Dealing with lexicon acquired from comparable corpora : validation and exchange

Estelle Delpech

Lingua et Machina

Béatrice Daille

Université de Nantes – LINA FRE CNRS 2729

Key words : terminology management, terminology extraction, comparable corpora, term-oriented terminology, TBX

1. Introduction

Recent years have seen a surge in research in multilingual terminology extraction. Second-generation tools have emerged, trying to tackle the drawbacks of term alignment from parallel corpora by using comparable corpora. There are two reasons for using comparable corpora instead of parallel corpora : (i) parallel corpora are scarce whereas comparable corpora are easily available, (ii) comparable corpora provide a way to observe language in use, contrarily to parallel corpora which are translations and bears influence from the original source text.

Techniques for the acquisition of terminology from parallel corpora were first introduced by (Rapp 1995; Fung 1997). These techniques rely on distributional semantics whose hypothesis is that words that are semantically close will tend to appear in the same contexts. Identification of term translations in comparable corpora requires three phases. The first phase consists in computing the context of each term in the source and target corpora. The context of a term *T* is represented by a vector indicating the number of times *T* co-occurs with each word *W* with in a given contextual window[1]. In the second phase, words in the source context vectors are translated into the target language by using a bootstrap bilingual dictionary[2]. In the third phase, the source and target vectors are compared[3] : the most similar the vectors, the most likely the

1 For instance : three words on its left and three words on its right.
2 The vectors are only partly translated due to the small coverage of the dictionary.
3 Using a similiarity measure such as the Cosine similarity, see (Rapp 1995) and (Fung 1997) for more details.

target and source terms are translations of each other. Finally, the output of the alignment algorithm is a list of one-to-many alignments : each source term is associated with an ordered list of candidate translations. The candidate translations are ordered from most to least probable. Results are evaluated by examining the best candidate translation (Top1), the ten best candidate translations (Top10) or the twenty best candidate translations (Top20).

The drawback of acquiring terminology from comparable corpora is that the acquired lexicons are not as reliable as those acquired from parallel texts. Lexicon acquisition from parallel texts outputs a one-to-one term alignment with high accuracy scores. For example, recent work on the matter (Lefever *et. al.* 2009) showed scores running from 85% to 90% accuracy. Conversely, systems that acquire lexicons from comparable corpora output one-to-many alignements : a source term associated to the set of its most probable target translations. As a consequence, the lexicons need to be post-editted before being injected in a termbase or in any other language processing module. For example, Fung (1998) shows a 80% precision on the Top20 candidates for single words alignments computed from large general language corpora (hundred millions of words or more). Dejean *et al.* (2002) find a 60% precision on the Top 20 candidates for single word terms using specialized language corpora of small size. Morin *et al*. (2007) indicate a 42% precision on the Top20 candidates for multi-word terms.

To our knowledge, existing term-alignment validation tools do not deal with this type of results. For example, the *iView* application from the *iTools* suite (Merkel and Foo 2007) presents its user with a list of one-to-one term alignements that have to be validated as correct alignments and as being domain specific. Side-information on term pairs consists of sample context sentences and some statistical data. Similarly, the commercial product *Araya Bilingual Term Extractor* (Waldhör 2006) also displays a list of one-to-one alignments with some statistical data. The *Xerox Terminology Suite*® provides quasi-exhaustive terminological records but no confidence score to help sort the candidate translations.

Terminology extraction from comparable corpora has raised the need for a new kind of terminology validation tool. This new kind of tool should be able to deal with one-to-many alignements or even many-to-many alignments if the source language term is a set of term-like sequences reflecting several variants of the same term. On top of that, we believe that sample

sentences and statistical data is important but not sufficient information to help the annotator or professional translator validate a term alignment. Last but not least, there should exist a data exchange format fitted for the exchange of automatically generated lexicon.

The paper is organized as follows : part 2 *A term-oriented annotation tool* describes the theoretical background that ruled the conception of our tool for the annotation of lexicon acquired from comparable corpora, part 3 *The term-alignment validation interface*introduces the validation tool's interface and part 4 *A TBX variant for automatically generated bilingual lexicons* proposes a TBX (Term Base eXchange) variant for the exchange of automatically extracted lexicons.

