The CASMACAT Project: The Next Generation Translator's Workbench

Daniel Ortiz-Martínez, Germán Sanchís, Francisco Casacuberta,
Vicent Alabau, Enrique Vidal, José-Miguel Benedí, Jesús González-Rubio,
Alberto Sanchís, and Jorge González

D. Sistemas Informáticos y Computación
Universitat Politècnica de València
Camino de vera s/n, 46022, Valencia, Spain
{dortiz,gsanchis,fcn,valabau,evidal,
jbenedi,jegonzalez,josanna,jgonzalez}@dsic.upv.es

Abstract. The goal of the CASMACAT (Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation) research project is to build the next generation translator's workbench to improve productivity, quality, and work practices in the translation industry. The CASMACAT project is co-funded by the European Union under the Seventh Framework Programme (FP7) Project and involves a total of four partners, namely, the University of Edinburgh (United Kingdom), the Polytechnic University of Valencia (Spain), the Copenhagen Business School (Denmark) and the Spanish translation company Celer Soluciones. The work to be developed during the project is distributed in eight workpackages and ranges from user interface studies and user modelling to new machine translation techniques that allow human users and machine translation systems to collaborate in order to obtain high quality translations.

1 Introduction

European integration and globalisation beyond it increases cross-border commercial, cultural, and political interaction. However, while the significance of political borders diminishes, the risk remains that the world will stay fractured by linguistic boundaries. The need to address each individual in a language that she speaks, and ideally in her native language, requires a huge amount of translation work.

While there have been significant improvements to machine translation technology, the vast majority of this work is targeted towards bulk translation that is good enough or fit for use. A user on the Internet is satisfied with a rough translation, if it fills her information need. Opposed to that is the demand for high quality translations by the marketplace: the translation of reports and announcements of multi-national organizations, marketing material and product descriptions of commercial companies, and many other localization needs. Such high quality translations are still almost exclusively provided by human translators.

Productivity of human translators can be increased with computer aided translation (CAT) tools: translation memories are standard in the translation industry, but post-editing machine translation output is only slowly becoming an increasingly used practice [8]. The current integration of machine translation technology into human translators work processes is often done overly simplistic, breaks their work practices, and it is widely resisted. Hence, the CASMACAT project will carry out in-depth study of translator behaviour to tailor the tools to the requirements of translators, and not the other way around. In the CASMACAT project a novel workbench that will increase the productivity of human translators is proposed by addressing their needs for the right type of assistance at the right time.

2 Project Goals

In recent years, there has been measurable progress in the field of machine translation, both in terms of quality and increased use, due to the widespread adoption of statistical methods. Vigorous research is carried out in academic and commercial research labs. As it will be described below, there has been some progress in aiding human translators, but the vast potential of creating a new workbench for human translators is mostly unfulfilled. The CASMACAT project[1] will try to generate such workbench by transferring the methods from the statistical machine translation community to the task of assisting human translators. Whereas the translation technology is ripe enough, design issues of the user interface and its acceptance by the translator have been widely neglected. The development of such tools must not simply follow technical possibilities, but it should be driven by a better understanding of the behaviour of human translators.

This is the goal of CASMACAT project: to carry out cognitive analysis that provides insight into the human translation process to guide our development of a new workbench for translators. The partners of the consortium are drawn from the leading groups in cognitive modelling of translators, statistical machine translation and computer aided translation. The commercial partner Celer Soluciones, a translation agency with significant experience in innovative computer aided translation tools, will evaluate the novel methods to provide feedback and guidance.

Currently, Europe is leading in providing translation services. To maintain and extend this leadership role, to a large degree driven by the linguistic diversity in a common economic and political union, European translation companies must continue to improve their work practices and adopt advanced technology. The CASMACAT projects aims to contribute to this process by developing novel types of assistance and bring them to the users.

2.1 Scope

Human translation is performed by different types of translators, tackles different text types, and deals with different language pairs.

[1] http://www.casmacat.eu/

- **Translators**: The needs of user communities will be addressed, ranging from professional translators to volunteer translators.
- **Text Types**: Much of what professional translators translate is repetitive, technical material. In contrast, volunteer translators are more commonly interested in generally accessible material, or technical content in their own area of expertise.
- **Language Pairs**: The methods developed in the CASMACAT project are generally language-independent. Work will be done on languages for which the necessary data resources are available and for which project participants have sufficient expertise.

2.2 Workbench

The CASMACAT project will develop a new open source workbench for human translators. All functionalities developed by the project will be integrated in a web-based online service which may also be installed locally on the desktop of a translator. The availability as web service will make it easy to integrate it into existing translation workflows. The academic partners of the consortium have previously developed such tools and will synthesize their experience.

