augustin, Gardez-Verlag.

Wahlster, W. (ed.), 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*, New York, Springer Verlag.

Wehrli, E. 1998. "Translating Idioms", *Proceedings of COLING-98*, Montreal, 1388-1392.

Wehrli, E. 2000. "Parsing and collocations" in D. Christodoulakis (ed.) *Natural Language Processing - NLP 2000*, Springer Verlag, 272-282.

Wehrli, E. 2003. "Translation of words in context" *Proceedings of MT-Summit IX*, New Orleans, 502-504.

Wehrli, E. 2004. "Un modèle multilingue d'analyse syntaxique" in A. Auchlin et al. (ed.) Structures et discours, Mélanges offerts à Eddy Roulet, Montreal, Nota Bene, 311-329.

Wehrli, E. 2006. "A Multilingual Approach to Natural Language Parsing", mimeo, LATL (given in appendix 1).