Project overview

Computational applications in the translation field suffer from the terminology bottleneck; this applies to both CAT (computer-assisted translation) and MT (machine translation): there is a lack of term-related resources (dictionaries, glossaries, etc.), especially for upcoming or rapidly developing technical domains. Moreover, current automatic approaches suffer from the scarcity of parallel corpora.

The TTC project explores term extraction from comparable corpora, i.e. from texts of the same domain (and possibly genre) in different languages which are not translations of each other. Such texts are available on the internet, as well as in companies. TTC develops techniques for the extraction of monolingual term candidates and their contexts for English, German, French, Spanish, Latvian, Russian and Chinese. A second step is term alignment: for each language pair treated, monolingual term candidates are compared to identify equivalent candidates. TTC explores different symbolic and statistical procedures for this purpose. The output of TTC are single-word terms, multi-word terms and their equivalents, as well as contextual data.

TTC develops software for the above-mentioned languages and tests it on the selected language pairs. The tools are provided as a standalone package (download and run) and as a Web Service. They include tools for corpus crawling and corpus management, for monolingual term candidate extraction and for equivalence candidate identification (term alignment). An integration with EuroTermBank and with selected tools for computer-assisted translation (CAT) and machine translation (MT) will be provided. Professionals from the translation, localization and/or documentation business interested in testing early prototypes should contact the coordinators.

www.ttc-project.eu

Consortium

University of Nantes (coordinator)	**LINA**
University of Stuttgart	**IMS**
University of Leeds	**CTS**
Sogitec Industries	**SOGITEC**
Syllabs SARL	**SYLLABS**
Tilde SIA	**TILDE**
Eurinnov	**EURINNOV**

Contact:
scientific-contact@ttc-project.eu

The project has received funding from the European Community's
Seventh Framework Programme (FP7/2007-2013)
under Grant Agreement Number 248005.

TTC
Terminology Extraction, Translation Tools and Comparable Corpora

www.ttc-project.eu

Aim

The TTC project aims at leveraging machine translation (MT) systems, computer-assisted translation (CAT) tools and multilingual content (corpora and terminology) management tools by generating bilingual terminologies automatically from comparable (non-parallel) corpora in seven languages: five EU languages (English, French, German, Spanish, Latvian) as well as Chinese and Russian, and twelve translation directions:

- Chinese-English, Chinese-French
- English-French, English-German, English-Russian, English-Latvian, English-Spanish
- French-German, French-Russian, French-Spanish
- German-Spanish
- Latvian-Russian

Objectives

The following are the main TTC project goals:

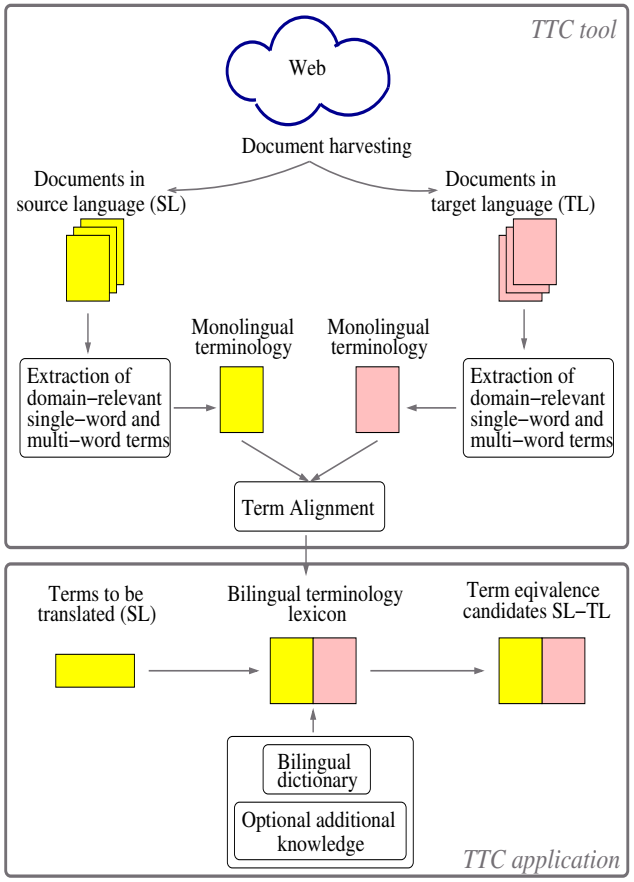
- compile and use comparable corpora, e.g. harvested from the Web;
- assess approaches that use a minimum of linguistic knowledge for term candidate extraction;
- define and combine different strategies for term alignment;
- develop an open Web-based platform including solutions to manage comparable corpora and terminologies, which will also be available for use with MT systems and CAT tools;
- demonstrate the operational benefits of the term extraction approaches on MT systems and CAT tools.

