TTC: Terminology Extraction, Translation Tools and Comparable Corpora

Helena Blancafort, Syllabs, Universitat Pompeu Fabra, Barcelona, Spain
Béatrice Daille, LINA, Université de Nantes, France
Tatiana Gornostay, Tilde, Latva
Ulrich Heid, IMS, Universität Stuttgart, Germany
Claude Mechoulam, Sogitec, France
Serge Sharoff, CTS, University of Leeds, England

The need for linguistic resources in any natural language application is undeniable. Lexicons and terminologies play indeed a central role in any machine translation tool, regardless of the theoretical foundations upon which the machine translation tool is based (e.g. statistical machine translation or rule-based machine translation). The EU project TTC ('Terminology Extraction, Translation Tools and Comparable Corpora') aims at leveraging machine translation tools, computer-assisted translation tools, and terminology management tools by automatically generating bilingual terminologies from comparable corpora in several European Union languages (English, French, German, Latvian and Spanish), as well as in Chinese and Russian. The TTC project will integrate developed and existing tools in an online platform including a tool to compile and handle comparable corpora, as well as a terminology management tool. The platform will be based on Web Services and will use reputable open solutions such as UIMA (Unstructured Information Management Architecture) and EuroTermBank.

1. Introduction

The need for linguistic resources in any natural language application is undeniable. 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 machine translation (MT) tool is based (e.g. statistical machine translation or rule-based machine translation, example-based translation, etc.).

The TTC[1] project focuses on the automatic or semi-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 multilingual corpora in specialized domains. Terminologies will include single word terms (SWT) and multi-word terms (MWT), as well as their variations.

Terminologies may 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, such as Europarl (Koehn 2005). 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 'sets of texts in different languages that are not translations of each other' (Bowker and Pearson 2002: 93). Moreover, TTC aims at developing tools for gathering and managing comparable corpora and terminologies: a focused web crawler, a tool to handle comparable corpora using the existing open source UIMA (Unstructured Information Management Architecture[2]) framework, and a terminology management tool. At the end of the TTC project, a platform will be set up to combine those tools and to compile and manage comparable corpora using

[1] www.ttc-project.eu/

[2] www.research.ibm.com/UIMA/

standards (TMF, TBX) and UIMA framework. The open terminology platform will support tasks such as terminology storage, search, editing and export, as well as the reuse of Eurotermbank[3]. Translation of technical documents for the domains of aerospace and IT documentation will be performed by end-users of the consortium using CAT and MT techniques to assess impact of the TTC project outputs. This three-year project has been granted financial support from the European Commission under the European Community's Seventh Framework Programme (FP7/2007-2013) and started in January 2010.

2. Main Goals

The main goals of the project are as follows:

a. Using comparable corpora

The TTC will be using comparable corpora and take advantage from the huge amount of multilingual textual data available on the web ('Web as a Corpus' approach, Kilgarriff and Grefenstette 2003). As corpora compiled from the Web can be inconsistent and too variable, the project develops a topical web crawler to gather comparable corpora from domain-specific Web portals or using query-based crawling technologies with several types of conditional analysis. The objective is to guarantee the monolingual comparability (i.e. the comparability between pairs of texts in the same language), as well as the interlingual comparability of the compiled corpora by identifying the most relevant criteria and measures, such as structural, modal or lexical criteria. Besides, the elaborated typologies should be easily adaptable to other languages.

b. Using a minimum of linguistic knowledge for candidate term extraction

Term candidates and their relevant context partners (e.g. collocations) will be extracted from corpora using available monolingual term extractors. However, the TTC platform plans to handle also under-resourced European languages, i.e. languages with less available tools and resources. Therefore, the scientific objective is the assessment of the minimal amount of language-specific linguistic knowledge which is needed to identify term candidates and extract terminologies. Moreover, as word translation is not always possible, we plan to extract MWT translations as well as to pinpoint terminological gaps, i.e. terms that do not exist in the target language and therefore need to be paraphrased. The extraction of MWTs and complex terms are crucial when dealing with specialized domains.

