Terminology Management: New Perspectives

Gilles Falquet, Jacques Guyot, Nizar Ghoula

1 Introduction

The initial goal of the Olanto Foundation[1] is to build and share a complete suite of professional computer aided translation (CAT) tools including a concordancer (bitext-based search engine); a statistical machine translation tool; a terminology database management system; and a translation memory management system.

These tools can be integrated with several Electronic Document Management Systems (EDMS) or with search engines. Despite the existence of a considerable number of open source tools in the CAT field, these tools remain complex and their integration incomplete. Thus, they do not meet the complete chain of needs commonly expressed by Translation Services and Language Service Providers. Additionally, they generally don't benefit from a robust distribution and support structure and some of them are not really scalable.

The Olanto tools (Ghoula & al., 2014) are currently in common use in companies and international organization that work with large corpora (hundreds of thousands of documents) and in many different languages.

In this paper we show that connecting terminological resources with other knowledge resources, such as text corpora, aligned texts, or ontologies can enhance their applicability and quality. We first show how the tools developed at Olanto leverage different types of knowledge resources and processes to provide new functionalities. Then we present a general model of interconnected knowledge resources and show how it can be employed one some complex tasks involving terminological knowledge.

2 Using external resources in the Olanto tools

Some of the Olanto tools take advantage of external resources, in particular multilingual aligned corpora, to enhance or improve their functionalities, as we will see on the following examples.

[1] olanto.org

The myCAT concordancer and its alignment component

MyCAT is a typical computer aided translation (CAT) tool that acts as a multilingual concordancer. When a user inputs a term or a whole sentence the tool finds all its occurrences in a text corpus, or all the occurrences of the longest possible subsequences of it. It shows the occurrence context (a sentence) and the translation of this context, if it exists in the translation memory.

To display the translation of a sentence, the tool must be able to find it in the translated document. In other words, we must have a *map* that relates each sentence of the source document to its translation in the target document. This problem, known as the sentence alignment problem, has been addressed in many theoretical and practical ways.

MyCAT uses a dictionary based technique to perform the sentence alignment. This technique is based on the hypothesis that a sentence S_b is probably a translation of a sentence S_a if S_b contains a large number of terms that are possible translations of some terms in S_a. Although this technique yields high quality alignments, it requires a bilingual dictionary. Moreover, this dictionary must have a sufficient coverage of the terms that appear in the corpus (domain terms)

What makes myCAT unique is that its automatic alignment feature is based on bilingual dictionaries that are built with Moses, a Statistical Machine Translation engine, from extracts of the translation memory. Thus the dictionaries is, by construction, well adapted to the domain vocabulary. The dictionary generation process consists in

1. Manually aligning a set of sentences of the translation memory (generally 1000 sentences)
2. Running the Ghiza tool, a part of Moses, that statistically generates a biligual dictionary for sequences of one to seven words.

The generated dictionary is not a real bilingual dictionary because the word sequences it contains do not all correspond to concepts, they are frequent word sequences. Nevertheless, experiments have shown that this type of dictionary is really efficient to perform dictionary based alignment.

myCAT also provides a Quote Detector (myREF) that compares a document to be translated with the complete corpus of previously-translated documents. It detects all parts of sentences (or full sentences or paragraphs) that may be quoted from other documents. It then displays and aligns the source and target versions of that reference document.

Performing quote detection on large documents in large corpora is an extremely compute intensive process (every sentence of the document must

be compared with every sentence of the whole corpus). Therefore we developed a highly efficient indexing system that considerably reduces the computation time.

The How2Say terminology explorer

A translator or terminologist can generally rely on the terms proposed in terminologies. Nevertheless, there are situations in which it becomes necessary to

- check if the proposed translations correspond to the actual use in a corpus
- find terms that are composed from given terms
- find synonyms that are specific to a corpus
- find translations for terms that are not yet in a terminology

It is also important that the user can have access to these information in the easiest way and interactively. These are precisely the objectives of the How2Say tool. Figure 1 shows a typical exploration with the interactive user interface of How2Say.

