

**ANNUAL EUOTRA WORKSHOP
BARCELONA JUNE 1993**

INTERACTIVE MT: A possible Eurotra development

Stelios Piperidis, Celia Roniotes, George Carayannis

Eurotra-EL, ILSP

In their description of Eurotra Hutchins and Somers, in "An Introduction to Machine Translation", 1992 claim a major shortcoming in Eurotra is "the too simplistic approach to semantics....This shallow semantics combined with deep and broad coverage of syntax has resulted in over-generation in analysis, and insufficient filtering by disambiguation...", "Some of this insufficiency of semantics could be minimised in a practical system by the incorporation of an interactive facility, and this has been recommended in reports to the Commission".

It is exactly this alternative that we propose to explore with this proposal on interactive MT. Basically, we are concerned with the design and construction of an interactive system using what resources we can from Eurotra. In conjunction with this we aim at incorporating some kind of knowledge modelling in order to render the interaction more effective.

Objectives:

The predominant idea behind interactive MT is the intervention of the human monolingual/bilingual user during the translation process. Thus, the system's inadequate performance could be compensated by the user whenever the machine flags that it has come across a problematic case, whose resolution is beyond its capabilities.

A. Melby " On human-machine interaction in translation" in "Machine Translation, Theoretical and Methodological Issues" S. Nirenburg, 1987, proposes three types of interaction:

pre-editing	SL users/controlled language
intra-processing	SL/TL users, TL questions (MAHT/HAMT)
para-processing	consulting list of not found words, up-grading dictionaries

The areas where the intervention of the user would enhance the system's performance and by definition stand as problematic for most MT systems are:

- **Ill formed input** , which can be viewed as a violation of the system's constraints. It may arise from real problems in the input or from actual system deficiencies. Thus, a spelling mistake might indicate that the system has inadequate lexical information, or an ungrammatical sentence might indicate inadequate grammar.

- **Ellipsis**, which is very frequent, not to say necessary, in particular sentential styles of many languages. It is not considered to be aberrant linguistic behaviour and requires special knowledge and processing techniques. E.g. anaphora resolution.

example:

I left my bag on the table and I forgot it.

άφησα την τσάντα μου στο τραπέζι και την/το ξέχασα.

- **Ambiguity**, where the system is unable to computationally derive the semantic interpretation of a sentence. Ambiguity resolution requires recourse to different sources of knowledge and information, the computational modelling of which

cannot be easily and cost-effectively achieved. Exemplary cases of ambiguity include:

Lexical Ambiguity: refers to polysemy (words with different meaning and common form, but are related), homonymy (words with different meaning and common form, but are unrelated) and categorial ambiguity (word with the same form but belonging to different grammatical categories).

example:

He sat near the bank (homonymy)

- financial institution
- river side
- small hill

Structural Ambiguity: which refers to attachment ambiguity (existence of more than one nodes to which a syntactic constituent may legally be attached, e.g. prepositional phrase attachment, subordinate clause attachment, gap-finding ambiguity (existence of more than one places to which an already moved constituent has been moved) and analytical ambiguity (existence of multiple legal analyses of a constituent, e.g. present participle/adjective or noun ambiguity, verb particle/preposition ambiguity).

example:

Mary saw the man in the park with the telescope

5 possible permutations

Semantic Ambiguity: which refers to the problem of determining the case slot that a particular syntactic constituent or a preposition flags.

example:

the author's description was accurate

Translational Ambiguity: which consists mainly of the problem of lexical ambiguity across languages.

example:

runway: διάδρομος προσγείωσης/απογείωσης

Interactive MT demands that a system:

- be aware of its deficient knowledge
- know when to engage in the interactive mode
- take into account the user's linguistic abilities and knowledge in relation to the texts he handles
- ensure the ease and speed of the user's reaction
- be able to recognise correct and incorrect reactions and respond accordingly

On the other hand interactive MT demands that a user:

- be either mono- or bi- lingual,
- must understand how to react
- must react correctly

The parameters to determine the halt of batch processing and the start of a consultation session with the user will be:

- processing time
- grading by the system of the translation to be output according to problems encountered in lexical, syntactic or semantic processing

The steering component of the user intervention will be an intelligent dialogue generator able to pose questions and receive answers from the user regarding knowledge the machine needs to carry on processing.

