HAL
open science

TTC Web Platform: from Corpus Compilation to Bilingual Terminologies for MT and CAT Tools

Helena Blancafort, Francis Bouvier, Béatrice Daille, Ulrich Heid, Anita Ramm

To cite this version:

Helena Blancafort, Francis Bouvier, Béatrice Daille, Ulrich Heid, Anita Ramm. TTC Web Platform: from Corpus Compilation to Bilingual Terminologies for MT and CAT Tools. Tralogy II. Trouver le sens : où sont nos manques et nos besoins respectifs ?, Jan 2013, Paris, France. 14 p. hal-00820331

HAL Id: hal-00820331

https://hal.science/hal-00820331v1

Submitted on 3 May 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TTC Web Platform[1]: from Corpus Compilation to Bilingual Terminologies for MT and CAT Tools

Helena Blancafort[(1)], Francis Bouvier[(1)], Béatrice Daille[(2)], Ulrich Heid[(3)], Anita Ramm[(3)],

[(1)] Syllabs (Paris), [(2)]LINA - Université de Nantes, [(3)] IMS - Universität Stuttgart

blancafort@syllabs.com; bouvier@syllabs.com; beatrice.daille@univ-nantes.fr; heid@ims.uni-stuttgart.de; ramm@ims.uni-stuttgart.de

Abstract

This paper describes the TTC Web platform, an online demonstrator to show the whole pipeline to compile bilingual terminologies out of comparable corpora gathered from the web using the tools developed in the TTC project *Terminology Extraction, Translation Tools and Comparable Corpora*. We present the whole chain which has been integrated into the platform, as well as their main components: a focused web crawler; a UIMA based tool for both monolingual term extraction and bilingual term alignment, tools for monolingual term extraction using both rule-based and probabilistic methods, and finally, an online terminology platform to edit the output of the TTC tools. The TTC tool chain is available for all the languages of the project: DE, EN, ES, FR, LV, RU and ZH. With respect to the potential users of the tools, in the first Tralogy conference we presented the different users and scenarios that were envisaged: from basic users to professionals of the MT industry. In this paper we will include the first feedback obtained from users during the second user workshop that was organized to demonstrate and test the tools with potential users and experts of the MT, CAT, and terminology management domain.

1 Introduction

Machine translation and computer-assisted translation still suffer from the terminology bottleneck. There is a lack of bilingual term-related resources, especially for new or upcoming domains. Today, bilingual terminologies can be generated automatically with tools such as the GIZA++ statistical machine translation toolkit (Och and Ney, 2003). These tools can generate bilingual terminologies taking as input a parallel corpus. However, these technologies suffer from the scarcity of domain-specific parallel corpora. With the availability of considerably amounts of monolingual text data in many languages, using web corpora for terminology work together with tools for automatic term extraction seems an interesting strategy to generate bilingual terminologies needed for machine translation (MT) and computer-assisted translation (CAT) tools. Here the idea is to search for monolingual data in two languages for the same topic, and then to apply tools that are able to align monolingual term extraction results and thus to provide bilingual terminologies.

The TTC project *Terminology Extraction, Translation Tools and Comparable Corpora*[2] focuses on the development of tools that generate bilingual terminologies automatically using as input monolingual text data collected automatically from the web. To illustrate the

[1] http://www.ttc.syllabs.com/

[2] www.ttc-project.eu/ The project has been granted funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n° 248347.

output of the tools developed during this 3-year project and to demonstrate the whole pipeline, TTC has set up an online demonstrator, the TTC Web platform. Using this web-based service, the user can compile bilingual terminologies out of comparable corpora with the tools developed in the project directly on the web site, without having to download or to install the tools. It is possible to run the tools both on proprietary data or on data crawled from the web via the platform. The resulting terminologies can be exported towards MyETB, an open terminology platform developed within TTC, and be edited.

The TTC platform is a complete online terminology that supports interactive translation work. The idea was to support the following scenarios:

- translation of texts in a new domain, requiring the provision of larger amounts of terminology (Bowker and Pearson 2002);
- translation of texts in a known domain, but with new terminology;
- correction of translations, possibly with terminological checking.

The service is available for all the languages of the project: DE, EN, ES, FR, LV, RU and ZH.

With a view to potential users, in the first Tralogy conference (Blancafort et al., 2011) we presented the different user types and scenarios that were envisaged: from basic users to professionals of the MT industry. In this paper, we present the TTC web platform and discuss the feedback obtained from users during a second user workshop that was organized to demonstrate and test the tools with experts of the MT, CAT, and terminology management domain.

The paper is structured as follows: after presenting the platform, we describe each module in detail and discuss the corresponding feedback obtained from translation and language technology professionals. The following modules are described: corpus compilation (section 2), monolingual terminology extraction (section 3), bilingual terminology extraction (section 4), as well as online terminology editing with MyETB (section 4). Finally, we draw some conclusions.