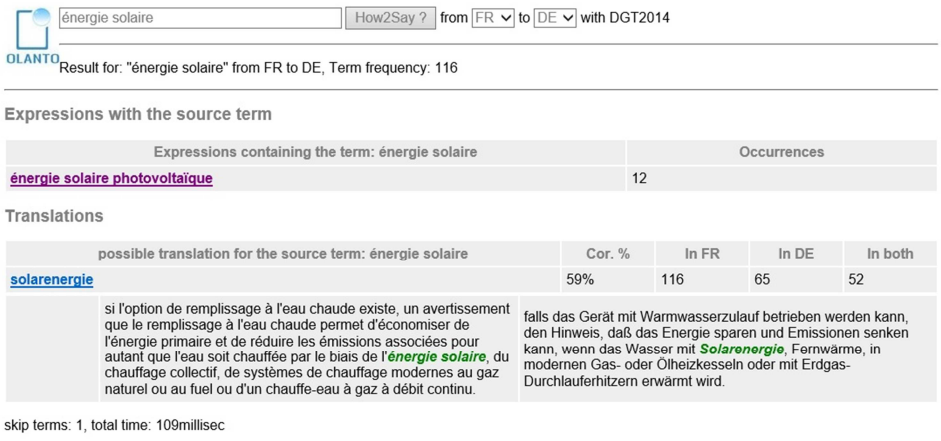


Figure 1: The How2Say user interface

In this case the user entered the French term *énergie solaire*. First the system found a term that contains *énergie solaire*, namely, *énergie solaire photovoltaique* in the French corpus. Then, by scanning the French-German aligned sentences it induced that *Solarenergie* is a possible translation (with 59% correlation) for *énergie solaire*.

How2Say can be seen as a virtual a virtual multilingual terminology. It does not *store* any term nor any translation but re-computes them interactively when the users request them. The algorithm to find potential translations proceeds as follows

It starts with a corpus of aligned sentences (possibly obtained with the myCAT alignment tool). The corpus is indexed to accelerate the term finding operations.

For a given term w in the source language A

1. find all the sentences in A that contain w
2. get the corresponding sentences in B (translations)
3. find the most frequent terms in the translation
4. for each frequent term u compute its correlation with w
5. retain the term(s) with the highest correlation

The correlation between a term w in A and a term u in B is a measure of how frequently u appears in the translation of a sentence that contains w, compared to how frequently it appear in the translation of a sentence that does not contain w. A high correlation indicates that u is probably a good translation of w.

It must be noted that How2Say is not a terminology extractor it does not try to recognize terms in a corpus but always starts with a user-provided term. Its main advantage is its simplicity, compared to training a fully automated translator. Therefore it can quickly adapt to any addition to the corpus (new documents, documents in a new language)

Validation of derived bilingual lexicons in myTerm

The same correlation-based technique is also utilized in the myTerm terminology management tool. In myTerm it is possible to create new bilingual lexicons by transitivity, e.g. combining a EN-FR lexicon and a FR-DE lexicon to obtain a new EN-DE lexicon. This may be the only way to create a bilingual lexicon for rare languages or non-common combinations. But It is well known that polysemy within both lexicons can produce associations between pairs of terms that do not make sense. For example, starting from the associations *time* → *temps* in EN → FR, *temps* → *Zeit* and *temps* → *Wetter*, in FR→DE, the composition produces two term associations: *time* → *Zeit* and *time* → *Wetter** for EN → DE.

The generated term pairs can be validated with a parallel corpus of the desired languages. For instance, going back to the example, we can find in a EN → DE corpus that the pair *time* → *Zei* has a high correletion wheras *time* → *Wetter* has a very low correlation and must be rejected.

3 A generic model for managing interconnected knowledge resources

Our experience in the development of the Olanto tools shows that complementing traditional terminological resources with other resources, in particular aligned sentences that form parallel texts, can improve their usability. With these additional resources it becomes possible to develop new terminology-related application or to improve the quality of existing ones. Similarly, our work on other projects showed that adding other types of resources, such as ontologies or folksonomies can also be extremely beneficial.

However, there is currently no terminology management tool that fully supports this type of extension. This is why we started developing a model of knowledge resources that covers the ontological, terminological and linguistic dimensions. In this model we consider that a network of interconnected resources is made of autonomous resources, enrichment (or interconnection) resources, and composite resources (Ghoula, 2014).

Autonomous resources

Autonomous resources are knowledge resources that can be used without reference to other resources, such as thesauri, terminologies, documents, corpora or ontologies. Autonomous resources can be categorized as either ontological, or terminological, or linguistic

Ontologies

From a very general perspective an ontology is a specification of some conceptualization of a domain. A conceptualization is an abstract model that represents the entities of a domain in terms of concepts, relations, and other modeling primitives. Most of the ontological languages specify the meaning of concepts with some form of explicit definition. Thus, an ontology is generally comprised of

- a representational vocabulary with different types of symbols (class names, relation names, etc.)
- a set of definitions that specify the meaning of the vocabulary

Each ontological language has its own types of symbols and definition expression language. For instance, in description logics the representational vocabulary consists of concepts, properties, and individuals; definitions are expressed as logical axioms that state, among others, equivalences, inclusions or exclusions between concepts as well as constraints on properties. The vocabulary of an ontology defined by UML class diagrams is made of classes, attributes, associations, etc. Definitions are graphically expressed by diagrams that can represent generalization/specialization or part/whole

constraints between classes, as well as constraints on the associations between classes.

