

Exploring Distributed MT¹

Oliver Streiter

Tel: 0049-681-3895126
email: oliver@iai.uni-sb.de

Antje Schmidt-Wigger

Tel: 0049-681-3895129
email: antje@iai.uni-sb.de

Ursula Reuther

Tel: 0049-681-3895128
email: ursel@iai.uni-sb.de

Catherine Pease

Tel: 0049-681-3895126
email: cath@iai.uni-sb.de

IAI

Martin-Luther-Straße 14
D-66111 Saarbrücken
Germany

¹Paper presented at the International Network Conference, 6-9 July Plymouth, United Kingdom

Exploring Distributed MT

Oliver Streiter, Antje Schmidt-Wigger,
Ursula Reuther and Catherine Pease

IAI

Martin-Luther-Straße 14
66111 Saarbrücken Germany
catler@iai.uni-sb.de

The enormous growth of the Internet has confronted nearly every agent operating in the Internet, human or otherwise, with documents written in unknown languages. In order to allow unrestricted access to these data, the availability of language processing tools within the Internet is a prerequisite. Furthermore, it will be absolutely necessary to interconnect these tools given the almost unlimited number of functionalities which have to be covered. In this paper we describe a practical experiment to connect two machine translation (MT) systems via the Internet, so that the total translation process is distributed across the sites and the Internet is part of the interface.

1 Introduction

The huge growth of the Internet has confronted nearly every agent operating in the Internet, human or otherwise, with documents written in unknown languages. In order to allow unrestricted access to these data, the availability of language processing tools, i.e. multilingual information retrieval, multilingual display, multilingual text generation, translation memories, terminological databases, lexicon servers and machine translation systems, is a prerequisite. Furthermore, it will be absolutely necessary to interconnect these tools given the almost unlimited number of functionalities which have to be covered (in terms of functionalities, language pairs, data formats etc).

Accordingly, Web applications are becoming more and more prominent in natural language processing (NLP) projects. Some of these projects are concerned with the development of standards for the representation of linguistic and meaning structures. Other projects focus more on technical aspects of the intergration of NLP tools.

In one of the most ambitious projects in this area a principled solution to these problems is sought: under the guidance of the United Nations University, multilingual text generation, information retrieval and multilingual display are combined to create future standards for meaning representation within Internet applications (cf. <http://unl.ias.unu.edu/> and <http://www.iai.uni-sb.de/UNL/unl-iai.html>).

Until recently, current single initiatives to offer NLP tools in the Internet have not yet been concerned with standards. Several Machine Translation providers for example are offering their services on the Web. These can be realized as a built-in, frozen system for Web browsers, as an online service

for small translation quantity or as a off-line, email based service for the translation of complete Web pages ([Clements96]).

Natural language processing (NLP) however requires very large linguistic resources in the form of lexicons, corpora and grammars. These resources are very time consuming wrt their development and maintenance, thus making the possibility of interacting different procedures and resources developed at different sites into an interesting and reasonable alternative. Until recently, however, it was necessary that one NLP-tool be frozen in its current state and then be linked to a second system. Further maintenance of the frozen system was impossible. With the development of the Internet the situation may change. Now it is possible to develop one Internet application which is based on resources and procedures coming from different sites, interacting in a way which is not directly visible to the user of the Internet application.

Thus, in the long term, different MT providers are forced and enabled to cooperate. This cooperation has to be based, if not on common standards, at least on common exchange formats, such as prepared in the European project for access to machine translation (OTELO <http://www.otelo.lu²>).

In the project presented here, "CAT2: Traducción Automática Multilingüal"³, two MT systems, located at the University of Mexico City (UNAM) and at the IAI in Saarbrücken have been linked, maintaining the two systems independently at different sites and connecting them via the Internet⁴. This means not only that the two systems are being developed and maintained at different sites, but the whole process of translation is distributed over two continents and at least two different sites. In the remainder of this article, we present the two systems which have been linked via the Internet, the architecture used for translation and the consequences of this architecture for the maintenance of the whole system.

2 The MT systems

CAT2 is an NLP formalism developed for the purpose of multilingual MT (cf. [Sharp and Streiter95]). Within this formalism different grammars and lexicons have been developed by different groups of researchers. A group of research institutions from Belgium, Luxembourg and Germany developed the ANTHEM system for the translation of medical diagnostic expressions (cf. [Ceusters et al. 94], [Streiter and Schmidt-Wigger95]). A consortium consisting of the IAI, the University of PARIS VII and the ERI in Cairo developed an English-German-Arabic MT component for medical classifications (cf. [Pease and Boushaba96]). At the UNAM

²The purpose of the LE-OTELLO project is to allow users of translation tools to have interaction of translation tools, i.e. machine translation systems, translation memories and terminological databases, provided by different vendors.