2. A term-oriented annotation tool

Literature on the theory of terminology (Bourigault et Slodzian 1999; Cabré 2003; L'Homme 2004) shows that there are two main conceptions of this field of research. One conception is said "concept-oriented" while the other is said "term-oriented". The difference between those two standpoints can be very briefly summarized as follows :

"Concept-oriented" terminology may be considered as the main stream and historical terminology theory. It is the theoretical standpoint represented in the International Organization for Standardization terminological norms. The aim of concept-oriented terminology is to modelize the knowledge of a domain by discovering its concepts and their relations; terms are considered as the mere linguistic expression of these concepts. The relation between a term and a concept is often seen as unequivocal and stable. This approach is well suited for prescriptive goals such as the creation of controlled-languages or language planning.

"Term-oriented" terminology rose in the 90's as a critic of concept-oriented terminology. It is deeply grounded in social sciences and linguistics. Its object of study is the term. This approach has shown that the definition of what is a term or not is rather subjective and highly depends on the final use of the terminology. This approach has also highlighted the fact that the term/concept relation is not always unequivocal and that terms do vary. Heavily relying on textual data, this approach seeks to describe terms usage and variation rather than prescribe it.

Contrarily to terminology management tools like *OpenTerminologyManager* (Waldhör 2002) or *Terminae* (Biébow and Szulman 1999), the annotation tool presented in this paper manipulates no such things as "concepts" and can be considered as a term-oriented and text-driven tool. It deals only with pairs of lexical units extracted from a corpus that are supposed to be correct translations of each other and thus considered as terms from a translation-aid point of view. It also provides the user with raw information extracted from a domain-specific corpus. This information can be extracted during the term alignment process and come as a side product of it or it can be retrieved from online public ressources like Wikipedia or Wiktionary. This information is intended to help the annotator grasp the *in-vivo* linguistic behaviour of the terms she or he has to annotate.

Such information includes :

- normalized form of the term (with the least flexionnal morphemes – usually the result of lemmatization)
- part-of-speech
- frequency or number of occurrences in the corpus
- definition
- collocations
- "soft" variants such as acronyms and orthographic variants
- stronger variants such as syntactic or morphosyntactic variants
- terms that have the same stem
- terms that appear in similar contexts
- contexts in which the term occurs : a sample paragraph, with a link leading to the original document

Although this kind of information would be of least interest in the construction of terminologies for localization or controlled-languages, we believe it is crucial in translation-aid applications. For example, concept-oriented approaches to terminology frequently seek to minimize (if not deny) term variation in order to control the use of terms. In translation-aid applications, the

detection of term variation and the harvesting of variants is very useful. While translating a text, a translator may stumble upon the variant of a term instead of the "authorized" canonical form of the term[4]. For that reason, term variants need to be taken into account during the validation process.

Another example is the use of definition to help the annotator or the translator understand the meaning of a term. While a definition will always remain useful, the meaning of a term can also be inferred from its contexts or by relating it to terms that appear in similar contexts. We see these two kinds of information (definition vs contexts and neighbouring terms) as complementary and think both should appear in a term-alignment validation tool.

In the annotation tool described in section 3, what is called a "term" is a lexical unit associated with an information record which parallels the terminological records found in most terminology management tools. It is characterized by a normalized form, a part-of-speech and a frequency. A term may have variants or may be in relation with other terms (either morphologically or semantically). A term variant is a plain lexical unit with no information record. The definition, collocations, relations to other terms, contexts and validated translations of a given term fully apply to its variants.

3. The term-alignment validation interface

As shown in Illustration 1, candidate term-alignments are displayed in the upper part of the interface. Because the accuracy of a term-alignment is not always clear-cut (two terms may be the exact translations of each other or vague equivalents), a candidate term alignment can receive one of these four labels : *correct*, *rather correct*, *rather incorrect*, *incorrect* or remain *unannotated*. The value of the annotation is expressed via a colour code, the darkest the colour, the more correct the alignment is. The use of a colour code helps catch at a glance the correctness of each alignment. If the right translation is not present among the candidate translations, the annotator can create an alignment with a new term.

4 For example, *aménagement de la forêt* and *aménagement forestier* are variations of the same term and both translate into English as *forest management* (Morin et al. 2004).

To help the annotator make a decision about the correctness of an alignment, the information records of the aligned terms are displayed in the lower part of the inteface. Each information record is divided into four sections :

- Tab 1 - Primary information : normalized form, part-of-speech, frequency, definition, collocations
- Tab 2 - Related terms : terms that share the same stem or that occur in similar contexts
- Tab 3 - Contexts : example of occurrences of the term and its variants in the corpus
- Tab 4 - Variants : soft (acronyms, orthographic variants) and strong variants (syntactic and morpho-syntactic variants).