Terminological, Lexical and semantic resources

This category comprises resources that organize knowledge according to some structure but that cannot be considered as ontologies. These resources can be term-oriented (terminologies, gazetteers, glossaries, dictionaries, lexicons, folksonomies) or classification-oriented (categorization schemes, subject headings, classification schemes, taxonomies) or relation-oriented (semantic networks, thesauri)

Linguistic resources

These resources are essentially written texts and transcriptions of spoken language productions that are encoded using a specific standard for text representation in digital form such as the Text Encoding Initiative (TEI). These texts are generally stored in documents that form text corpora.

Enrichment resources

We define enrichment resources as knowledge resources that interconnect elements from one or multiple autonomous resources. Enrichment resources are the result of applying an automated or manual process involving a set of elements that can be entities or resources.

Alignment resources contain links that represent semantic equivalence (or subsumption) between entities of the aligned resources. For instance, an alignment between two sets of sentences (linguistic resources) represents the equivalence of meanings. An alignment between two terminologies or ontologies connects concepts that are considered as equivalent.

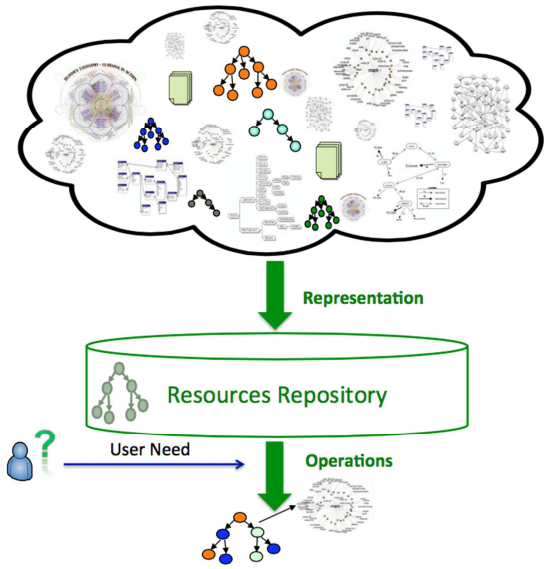
An annotation resource enriches an existing resource by associating its elements to descriptors that make their description more precise. For instance, words in a text can be associated to concepts in an ontology to disambiguate their meaning. Sentence can be associated to their discursive function (definition, hypothesis, argumentation, …)

Combined Resources

These resources combine a set of autonomous resources with some enrichment resources into standalone resources. For example, a parallel corporus or a comparable corpus is a resource of this kind since it contains documents (autonomous) and alignments between their content (enrichment). Semantic hypertexts are also combined resources combining linguistic resources indexed by terms or concepts from terminological or ontological resources (e.g., Wikipedia). Large biomedical ontologies or thesauri often result from merging different vocabularies or terminologies using alignments between them.

Operations on resources

The goal of a knowledge repository is not only to store and retrieve resources but also to combine them to produce new resources that satisfy the user needs, as shown on Figure



.

Figure 2. Principles of a knowledge resource repository

Although there are many different types of knowledge resources, we have observed that a small set of generic operations is sufficient to describe complex processes on these resources (Ghoula & al., 2010).

Representation operators essentially translate from one representation model to another. For instance, a formal ontology in the OWL language can be transformed into a simple hierarchy of concepts.

Derivation and combination operators serve to

- extract a new resource from an existing one by selecting a part of its content (selection)
- aggregate several resources into a new one (merge)
- obtain a new resource by composition (e.g. obtain a bilingual A-C lexicon by composing an A-B lexicon with a B-C lexicon, or compose two ontology alignments)

Enrichment operators create enrichment resources (alignments or annotations).

These operators are generic and specific algorithms must be devised depending on the resource type to process. For instance, the alignment of text sentences and the alignment of ontologies require their own specific algorithms.

4 Using knowledge resources to perform complex terminological tasks

As shown (Ghoula 2014) the above-defined operators can be combined to perform sophisticated tasks on knowledge resources. This section shows two examples of complex tasks that are carried out with these operations.

