Terminology Extraction, Translation Tools and Comparable Corpora: TTC concept, midterm progress and achieved results

Tatiana Gornostay, Anita Gojun, Marion Weller, Ulrich Heid, Emmanuel Morin, Beatrice Daille, Helena Blancafort, Serge Sharoff, Claude Méchoulam

Terminology Extraction, Translation Tools and Comparable Corpora: TTC concept, midterm progress and achieved results

Tatiana Gornostay[a], Anita Gojun[b], Marion Weller[b], Ulrich Heid[b], Emmanuel Morin[c], Beatrice Daille[c], Helena Blancafort[d], Serge Sharoff[e], Claude Méchoulam[f]

Tilde[a], Institute for Natural Language Processing of University of Stuttgart[b]

Laboratoire Informatique de Nantes Atlantique of University of Nantes[c], Syllabs[d], University of Leeds[e], Sogitec[f]

E-mail: scientific-contact@ttc-project.eu

Abstract

The TTC project (Terminology Extraction, Translation Tools and Comparable Corpora) has contributed to leveraging computer-assisted translation tools, machine translation systems and multilingual content (corpora and terminology) management tools by generating bilingual terminologies automatically from comparable corpora in seven EU languages, as well as Russian and Chinese. This paper presents the main concept of TTC, discusses the issue of parallel corpora scarceness and potential of comparable corpora, and briefly describes the TTC terminology extraction workflow. The TTC terminology extraction workflow includes the collection of domain-specific comparable corpora from the web, extraction of monolingual terminology in the two domains of wind energy and mobile technology, and bilingual alignment of extracted terminology. We also present TTC usage scenarios, the way in which the project deals with under-resourced and disconnected languages, and report on the project midterm progress and results achieved during the two years of the project. And finally, we touch upon the problem of under-resourced languages (for example, Latvian) and disconnected languages (for example, Latvian and Russian) covered by the project.

Keywords: language resources, under-resourced languages, disconnected languages, terminology extraction, comparable corpora, computer-assisted translation, machine translation

1. TTC concept and main objectives

The TTC project (Terminology Extraction, Translation Tools and Comparable Corpora) [1] has contributed to leveraging:

- computer-assisted translation (CAT) tools,
- machine translation (MT) systems,
- and multilingual content (corpora and terminology) management tools

by generating bilingual terminologies automatically from comparable corpora in five EU languages belonging to three language families: Germanic (English and German), Romance (French and Spanish), and Baltic (Latvian) as well as outside the European Union: Slavonic (Russian) and Sino-Tibetan (Chinese).

TTC is a three-year project and its main concept is that parallel corpora are scarce resource and comparable corpora can be exploited in the terminology extraction task. The main TTC objectives are as follows:

- to compile and use comparable corpora, for example, harvested from the web;
- to assess approaches that use a minimum of linguistic knowledge for monolingual term candidate extraction from comparable corpora;
- to define and combine different strategies for monolingual term alignment;
- to develop an open web-based platform including solutions to manage comparable corpora and terminology which are also supposed to be available for use in CAT tools and MT systems;
- to demonstrate the operational benefits of the terminology extraction approaches from comparable corpora on CAT tools and MT systems.

2. Parallel vs. comparable corpora

In the end of the 20th century, in natural language processing there was observed a paradigm shift to corpus-based methods exploiting corpora resources (monolingual language corpora and parallel bilingual corpora) with the pioneer researches in bilingual lexicography (for example, Warwick and Russell, 1990) and machine translation (for example, Sadler, 1990).

A parallel corpus is a collection of texts which is translated into one or more languages in addition to the original (EAGLES, 1996). As a rule, parallel corpora are available for certain language pairs, usually including English. This occurs due to the fact that most of natural language processing tools are tailored for English or major European languages (Singh, 2008) in certain domains, for example, the legal domain. The two largest multilingual parallel corpora in the legal domain are:

- the Europarl corpus that covers the language of debates in the European Parliament (Koehn, 2005) and biased to the legal domain;
- the JRC-Aquis corpus that is a huge collection of the European Union legislative documents translated into more than twenty official European languages and includes such rare language combinations as, for example, Estonian-Greek and Maltese-Danish, however still biased to the legal domain (Steinberger et al., 2006).

