TTC: Terminology Extraction, Translation Tools and Comparable Corpora

European Community
Seventh Framework Programme (FP7/2007-2013)
STREP
248005
http://www.ttc-project.eu/

List of partners
Université de Nantes, France (coordinator)
Universität Stuttgart, Germany
University of Leeds, United Kingdom
Sogitec Industries, France
Syllabs SARL, France
Tilde SIA, Latvia
Eurinnov, France

Project duration: January 2010 — December 2012

Summary

In scientific domains, resources like parallel corpora and bilingual dictionaries are often not available. As a consequence, translators spend a lot of time to create and manage terminology lists. Similarly, the lack of parallel data makes it difficult to build statistical machine translation systems.

The project TTC aims at providing data for machine translation systems, computer-assisted translation tools, and terminology management tools by automatically generating bilingual terminologies from comparable corpora. The project covers several languages of the European Union (English, French, German, Latvian and Spanish), as well as Chinese and Russian. To this end, a tool chain for compiling document collections, for terminology extraction and for bilingual term alignment is being developed, which concludes with exporting terminology data into CAT tools and MT systems.

Domain-specific corpora of several languages are collected by using a focused crawler. They then undergo pre-processing (tokenizing and POS-tagging): the project relies on flat linguistic analysis as it is available for most languages. For each language covered by the project, monolingual term extraction is performed. A part of the term extraction step consists in the identification of term variants, which provide valuable information for terminologists and can also help to deal with data sparsity.

The extracted terms are then grouped into bilingual term equivalent pairs (term alignment) using different approaches (context vector based term alignment and lexical alignment strategies). The resulting bilingual term lists can then be fed into translation systems and CAT tools. The use of terminology in machine translation tasks is also regarded as a form of extrinsic evaluation of the output of the tool chain.