Term Translations in Parallel Corpora: Discovery and Consistency Check

Dan Tufiş

Research Institute for Artificial Intelligence of the Romanian Academy & University „A.I. Cuza" of Iaşi
Calea „13 Septembrie", no. 13, PO 050711, Bucharest
tufis@racai.ro

Abstract

The paper describes a method for identifying term translations in parallel corpora, developed within the FF-POIROT European project. This project aims at building multilingual (Dutch, Italian, French and English) resources in the financial/legal domain that may be used in knowledge and information systems by investigative bodies, and law enforcement in order to detect, investigate or help prevent instances of actual or attempted financial fraud. The methodology builds on our word alignment procedure based on translation equivalents extracted from parallel corpora. When a validated list of multiword terms is available in one language, the procedure provides the translations in any of the languages present in the parallel corpus. Given that a term is usually semantically non-ambiguous, the found translations of different occurrences of the same term should be the same (modulo inflectional variations). If this is not the case, one might suspect a non-systematic translation of the original term. When a man-made term list is not available, the system tries to discover the term candidates extracting sequences of words that appear together more frequently than expected by chance. By the procedure mentioned before, the candidate terms occurrences in one language are linked to their translation equivalents in the other languages.

Introduction

Designing a terminological database follows a methodology which terminologists are well aware of, and used to. Populating it requires, beyond the terminologists, domain experts who can decide which are the terms, what are definitions and what are the relations among them. In a multilingual environment the population of the terminological database is made more difficult with at least the following tasks: identifying the terms in the other languages, mapping them one another and ensuring uniform cross-lingual interpretation, that is consistent definitions in any language for the terms that are mapped as equivalents. It is common knowledge that terminological consistency over a large collection of thematic documents (very likely to be authorised by different organisations/people) is hard to achieve even when only one language is concerned. When multilingual dimension of a text collection is considered, ensuring the terminological consistency is too much hard a task for manual fulfilment and as such, economically worth paying attention to. However, one can easily identify areas where terminology inconsistency, especial in a multilingual setting may be extremely harmful. Combating frauds in a multi-national and multi-language geographical area such as Europe is a very good example where a common understanding/ (ontology supported) and a coherent and consistent linguistic realisation (multilingualy lexicalised ontology supported) of relevant documents are definitely objectives for the achievements of which the cost is not the primary judgement criterion. However, as these goals may be, to a large extent, automatised, incurring a dramatic cost decrease, focussing the research efforts along these lines is not surprising.

One of the goals of the FF-POIROT project (http://www.starlab.vub.ac.be/research/projects/poirot/) is to ensure that the multilingual terminological data-base reflects the consistent translations as used in the regulatory documents in various languages of the project. For this purpose, since one of the targeted domains of our project is in the area of VAT, we compiled a parallel corpus for the languages of the project based on the VAT 6th directive of EEC (Directive 77/388/EEC of 17 May 1977).

Preprocessing of the Parallel Corpus

The construction of the parallel corpus, assumed cleaning it up (the texts were exported from PDF documents), tokenization, tagging, lemmatization and sentence alignment. Tokenization for French and English were performed by Multext segmenter (MtSeg). Tagging was done with different taggers and tagsets for the three languages: XEROX's tokenizer and tagger for Dutch, ISSCO's tagger for French and TnT for English. The first two taggers also lemmatized the respective texts, while for English we used our own lemmatizer. The tagsets in the three parts of the parallel corpus were quite different and therefore we designed a mapping (loosing information) to a common denominator. The sentence alignment (achieved by a modified version of Gale and Church's (1993) *CharAlign*). The final pre-processing step was turning the vertical three-column texts and the alignment indexes produced by the previous steps into a simplified cesAna (http://www.cs.vassar.edu/CES/dtd2html/cesAna/) document. The VAT corpus we created is in itself an extremely useful multilingual resource. The cesAna-like mark-up is suggested below:

```
<text id="Ozz."><body>
 <tu id="Ozz.1">
  <seg lang="en">
   <s id="Oen.1">
    <w lemma="the" ana="Dd">THE</w>...
   </s>
  </seg>
  <seg lang="nl">
   <s id="Onl.1">
    <w lemma="richten#lijn" ana="Nc">Richtlijn</w>...
   </s>
  </seg>
  <seg lang="fr">
   <s id="Ofr.1">
    <w lemma="directif" ana="Adj_sg">Directive</w>...
   </s>
  </seg>
 </tu>
...
</body></text>
```

Figure 1: Encoding of the VAT-parallel corpus

The table below gives an overview of the small VAT corpus (only three languages included).

