HAL
archives-ouvertes.fr

Building Bilingual Terminologies from Comparable Corpora: The TTC TermSuite

Béatrice Daille

▶ **To cite this version:**

Béatrice Daille. Building Bilingual Terminologies from Comparable Corpora: The TTC TermSuite. 5th Workshop on Building and Using Comparable Corpora with special topic "Language Resources for Machine Translation in Less-Resourced Languages and Domains", co-located with LREC 2012, May 2012, Istambul, Turkey. hal-00819594

HAL Id: hal-00819594
https://hal.archives-ouvertes.fr/hal-00819594

Submitted on 1 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.

Building bilingual terminologies from comparable corpora: The TTC TermSuite

Béatrice Daille

University of Nantes
LINA, 2 Rue de la Houssinière
BP 92208, 44322 Nantes, France
beatrice.daille@univ-nantes.fr

Abstract

In this paper, we exploit domain-specific comparable corpora to build bilingual terminologies. We present the monolingual term extraction and the bilingual alignment that will allow us to identify and translate high specialised terminology. We stress the huge importance of taking into account both simple and complex terms in a multilingual environment. Such linguistic diversity implies to combine several methods to perfect accurately both monolingual and bilingual terminology extraction tasks. The methods are implemented in TTC TermSuite based on a UIMA framework.

Keywords: comparable corpora, terminology extraction, alignment, language for special purpose

1. Introduction

The need for lexicons and terminologies is overwhelming in translation applications because of cross-linguistic divergences and mismatches that arise from the perspective of the lexicon. For scientific domains, terminological resources are often not available or up-to-date, especially for emerging domain; moreover, the languages that are covered are often limited to 2 or 3 languages of which one is English. Previously translated texts could be used to create such linguistic resources such as the GIZA++ statistical machine translation toolkit (Och and Ney, 2003). But, there is no parallel corpora for most specialized domain and most pairs of languages. To tackle the drawbacks of term alignment from parallel corpora, comparable corpora seem to be the right solution to solve textual scarcity: as monolingual productions, they are authentic texts out of translations, and the babel web ensures the availability of large amounts of multilingual documents. The TTC project relies and this hypothesis and its aim is to perform terminology extraction from comparable corpora and to demonstrate the operational benefits on MT systems.

To build high-specialized terminologies, terms are extracted monolingually from the comparable corpus. To collect close candidate terms across languages, it is necessary to use a term extraction program that is able to handle both simple and complex terms (Kageura, 2002) and able to deal with terminology variation. Once monolingual candidate terms are extracted from the two parts of the bilingual comparable corpora, the alignment program which task is to propose for a given source term, several candidate translations should be able to handle both simple and complex terms. Within this context, we present TTC TermSuite, a terminology mining chain that performs both monolingual and bilingual terminology extraction from comparable corpora for seven languages.

2. Monolingual terminology extraction

To build high-specialized terminologies, terms are extracted monolingually from the comparable corpus. To collect close candidate terms across languages, it is necessary to use a term extraction program that applies the same method in the source and in the target languages. To work at the multilingual level, we have to reconsider the rough distinction between simple and complex terms to take into account morphological compounds. Morphological compounds are identified by tokenisation programs as single-word terms but for some languages such as German, they look quite similar to multi-word terms. The translation of MWTs is the most need as they constitute around 80% of the domain-specific terms, see for example Nakagawa and Mori (2003) for Japanese language. For German language, morphological compounds appear to be much more frequent than MWTs: 52% of nouns were reported by Weller et al. (2011) on the renewable energy TTC corpus[1].

Compound consists of the concatenation of two or more lexemes to form another lexeme. We distinguish 2 types of compounds: neoclassical compounds and native compounds. The first one are built with at least one neoclassical element such as *patho*, *bio-*, *-logy* (Bauer, 1983); the second are built with words of the native language such as *windmill*. Neoclassical compound could be identified thanks to a list of combining forms and dictionary look-up (Harastani et al., 2012) and native compounds by a splitting algorithm which is combined with a dictionary look-up (Weller and Heid, 2012).

The terminological occurrences that are extracted are SWTs and MWTs whose syntactic patterns correspond either to a canonical or a variation structure. The patterns are expressed using MULTEXT part-of-speech tags and are provided for all TTC languages. The main patterns whatever is the language are N and A for SWTs. For French and Spanish, the main patterns of MWTs are N N, N S:p N and N A. The variants handled are graphical, morphological, and syntactic. The three types of terms face up variants even if some are more likely to concern one main type

[1] `http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html`

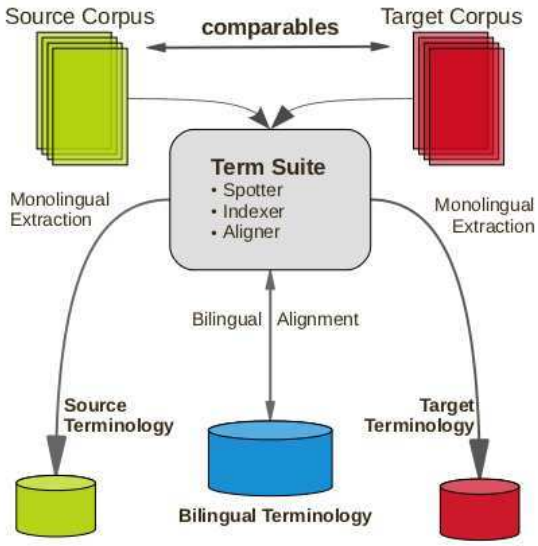

Figure 1: TTC TermSuite Architecture

such syntactic variants for MWTs. Monolingual terminology extraction and variant detection for multi-word terms were evaluated by Gojun and Heid (2012) for German language on the security domain. They gave a recall of 65% and were able to increase the existing terminology of the domain with new terms by 25%.