Translog [3, 4, 2] is the leading tool for analyzing text production processes developed by CBS, which has been used in a large number of translation process studies [9]. MIPRCV-IMT is a reference implementation of novel techniques in interactive-predictive machine translation (or interactive translation prediction from now on) that were developed by UPVLC [1, 7, 6]. Caitra [5] was developed by UEDIN in order to explore new types of assistance for human translators.

The new CASMACAT workbench will integrate all relevant functionalities of the pre-existing tools. All new features explored in the project will be implemented within the new system.

2.3 Cognitive Analysis

An important objective of the CASMACAT project is to gain insight into the cognitive processes involved in human translation. How large are the text segments actively considered by a translator (the whole text, individual sentences, or only subsentential segments of limited length)? What are the subtask that a translator spends most time on (e.g., understanding the source text, looking up unknown words, investigating lexical translations, syntactic restructuring of the sentence, ensuring fluency of the output)? How does translation differ from well-studied simpler cognitive processes such as reading and text production?

The cognitive analysis will inform the design of the CASMACAT translation workbench in a range of ways. It will determine what types of assistance are offered to the translator, what information should be displayed on the screen, and what information should be hidden as it would be distracting. The different versions of the user interface will be evaluated in user studies using eye-tracking and other commonly used methods in cognitive analysis.

2.4 Advanced Computer Aided Translation

CASMACAT will use well-established statistical methods and explore novel approaches in order to generate and disambiguate translation proposals. Dynamically generated translation options will be sent to and visualized in an interactive translation assistance tool. Input from the user will be given by means of keyboard, mouse or electronic pen (e-pen). Two basically different approaches to computer assisted translation (CAT), Interactive Translation Prediction and Interactive Editing, will be developed, compared and evaluated and a cognitive model of the translator will be developed to predict the translators performance.

A novel reworking of the idea of interactive translation prediction (ITP) will allow for the construction of systems that produce high-quality results by placing a human operator at the centre of the production process. The ITP paradigm embeds a statistical MT engine within an interactive editing environment. The human serves as the guarantor of high quality; the role of the automated systems is to ensure increased productivity by proposing well-formed extensions to the current target text, which the operator may then accept, correct or ignore. Interactivity allows the system to take advantage of the human-validated portion of the text to improve the accuracy of subsequent predictions.

3 Project Partners

The project is made up of 4 partners from 3 European countries: 3 universities (United Kingdom, Denmark, Spain) and 1 company (Spain):

1. University of Edinburgh, United Kingdom
2. Polytechnic University of Valencia, Spain
3. Copenhagen Business School, Denmark
4. Celer Soluciones, Spain

The partners are complementary due to their different scientific expertise statistical machine translation, computer aided translation, open source software development, cognitive studies of natural language processing, translation process research, and delivery of translation services. Each partner's expertise and their research interest are a good fit with their assigned tasks.

- **Academic Partners**
  1. **UEDIN**: University of Edinburgh, School of Informatics[2]
     Drs. Philipp Koehn, Frank Keller
     Areas of Expertise: statistical machine translation, open source development, machine learning, computer aided translation, evaluation, cognitive modeling of language processing.
  2. **UPVLC**: Universitat Politécnica de Valéncia, Technical Institute for Informatics[3]

[2] http://www.inf.ed.ac.uk/
[3] http://www.iti.upv.es/

Drs. Francisco Casacuberta, Enrique Vidal
Areas of Expertise: statistical machine translation, interactive machine translation, finite state approaches to machine translation, handwritten text recognition.

3. **CBS**: Copenhagen Business School, Department of International Language Studies and Computational Linguistics [4]
Drs. Michael Carl, Arnt Lykke Jakobsen, Jakob Elming, Christopher Teplovs
Areas of Expertise: translation process research, statistical machine translation, data visualisation, web-design and collaborative web applications.

– **Company**
1. **CS**: Celer Soluciones[5]
Roberto Silva, Enrique Diaz de Liaño
Areas of Expertise: translation services, evaluation and use of novel computer ai

4 The CASMACAT Workbench

The CASMACAT Workbench allows users to enter documents in a source language, and then receive assistance in translating them into a target language. The assistance is based on information from machine translation systems.