In view of the quantity of multilingual information that grows exponentially and the need of its translation, parallel corpora can hardly be exploited for facilitating CAT and MT mostly due to their scarceness and limited language

[1] http://www.ttc-project.eu

and domain coverage. This is a well-known and acknowledged fact by the community and it poses a restrictive problem for various translation tasks, be it performed by a human, for example, human and CAT, or a machine and data-driven approaches to MT, for example, statistical machine translation (SMT). Thus one of the main tasks of contemporary natural language processing and corpus linguistics theory and practice is to reduce a linguistic gap between those language pairs that lack cross-language parallel resources and a potential solution to this task is to exploit comparable corpora.

A comparable corpus is a collection of similar texts in more than one language or variety (EAGLES, 1996) and it was introduced to the community in the late 90-ies (Rapp, 1995; Fung, 1995). Since that time, comparable corpora have been actively exploited in different research areas and MT in particular.[2]

The TTC project researches the way in which comparable corpora can be exploited in the terminology extraction task and leveraging translation (CAT and MT) and content (corpora and terminology) management tools.

3. TTC terminology extraction workflow

The TTC multilingual terminology extraction workflow consists of several processing steps (Figure 1).

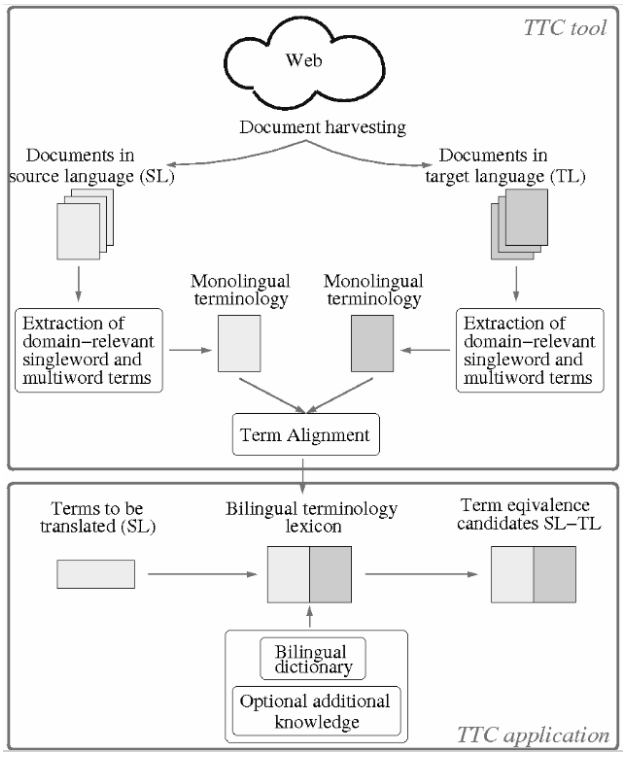

Figure 1. TTC terminology extraction workflow

3.1 Comparable corpora collection

For each TTC language, two domain-specific monolingual corpora have been collected in the wind energy and mobile technology domains.[3] To compile the corpora, we used the focused web crawler developed within the project (Groc, 2011) fed with parallel term seeds in all of the TTC languages. Automatically collected noisy corpora then were manually revised by linguists to get the specialized corpora in the two domains. The size and the quality of the revised corpora vary a lot from language to language. To reach the size of 300 000 running words per domain and per language, the revised corpora were extended with documents manually collected from the web.[4]

To be used in the terminology extraction task, the collected corpora undergo three pre-processing steps:

- tokenization: annotation of word boundaries,
- tagging: annotation of part-of-speech (POS) tags,
- and lemmatization: annotation of lemmas.

3.2 Monolingual terminology extraction

The terminology extraction process in TTC consists of three steps.[5] During the first step, term candidates – single word terms (SWT) and multi-word terms (MWT) – are extracted from the domain-specific corpora collected from the web. The extraction is based on a set of Part-of-Speech patterns (defined for all of the TTC languages) which describe different types of linguistic units, such as nouns (SWT) and adjective + noun, noun + noun, adjective + noun + noun (MWT), etc. During the second step, domain-relevant term candidates are identified. Within the project, we use a frequency-based notion of domain specificity as defined in Ahmad (1992). The final step includes the identification of term variants which may be both synonymous (for example, graphical: *Wind-Energie* ↔ *Windenergie* in German) and related (for example, syntactical: *vēja enerģija* ↔ *vēja un saules enerģija* in Latvian). [6] The output of the extraction component is a list of term candidates sorted descending by their domain specificity values.

3.3 Bilingual terminology alignment

During the next processing step within the TTC terminology extraction workflow, source language and target language monolingual terminologies extracted from comparable corpora are aligned to each other. The result of the alignment step is bilingual domain-specific terminology.