LANGUAGE	EN	FR	NL
No. of occurrences	41722	45458	40594
No. of word forms	3473	3961	3976
No. of lemmas	2641	2755	3165

Table 1: The "VAT" corpus overview

There are three scenarios which are covered by our methodology:

a) there exist comparable indexes of terms in a given area for one or more language concerned; here the main task is to map the existing terms in the indexes and identify the corresponding terms for languages where such lists are not available.

b) the existing index for an area of interest in two or more languages considered in the multilingual environment are uneven; here beside the previous task one has to consider automatically ballancing of the uneven indexes; under this scenario a special case is representing by dealing with new coined terms, missing from all the indexes;

c) there is no authorised index of terms for the selected area in any of the languages concerned; this situation, less likely, is the worst case where term discovery (and extraction) is a prerequisite process, after which either of the previous scenarious are applicable; term identification and extraction can be done manually (slow and expensive but presumably accurate) or automatically (fast and cheap but certainly less accurate than when done by human experts).

Word-level Alignment of the Parallel Corpus

Based on our previous translation-lexicon extraction program (Tufiş and Barbu, 2002), called TREQ, we developed a highly accurate word-aligner, TREQ-AL (Tufiş et al. 2003). In the shared task on word-alignment, organised on the occasion of NAACL-HLT conference in Edmonton, May 2003, TREQ-AL (after few bug-fixes) obtained the best score in word-aligning a Romanian-English parallel text (provided and evaluated by the organisers). TREQ-AL together with a collocation extractor (based on Ted Pedersen's NSP-v0.53) have been incorporated into a versatile parallel text mining system which we used for finding multilingual translations for the VAT terminology. Since both TREQ and TREQ-ALL are extensively described in the above mentioned papers, we summarise here the basic technicalities: the first draft of the alignment is done based on a improved variant of Melamed's (2001) competitive linking that considers besides coocurence scores (in our case, log-likehood) among words of compatible grammar (POS) and meta-grammar categories (sets of POSes), indirect association filters, also word similarities (cognate scoring) and their relative distances; the final alignment consider the heuristics according to which the words in a chunk of one part of the bitext usually are aligned to words belonging to a chunk in the other part of bitext (see Tufiş et al., 2003; Tufiş et al., 2004). The content words at each end of an alignment link make a translation equivalence pair (TEP) and all TEPs make a translation equivalence lexicon. The snapshot in the appendix exemplifies the graphical interface to TREQ(-ALL).

Reference Multiword Term List Alignment

Once a bitext is word aligned and a given list of terms in the hub language is available before-hand, the multilingual term extraction is very simple: each reference term is located in the alignment units and the sequence of words in the target language that is delimited by the leftmost and rightmost translation equivalents is taken to represent the possible translation of the hub language term. All the possible translations, extracted from different alignment units are then processed to identify the longest common sequence that also observe some constituency restrictions (expressed as regular expressions). It may happen that different translations will be extracted (some candidates share a common sequence and other candidates share a different common sequence).

If term lists are available for more languages, applying this procedureeach time with a different term list ensures:

-automatic interlingual checking of term-lists consistency;

-automatic importing the translation equivalence terms from one of the term list to the others, thus bringing all the languages on the same level of conceptual terminological coverage.

An expert in the VAT area manually extracted a list of English terms that appeared in the VAT corpus. This list contains 1043 terms, the words of each term being in inflected forms. We lemmatized the words in each term and eliminated the duplicates. In the VAT corpus built as described above, there remained only 900 terms. For most of the entries (834) there were identified translations in both target languages. For instance, if we have the English term **Community transit procedure** the equivalent term in French is **procédure de transit communautaire**.

In our example the equivalents are:

Community ⇔ communautaire (cross-part of speech equivalents);Transit ⇔ transit; Procedure ⇔ procédure.

The leftmost and the rightmost French translation equivalents for a constituent of the English term are **procédure** and **communautaire** respectively. Therefore, the entire French string delimited by these translation equivalents (notice that the French **de** has no equivalent in English) is taken as a candidate term for the English term.

Of course, it is not the case that everything between left most and right most translated words can be a term. There can appear a lot of other words that have nothing to do with the starting term. So, we write down an algorithm that calculates a score for each expression (or candidate) that seems to be an equivalent for the starting term.

All candidates are extracted from the corpus following the mapping described above. Then, these candidates are used to generate a list with the longest common strings that appears between candidates. We calculate a score for each longest common string, taking into account all content words translated and not translated, using the following formula: $DICE = 2*N_{1,2}/(N_1+N_2)$, where N_1 is the number of content words of the source language term, N_2 is the number of content words of the target language term, and $N_{1,2}$ is the number of content words in the source language term, which could be aligned to content words in the target language term. The candidate with the highest score is the one equivalent to the given expression. There may be more candidates with the same highest score. In this case we take into account their frequency and choose the most frequent one. If they have the same frequency, unless a human valuator decides the choice is random.