Figure 1 shows a diagram of the architecture of the CASMACAT Workbench. According to the figure, there are five major components:

– **Editor**: is the interface between the user and the functionalities provided by the CASMACAT workbench. It communicates with two different components: the GUI server and the translation server. Input can be given via keyboard, mouse or e-pen. In addition to this, the behaviour of the translator can be studied by means of an eye-tracking system.
– **GUI server**: serves web pages to the editor interface. Handles information useful for cognitive analysis of the translation process, including logging and replay information. Such information is stored into the workbench databases.
– **Translation server**: provides translation services including regular MT and interactive MT by means of a specific application programming interface (API). In addition to this, it is also used for authentication and document management purposes (document preprocessing, storing of translation information per each source document, etc.).
– **HTR server**: handles user interactions by means of an e-pen. It offers a specific API to the CASMACAT editor.
– **Database**: two databases have been identified. The first one should be visible to the translation server and store user information, documents, partial or total translations of the documents. The second one should be visible to the GUI server and store replay information.

[4] http://uk.cbs.dk/research/departments_centres/institutter/isv/
[5] http://www.celersol.com/?idioma=en

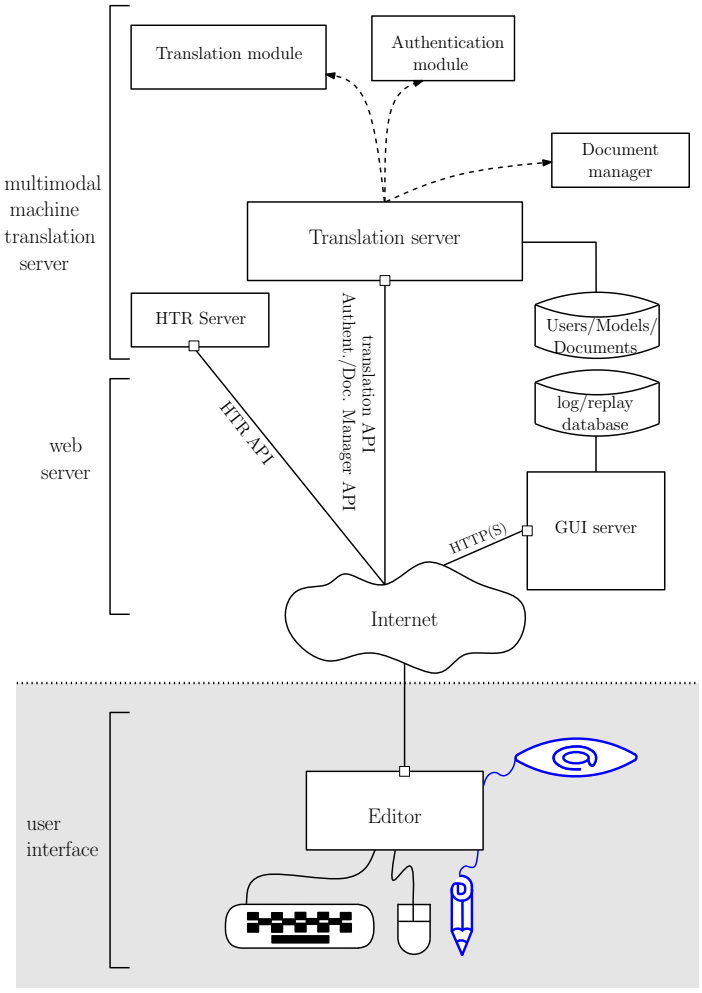

Fig. 1. CASMACAT workbench architecture.

5 Workpackages

The CASMACAT project is structured in a total of eight workpackages. In the rest of this section we briefly describe each workpackage and provide some information about the timing of tasks.

5.1 Workpackage Description

WP1: User Interface Studies, Cognitive and User Modeling
This WP lays the empirical foundations for the development of the CASMACAT

workbench. A series of experiments will establish basic facts about translator behaviour in computer-aided translation, investigating the usefulness of visualisation option in post-editing and interactive translation, for different types of text and for translators with different degrees of expertise.

WP2: Interactive Translation Prediction
This work-package is devoted to investigating innovative approaches to human-MT interaction covering three important dimensions: effectiveness, efficiency and communication. Improved interaction protocols and more efficient search procedures will allow the systems to supply more useful and faster predictions. Finally, sophisticated interaction modalities will enrich the information supplied to the system and will facilitate the generation of user feedback, thereby enhancing the overall performance and usability. Some of the topics that will be studied withing WP2 include alternative search and machine learning criteria from prediction, use of e-pen as additional input modality, use of new SMT models for ITP, etc.

WP3: Interactive Editing
In this work package, three main lines of research are pursued: (1) enriched post-editing, where augmented information about the machine translation output aids the post-editor to highlight possible weaknesses of the translation, (2) authoring assistance, where knowledge about the document, prior translations, and general knowledge about language use helps the editor to find better translations, and (3) automatic reviewing, where a second pass over the translation is aided by automatically detecting untranslated or added material, as well as ensuring consistent translation of terminology.