The manual annotation process being a long and tedious task, all main actions are accessible via keyboard short cuts.

The annotation tool was developped using PHP/MySql and Ajax. It is available online and can be freely tested at [http://62.193.49.219/Metricc/InterfaceValidation/]. First beta-testers were quite happy with it and formulated several suggestions. Next improvements will be :

- implementation of TBX import / export functions
- hypertext navigation inside the information record : for example, a click on a related term should display its information record
- advanced search options (boolean operators, regular expressions)
- term filtering and term ordering functionalities

4. A TBX variant for automatically generated bilingual lexicons

This section adresses the issue of data exchange. Currently, there exists no standard for the exchange of automatically generated bilingual lexicons. Such a format should allow the encoding one-to-many alignments. Well-known formats for the exchange of terminological data include OLIF (Open Lexicon Interchange Format) by Lieske (2001), Geneter (GENEric model for TERminology) by Le Meur (1998) and TBX (TermBase eXchange) developped by the International Standards Organization (2008).

The TBX format perfectly fits with our goal because it is modular. It includes two modules : (i) a core XML structure, defined by a DTD and (ii) an XML formalism which identifies a set of data-categories and their constraints. The data-categories and constrainsts can be customized to create a TBX variant.

The core module is an XML structure compliant with the TMF (Terminological Markup Framework) meta-model (ISO, 2001). This meta-model organizes a termbase on three levels :

1. a concept-level, materialized by the xml tag <termEntry>
2. a language level, materialized by the xml tag <langSet>
3. a term-level, materialized by the xml tag <ntig> or <tig>

A <termEntry> includes one or several <langSet>, which in turn includes one or severel <ntig> or <tig> where the terms are encoded along with their linguistic information (part-of-speech, frequency, etc). Terms belonging to the same <termEntry> are considered as synonyms. As a consequence, terms belonging to different <langSet> but to the same <termEntry> can be considered as translations of one another. This is how term alignments will be encoded in TBX format : as two terms (<ntig>) under the same <termEntry> and belonging to different <langSet> (see Illustration 3 for an example).

The XML formalism used to identify a set of data-categories and their constraints is called XCS (eXtensible Constraint Specification). There exists a default TBX terminological markup language that uses the data-categories and constraints defined in the ISO norm 1260 (ISO, 1999). Obviously, the TBX default data-categories and constraints were not designed for the exchange of automatically generated lexicon. The TBX variant proposed here uses a subset of the default data-categories (*partOfSpeech, frequency, usageNote, corpusTrace, termType, reliabilityCode*) and three additional data-categories : *termDefinition, relatedTerm* and *termReference*.

Illustration 2 shows the TBX encoding of a term and its information record. The term is

phonème (phoneme) and has one variant *phon.* which is an abbrevation. Two data categories had to be added in order to encode the whole information record :

- *termDefinition* : this data-category is used to encode the definition of each term. It appears at the term level (<ntig> or <tig> tag).

  The default TBX data category *definition* could not be used for that purpose because its definition[5] in the ISO norm 12620 (ISO, 1999) states that it applies only to concepts, not to terms. Now, the annotation tool presented here is designed following a term-oriented approach. This means that we consider that even if two terms are correct translations of each other, they do not automatically fall under the same definition, as they may be mere equivalents or the meaning of a term in one language may be richer than the meaning of its translation.

- *relatedTerm* : this data-category is used to indicate terms that appear in the same contexts or that share the same stem. Again, the ISO norm 12620 (*ibid.*) provides data categories for relations between concepts, not between terms. This is why this category was added.

Illustration 3 shows the source term *phonème* (phoneme) with its set of candidate target translations. The data category *reliabilityCode* is used to indicate the score the extraction program gave to the alignment of the source and target term.

Like source terms, each target term comes with its information record. However, a target term might be repeated inside the TBX document because it may have been considered by the extraction program as a potential translation for several source terms. Repeating the whole information record each time a target terms appears in the document would be redundant and heavy to process. This is why a third data category has been added.This data category is called *termReference* and only occurs at the term level with a term target. It is used to refer to a term that has already occurred in the TBX document without having to repeat the whole information record again.