c. Defining and combining different strategies for term alignment

TTC's objective concerning term alignment consists in improving methods for term alignment from comparable corpora, especially for specialized domains and MWTs. For this purpose, we intend to define, combine and evaluate lexical, contextual and corpora strategies. More specifically, we explore compositional methods (use of synonyms and variants, computation of interlingual representations), introduce depth in the context vectors used to express lexical contexts, and improve the adequacy of specialized comparable corpora with other textual resources such as parallel corpora or general language corpora (Morin et Daille 2009; Daille 2007).

d. Developing an open platform for use with MT and CAT tools

The TTC consortium develops complementary tools in order to manage comparable corpora and generated terminologies based on the UIMA framework and EuroTermBank. All developed and existing tools will be integrated into a demonstration platform offering all

[3] www.eurotermbank.com.

necessary Web Services. This platform will be easily integrated into various translation processing chains (including CAT translation, rule-based MT and statistical MT). The French TTC partner LINA from the Université de Nantes is building a French-speaking UIMA community in the domain of Natural Language Processing. Some services have already been set up and are currently available on the UIMA[4] portal. UIMA wrappers for term extraction and alignment tools as well as UIMA collection management tools will be developed during the project (Hernandez et al. 2010).

e. Demonstrating the operational benefits on MT tools and CAT tools

The main purpose of the TTC project is to propose solutions for translation in specialized domains and for languages with scarce linguistic resources. Besides, one of the objectives is that the proposed solutions can be rapidly used in an operational environment. As a consequence, the benefits of the developed solutions will be assessed regarding the gain in productivity to generate the terminologies, but also regarding the quality of translation, and the amount of invested linguistic knowledge – so as to reduce the gaps in language coverage by requiring as little knowledge as possible.

The impact on statistical MT will be evaluated as well. The work outlined in this project will leverage statistical MT performance in two ways. First, it will reduce the number of OOV words (out-of-vocabulary words, i.e. SWTs and MWTs which do not appear at all in the parallel data on which the statistical MT system is trained). Secondly, it will automatically determine additional translation candidates for SWT and MWT which occur either infrequently in the parallel corpora or which occur with a sense that is only correct in the domain of the parallel corpus and not in the domain of the translation to be performed.

3. Overall strategy of the work plan

The TTC project will develop a platform for automatically generating specialized monolingual and multilingual terminologies using comparable corpora and relying on existing open technologies (UIMA) and standards (TMF, TBX). The next figure illustrates the architecture of the expected multilingual terminology mining chain.

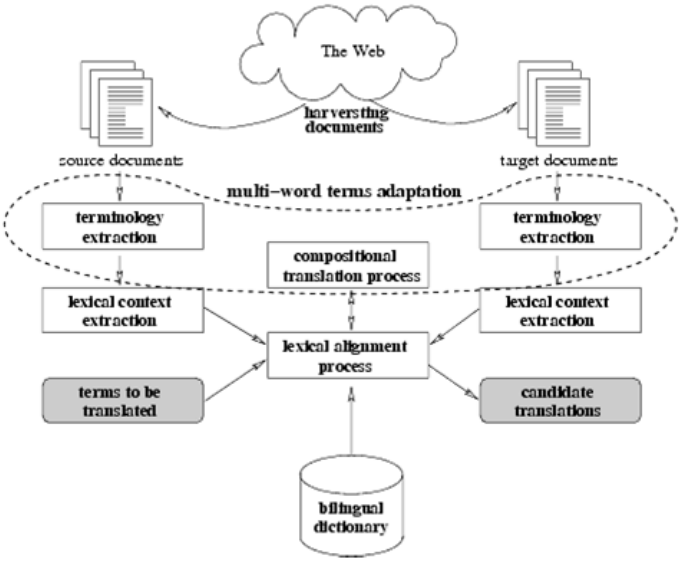

Figure 1: Architecture of the multilingual terminology mining chain

[4] www.uima-fr.org

Generated terminologies will be plugged into existing MT and CAT tools in order to improve their performance, especially for specialized domains and/or under-resourced languages.