2 Overview of the TTC Platform

2.1 The Components of the Platform

The TTC platform is a web-service to run terminology tools on monolingual, as well as comparable corpora from two languages. The platform has a modular architecture with

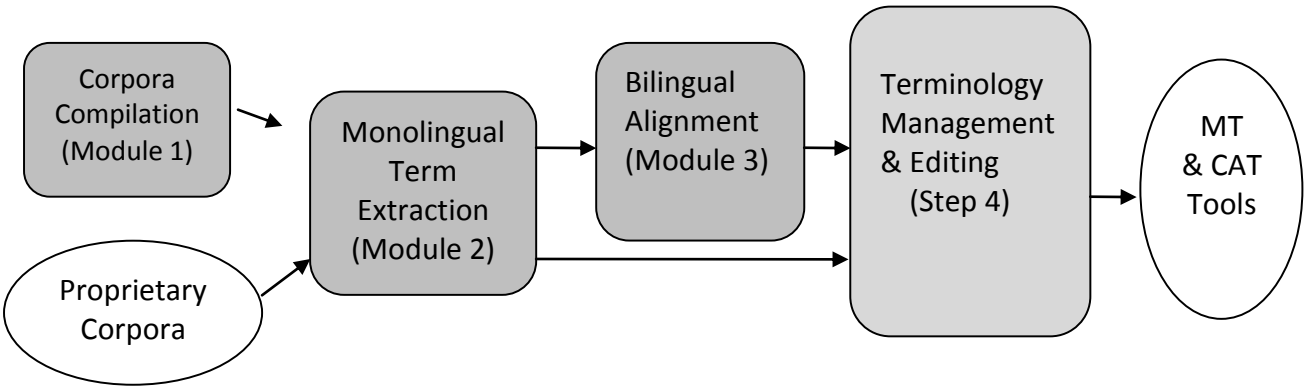

Figure 1: Modules of the TTC Web Platform and workflow

three main components. Each component represents a step in the workflow to generate bilingual terminologies from comparable corpora. The output of this workflow can be imported into a machine translation or computer-assisted translation tool. The figure above illustrates the TTC pipeline:

In the following, we briefly describe each module.

1) **Corpus compilation** with **Babouk** (Groc 2011). Babouk is a focused crawler used to collect documents from the web starting from user-defined seed terms (domain-specific terms).
2) **Monolingual terminology extraction** with **TTC TermSuite** (Rocheteau and Daille 2011, Daille 2012) and **SylTools,** Syllabs tools for monolingual terminology extraction. TTC TermSuite extracts single (SWT) and multi-word terms (MWT), and includes the identification of term variants. This identification of term variants is an innovative feature with respect to existing tools in the domain. SylTools includes knowledge-rich and knowledge-poor tools for both SWT and MWT extraction. Here the innovative feature is the use of knowledge-poor methods, that can relatively easily be adapted to less resourced languages. TTC TermSuite is UIMA-based and available for free download as well.
3) **Bilingual alignment** with **TTC TermSuite Aligner**. TTC TermSuite Aligner is the module of TTC TermSuite to generate bilingual terminologies suggesting several translation candidates for a single source language term. The innovation of this tool is the possibility of running bilingual alignment on comparable corpus and not only on parallel corpora, which is the case in existing tools. To our knowledge, commercial tools only handle parallel corpora. The required input is the output of the previous step 2.
4) **Terminology Management and Editing.** The terminological output obtained with the TTC tools can be exported via the platform directly to MyETB, the open terminology platform to edit and handle TTC terminologies. MyETB is integrated into EuroTermBank[3]. This is the last step of the TTC pipeline. Finally, the user can export the output from MyETB and use it as input for automatic translation, both rule-based and statistic-called, or for CAT tool systems.

2.2 Specific properties

In the TTC platform, each module can be used separately; this means that the tools for monolingual term extraction can be used directly on proprietary data without using the crawler. Each module is independent and the results obtained with each one can be downloaded or used as input for the subsequent step or for another application. Input and output formats follow standard conventions in the terminology domain, namely the TBX format, as well as the UIMA format XMI and the less verbose tab-separated format CSV/TSV. The aim of the platform is to reproduce the whole terminology preparation pipeline without users having to install any tool.

The modular architecture makes it possible to use the TTC platform for different needs, because there are different group of users and different usage scenarios. We identified three main categories of scenarios which needed to be addressed, and several factors to be taken into account: (1) the types of language activities to be carried out with the help of the TTC tools, (2) the situation of users with respect to the availability of language resources as an input for the TTC tools, and (3) the profile of users depending on their

[3] www.eurotermbank.com

level of expertise regarding the use of translation tools. These factors were considered when designing the tools. Due to different profiles of users, it seemed important to make the tools available to the users in two different procedures: web-based (as described above) and via the “download-and-run” procedure. Moreover, to assure that different kinds of users can use the platform smoothly, each module offers several optional parameters. This means that the tools can be used with the default setting by users that are less familiar with crawling and terminology extraction tools, whereas advanced users can configure the tools according on their needs.