³This project has been jointly sponsored by CONACYT (Consejo Nacional des Ciencia y Tecnología in Mexico) and the Forschungszentrum Jülich GmbH in Germany within the frame of a mutual agreement on cooperating in the domain of research and technology.

⁴These two systems have both been developed using the CAT2 formalism, thus making it easier to link them, both in terms of the structures, as they use the same formalism and in terms of the representation of information, as they have the same underlying philosophy.

(Universidad Autónoma de México), a bi-directional English-Spanish MT system has been developed mainly for the purpose of translating newspaper articles (cf. [Sharp and Streiter95]). A German-English-French MT component is under development at IAI Saarbrücken, which, similar to the UNAM system, is not tuned to special subject domains (cf.[Streiter96]) (<http://www.iai.uni-sb.de/cat2/cat/en/trans.html>). Systems can be started from a Web page, where the translation is effected either interactively or in batch mode. The translation service is free. In order to offer new language pairs to the user and to explore the possibility of distributed MT, UNAM and IAI decided to link their MT systems while keeping them independent for further development at their respective sites.

3 Linguistic Interface

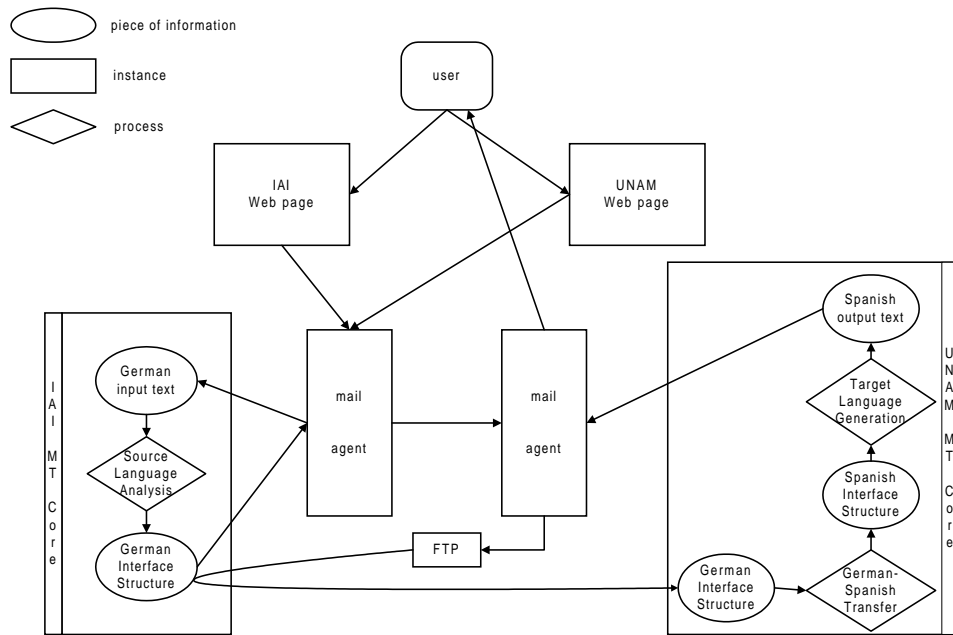
In order to link the two MT systems, a linguistic interface had to be written. This interface has to cope on the one hand with differences between languages (e.g. between Spanish and German) and on the other hand with differences between the conception of the two MT systems. The interface consists of three sub-components, the lexical transfer rules, the feature transfer rules and complex translation rules⁵. While the lexical transfer rules mainly treat differences between the languages (e.g. the lexicon), the latter two mainly treat differences between the two MT components. Whenever it was possible, feature transfer and structural transfer were avoided by a mutual harmonization of linguistic structures. In the future, such linguistic interfaces should be reduced to a minimum by the adoption of standards for linguistic representations as they are developed in different European projects (PAROLE⁶: <http://guagua.echo.lu/langeng/en/le2/le-parole/summary.html>, LRE-EAGLES: <http://guagua.echo.lu/langeng/en/lre1/eagles.html>, MULTILEX: <http://albion.ncl.ac.uk/esp-syn/text/5304.html>).

4 Architecture of the Translation

The translation request can be started from UNAM or IAI Web pages. For this purpose, both project partners added to their Web page the (until then not covered) language pairs Spanish-German, Spanish-French, German-Spanish and French-Spanish. Both Web pages function according to the same principles: If the source language (SL) is Spanish, the input text and the user's email address are sent to the UNAM translation module, if the SL is German or French, the text and the user's reply address are sent to the IAI translation module, as exemplified in the diagram below for the language pair German \rightarrow Spanish. As an alternative, direct email access to the email agents is possible as well.

⁵More information about this linguistic interface can be found in [Streiter et al. 98].

⁶Lexical resources within the PAROLE project, for example, are being collected and standardised for 20 languages in Europe.