We have proposed to increase the coverage by automatically aligning neoclassical compounds that are extracted from bilingual comparable corpora. Neoclassical

[2] See, for example, the FP7 ACCURAT project research on collecting and using comparable corpora for statistical machine translation (Skadiņa et al., 2012).

[3] TTC comparable corpora are available for download on the website of the University of Nantes under the following link: http://www.lina.univ-nantes.fr/?Linguistic-Resources-from-the.html.

[4] For more information about the TTC domain-specific comparable corpora collected from the web and manually revised, see the project deliverable D2.5 under the following link: http://www.ttc-project.eu/images/stories/TTC_D2.5.pdf.

[5] The project deliverable "D3.4 Set of tools for monolingual term candidate extraction: single and multiword terms and context properties, for example, collocations".

[6] We rely on the set of term variants described in Daille (2005).

compounds are terms that contain at least one neoclassical element (prefix, suffix, and/or root), for example, a term *neuropathy* contains two neoclassical elements *neuro* and *pathy*. For that purpose, a language independent method has been proposed for extracting and aligning neoclassical compounds in two languages. According to this method, neoclassical compounds in the source language are translated compositionally into neoclassical compounds in the target language. For example, the French term *neuropathie* is translated into English by finding the equivalent of each component individually: *neuro* → *neuro* and *pathie* → *pathy* and combining these equivalent parts in order to obtain the English translation *neuropathy*. It should be noted, that this translation has to be found in the corpus in order to be considered as correct.

A tool has been developed in order to extract and align neoclassical compounds between two languages from comparable corpora.[7] Experiments were carried out on the following pairs of languages (in both directions): English ↔ French, English ↔ German, and English ↔ Spanish. The results have demonstrated a high precision for all of the translation directions participated in the evaluation. For example, 100 aligned terms were obtained for English↔French with a precision of 98% from the TTC comparable corpora in the wind energy domain.

4. TTC usage scenarios

The resulting bilingual domain-specific terminology can be used as an input to CAT tools and MT systems.[8]

4.1 CAT usage scenario

The extracted bilingual terminology can be integrated into CAT tools which are used by human translators. CAT tools provide the user with target language equivalences and the translator can choose an optimal translation for a source language term. Within the TTC project we evaluate two usage scenarios with CAT involving the English → French language pair in the aeronautic domain and the English → Latvian language pair in the mobile technology domain. The results will be reported by the end of the third year of the project (December 2012).

4.2 MT usage scenario

The output of the TTC term alignment tools can be fed into MT systems as an additional bilingual resource. We explore possibilities of integrating bilingual terminology and domain-specific target language texts (language model data) into statistical machine translation (SMT). First experiments showed that SMT systems using domain-specific texts and bilingual term lists produced by the TTC tools provide better translations than SMT systems without access to these additional knowledge sources (Weller, 2012).

5. TTC & under-resourced languages

One of the TTC languages is Latvian – an under-resourced language of the European Union with approximately 1.5 million native speakers worldwide. For Latvian, the main basic language resources and tools, for example, corpora, lexicons, morphological analysers, etc., are available for processing and evaluation purposes (Skadiņa et al., 2010). More advanced language resources and technologies (for example, discourse corpora, techniques for semantic processing, etc.) are being researched and prototypes are available for some of them.

The resoursefulness of the Latvian language is far from the goal since there is a noticeable gap in language resources and tools of the Latvian language which are a prerequisite of the sustainable development of the language. There are various grammatical characteristics of the Latvian language that make it much more difficult for automatic processing and the two of them (which are most conspicuous and identified as most problematic) are rich inflection and relatively free word order.

Nevertheless, a significant progress has been made in MT for the Latvian language. At the same time, its performance depends on the availability of language resources to a great extent, data-driven approaches in particular. Thus the most researched and developed language pairs in the aspect of SMT are English → Latvian and Latvian → English (Skadiņš et al., 2010). The Latvian-Russian MT is ensured by the rule-based system (Gornostay, 2010).

Nowadays, MT is not anymore considered as a competitor by translators and the task of MT domain adaptation has gained a wide interest. However, for under-resourced languages, the problem of the availability of parallel and even comparable texts still remains an issue. Thus, the Latvian comparable corpus collected within TTC has the smallest size out of the seven TTC languages (cf. 220 823 running words in the Latvian wind energy corpus and 313 954 – in the English wind energy corpus, 314 954 – in the French wind energy corpus, and 358 602 – in the German wind energy corpus). The task of obtaining more corpora for the domain adaptation of the English-Latvian SMT system is currently under consideration within the TTC MT usage scenario.