Results and impacts

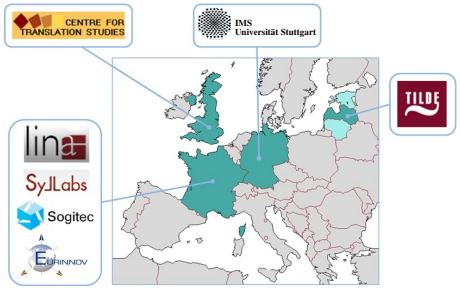
The expected results of the TTC project include the following:

- domain-specific resources (lemmatized and POS-tagged comparable corpora, monolingual and bilingual terminologies) for subdomains of renewable energy and computer science, for the seven project languages;
- a focused crawler to compile comparable corpora;
- a toolkit to handle comparable corpora and to extract terminologies;
- generic methods and tools for automatic terminology extraction and alignment, as well as algorithms for recognizing term variants, and for morphological analysis and grouping of terms;
- an open terminology platform.

Multilingual terminology mining chain



The TTC consortium



Multilingual terminology mining chain

The TTC tool consists of several processing steps.

- Compilation of domain-specific comparable corpora:
  For each language pair, two domain-specific corpora are built, one for the source language (SL) and one for the target language (TL). The corpora consist of documents extracted by *crawling the web*, i.e. by gathering documents from the Web with content related to a specific domain. The extracted corpora need not be parallel, i.e. translations of each other, but they can rather be composed of similar, comparable documents. Therefore, they are called *comparable corpora*.
- Extraction of domain-relevant terminologies:
  The terminology extraction process involves two steps. Firstly, from a SL and a TL corpus, words (single-word terms) and word sequences (multi-word terms) such as "noun + noun", "adjective + noun" are extracted. From the set of the extracted candidates, the *domain-relevant terms* are then identified resulting in two *monolingual terminology lists*.
- Bilingual term alignment:
  SL and TL terminologies extracted from comparable corpora are aligned to each other. The result of the alignment step is a *bilingual domain-specific terminology lexicon*.

The resulting bilingual terminology lexicon can be used as an input to CAT tools and MT systems.

- CAT tools:
  The extracted bilingual terminology lexicon can be integrated into computer-assisted translation (CAT) tools which are used by human translators. The CAT tools provide the user with TL equivalences. The translator can then choose an optimal translation for a SL term.
- MT systems:
  MT systems can benefit from the bilingual terminology lexicon since they often have dictionaries with a limited number of specialised terms. We expect that an additional domain-specific lexicon will improve the automatic translation of specialised documents.

The TTC project consortium

The TTC project consortium brings together seven partners from four different countries: France, Germany, United Kingdom and Latvia. The project coordinator is the French laboratory LINA (Laboratoire d'Informatique de Nantes) from Université de Nantes. Further academic partners are CTS (Center for Translation Studies at the University of Leeds and IMS (Institut für Maschinelle Sprachverarbeitung) from Universität Stuttgart. Industrial partners are the French companies Sogitec and Syllabs and the Latvian company Tilde. Management is handled by Eurinnov, a French consultancy company, specializing in the support of innovative SMEs and research institutions.