Every MT core, however, effects only one part of the whole translation process. The analysis of the SL is affected in this example by the IAI MT core. The result of this analysis is the German Interface Structure. The Interface Structure and the user's reply address are then stored as an object file and transferred via ftp by the partner system, after it has been informed via email of the successful analysis of the text. At the target site the objects are loaded by the MT core, transferred to the target language and generated into Spanish. Finally, the generated output text is sent to the user's reply address.

This architecture, however, implies that the transfer component which operates between the two Interface Structures has to be available within two MT cores at two different sites, which makes the updating of the transfer component complicated. It would be preferable to have one copy of the transfer component at one site only (e.g. Mexico) and to receive this transfer component either within a server architecture or attached to the (Spanish) object to be translated at IAI. As mentioned above, within future applications, however the transfer component could be rendered completely unnecessary through measures of standardization and generally accepted pivot languages.

5 Testing Alternatives

Adding new language pairs in the way described led quickly to a stable test version which can handle almost the same amount of structures as can be handled by the two different systems, even if the linguistic modelling is somewhat different.

To get a useful translation service, however, lexical work on transfer rules continues to represent a major part of the work to be done; work which

becomes even more complicated by the maintenance and updating of the two identical transfer components at the two sites. To circumvent this work, it proved a tempting alternative to use modules of a language common to both systems, i.e. English, as a pivot. This sort of pivot-based approach in distributed MT environment has already been proposed in [Schubert88]. Contrary to the proposal of [Schubert88], we do not use an (English) surface string as pivot, but the (English) interface structure, which is less ambiguous than the related surface string. We thus linked the modules of IAI to the modules of UNAM via the respective English interface structure. Within this approach only feature translation and structure translation has to be accounted for. Lexical transfer is no longer needed because both interface structures are filled with identical, English lexical nodes.

To ensure the feasibility of this approach, we attempted to generate one single lexicon which would be used by both systems. This was done by providing an interface for feature conversion for the English lexical entries of both institutions into an 'interlexical' representation, which is intended to be a system independant description of the semantic and syntactic information contained in the entries of the different lexicons. This approach of resource databases containing a multifunctional lexicon, or, as an alternative, the approach of a common exchange format (see p. 3) for lexical resources, is a central issue in today's MT development ([Thurmair97]). The exchange format adopted within this project could easily be converted into one of the arising standards such as the OLIF lexical interchange format developed within OTELO.

6 Conclusions

We have shown that the use of the Internet offers new perspectives for the development and use of linguistic resources in general and MT applications specifically. Based on simple Internet facilities such as email and ftp we installed and tested a translation service which covered new translation pairs (German-Spanish and French-Spanish) where the whole process of translation is distributed over two continents. In the experiments made, the architecture based on an English disambiguated interface structure proved to be most promising. The unfortunate fact that UNAM closed its translation service some months after the end of the project does not invalidate our approach but shows that for such applications not only technical solutions and standardized exchange formats are needed, but more importantly also long-term political and administrative stability in order to convert the new technological possibilities into real-life applications.

References

- [Ceusters et al. 94] Werner Ceusters, Guy Deville, Emmanuel Herbigniaux, Pierre Mousel, Oliver Streiter, and Geert Thienpont. 1994. The ANTHEM Prototype. IAI WP 31. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Clements96] David Clements. 1996. The value of online MT in the age of the 'cyber society'. Proceedings of the Second Conference of the Association for Machine Translation in the Americas.
- [Pease and Boushaba96] Catherine Pease and Abd Al-Aziz Boushaba. 1996. ARAMED. Extension and integration of Arabic lingware components in a unification-based MT system for the field of medical terminology and classification. In *First KFUPM Workshop on Information & Computer Science (WICS)*, Dhahran, June 9.
- [Schubert88] Klaus Schubert. 1988. The architecture of DLT - interlingua or double direct? In Dan Maxwell, Klaus Schubert, and Toon Witkam, editors, *New Directions in Machine Translation*. Foris Publication, Dordrecht - Holland.
- [Sharp and Streiter95] Randall Sharp and Oliver Streiter. 1995. Applications in Multilingual Machine Translation. In *Proceedings of The Third International Conference and Exhibition on Practical Applications of Prolog, Paris, 4th-7th April*. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Streiter and Schmidt-Wigger95] Oliver Streiter and Antje Schmidt-Wigger. 1995. The integration of linguistic and domain specific knowledge: CAT2 within ANTHEM. In *Proceedings of the Conference on Health Telematics95*, pages 387–392, Ischia, July 2-6. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Streiter et al. 98] Oliver Streiter, Antje Schmidt-Wigger, Ursula Reuther, and Catherine Pease. 1998. Experiments in Distributed MT. In *Workshop on Distributing and Accessing Linguistic Resources, Granada, Spain, May 27*. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Streiter96] Oliver Streiter. 1996. *Linguistic Modeling for Multilingual Machine Translation*. Informatik. Shaker Verlag, Aachen.
- [Thurmair97] Gregor Thurmair. 1997. Exchange interfaces for translation tools. In *MT SUMMIT 97*.