3. Bilingual terminology alignment

Once source and target terminologies are extracted form monolingual corpora, the alignment step could be set up. The output is a bilingual domain-specific terminology lexicon where for one source term you need to translate, you will obtain several candidate translations ranked form the most likely to the less. The method to align a source term with a target term relies on the hypothesis that a word and its translation tend to occur in similar contexts within a comparable corpora. The context of a word is expressed thanks to co-occurrences appearing in a context window. The co-occurrences are translated using a general bilingual language dictionary in the target language and compare to existing contexts of target words. The context-based projection approach proposed by (Rapp, 1995) for aligning words from bilingual comparable corpora is the gold standard. Using this approach, a precision of 60% is obtained for the translation of SWTs by examining the first 20 candidates translations using specialized language corpora of small size (0.1 million-word English-German corpus in (Déjean et al., 2002) and 1.5 million-word French-Japanese corpus in (Morin et al., 2010). But results drop significantly for MWTs, a precision of 42% of the 20 first candidates in a 0.84 million-word French-Japanese specialized language corpus (Morin et al., 2010). It is thus necessary to use another method.

For MWTs, it is possible to exploit the compositional property that characterizes half of MWTs - 48.7% have been reported by (Baldwin and Tanaka, 2004) for English/Japanese N N compounds. A compositional translation approach will translate each word of the MWT individually using a bilingual dictionary, and then appropriately piecing together the translate parts. It is possible to implement the composition approach at the morpheme or at the word level (Baldwin and Tanaka, 2004). For neoclassical compounds, we apply the compositional approach at the morpheme level making the assumption that most neoclassical compounds in a source language translate compositionally to neoclassical compounds in a target language. For example, the translation of the English noun *hydrology* in French is *hydrologie*, which can be interpreted by the combination of the translation of the composing elements, *hydro* (water): Fr *hydro* and *logy* (study): Fr *logie*. For MWTs, we apply the compositional approach at the word level. For example, the translation of the French MWT *fatigue chronique* is obtained by translating both *fatigue* and *chronique* into *fatigue* and *chronic* using a bilingual dictionary look-up.

4. TTC TermSuite

TTC TermSuite[2] is designed to perform bilingual term extraction from comparable corpora in five European languages: English, French, German, Spanish and one under-resourced language, Latvian, as well as in Chinese and Russian. The general architecture is presented in Figure 1. TTC TermSuite is based on the UIMA framework which supports applications that analyze large volumes of unstructured information. UIMA was developed initially by IBM (Ferrucci and Lally, 2004) but is now an Apache project[3].

4.1. General architecture

The architecture could be described from the point of view of the hierarchy of treatments or from the point of view of the data workflow. TTC TermSuite is a 3-step functional architecture that is driven by the required inputs and provided outputs of each tool. The bilingual term alignment (step 3 ALIGNER) requires processes of monolingual term extraction (step 2 INDEXER), itself requiring text processing (step 1 SPOTTER). The spotter applies a shallow pre-processing of the monolingual corpora, performing tokenization, part-of-speech tagging, stemming and lemmatization. The workflow is summarized in Figure 2: at the first step, we treat one document by one. If we get n documents, we will obtained n documents linguistically analyzed through the spotter; From this set of documents, we perform monolingual term extraction using the indexer which output is a terminology file; The last step is the alignment that requires one source and one target terminology files and proposes as output a bilingual terminology file.

4.2. Monolingual term extraction

Monolingual term extraction consists in processing a monolingual corpus document by document and in providing its terminology. It involves:

1. the recognition and the indexing of both single-word and multi-word terms;
2. the computing of their relative frequency and their domain specificity;