5 ISO12620-A0501: *" **Description**: A statement that describes a concept and permits its differentiation from other concepts within a system of concepts."*

5. Conclusion

We have described an interface specialized in the annotation of lexicons automatically extracted from comparable corpora. The need for such a tool is justified by the fact that bilingual lexicons extractors that deal with comparable corpora output one-to-many alignments that are not processed by traditional term-alignments validators. We have also argued that a validation interface should provide its user with more information than sample sentences and statistical data, especially in the field of assisted translation. This additional information should be corpus driven, so as to enable the annotator to understand the *in-vivo* linguistic behaviour of the terms he or she has to align. Finally, we have suggested a TBX variants that enables the encoding and exchange of one-to-many alignments.

6. Acknowledgements

This work was supported by the French National Research Agency (ANR), funding n° ANR-08-CORD-009. We also would like to thank the company Lingua et Machina for its support (http://www.lingua-et-machina.com).

References

Bourigault, D. and Slodzian, M. (1999) 'Pour une terminologie textuelle', *Terminologies Nouvelles*, n°19, 29-32.

Biébow, B. and Szulman, S. (1999) 'TERMINAE : a method and a tool to build a domain ontology', in Benjamins, V.R., Fensel, D. and Pérez, A.G. (eds) *Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure*, 25-30.

Cabré, T. C. (2003) 'Theories of Terminology', *Terminology*, 9(2): 163-199.

Déjean, H. and Gaussier E. (2002) 'Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables', *Lexicometrica, Alignement Lexical dans les Corpus Multilingues*, 1-22.

Fung, P. (1997) 'Finding Terminology Translations from Non parallel Corpora', in *Proceedings of the 5th Annual Workshop on Very Large Corpora* (VLC'97), Hong Kong, 192-202.

Fung, P. (1998) 'A statistical view on bilingual lexicon extraction : From parallel corpora to non-parallel corpora', in *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA'98*), Langhorne, United States, 1-17.

ISO (1999) 'ISO norm 12620 : Computer applications in terminology Data categories', International Organization for Standards, Genève.

ISO (2001) 'ISO norm 16642 : Computer applications in terminology Terminological markup framework', International Organization for Standards, Genève.

ISO (2008) 'ISO norm 30042 : Systems to manage terminology, knowledge and content – TermBase eXchange (TBX)', International Organization for Standards, Genève.

L'Homme, M.-C. (2005) 'Sur la notion de « terme »', Meta : journal des traducteurs / Meta : Translator's Journal, 50(4): 1112-1132.

Lefever, E., Macken,L. and Hoste, V. (2009) 'Language-independent bilingual terminology extraction from a multilingual parallel corpus', in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, 496-504.

Le Meur, A. (1008) 'GENETER: a generic format for the distribution and reuse of heterogeneous multilingual data', in *LREC 1998 Proceedings*, Grenada,

Lieske, C., Mc Cormick, S. and Thurmair, G. (2000) ' The Open Lexicon Interchange Format (OLIF) comes of age', in *Machine translation in the information age, MT summit Nº8*, Santiago de Compostela , Spain, 211-216.

Merkel, M. and Foo, J. (2007) 'Terminology Extraction and Term Ranking for Standardizing Term Banks ', in *16th Nordic Conference of Computational Linguistics*, Tartu, Estonai, 349-354.

Morin, E., Daille, B., Takeuchi, K. and Kageura K. (2007) 'Bilingual Terminology Mining -- Using Brain, not brawn comparable corpora', in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)* Prague, Czech Republic, 664-671.

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

Waldhör, K. (2006) 'Araya Bilingual Term Extraction', *User Manual. Heartsome Europe GmbH*,

available online at [http://www.heartsome.de/en/xliffug_en.pdf], accessed March 9th, 2010.

Waldhör, K. (2002) 'OpenTerminologyManager – a Web and Standards based OpenSource Terminology Management Tool', in *Workshop on International Standards of Terminology and Language Ressources Management , LREC 2002.*

Illustrations

Illustration 1 should appear on page 5, line 19.

Illustration 2 should appear on page 8, line 15

Illustration 3 should apper on page 8, line 27.