2.3 User Feedback

As mentioned before, in the first year of the project we organised a first user workshop to validate the goals of TTC and assure that the tools developed respond to real user needs; we also run a survey to learn more about users’ needs (Gornostay 2010). This was important to get user feedback about the targeted tools, their functionalities and the quality and usability of the automatically computed monolingual and bilingual term lists. Following this concern about the users’ real needs, at the end of the project, we thus organized a second user workshop to which researchers, terminologists, CAT and MT tool developers were invited. Guests from different fields of activity provided their opinions about the tools from different points of view.

The workshop contained three main parts: (i) theoretical background, (ii) demonstrations and (iii) discussion. Firstly, the theoretical background of the methods implemented in the project was sketched. After that, we demonstrated the developed tools. The workshop guests were invited to follow the instructions of the presenters on how to use the TTC tools. The workshop participants also got samples of extracted term lists which they then evaluated. The feedback about the tools has been collected in the form of a questionnaire which included a number of questions about each of the presented tools. The users were asked to rate the functions of each of the tools which made the compilation of the answers quite simple.

Based on the analysis of the questionnaire, the tools and their functions were discussed with the guests. The users were encouraged to express their opinions about the tools, and also to make proposals on improving and adapting the tools to their needs. In the following sections, we give summaries of the discussions on each of the TTC tools.

3 Step 1: Corpus Compilation

The first step of the TTC pipeline is to compile the corpus that will serve as an input to the tools for automatic term extraction. Here the user can gather a monolingual corpus for a specific domain using Babouk, (Groc 2011), a focused crawler. A focused crawler collects data from the web and only keeps domain-specific documents. The user defines the domain by giving a number of domain-related terms (seed terms). As the volume of documents available on the web is enormous, the focused crawler has to minimize the number of web pages explored and follow links that are most relevant to the specialized domain defined by the user.

3.1 Tool Functions and Technology

To start, Babouk takes as input a list of seed terms or seed URLs specified by the user, and then randomly combines the seeds to create queries and submit those queries to a search engine. The top *N* results *for each query* are then used as seed URLs. The first step to expand the seeds is based on Bootcat (Baroni et al., 2004).

In order to provide only domain-relevant pages, Babouk uses a weighted-lexicon-based thematic filter, which computes the relevance of the web pages. This lexicon is built during

the first iteration of the crawling process: first, the user gives some seed terms, then new terms are extracted from documents to extend the lexicon with new terms, and finally, the new lexicon is weighted automatically using a web-based method. Other filters applied to discard documents are: document minimum and/or maximum size as well as the readability of the document. Several filters can be parameterized by the user, but as mentioned above, the tool can be used with default settings to ensure its accessibility to basic users.

A frequent question from the user side is the number of seeds needed, as well as the degree of domain-specificity. It is not easy to answer this question, because it is hard to evaluate the specificity and adequacy of a crawled corpus; it depends on the needs and expectations of the user, as well as on the way he/she wants to use it. Our experience shows that it is better to use seed terms rather than seed URLs. Experiments during the project to evaluate the recall of the terminology extracted out of a crawled corpus with respect to reference term lists showed that a short list of low or medium specialized terms is enough, e.g., a list of 5 to 10 terms. For all languages with the exception of German, 5 to 10 seeds give a large corpus. A large list of seed terms does not improve the results. Using more seeds can lead to a topic shift, especially when they are quite ambiguous. When using very highly specific terms, there is a risk of obtaining a small corpus. Here again, it depends on the expectations of the user, whether the priority is a large corpus or a small but very specific one. If the aim is to use the corpus for the automatic generation of bilingual terminologies, a big corpus is needed. Following the feedback of the second user workshop, a new feature was added, namely including positive and negative seeds. This means that the user can determine mandatory seeds that should appear in all crawled documents, as well as “negative seeds”, i.e., words that should not appear in the crawled data.

The output of the crawler consists of a folder with all documents converted into txt, with utf8 encoding, the source file, an xml metadata file based on Dublin Core, and an html file, so that the user can navigate through the files and have a quick look at the crawled data.

The most remarkable output of the project using Babouk is the delivery of a comparable corpus in the domains of wind energy and mobile communication for all languages of the project made available to a wider public[4]. The corpus size varies from 300.000 to 400.000 words, depending on the domain and language. Moreover, Babouk is currently used to improve the translation quality of a statistical machine translation tool from German into French. In this case, the corpora are used to improve the language model of the target language.

3.2 Limits of the Crawler and User Feedback

The workshop participants saw Babouk as a very useful tool for collecting domain-specific texts from the Web. Babouk offers a number of possibilities to customize the search for documents on the Internet, which were all found to be useful.

An important aspect of the tool's usability that was discussed during the workshop and also in previous consortium meetings is the quality and inspection of the crawling results. The user has to download the crawled data in order to estimate its adequacy. The workshop participants said that it would be good to inspect the crawled data before actually downloading it. At the moment, this is not possible. Instead, the user can be provided with some statistics about the collected texts, e.g. text type, text length, number of types/tokens, etc.