6. TTC & disconnected languages

Among the so-called "well-researched" language pairs as English-French / German / Latvian / Chinese, French-German / Spanish / Russian and German-Spanish, other TTC working language pairs are Latvian-Russian and Chinese-French which pose the problem of "disconnected languages". In this situation we deal with two major, or state, languages for which a relatively large amount of monolingual language resources are available but they lack cross-language resources due to their cultural / historical / geographical disconnection.

Despite of the long history of the Latvian and Russian language relationships and their relative similarity

[7] For more information about the Neo-classical MWT detection program for English/French/German, see the project deliverable D4.1 under the link:
http://www.ttc-project.eu/images/stories/TTC_D4.1.pdf.

[8] For more information about TTC usage scenarios see Blancafort et al. (2011).

(Gornostay, 2010), there is a considerable lack of Latvian-Russian parallel resources available for research, for example, SMT training and domain adaptation or terminology resource compilation. Within the TTC project, the Latvian-Russian language pair is currently under consideration and the evaluation results of the bilingual terminology extraction for these languages will be reported by the end of June, 2012.

7. Conclusion

TTC is at the beginning of its third year now and so far the project has made significant progress towards the main scientific and technological objectives for the first two years of the project (TTC Annual public report, 2010; 2011).

8. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248005.

We would like to thank Inguna Skadiņa for her time and attention to this abstract.

9. References (formatting example)

Ahmad, K. (1992). What is a term? The semi-automatic extraction of terms from text. In *M. Snell-Hornby, F. Poechhacker and K. Kaindl (eds) Translation studies: an interdiscipline*, pp. 267-278.

Daille, B. (2005). Variants and application-oriented terminology engineering. In *Terminology*, Vol. 11, pp. 181-197.

EAGLES (1996). Preliminary recommendations on corpus typology. Electronic resource: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceedings of the Association for Computational Linguistics*, pp. 236-243.

Gornostay, T. (2010). Latvian-Russian Machine Translation in the System of Social Communication, PhD thesis.

Groc, C. de (2011). Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *Proceedings of the IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Lyon, France, August 2011.

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket, Thailand, 2005.

Rapp, R. (1995). Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.

Sadler, V. and Vendelmans, R. (1990). Pilot implementation of a bilingual knowledge bank. In *Proceedings of Coling-90: Papers presented to the 13th International Conference on Computational Linguistics*, Helsinki, 20-25 August, 1990, Vol. 3, pp. 449-451.

Singh, A.K. (2008). Natural Language Processing for Less Privileged Languages: Where do we come from? Where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, Hyderabad, India, January 2008, pp. 7–12.

Skadiņa, I., Aker, A., Glaros, N., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiljevs, A. and Babych, B. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation LREC 2012. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Skadiņa, I., Auziņa I., Grūzītis N., Levāne-Petrova K., Nešpore G., Skadiņš R., Vasiljevs A. (2010). Language Resources and Technology for the Humanities in Latvia (2004–2010). In *Proceedings of the Fourth International Conference Baltic (HLT 2010)*, IOS Press, Frontiers in Artificial Intelligence and Applications, Vol. 219, pp. 15-22.

Skadins, R., Goba, K., Sics, V. (2011). Improving SMT with Morphology Knowledge for Baltic Language. In *Proceedings of the Research Workshop of the Israel Science Foundation – Machine Translation and Morphologically-rich Languages*, January 23-27, 2011, Haifa, Israel.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy, 24-26 May 2006, pp. 2142-2147.

TTC Annual public report. (2010). Electronic resource: http://www.ttc-project.eu/images/stories/TTC_Annual_public_report_2010.pdf.

TTC Annual public report. (2011). Electronic resource: http://www.ttc-project.eu/images/stories/TTC_Annual_public_report_2011.pdf.

Warwick, S. and Russell, G. (1990). Bilingual concordancing and bilingual lexicography. In Proceedings of the 4th International Congress EURALEX, Spain, 1990.

Weller, M. (2012). TTC: Terminology Extraction, Translation Tools, and Comparable Corpora. In Proceedings of the 16th Annual Conference of the European Association for Machine Translation, May 28-30, Trento, Italy, 2012. (submitted)

Blancafort, H., Heid, U., Gornostay, T., Mechoulam, C., Daille, B., Sharoff, S. (2011) User-centred Views on Terminology Extraction Tools: Usage Scenarios and Integration into MT and CAT tools. TRALOGY 2012: Translation Careers and Technologies: Convergence Points for the Future", March 3-4, Paris, France. Electronic resource: http://www.ttc-project.eu/images/stories/TTC_Tralogy_2011.pdf.