Enriching ontologies with new labels

As a practical scenario for using the repository, let's consider that an ontology designer wants to enrich an ontology O with terms and definitions of a terminology T. This can be achieved by performing the following sequence of operations

1. Translate T and O into a common formalism (e.g. into the OWL ontology language) and obtain T' and O'
2. Align T' and O'
3. Filter out the unnecessary information from O'
4. Merge T' and O' to produce the enriched ontology

Checking the semantic consistency of a thesaurus

Thesauri are common knowledge resources that play an important role in document classification and information retrieval. However, their usability is often limited due to their semantic vagueness or inconsistency. In particular, the broader/narrower term relation (BT/NT) may not correspond to a generic/specific or whole/part relation, as recommended by the ISO 25964 standard.

The following thesaurus validation technique takes advantage of several knowledge resources: the WordNet lexical ontology, the DOLCE top-level ontology, a DOLCE-WordNet alignment. It is based on techniques proposed in (Lacasta & al., 2013)

1. Align the thesaurus entries with entries (synsets) of the wordnet lexical ontology (this is done with an automated alignment algorithm that takes into account the entry labels, their lexical structure, and their context)
2. Use an wordnet-DOLCE alignment to align the thesaurus entries with DOLCE concepts (a direct thesaurus to DOLCE alignment is not feasible because DOLCE contains only high-level concepts)

3. When two thesaurus entries have a BT/NT relation, look for the corresponding concepts in DOLCE and check the semantic relations that may hold between these concepts. If one of these relations is compatible with a generic/specific or part/whole meaning then accept the BT/NT relation otherwise a potential error is reported

Example. In the Urbamet thesaurus, *accident* is narrower than *car*. As shown on Figure 3, the validation process will maps *car* to *physical object* in DOLCE and *accident* to *event*. Since the only possible relation between a physical object and en event is *participant-in*, which is not a generic/specific or whole/part relation. Therefore the relation between car and accident in the thesaurus is flagged as a potential problem.

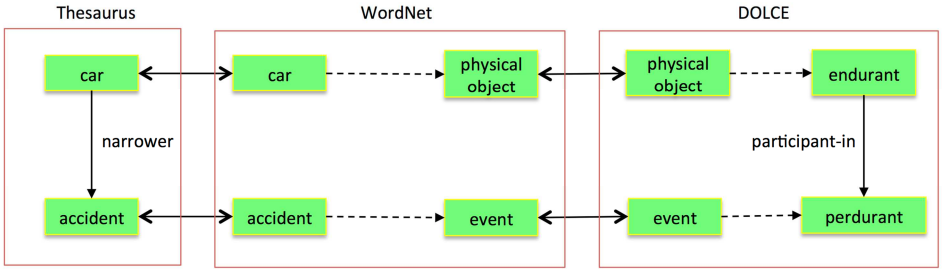


Figure 3. The relation validation process for the car NT accident relation

This validation technique has been implemented and it is currently tested on various thesauri. If the technique is sufficiently efficient to improve the quality of thesauri this means that a large number of existing thesauri could be improved and then transformed into ontologies.

5 Conclusion

In this paper we propose a comprehensive model of knowledge resource repository that grew up from our experience in developing the Olanto tools. This model can represent heterogeneous knowledge resources and operations on these resources. This model, and its implementation in modern knowledge resource repositories open new perspectives in the design and implementation of sophisticated new applications to perform terminological tasks.

Bibliography

Ghoula, N., Guyot, J., Falquet, G. (2014) Terminology Management Revisited. Translation and the Computer 36. London.

Ghoula, N., Falquet, G., Guyot, J. (2010). TOK: A meta-model and ontology for heterogenous terminological, linguistic and ontological knowledge resources. In Proc. ACM/IEEE Web Intelligence Conf., Toronto. 2010.

Ghoula, N. (2014) An ontology-based repository for combining heterogeneous knowledge resources PhD dissertation, Université de Genève. http://archive-ouverte.unige.ch/unige:45148

Lacasta, J., Nogueras-Iso, J., Falquet, G., Teller, J., Zarazaga-Soria, F.J. (2013) Design and evaluation of a semantic enrichment process for bibliographic databases. *Data & Knowledge Engineering* 88:94–107. DOI: http://dx.doi.org/10.1016/j.datak.2013.10.001)

Autor_adresse

Prof. Gilles Falquet
Université de Genève
Centre universitaire d'informatique
7 route de Drize
CH-1227 Carouge, Switzerland
Gilles.falquet@unige.ch

Dr. Jacques Guyot
The Olanto Foundation
10 Chemin de Champ-Claude
CH-1214 Vernier, Switzerland
jacques@olanto.org

Dr. Nizar Ghoula
Yellow pages
Montréal, QC, Canada
nizarghoula@gmail.com