---

[4] The corpus is free for download under the following link : http://www.lina.univ-nantes.fr/?Linguistic-Resources-from-the.html

The quality of the corpora is a frequent point of discussion when working with web-based data. For terminological work, highly specialized documents are required. However, the documents that are identified by a search engine are very often documents from web pages that are ranked at the top by the search engine, but that are not necessarily very specialized. This means that for technical domains, we do not necessary get technical documents written by specialists for a specialized audience, as it is often the case in terminological work, but rather documents written for the general public with a less specialized language. In English, it is easier to get specialized documents, because scientific papers are very often written in this language. For other languages, it is more difficult to find scientific papers for specific domains.

On the other hand, the quality of the corpus depends on the specific need of the users. As we have mentioned before, the TTC platform has users with different profiles and needs. Our impression is that the tool is very useful when the translator has no material about the domain for which he/she is looking for information. This concerns scenarios where the translator tackles a new domain, where the translation job is defined without providing any material to the translator, or when the domain is a new domain for which reference terminologies do not exist.

To improve the quality of the corpus, the workshop participants suggested taking advantage of the metadata information that the crawler provides, but that is so far not used to evaluate the domain-specificity of the corpus. They suggested adding parameters so that the user can define constraints on the available metadata (e.g. author, publication date and the like).

Concerning the required input and the selection of seed terms, the participants claimed that for some domains it might be easier to come up with suitable seed words than for others. If the initial set of seed words does not lead to acceptable crawling results, one might have to repeat the crawling using other seed words. However, in a real-use scenario, there is often a lack of time for running the crawler several times in order to optimize the crawling results.

One of the hurdles we encountered with the crawling task is the dependency on a web search engine. The success of the crawling depends highly on the search engine used, its performance and also its availability. Babouk is currently using the Microsoft Bing web service, which has recently restricted the access to the free model. This means that the number of crawls that we can launch per month is less than around 20 crawls. These limits are beyond our control for the moment. For commercialisation of the tool, we would thus have to deal with paid-for services of the Microsoft Bing API to assure that the user will be able to crawl with fewer restrictions.

4 Step 2: Monolingual Terminology Extraction

Monolingual term extraction is the process by which term candidates are extracted from a monolingual corpus. All the tools for monolingual terminology extraction implemented in the TTC web platform take as input a domain specific corpus in UTF-8 text format and output list of terms candidates ranked by a specific score (e.g. absolute/relative corpus frequency, domain-specificity, etc.). Here it is possible to upload a corpus or to launch the job on the output of Babouk, that is, on the crawled corpus.

Monolingual terminology extraction can be performed with two different tools: TTC TermSuite and SylTools. TTC TermSuite (Rocheteau and Daille 2011), is a UIMA-based open-source tool that includes several modules for both monolingual and bilingual terminology extraction. SylTools propose a probabilistic tool.

All TTC tools extract both SWT and MWT. The main difference between TTC TermSuite and SylTools is the fact that TTC Term Suite handles term variation, which means that term candidates can be output with their corresponding term variants.

Morphological, syntactical and graphical variants (Daille, 2005) are handled. Concerning the word categories, both tools provide nouns and adjectives. With TTC TermSuite, the user can configure the tool to also extract other word categories, such as adverbs and verbs. TTC Term Suite is rule-based, which means that the terminology patterns are hand-crafted. With SylTools, we can run both rule-based terminology extraction and probabilistic terminology extraction (for Spanish, English, Latvian, German and French). Here, the aim is to compare results depending on the method used, i.e. to compare knowledge-rich and knowledge-poor approaches for terminology extraction. All tools can export the extracted data in the TBX format. Additionally, TTC TermSuite also exports the terminology in UIMA-compliant XMI format.

In the next two sub-sections, we will describe SylTools, as well as TTC TermSuite for monolingual extraction.

4.1 Description of SylTools

SylTools provide terminology extraction with two different tools: a standard symbolic term extractor (hand-written NP rules) and a probabilistic tool. To rank the term candidates, the method proposed by (Ahmad et al., 1992) is used. To compute the relative frequency, we work on lemmas with the exception of Latvian, as the tagger we used does not lemmatize.

4.1.1 Description of SylTools Probabilistic Term Extraction Tool

The probabilistic tool is a knowledge-poor tool that has been developed to assess whether a probabilistic method can obtain useful results in comparison to a knowledge-rich method. This is especially interesting for languages for which a POS tagger is not available, which can be the case for an under-resourced language or in the case of a commercial development of a tool. In an industrial framework, when developing a new language, it is not always possible to use an existing tagger, e.g., due to the restrictions of uses for commercial purposes of known open-source tools, or due to the high prices of the tools. Therefore, it is interesting to develop a tool with less knowledge. While a knowledge-rich tool like the one presented above uses a morpho-syntactic lexicon, a POS tagger and hand-written rules to identify term candidates, the probabilistic tool just needs a big raw corpus, as well as a small corpus with manually annotated sentences (noun phrases). This small corpus can be annotated by a linguist in one day only.