WP4: Adaptive Translation Models
Human interaction offers another unique opportunity to improve the performance of the ITP systems by tuning the translation models. This WP tackles the problem of how text validated by the human translator can be used to adapt the system to changing environment.

Techniques based on on-line learning, cache-based approaches, Bayesian adaptation and/or reinforcement learning will be explored for their full integration with the statistical translation model training.

WP5: Integration
All the methods developed by the CASMACAT project will be integrated into a new translators workbench, the CASMACAT workbench. Its development will take place throughout the project in three distinct stages: specification of requirements, implementation of core functionality, and integration of novel methods.

WP6: Evaluation
In this work package, the CASMACAT workbench will be exposed to a wider

community of users and engage the localization industry to gain wider adoption. In addition to this, field trials will carried out to study the use of the workbench in a real-world environment, promote it to community translation platforms and the language service industry.

WP7: Dissemination
This workpackage is devoted to explore ways to reach the different type of potential users of the CASMACAT workbench as well as to publish scientific results.

WP8: Management
This workpackage is devoted to the different aspects of project management.

5.2 Timing of Tasks

CASMACAT is a 3-year project from November 2011 to October 2014. The workplan is strongly focused on the development of the CASMACAT workbench (WP5). Specifically, three different releases will be developed:

1. **Initial release** (month 6): restricted to the capability of post-editing machine translation output provided by a static system and logging of translator activity.
2. **Beta release** (month 18): which integrates advanced interactive machine translation methods and a number of different features, including the use of confidence measures, integration with translation memory, online learning for interactive prediction, support for eye tracking, etc.
3. **Final release** (month 30): which integrates all remaining advances of the CASMACAT project.

6 Related Projects

The CASMACAT project is funded in the same cycle with two related projects which also broadly aim at improving the interaction of machine translation technology and human translators. The MATECAT (Machine Translation Enhanced Computer Assisted Translation) project[6] mainly aims at the improvement of statistical machine translation technology to more closely suit the needs of translators working in a post-editing workflow. The ACCEPT (Automated Community Content Editing Portal) project [7] aims at improving the role of machine translation in community and volunteer translation platforms.

While the specific aims of the three projects are quite different, there are a few tasks where they pursue similar goals and where co-ordination between the three projects is essential to obtain synergy and not duplication of efforts. It

[6] http://www.matecat.com/matecat/the-project/
[7] http://www.accept.unige.ch/index.html

is expected a especially close co-ordination between the MATECAT and CASMACAT project, including joint project meetings and joint dissemination events.

In addition to the above mentioned MATECAT and ACCEPT research projects, there is also another project which is not completely focused on machine translation but is also related to the CASMACAT project, the TRANSLECTURES (Transcription and Translation of Video Lectures) project [8]. The aim of TRANSLECTURES is to develop innovative, cost-effective solutions to produce accurate transcriptions and translations in the VideoLectures.NET web portal, a free and open access educational video lectures repository.

References

1. Alabau, V., Ortiz-Martínez, D., Sanchis, A., Casacuberta, F.: Multimodal interactive machine translation. In: ICMI-MLMI '10: Proceedings of the 2010 International Conference on Multimodal Interfaces. Beijing, China (2010)
2. Carl, M., Jakobsen, A.L.: Towards statistical modelling of translators activity data. International Journal of Speech Technology 12(4) (2010), http://www.springerlink.com/content/3745875x22883306/
3. Jakobsen, A.L.: Logging target text production with Translog. In: [**?**]. pp. 9–20 (1999)
4. Jakobsen, A.L.: Investigating expert translators processing knowledge. In: [**?**]. pp. 173–189 (2005)
5. Koehn, P.: A web-based interactive computer aided translation tool. In: Proceedings of the ACL-IJCNLP 2009 Software Demonstrations. pp. 17–20. Association for Computational Linguistics, Suntec, Singapore (August 2009), http://www.aclweb.org/anthology/P/P09/P09-4005
6. Lagarda, A.L., Civera, J., Juan, A., Casacuberta, F.: Interactive pattern recognition and human language technology for di gital audiovisual content processing. Journal of Machine Learning Research 11: Workshop on Applications of Pattern Analysis, 103–110 (09 2010)
7. Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.: Online learning for interactive statistical machine translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT). pp. 546–554. Los Angeles (Jun2 2–4 2010)
8. Plitt, M., Masselot, F.: A productivity test of statistical machine translation post-editing in a typical localization context. The Prague Bulletin of Mathematical Linguistics 93, 7–16 (2010)
9. Schou, L., Dragsted, B., Carl, M.: Ten years of Translog. Copenhagen Studies in Language 38, 37–48 (2009)

[8] http://llach.dsic.upv.es/ translectures/