Through ongoing dialogues between machine and human the system can ask the user either to explicitly provide the solution or to supply further information which will enable it to reach a solution. With this sort of help from the user the system can correct ill-formed input sentences, update its lexicon, make predictions on its own about unknown words, resolve several types of ambiguity and finally provide correct/acceptable translations in the target language.

It can be expected that the system, interfaced with a simple generation mechanism, can generate a question or a series of questions to pose to the user. It could also, prompt the user to choose between possible translation options. Of primary importance is the determination of questions to be presented to the user as well as to assign a degree of priority to them.

The type of questions as well as the sequence in which they will be posed will be determined by two parameters:

- a. the personalised user profile
- b. the particular problem to be solved

The **personalised user profile** will be constructed by a separate module in charge of eliciting information about the user's linguistic, source and target language knowledge in this respect. According to this profile, the system will know whether it is worth asking the user a bilingual question or not. It will also be able to decide on the priority that each certain type of question will be assigned. Priority values will then be passed on to the query generation module.

Stereotypical questions will be presented when ill-formed input or elliptic cases are encountered prompting the user to reconstruct the respective sentences. They will also be used to introduce multiple options for anaphora resolution and disambiguation of translational lexical ambiguities.

Factual questions will have a simple Vaux-NP-V-NP form with the verb either in the simple present or past tense and the subject NP either a proper noun or a

definite NP. The lexical acquisition phase of the query generator will derive from semantic concordances incorporated in the semantic module of the system coupled with a preference strategy deriving from bigram and trigram probabilities.

Ontological questions will have the form of a copular sentence with a non-terminal node of the ontological hierarchical structure in the predicate.

Anaphora resolution will be triggered by a stereotypical question with multiple options comprised by the elements of the stack compiled by the possible referents of the pronoun or the NP. The length of the stack is to be determined by a window defining the limits of search for possible referents.

Translational lexical ambiguities will be similarly treated with the multiple options comprised by the multiple lexical entries deriving from one-to-many lexical mappings.

- Thus, for the structurally ambiguous sentence:

Mary saw the man in the park with a telescope

questions to be generated might be:

Were you in the park? factual
Do you have a telescope? factual

- For the lexically ambiguous sentence:

He sat near the bank.

questions to be generate:

Is the bank an institution? ontological

- For anaphora resolution in

I left my bag on the table and I forgot it.

questions to be generated:

What does 'it' refer to? stereotypical
- the bag?
- the table?
- the whole event?

or if the user is bilingual:

What should it be translated to?
- την?
- το?

The design of the interactive environment,, including the user interface and the dialogue generator mechanism will be based on a multi-agent, multi-expert approach involving linguistic, cognitive, statistical and artificial intelligence

expertise, thus illuminating the necessity for synergy between the approaches as well as their complementarity.

Particular emphasis will be given to the structure of the user interface of the system ensuring high level ergonomics and conviviality on the part of the user. It has to combine both keyboard input and mouse/trackball selection. Mouse selection is expected to be employed in the user profile construction and in multiple choice question answering, while by keyboard input the user will be able to issue the correct version of ill-formed input and equivalent phrases to elliptical phenomena.

Resources

In an attempt to avoid duplication of work and focus on the interactive components, the intended system will be structured on the grammatical and lexical resources of the EUROTRA system. This decision follows from two observations:

- a. the high quality of these grammars
- b. their broad coverage of linguistic phenomena of both languages to be involved (English and Greek).

Particular modifications, additions, deletions and prioritisations may prove necessary for both the grammar and the lexica, depending on the specific sublanguages the system will be applied to which will be user specified.

We also expect to import into the system architecture features such as :

- pre-processing components such as deformatting/reformatting facilities, taggers, spell/syntax checkers, translation memory, etc.
- links with on-line term bases
- links with on-line dictionaries/glossaries

Expected implementation of results

The direct outcome of this project will be an interactive MT system, usable within a wider translation platform involving such supporting facilities as: term bank access, information access and retrieval. The system to be produced achieves the following results:

- fast and high quality translation
- consistency in terminology usage
- elimination of repetitive translational tasks
- high-level system ergonomics
- decrease of the funds invested in translation on the part of the users