To train a POS tagger, we used a method based on (Clark, 2003)[5], as well as on Conditional Random Fields (CRFs) (Lafferty et al., 2001). As an input to train our pseudo POS tagger, we used large monolingual tokenized corpora (ranging from 500 million words in English to 50 million words in Spanish). To train the probabilistic CRF-based noun-phrase extractor, we used manually annotated corpora in each language with 300 to 600 sentences.

4.2 TTC Term Suite

TTC TermSuite is a tool to automatically generate bilingual terminologies using comparable corpora in all the languages of the project. TermSuite is open-source, based on the UIMA framework and available for download on the Google code page of the project[6]. To simplify the use of TTC TermSuite on the TTC web platform, the two components called Spotter and Indexer have been merged into one single module for

[5] Available on Alexander Clark's Web page: http://www.cs.rhul.ac.uk/home/alexc/pos2.tar.gz

[6] http://code.google.com/p/ttc-project/

monolingual term extraction. The user gives as input a corpus and obtains as output a list of monolingual term candidates in several formats. The main features of the monolingual term extractor of TTC TermSuite are:

1) Recognition and indexing of both single-word and multi-word terms;
2) Computing their relative frequency and their domain specificity to rank the term candidates;
3) Detection of neoclassical compounds;
4) Grouping of term variants.

4.3 Output of Monolingual Terminology Extraction

The generated terminologies can be downloaded in several formats, and also be visualized with a results viewer via the interface. The figure below illustrates the output of the viewer. The results obtained with the tools for monolingual terminology extraction vary depending on the language. The probabilistic approach has achieved good results and is a valid and a useful method to develop a monolingual extractor quite rapidly from scratch, which is very encouraging for scenarios for which we do not have any resources for the language, such as POS-taggers and lexicons. To illustrate this, we computed the recall of the terms in a reference term list of the domain of wind energy published by the Danish Wind Association[7] on a corpus compiled with Babouk: with the SylTools rule-based tool in Spanish we obtained a recall of 72.39%, while using the probabilistic tool we got a score of 69.40%.

Id	Term	Term pilot	Part of speech	Pattern	Complexity	Spec.	Occurr.	Freq.	Forms list	
270596	turbine	turbine	Noun	Noun	Single-word	285	2807	0.0121	Form	Occurr.
									turbine	1791
									turbines	1016
917089	wind turbine	wind turbine	Noun	Noun-noun	Multi-word	188	1850	0.0080	Form	Occurr.
									wind turbine	1148
									wind turbines	702
121371	blade	blade	Noun	Noun	Single-word	145	1425	0.0062	Form	Occurr.
									blade	1019
									blades	406

Figure 2: Output of TTC Term Suite on the TTC Web Platform

4.4 User Feedback

To illustrate and to evaluate the TTC tools for monolingual terminology extraction, the participants were asked to check samples of data extracted with different tools developed within the project. The general feeling was that the output was good and that it could be successfully used in a real working situation. The TTC tools provide not only the plain term lists, but also additional information about the terms. Within TermSuite, this additional

[7] http://wiki.windpower.org/index.php/Glossary

information can either be hidden or displayed, according to the user's interest. For a quick checking of term candidate lists, however, the users prefer to have as little information as necessary. Furthermore, they expressed a clear need for example sentences, which would significantly alleviate the decision whether a term candidate is domain-specific or not. The current version of the TTC tools unfortunately do not provide this kind of information, or, in general, an access to the corpus from which the term candidates are extracted.

Contrary to our expectations, the users rated the monolingual term variants as rather not important to have a look through. Since our tools work with data from the web, most variants are arbitrary forms of the corresponding base term. As such, they are not interesting for the user, as they do not occur in the texts used and produced in a company for which the terminology is being collected. However, the identification of variants found indeed in the texts from a company would make it possible to proscribe term variants which should actually not be used. Such forms (variants) can then be marked as "dispreferred". Variants are crucial to improve the results of the ranking of the candidate terms: the automatic evaluation using reference lists shows that, for example, in Russian, the identification of variants improves the precision by 100% on the 100 first candidates.

The term extraction process relies on lemmas of the words occurring in the corpus. Due to this fact, the resulting term candidate lists contain lemmatized entries, e.g. "erneuerbar Energie" (renewable energy) instead of the inflected form "erneuerbare Energie" where the agreement between adjective and noun is considered. For morphologically rich languages such as Russian and Latvian, lemmatized MWTs could be hard to understand. Therefore, the users would prefer to have inflected forms rather than lemmas. So, the occurrences of the candidate terms and a pilot form that is the most frequent occurrence are included in the TBX output. The TSV output lists contain the pilot forms and not the lemma forms.