The work breakdown structure in figure 2 shows the organization of the tasks that will be accomplished during the project.

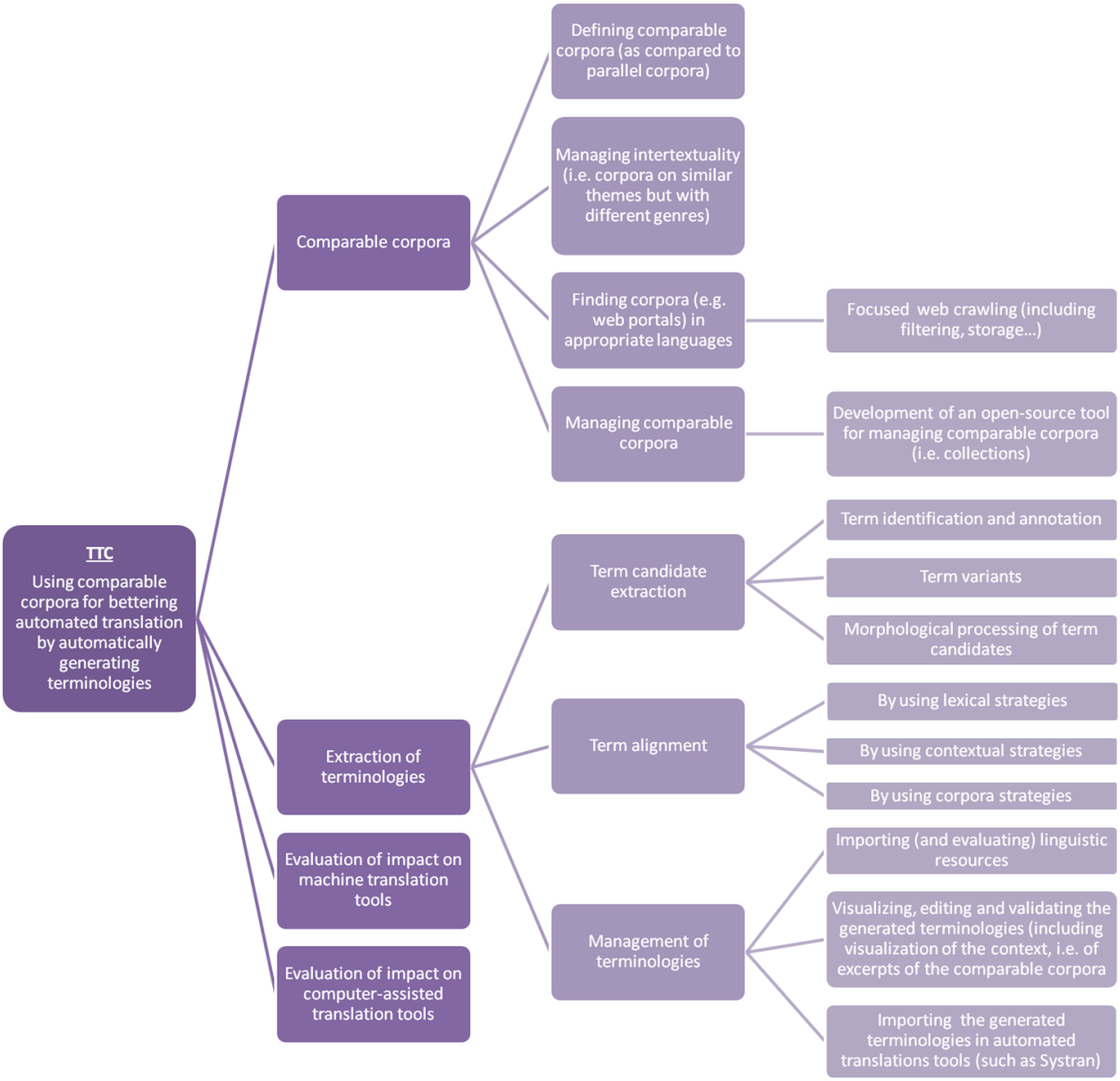

Figure 2: Work breakdown structure

Currently, requirements are being defined. This is being done in close collaboration with end users and by consulting external stake-holders. A survey about terminology and corpora practices in the translation and localization industry has been launched recently[5]. First results show today's demands from translation and localization professionals: consolidated accesses to terminology resources and consolidated terminology platforms, consistency check and support features, corpora and domain oriented solutions, as well as user-friendly interfaces (Gornostay 2010). Moreover, ongoing work includes the definition of functional

[5] The survey can be accessed here: http://www.visidati.lv/aptauja/345124026/

specifications and exchange formats according to international standardization initiatives such as ISO TC-37 and the EU project CLARIN[6].

4. Expected Results and Impacts

The expected results and impacts of TTC include:

- Domain-specific resources for renewable energy and computer science in seven languages (monolingual as well as bilingual aligned terminologies).
- Comparable corpora, lemmatized and POS-tagged.
- Generic methods and tools for automatic extraction of terminologies and alignment algorithms, as well as for recognizing term variants, for term inflection and morphological analysis.
- Focused crawler to compile comparable corpora.
- Open source suit tools to handle comparable corpora and to extract terminologies.
- Online terminology platform.
- Enhancement of existing on-line tools for MT (MOSES).

5. Consortium

The project consortium brings together seven partners from four different countries (France, Germany, Latvia and UK). Project leader is French laboratory LINA (Laboratoire d'Informatique de Nantes) from Université de Nantes. Further academic partners are the Centre for Translation Studies (CTS) from University of Leeds, as well as Institut für maschinelle