Statistical and hybrid machine translation between all European languages

Publishable executive summary

01 September 2006 – 30 November 2007

Acronym: EuroMatrix
Contract N°: IST 034291
Duration: 30 months from 01/09/2006

Coordinator:
Prof. Dr. Hans Uszkoreit
Saarland University
D-66123 Saarbrücken
uszkoreit@coli.uni-sb.de

Scientific coordinator:
Dr. Philipp Koehn
Edinburgh University
pkoehn@inf.ed.ac.uk

PROJECT FUNDED BY THE EUROPEAN COMMUNITY UNDER THE SIXTH FRAMEWORK PROGRAMME FOR RESEARCH AND TECHNOLOGICAL DEVELOPMENT.

Information Society
Technologies

Contractors involved

Saarland University Computational Linguistics	Germany
The University of Edinburgh School of Informatics	United Kingdom
Univerzita Karlova v Praze/ Charles University Prague Institute of Formal and Applied Linguistics	Czech Republic
CELCT Center for the Evaluation of Language and Communication Technologies	Italy
MorphoLogic	Hungary
Group Technologies AG	Germany

Summary description of project objectives

EuroMatrix aims at a major push in machine translation technology applying the most advanced MT technologies systematically to all pairs of EU languages. Special attention is being paid to the languages of the new and near-term prospective member states. As part of this application development, EuroMatrix designs and investigates novel combinations of statistical techniques and linguistic knowledge sources as well as hybrid MT architectures.

EuroMatrix addresses urgent European economic and social needs by concentrating on European languages and on high-quality translation to be employed for the publication of technical, social, legal and political documents.

EuroMatrix aims at enriching the statistical MT approach with novel learning paradigms and experiment with new combinations of methods and resources from statistical MT, rule-based MT, shallow language processing and computational lexicography/morphology.

EuroMatrix has the following concrete objectives:

- Translation systems for all pairs of EU languages, with a special focus on the languages of new and near-term prospective member states
- Efficient inclusion of linguistic knowledge into statistical machine translation
- The development and testing of hybrid architectures for the integration of rule-based and statistical approaches
- Organization, analysis and interpretation of a competitive annual international evaluation of machine translation with a strong focus on European economic and social needs
- The provision of open source machine translation technology including research tools, software and data
- A systematically compiled and constantly updated detailed survey of the state of MT technology for all EU language pairs based on the developed systematic translation between all EU languages, the comparative MT evaluations and an inventory of available and needed tools, components, lingware and data.

Work performed

The partners have successfully completed a range of key tasks in the first half of the project, both in advancing the theory, developing implementations, organizing evaluations, and disseminating results to a wider research community and to potential users.
The project pursued the integration of statistical and linguistic knowledge for improved machine translation in several ways, involving all the academic and industrial partners.

Members of the consortium developed the open-source implementation of Moses, a statistical machine translation system that incorporates factored statistical translation models, and applied it to tasks where linguistic representations could be used to improve translation quality, especially for the translation into morphologically rich languages. A high level of support was given to a rapidly growing set of users of Moses within and beyond the project. A mathematical model for tree transfer of deeper linguistic representations has been designed and a first version of an MT system based on this model has been implemented and tested. Several partners designed, implemented and evaluated different ways to combine modules of statistical and rule-based translations engines into hybrid architectures.

All activities towards improved MT methods are backed by the organisation of an open evaluation campaign by an independent entity within the consortium, which is dedicated to the evaluation of natural language methods. The EuroMatrix project successfully staged a dry-run evaluation in Spring 2007 to prepare the similar evaluations planned for Spring 2008 and early 2009.

Results achieved so far and expected end results

During the first half of the project, the EuroMatrix project has made tremendous progress towards all its ambitious goals, not only in terms of advancing the state of the art in machine translation methods, but also concerning the dissemination of the technology to foster research in academia and the commercial sector.

The open-source SMT system Moses, developed by EuroMatrix consortium partners, has already found very widespread use within and beyond the consortium, including small and medium sized companies in Europe and big European organisations. Given the high level of support provided to the users, Moses is by now the canonical platform for work in the SMT paradigm and will attract additional development effort even from outside of the project.

The investigation of MT approaches based on deeper linguistic representations and the implementation of hybrid combinations of statistical and rule-based knowledge already provided promising first results and very valuable insights into the relevant research questions, which will be exploited in further work within the project. We expect that these activities will result in a significant improvement of translation quality for language pairs where both rule-based MT technology and parallel corpora exist.

The distribution of parallel corpora and annotations such as sentence and word alignments has provided significant support to research and development of machine translation and language processing technology for European languages, as did the organisation of a first evaluation campaign and associated research workshop that attracted a very large audience. Not only does the evaluation campaign organised by EuroMatrix constitute a remarkable service to the community; by integrating innovative approaches at every stage it also provides a valuable basis for further research efforts on evaluation techniques.

Whereas the dry-run evaluation campaign involved translation from and into Czech, further activities will include translation from and into Hungarian and between languages other than English. Work on the use of parallel corpora in all 23 EU languages has started and will provide a complete coverage of all 506 pairs of EU languages by the end of the project.

The successful organisation of the first Machine Translation Marathon in a sequence of three similar planned events additionally helped to spread the knowledge about the work done within EuroMatrix to a wider community. We expect that the second and third Machine Translation Marathons will involve even more participants from Eastern Europe and will thus have an even bigger impact on language pairs where new technology is most urgently needed.

Intentions for use and impact

The motivation behind the EUROMATRIX project is to build a foundation for machine translation research for all European languages. By organising annual competitions and workshops, we want to create a forum for researchers to gather and exchange ideas. By providing tools and resources, we want to engage as many researchers as possible. By engaging in research in promising directions in the field of statistical and hybrid machine translation, we want to extend the state of the art in machine translation systems.

The EuroMatrix project has a strong commitment to make implementations and resources open source and freely available. We expect that the technology developed within the project will be widely applied by researchers, industry, and public institutions, and we already see strong evidence for this kind of dissemination.

The second Machine Translation Marathon will be accompanied by a conference "TRANSLINGUAL EUROPE" which aims to inform invited representatives of industry, commerce, research and administration about recent progress in translation technology. In order to determine the requirements and the state of the art with respect to the European languages, the results of the evaluation campaign and a survey of available products and resources will be presented and discussed. A third theme of the conference is the discussion of opportunities and challenges for European research, development and technology transfer in this important application area of information technology.

Cooperation with related research and evaluation efforts on a European and international level will furthermore facilitate the comparison of the performance of our systems on a larger scale. Judging whether the translation quality of our systems is suitable for a particular purpose is outside the scope of this project, but would be a suitable topic for subsequent undertakings.

Publishable results

The project has so far generated more than a dozen peer-reviewed conference and workshop publications as well as six Msc/Diploma theses. The deliverables of the scientific work packages are all intended for public distribution and will be made available for download from the project website. The project has generated significant datasets and contributed to an open-source implementation of a statistical MT system which are all available for free download for interested users.