The TTC tools do not include a component for special handling of abbreviations. The users pointed out that there is a need for such a component. Further discussion also confirmed that it is very hard to decide whether a specific term belongs to the domain of interest or not. Our experimental domain of wind energy is highly related to the domain of renewable energy: a fact which is also reflected in the texts we collected and used for the extraction experiments. Many term candidates could be assigned to the domain of renewable energy, however, it is hard, even for a human, to decide from which of the two domains a specific term candidate comes from.

5 Step 3: Bilingual Terminology Alignment

This module is part of TTC TermSuite and is used to obtain bilingual terminologies by aligning monolingual terminologies. It takes as input the monolingual term lists generated with the previous module (i.e. TTC TermSuite), and proposes for each source term one or more translation candidates. Translation candidates are ranked by a score, the user can configure the maximum number of translation candidates per term. The platform includes lexicons for several language pairs, but the user can still upload his/her own lexicon.

TTC TermSuite Aligner handles SWT and MWT, as well as neoclassical compounds. Two different strategies with regard to the nature of the terms are adopted: the distributional method (context-based projection approach) for SWT and compositional method for MWT and neoclassical alignment The context-based projection (distributional method) approach relies on the hypothesis that a word and its translation tend to occur in similar contexts within two parts of a comparable corpus and is based on a similarity of these contexts. The compositional method assumes that the correct translation of a MWT in a source language can be obtained by translating its components individually using a

general bilingual dictionary. The translations are then combined with each other and finally compared with the lists of the target language term candidates in order to obtain only those translations candidates which occur in the target language corpus. The compositional method has been evaluated for both neoclassical compounds and MWT. The translations obtained are highly reliable: for neoclassical compounds, precision is between 97% and 100% for the wind energy domain and for 3 language pairs (FR-ES, FR-EN and FR-DE) when evaluating only the first translation candidate (Harastani et al. 2012). For MWT, precision is between 79% and 96% for FR-EN and FR-DE in a medical domain when evaluating the top 5 translation candidates (Morin et Daille, 2012).

5.1 Input and Output of the Aligner

To perform bilingual alignment, it is necessary to first run the monolingual term extraction using TTC TermSuite on two monolingual corpora (source and target language). The aligner requires a bilingual lexicon, as well as the list of terms to be translated. If no list is provided, the tool aligns all the terms found in the document. TTC TermSuite Aligner outputs the generated bilingual terminology in three different formats: TSV, XMI and TBX. The TBX file can be sent to MyETB, the open terminology platform, where the bilingual terminology output can be edited. The user can view the results of the alignment on the platform via the viewer. The figure below shows for example nine translation candidates suggested for the MWT “direction of rotation”. For each translation candidate, the alignment score (called "similarity score"), as well as the target ID is given.

direction of rotation 1058019

Term candidate	Similarity	Target Id
sens de rotation	0.1395	1632878
axe de rotation	0.1395	2869380
axe suivant le rotation	0.1047	1997700
axe de rotation perpendiculaire	0.1047	3703232
pale sens de rotation	0.1047	5433183
sens de rotation horaire	0.1047	6034278
axe de rotation vertical	0.1047	6097648
rotation sens du aiguille	0.1047	6832315
rotation sens	0.0930	6249812

Figure 3: Suggested translation candidates for one MWT

5.2 User feedback

The users showed great interest in automatic term alignment. While for MWTs and neoclassical terms, the alignment component outputs only a few alignment candidates, the list of alignment candidates for SWTs may be much longer which means that the users have to check much more data to find the correct alignment candidate for each source language term. Due to time constraints that translators, interpreters and terminologists have, the manual checking of the alignment could be thus somewhat problematic. Half of the users said that they would only check the top 4-5 alignment candidates for each source language term which is rather insufficient for the SWT alignment. At the moment, the TL variants are considered to be potential alignment candidates. The users, however, would only like to get the alignment to the target language base term, which is then linked to its variants.

The users manually checked a sample of the alignment for DE and EN in the domain of wind energy. Similarly to the evaluation of monolingual term lists, they had different

opinions about the correct term equivalences. Terms can have several correct translations: which one is the correct one in the given domain can be in many cases determined by having a look at the target language sentences including these translation candidates. The users pointed out that such kind of corpus access is required in order to identify correct term pairs. As already noted previously, for now the TTC Web platform does not support access to the corpus. The access to the corpus is supported by TermSuite through its graphical interface.

The alignment component allows users to set numerous parameters for the computation of the alignment. The users showed little interest in changing the alignment settings: they usually use default values stored within the tool.

6 Step 4: Terminology Management and Editing

MyETB is a tool developed in TTC for terminology storage and processing (import, search, editing, and export). MyETB is based on the EuroTermBank (ETB) platform. It integrates with all the features currently provided by ETB.