Sprachverarbeitung (IMS) from Universität Stuttgart. IMS contributes to research on monolingual term extraction (MWT and SWT), on morphology and syntax. CTS works on corpus collection and term extraction activities, and leads the evaluation of the impact of TTC results on MT tools. Industrial partners are the French companies Sogitec and Syllabs, as well as the Latvian company Tilde. Syllabs works on the focused crawler, Tilde is responsible for the terminology platform. As an end user of translation and terminology tools in the aerospace domain, Sogitec will perform the evaluation of TTC impacts on computer-assisted translation tools. Finally, Eurinnov is responsible for administrative, financial and legal management as well as scientific and technical coordination.

[6] www.clarin.eu

References

Bowker, L.; Pearson, J. (2002). *Working with Specialized Language. A practical guide to using corpora*. Routledge: London/New York.

Daille, B. (2007). 'Variations and application-oriented terminology engineering'. In Ibekwe-SanJuan, F.; Condamines, A.; Cabré Castellvì, M.T.. (eds., 2007). *Application-Driven Terminology Engineering*. John Benjamins. 163-177.

Gornostay, T. (2010). 'Terminology Management In Real Use'. In *The 5th International Conference in Commemoration of Rajmund Piotrowski (1922-2009) 'Applied Linguistics in Science and Education'*. Saint-Petersburg.

Hernandez, N.; Poulard, F.; Vernier, M.; Rocheteau, J. (2010). 'Building a French-speaking community around UIMA, gathering research, education and industrial partners, mainly in Natural Language Processing and Speech Recognizing domains'. (to appear in *New Challenges for NLP Frameworks*, workshop at LREC 2010 : Malta).

Kilgarriff, A.; Grefenstette, G. (2003). 'Web as Corpus: Introduction to the Special Issue'. In *Computational Linguistics* 29 (3). 333-347.

Koehn. P. (2005). 'Europarl: A Parallel Corpus for Statistical Machine Translation'. In *MT Summit* 2005.

Morin, E.; Daille, B. (2009). 'Compositionality and lexical alignment of multi-word terms'. In *Language Resources and Evaluation* 44 (1-2). Springer Netherlands.