TTC project: Public Fact Sheet

TTC - Terminology Extraction, Translation Tools and Comparable Corpora

Project duration: 1st of January 2010 to 31st of December 2013 (36 months)

The TTC project aims at leveraging machine translation tools (MT tools), computer-assisted translation tools (CAT tools) and multilingual content management tools by automatically generating bilingual terminologies from comparable corpora in five European languages (English, French, German, Spanish and one under-resourced language, Latvian), as well as in Chinese and Russian.

Objectives of the project

The need for linguistic resources (terminologies, lexicons, translation memories, etc.) is overwhelming in any natural language application, but the problem is especially difficult for translation applications because of cross-linguistic divergences and mismatches that arise from the perspective of the lexicon. Lexicons and terminologies play indeed a central role in any machine translation tool, regardless of the theoretical foundations upon which the MT tool is based (statistical machine translation, rule-based machine translation, example-based translation...). Computer-assisted translation tools heavily use terminologies and translation memories to assist the human translator in the translation process, e.g. to create and maintain terminologies. Another functionality of computer-assisted translation is a dictionary based generation of rough translations. In fact, some advanced computer-assisted translation solutions include controlled machine translation.

Besides MT, automatic translation of terminologies from one controlled vocabulary into another is essential to the integration and the use of diverse information systems. Bilingual terminologies for several languages are adaptive and interoperable solutions for managing multilingual content and communication.

Briefly the TTC project work aims at:

- Compiling and using comparable corpora;
- Using a minimum of linguistic knowledge for candidate term extraction;
- Defining and combining different strategies for term alignment;
- Developing an open platform for use with MT and CAT tools including solutions to manage comparable corpora as well as terminologies ;
- Demonstrating the operational benefits on MT tools and CAT tools.

All these target outcomes have similar impacts, i.e. bettering translation in order to overcome language barriers through technological means. Final outcomes of the TTC project aim at improving translation activities from industry documentation to multilingual content management.

Project description

The TTC project focuses on the automatic acquisition of aligned bilingual terminologies for computer-assisted translation and machine translation. To do this, important steps of the project are the

automatic extraction of monolingual terminologies and the bilingual alignment of the extracted terminologies from a large set of multilingual corpora.

Such terminologies could be extracted from parallel corpora, i.e. from previously translated texts, but such corpora are scarce. Previously translated data is still sparse and only available for some pairs of languages and few specific domains. Thus, no parallel corpora are available for most of specialized domains, especially for emerging domains (such as renewable energy). As a consequence, the project develops methods and tools for automatic extraction of terminologies from comparable corpora, i.e. from corpora corresponding to a same domain, but not necessary being a translation from each other. It also develops tools for gathering (topical web crawler) and managing these comparable corpora and for managing terminologies.

At the end of the TTC project, a platform will be set up to compile and manage comparable corpora using standards (TMF, TBX) and the existing open source UIMA framework. An evaluation and a validation of this work will be done by the consortium on CAT tools and Machine Translation tools. Translation of technical documents for aerospace and IT domain will be done using CAT and MT techniques to assess impact of the TTC project outputs.

Expected results and Impacts

The TTC project develops generic methods and tools for automatic extraction of terminologies and alignment algorithms including adaptors to domains and languages, in order to break the lexical acquisition bottleneck in both statistical and rule-based machine translation.

It will also develop or adapt tools for gathering and managing these comparable corpora and for managing terminologies. In particular, a topical web crawler and an open terminology platform will be developed. The platform will allow to create thematic corpora given some clues (such as terms or documents on a specific domain), to extract monolingual terminology from such corpora, to create a comparable corpus in a target language from a corpus in a source language, to align bilingual terminologies, to choose the tools to apply for terminology extraction, to expand a given corpus and to export monolingual or bilingual terminologies in order to use them easily in automatic and semi-automatic translation tools.

Impact of translation tools and methods resulting from the TTC project will be evaluated in four domain of application:

- Computer Assisted Translation (CAT) tools;
- Machine Translation (MT) tools;
- Terminology management tools.

Participant role in the project

UN-LINA: Besides the role of coordinator, UN-LINA and the Natural Language Processing team will work on comparable corpus compilation, terminology extraction, bilingual term alignment, architecture for terminology acquisition and the definition and development of a platform based on UIMA to handle comparable corpora and integrating NLP tools for morphosyntactic tagging and terminology extraction.

IMS-STUTTGART: IMS will apply its competences in computational linguistics to TTC project. IMS will especially contribute to tasks related to monolingual term extraction (MWT and SWT). Finally, IMS will bring its expertise on NLP tools in the area of morphology and syntax.

UL-CTS: The Center for Translation Studies from Leeds University will play an important role in corpus collection and term extraction activities. UL-CTS will lead the evaluation of the impact of TTC results on MT tools.

SOGITEC: As an industrial partner, SOGITEC will provide its end user vision to help the consortium in the definition of requirements and specifications. SOGITEC will perform the evaluation of TTC impacts on computer-assisted translation tools.

SYLLABS: Syllabs R&D team will bring its skills on the development and optimization of multilingual linguistic resources such as corpora and lexica for NLP. Syllabs will also work on the focused web crawler of multilingual data.

TILDE: TILDE is a company specializing in language technologies whose role in the TTC project is to develop the terminology management platform, work on procedures for term extraction and term alignment and finally optimize the process of translation in localization services. Furthermore, TILDE is responsible for dissemination activities.

EURINNOV: EURINNOV role in the project is related to management and dissemination activities of the TTC project. EURINNOV ensures support to Administrative, Financial and Legal management as well as Scientific and Technical coordination.

Consortium

Partner	Short name	Country
Université de Nantes (coordinator)	UN-LINA	France
Universität Stuttgart	IMS	Germany
University of Leeds	UL-CTS	UK
Sogitec Industries	SOGITEC	France
Syllabs SARL	SYLLABS	France
Tilde SIA	TILDE	Latvia
Eurinnov	EURINNOV	France