[2] http://code.google.com/p/ttc-project
[3] http://uima.apache.org

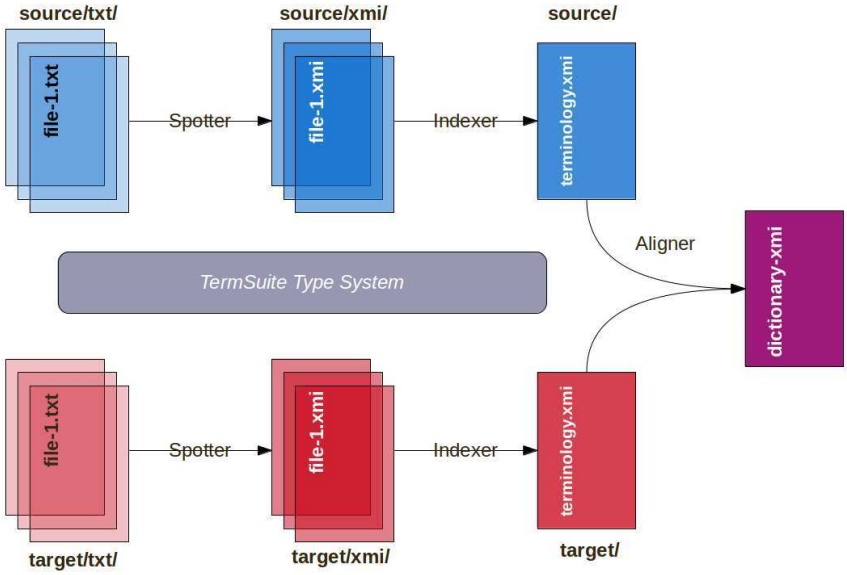

Figure 2: TTC TermSuite Workflow

3. the detection of neoclassical compounds above the set of single-word terms;
4. the grouping of term variants;
5. the filtering of some candidates using thresholds that could be expressed on the relative frequency or the domain specificity.

The term variant grouping functionality takes place once terms have been annotated as single-words or multi-words, and once single-word terms have been flagged as thus or as neoclassical compounds. After the collecting of term-like units, the TTC TermSuite organizes them that result in clusters of candidate terms. The clustering adopts different strategies that depend on the nature of the variation: graphical term variants are detected using edit distances, morphological variants using monolingual lists of affixes, and syntactical term variants using pattern rules over feature structures.

4.3. Bilingual term alignment

Bilingual term alignment adopts different strategies with regards to the nature of terms: for a SWT, it is the context-based projection approach; for neoclassical compounds and MWT compositionality-based method approaches are launched. The alignment of neo-classical compounds were evaluated on the En-Fr, En-De and En-Es pairs of languages on the TTC renewable energy corpus and showed a high precision for all pairs of languages (Harastani et al., 2012). For example, 100 aligned terms were obtained for the En-Fr pair with a precision of 98%. SWTs and MWTs have not been yet evaluated but as state of art methods have been implemented, we foresee for SWTs to reach a precision of around 60% on the first 20 translations, and for MWTs a precision of 68% for a recall of 40% (Morin and Daille, 2009). However, the combination of the two main strategies: context and compositionality-based methods should increase the overall performance. The coming evaluation of TTC TermSuite will hopefully confirm these numbers.

5. Conclusion

TTC term extraction techniques rely on low-level annotated corpora where sentence boundaries, word classes and lemmas are annotated. Patterns are used to extract term candidates: simple and complex terms are handled as well as their variants. Several statistics are computed that could be used to filter the list of monolingual candidate terms. The alignment combined compositional and context-based methods to treat both simple and complex terms. The bilingual terminology building is implemented in TTC TermSuite based on the UIMA framework for English, French, German, Spanish, Latvian, Chinese, and Russian.

6. Acknowledgement

The research leading to these results has received funding from the European Communitys Seventh Framework Programme (*/*FP7/2007-2013*/*) under Grant Agreement no 248005.

7. References

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In *ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31.

Laurie Bauer. 1983. *English word-formation*. Cambridge university press.

Hervé Déjean, Fatia Sadat, and Éric Gaussier. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224, Tapei, Taiwan.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10:327–348, September.

Anita Gojun, Ulrich Heid, Bernd Weissbach, Carola Loth, and Insa Mingers. 2012. Adaptating and evaluating a generic term extraction tool. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istambul, Turkey, May. ELDA.

Rima Harastani, Béatrice Daille, and Emmanuel Morin. 2012. Neoclassical compound alignments from comparable corpora. In Alexander F. Gelbukh, editor, *CICLing (2)*, volume 7182 of *Lecture Notes in Computer Science*, pages 72–82. Springer.

K. Kageura. 2002. *The dynamics of terminology: a descriptive theory of term formation and terminological growth*. Terminology and lexicography research and practice. J. Benjamins Pub.

E. Morin and B. Daille. 2009. Compositionality and lexical alignment of multi-word terms. In *Language Resources and Evaluation (LRE)*, volume 44 of *Multiword expression: hard going or plain sailing*, pages 79–95. P. Rayson, S. Piao, S. Sharoff, S. Evert, B. Villada Moirón, springer netherlands edition.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2010. Brains, not brawn: The use of "smart" comparable corpora in bilingual terminology mining. *TSLP*, 7(1).

Hirosi Nakagawa and Tatsunori Mori. 2003. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

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

Reinhard Rapp. 1995. Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

Marion Weller and Ulrich Heid. 2012. Simple methods for dealing with term variation and term alignment. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istambul, Turkey, May. ELDA.

Marion Weller, Anita Gojun, Ulrich Heid, Béatrice Daille, and Rima Harastani. 2011. Simple methods for dealing with term variation and term alignment. In Kyo Kageura and Pierre Zweigenbaum, editors, *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, pages 87–93, Paris, France, November. INALCO.