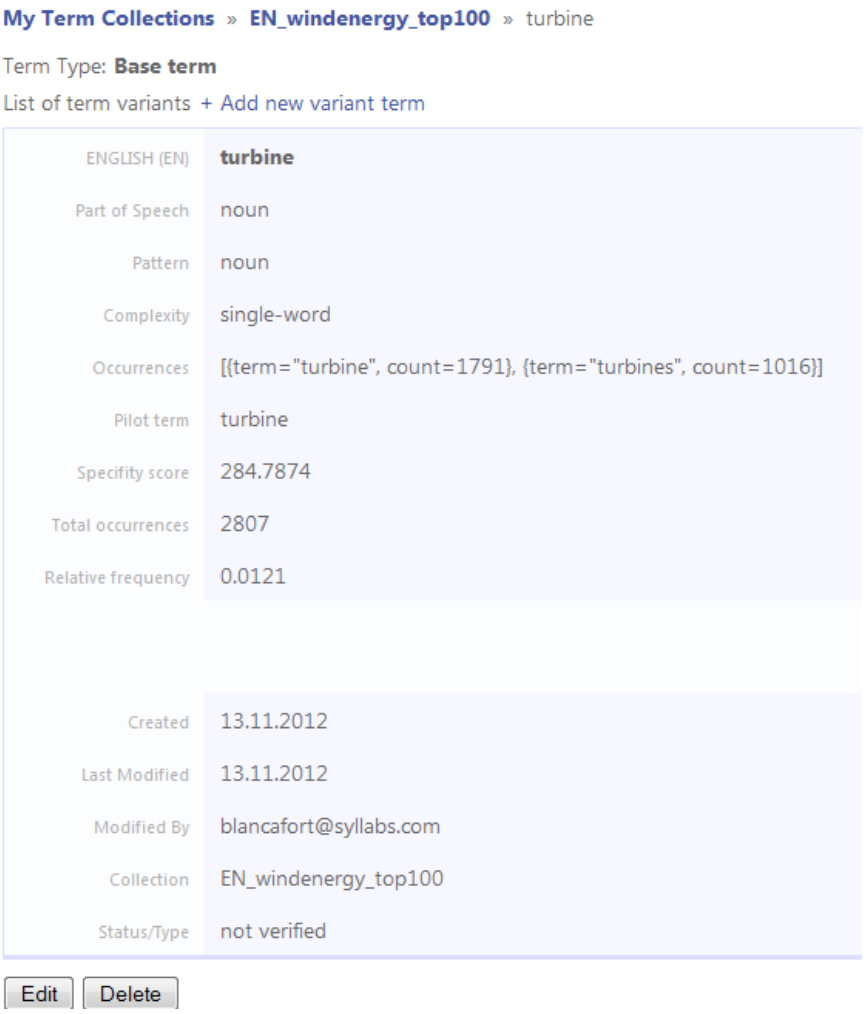

Figure 4: Sample output of TTC Term Suite exported to MyETB

A range of new TTC-specific features have been implemented. ETB is a rich source of consolidated public terminology that provides access to content that is stored in its central database as well as a number of interlinked terminological databases in a federation model (Auksoriute et al., 2006). MyETB has been integrated into the TTC demonstration platform via an API, which means that the user can directly export the terminologies obtained with the TTC platform towards MyETB. With MyETB, the user can edit the

terminologies obtained with TTC. For example, it is possible to export the manually edited and revised terminologies, so that they can be used in translation tools, both MT and CAT tools. The figure above shows the entry for the SWT “turbine” after automatic export into MyETB. The user can edit this entry, that is, modify each attribute, if necessary, delete the entry or just validate the term to give it the “verified term” status.

6.1 User feedback

In general, it was appreciated to have a tool to edit the terminology collections that are automatically produced by the TTC tools, and the possibility to send these terminology collections directly to MyETB via the platform. The workshop participants made some suggestions to add functions such as to accelerate the task of correcting the TTC terminologies. Correcting the TTC output means manual checking of the terms and deciding whether a term is domain-specific or not. This checking must be performed quickly, so the users would appreciate to have simple lists of terms with check boxes for "good", "not good" and "don't know". Terms marked as "not good" would then be automatically removed from the list. The category "don't know" would allow the user to decide later on the status of the terms for which the decision cannot be made right away. For this kind of quick checking of the automatically extracted term lists, the users prefer to have information displayed as necessary. For other tasks however, the access to the example sentences mentioned previously would be very welcome. This shows that the functions that a tool should provide depend highly on the users' needs and profile, as well as on the amount of time they can spend on terminological work.

MyETB should provide a function for exporting the checked terminology data in a format needed by TermSuite, especially for its alignment component. The data should be in general downloadable in more formats. In particular, MyETB should allow the exportation of simple term lists (glosses) which are often needed by interpreters.

7 Conclusion

This paper describes the TTC web platform developed in the TTC project to compile monolingual and bilingual terminologies from comparable corpora. This platform is a demonstrator of the TTC pipeline. Via the platform the user can test the whole pipeline: from corpus compilation to bilingual alignment, and edit the result on the open terminology platform MyETB. Finally, the generated terminologies can be used to enhance machine translation and CAT tools.

