EUROTRA, History and Results

Bente Maegaard
Center for Sprogteknologi, Njalsgade 80, DK-2300 Copenhagen S
Phone: +45 35 32 90 90, Fax: +45 35 32 90 89
E-mail: bente@cst.ku.dk

EUROTRA is the European Community machine translation research programme 1982-92. Among various results is the PaTrans system which translates patent texts from English to Danish.

1. The EUROTRA programme

EUROTRA is the European Community machine translation research programme. The Community started the programme in 1982, with the goal of creating an advanced system for automatic translation, capable of treating all the official working languages of the community – at that time Danish, Dutch, German, English, French, Italian, and later also Greek, Spanish and Portuguese.

The result of the programme, which ran until 1992, was 1) a prototype of such a multilingual translation system. The EUROTRA prototype has proven to be a basis for the development of practical systems with state-of-the-art performance. Secondly, EUROTRA gave rise to a promotion of research and European collaboration in the fields of MT, machine translation, and NLP, Natural Language Processing, in all 12 community countries. At its peak, the project counted more than 200 researchers.

EUROTRA was a research project with an important implementation component. For all languages, the project produced a large grammar and a general language dictionary as well as a term dictionary (for telecommunications), and for around half of the language pairs the project produced translation modules consisting of bilingual dictionaries and translation rules. These resources have been of great value in several other EU projects; for many languages they constituted the largest implemented NL resources.

The EUROTRA project was in several countries supplemented by national research, e.g. the German “Begleitforschung”. This gave extra impetus to the development of the field.

Europe's multilinguality is a richness (as well as a burden). The EUROTRA investment, which was shared between the EC and the national governments, plus the additional national efforts in some countries, built up an expertise which is being exploited in numerous ways, in MT and other NLP applications. Europe has become clearly competitive with countries outside.

At the political level, the project also increased awareness of the possibilities lying in language technology. The Danzin-report on the project recommended the inclusion of a specific action line, Linguistic Research and Engineering (LRE) in the 3rd Framework Programme, and this was followed by Language Engineering (LE) in the 4th Framework Programme.

2. Results building on EUROTRA

Most important, however, is the fact that the EUROTRA results have been used to produce machine translation: In Denmark, the PaTrans system was produced for the translation company Lingtech which specialises in the translation of patent texts. PaTrans was built on the basis of EUROTRA technology; it translates chemical texts from English to Danish. It has been in use since May 94 and currently saves 60-75% of Lingtech's translator costs. The system is being tailored to treat mechanical texts as well, and other language pairs are under consideration.

The PaTrans system is a fully developed production system. It contains a document handler which ensures that the format and layout are preserved under translation. Secondly, it handles pre-editing marking, such as 'not-to-be-translated', 'title' etc. A conversion programme converts the layout information of a WordPerfect file into SGML codes. The document handler contains a module for the automatic recognition of chemical formulae, since they appear in large amounts.

The system grammar is specifically designed to cater for the text type and subject field relevant sublanguage. This means that in some cases the EUROTRA grammar was extended, e.g. with the description of lists, in other cases it was simplified, e.g. the treatment of modality could be simplified compared to the more general EUROTRA treatment. The vocabulary is divided into the general dictionary and term dictionaries. The general vocabulary covers those words and compounds which occur in patent texts; it is being updated by CST at a regular basis, but is still rather small, around 6000 entries: The sublanguage seems to be pretty restricted wrt. general language vocabulary.

The term dictionaries on the contrary are maintained by the customer. A special term encoding tool, PaTerm, was developed which enables the (non-computational linguist) user to encode terms efficiently. A hierarchy of term bases was created in order to enable to user to specify the priority in which the term bases should be used by the system. The user is encouraged to create a term base for each subject field, in order to keep different meanings of a term belonging to different subject fields apart (e.g. *composition* in chemistry vs. in music).

At the software side, two important changes and additions were made: First of all, the parser was streamlined and made faster. Secondly, a failsoft mechanism was introduced. The failsoft mechanism provides a result even in cases where words are missing, or where the sentence structure is not covered by the grammar, i.e. the sentence is ungrammatical wrt. the grammatical description. In such cases, the parser cannot produce a well-formed structure at a certain point in the processing.

The failsoft mechanism collects as much as has been produced and outputs it. Sentences which have been through the failsoft component are marked (by /- and -/) in the output, so that the post-editor can pay special attention to them.

Obviously, PaTrans produces raw translation which has to be post-edited. In the following short excerpts we show the first two sentences of a patent translated at Lingtech.

Original:

> The present invention relates to a process for producing lube oil. More specifically, the present invention relates to a process for producing lube oil from olefins by isomerization over a silicoaluminophosphate catalyst.

Raw translation:

> Den foreliggende opfindelse angår en fremgangsmåde til at fremstille smøreolie. /- Mere specifikt, foreliggende opfindelse angår en fremgangsmåde til at fremstille smøreolie fra olefiner med isomerisering i løbet af en silicoaluminophosphatkatalysator. -/

Post-edited translation:

> Den foreliggende opfindelse angår en fremgangsmåde til at fremstille smøreolie. Mere specifikt angår den foreliggende opfindelse en fremgangsmåde til at fremstille smøreolie ud fra olefiner ved isomerisering over en silicoaluminophosphatkatalysator.

The first sentence is totally correctly translated. Note, that the English present participle *producing* has been correctly transformed into a Danish infinitive *at fremstille.* The second sentence has been treated by the failsoft component. This is the reason that the constituent order is not correct: the verb should be the second constituent, after *Mere specifikt.* Some prepositions have to be changed as well.

The above is a pretty true example of the quality of the output. Most sentences need some post-editing, but often only minor corrections are necessary. The proof of the quality is that Lingtech saves 60-75% of their translator cost as mentioned above. Currently, work is ongoing to improve the quality of output, in particular by improving the failsoft output.

3. Conclusion

PaTrans has a very good performance. It was only possible to obtain this performance because we could build on the existing results from EUROTRA, and because we could concentrate on a specific text type and subject field, - it should be noted however, that the text type is not a very restricted one, and consequently similar systems can be made for other text types.

4. Postscript

I have concentrated on PaTrans, because it is the most advanced of the EUROTRA spin-offs, and because this is the product I know best. Other EUROTRA teams are working in similar directions, however: IAI, Saarbrücken, Germany, created the EUROTRA version CAT2 already while the project was running. The CAT2 system and associated linguistic resources have been used in two industrial projects, and in several EU research projects. Additionally, the English, Spanish and Italian EUROTRA groups are working on a translation system, based on EUROTRA technology and resources, in the TRADE project under LRE. This project also includes two user organisations, a Spanish software house and an Italian social security organisation.

Literature

The EUROTRA technical specifications and formal specifications are presented in Copeland, Durand, Krauwer, Maegaard (ed.): *Studies in Natural Language Processing,* Vol. 1 and 2, CEC, Luxembourg, 1991.

Additionally, Vol. 3-8 treat other aspects of EUROTRA and related research.