Illustration 1: The term alignment validation interface

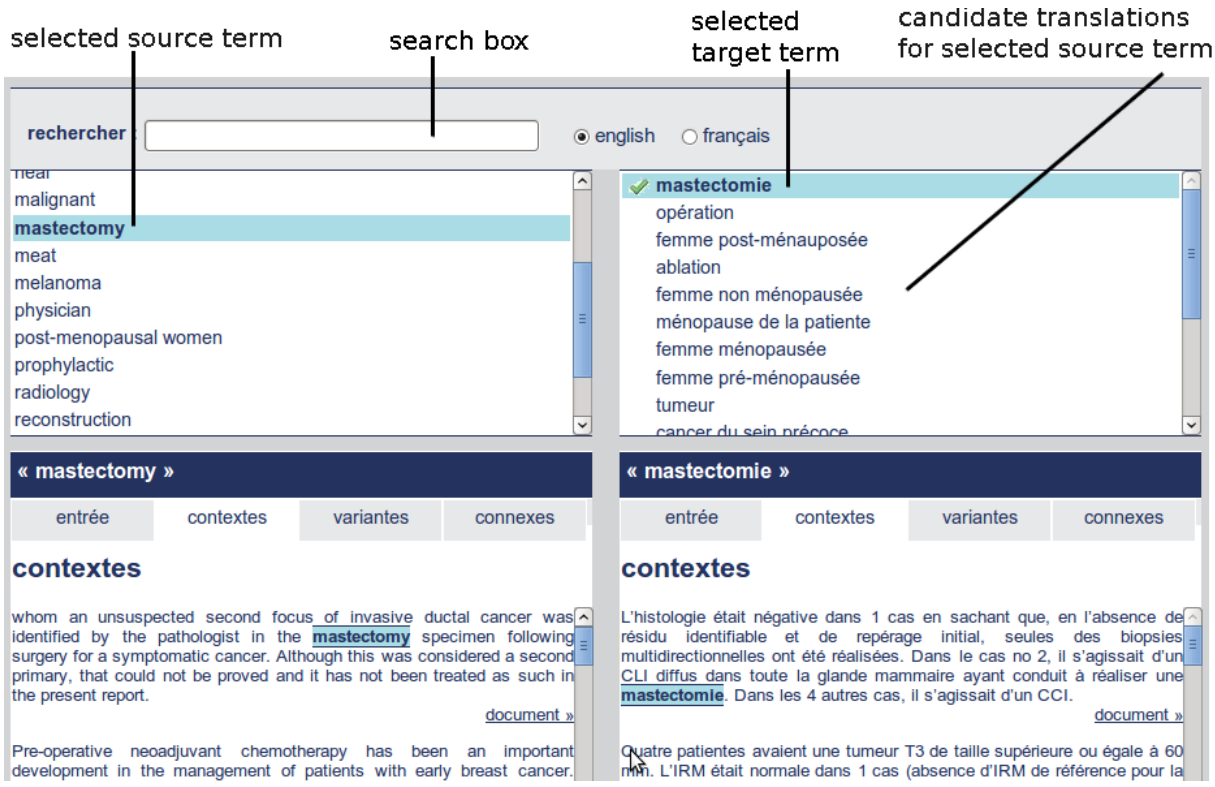

Illustration 2: A term and its information record in TBX Format

```
<ntig id="fr1">
      <termGrp>
            <term>phonème</term>
            <termNote type="termType">entryTerm</termNote>
            <termNote type="partOfSpeech">NOUN</termNote>
            <termNote type="frequency">commonlyUsed</termNote>
            <termNote type="usageNote">réalisation dun phonème, phonèmes
distincts,distribution d'un phonème</termNote>
            <termNote type="relatedTerm" target="fr2">morphème</termNote>
      </termGrp>
      <descrip type="termDefinition">plus petite unité discrète ou
distinctive que l'on puisse isoler par segmentation dans la chaîne
parlée</descrip>
      <xref type="corpusTrace" target="file://fr1.html">contextes</xref>
</ntig>
<ntig id="fr1a">
      <termGrp>
            <term>phon.</term>
            <termNote type="termType" target="fr1">variant</termNote>
            <termNote type="partOfSpeech">nom / abbréviation</termNote>
            <termNote type="frequency">commonlyUsed</termNote>
      </termGrp>
</ntig>
```

Illustration 3: Candidate biterms in TBX Format

```
<termEntry>
      <langSet xml:lang="fr">
            <ntig id="fr1">
                  <termGrp>
                        <term>phonème</term>
                  </termGrp>
            </ntig>
      </langSet>
      <langSet xml:lang="es">
            <tig>
                  <term>fonema</term>
                  <ref type="termReference" target="es1"></ref>
                  <descrip type="reliabilityCode">9</descrip>
            </tig>
            <tig>
                  <term>fono</term>
                  <ref type="termReference" target="es2"></ref>
                  <descrip type="reliabilityCode">5</descrip>
            </tig>
            <tig>
                  <term>alófono</term>
                  <ref type="termReference" target="es3"></ref>
                  <descrip type="reliabilityCode">3</descrip>
            </tig>
      </langSet>
</termEntry>
```