The presentation of the TTC web platform to a community of potential users in the second user workshop organized within the project validated the TTC pipeline and the interest of having such a web platform. Overall, the workshop participants were positively impressed by the output of the project and the tools. A modular architecture and a user-friendly online interface enable different type of professionals to use the tool for different purposes, e.g. interpreters looking for a quick overview of the terminology of a specific domain, or translators using machine translation tools to speed up the translation process. The platform seems especially useful for translation professionals dealing with domains for which corpora and terminologies are not available, which is the case for upcoming domains or for language and language pairs with fewer resources. For these scenarios, using the platform to gather domain-specific data from the web and build terminologies automatically is a good alternative. Because the quality of the web data is unpredictable, depending on the scenario it was also appreciated to be able to use the tools with proprietary data. The actual objective of the project is to evaluate the impact of using the

TTC generated terminologies in machine translation and CAT tools. Experiments are currently carried out.

References

Katan, David (2009), « Translation Theory and Professional Practice : A Global Survey of the Great Divide", Hermes, Journal of Language and Communication studies, Department of Language and Business Communication, Aarhus School of Business, Université d'Aarhus, Danemark

Ahmad, K., Davies, A., Fulford, H. & Rogers, M. (1992), "What is a Term? The semi automatic extraction of terms from text" in "Paper presented at a conference held at the University of Vienna, Institut für Übersetzer- und Dolmetscherausbildung, Translation Studies - An Interdiscipline".

A. Auksoriute, I. Belogrīvs, A. Bielevičiene, A. Blaudums, J. Bordāns, T. Borkowski, J. Borzovs, A. Braasch, E. Cauna, A. Dravniece, N. Dudlauskiene, C. Galinski, L. Henriksen, H.-J. Kaalep, A. Kalniņš, D. Kierzkowska, H. Kilgi, B. Kis, A. Liedskalniņš, B. Maegaard, A. Mitkevičiene, S. Olsen, C. Povlsen, U. Priede, G. Proszeky, I. Puksts, M. Raguz, I. Raupach, R. Salupere, K.-D. Schmitz, K. Siwek, R. Skadiņš, V. Skujiņa, R. Stunžinas, G. Tardy, A. Uritam, A. Vasiļjevs, J. Zabielaite (2006), "Towards Consolidation of European Terminology Resources: Experience and Recommendations from EuroTermBank Project". Riga, Latvia.

Baroni, Marco and Bernardini, Silvia (2004), "BootCaT: Bootstrapping Corpora and Terms from the Web". In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal, 2004.

Clark, Alexander (2003), "Combining distributional and morphological information for part of speech induction". In *Proceedings of EACL 2003*, Budapest, Hungary.

Daille, Béatrice (2005), "Variations and application-oriented terminology engineering". *Terminology*, 11(1):181-197.

Daille, Béatrice (2012), "Bilingual Terminologies from comparable corpora: the TTC TermSuite". In *Proceeedings of The Fifth Workshop on Building and Using Comparable Corpora (BUCC 2012),* LREC 2012. Istanbul, Turkey.

Blancafort, Helena; Heid, Ulrich; Gornostay, Tatiana; Méchoulam, Claude ; Daille, Béatrice; Sharoff, Serge (2011), "User-centred Views on Terminology Extraction Tools: Usage Scenarios and Integration into MT and CAT Tools". Tralogy Conference *Translation Careers and Technologies: Convergence Points for the Future*, Paris, France.

Blancafort, Helena and Gornostay, Tatiana (2010), "Calling Professionals: Help Us to Understand Your Needs! TTC Survey 2010 results". (presentation published on the project website[8]).

Bowker, Lynne and Pearson, Jennifer (2002). Working with Specialized Language: a practical guide to using corpora. Routledge.

[8] http://www.ttc-project.eu/images/stories/TTC_Survey_2010.pdf

Gornostay, Tatiana (2010), “Terminology Management in Real Use”. The 5th International Biannual Conference *Applied Linguistics in Research and Education* in Memoriam Rajmund Piotrowski (1922-2009), Saint-Petersburg, Russia.

Groc, Clément 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.

Harastani, Rima; Daille, Béatrice and Morin, Emmanuel (2012), “Neoclassical Compound Alignments from Comparable Corpora”. In *Proceedings of the 13th International Conference of Computational Linguistics and Intelligent Text Processing* (CICLING 2012), p. 72-82. Lecture Notes in Computer Science 7182, Springer Verlag.

Daille, Béatrice and Morin, Emmanuel (2012), “Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique“. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2: TALN, ATALA/AFCP, 2012: 141-154. Grenoble, France, 2012.

Och, Franz Josef and Ney, Hermann (2003), “A systematic comparison of various statistical alignment models”. *Computational Linguistics*, 29(1):19–51.

Rocheteau, Jérôme and Daille, Béatrice (2011), “TTC TermSuite: A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora”. IJCNLP, Chiang Mai, Thailand.