The validation of the extracted terms in case of multiple translations of a witness term demonstrated that even official multilingual documents (EC 6th Directive on VAT) are not consistent in their cross-lingual terminological use of the terms. For the list of English terms manually extracted by the VAT expert, the algorithm found term translations in 92.6% of cases. In a detailed error analysis project report we showed that practically all missing translations were due to pre-processing errors: spelling errors, wrong tokenization, tagging errors and very few wrong or missing terms due to sparseness of data. A native bilingual (FR-NL) speaker terminologist, member of the consortium, with an excellent command of English analyzed the found translations and reported the precision of 72%. We rerun the experiment, providing, when multiple possible translations, up to the best first 3 candidates. Not surprisingly, the precision got as high as 92.2%. Most of the difference was due to variations in translating some words of the same term (bad practice for terminology translation).

Mining for Multiword Terms

When a hub-language list of terms is not available in advance, our system produces a list of potential reference terms. A recursive program computes (based on log-likelihood scoring) the bigrams showing collocational scores above expectation. At each step (the number of iteration steps is a user-supplied parameter) the programs joins the pairs of words that show an association score higher than a pre-established threshold (another user-supplied parameter) turning the respective bigrams into a single token. This way, with n iteration steps, it is possible to identify terms containing up to 2^n words. The stop words are systematically skipped. The procedure is language independent and it can be applied to each of the monolingual part of the parallel corpus, thus obtaining term candidates in each language. Unless some linguistically motivated filtering is performed the candidate term lists would contain a lot of noise. The linguistic filters might take into account language specific constituent structures such as: a term must be a phrasal structure such as a noun phrase, a verb phrase, a clause, a term should not contain numbers, personal pronouns or conjunctions etc. With respect to the last restriction we assume that the conjunction separates two terms (that is "bank financial and insurance transaction" is considered to be a conjunction of two terms "bank financial transaction" and "insurance transaction"). We also used language specific (English) pruning rules (e.g. no leading or tailing determiners or prepositions: "5000 European units of account" becomes "European units of account", "a taxable person" becomes "taxable person", "another member state" becomes "member state"). We used a set of 18 constituency restriction rules (probably incomplete and sometimes over-restrictive; this is subject to further investigation).We applied these filtering rules on the list for English terms manually extracted by the human expert in VAT area and 398 of the 900 terms in our corpus did not pass the constituency restrictions. Here are some of these terms (the words inside a term are lemmatized) in which the offending constituent is underlined:
<5,000 European unit of account>; <amusement or entertainment>; <another place of destination>; <in one calendar year>; <in pursuance of an order>; <not establish>; <on an occasional basis>; <part thereof>; <price be reduce> etc. We eliminated them, plus another 145 single word terms (our algorithm looks only for multiword terms). The remaining 357 terms represented our Gold Standard (GS), used to evaluate the automatic term extraction algorithm.
The result of the evaluation is summarised in the table 2.

Total number of extracted terms	Terms in GS, found	Terms in GS, found as sub-terms	Terms in GS, not found	Terms not in GS, but likely to be correct
1977	144	79	134	≈1500

Table 2: Term extraction evaluation

The number in the last column (terms not in GS, but likely to be correct) is very subjective and should be confirmed by an expert in the area. It was estimated by counting 254 collocations which are definitely not terms (such as *<so far>, <not elsewhere>, <taxable person until expiry>, <following islands>*, etc.). Here are some examples of multiword expressions our algorithm found but which are missing from the man-made term index (for readability reasons, we did not lemmatize these word sequences): *<acquisition of the right to dispose as owner of movable tangible property>; <charges due outside the importing member state>; <Community transit procedure>; <electronically supplied services>; <identification number for value added tax>; <immovable property acquired as capital goods>; <person liable to pay the tax>; <products subject to excise duty>; <recapitulative statement of the acquirers identified for value added tax purposes>; <sale of goods on deferred terms>; <subsidies directly linked to the price of such supplies>; <tax exemption>; <temporary importation with full exemption from import duties>; <unit price exclusive of tax>; < valuations of movable tangible property>; <work on movable tangible property>* etc.
The terms extraction procedure may be followed by the alignment procedure described in the previous section. The terms that are found both ways (by the collocational method and by the translation equivalence method) are supposed to be the terms of interest.

Implementation and Conclusions

We have developed a system for finding translations in parallel corpora of a multiword terms glossary. If such an authoritative list of terms is not available, the system generates a list of candidates that should be validated by a domain expert. The system has a friendly interface, combining certain technologies like DTHML, XML, and XSL with languages HTML, JavaScript, Perl, PerlScript. This application runs under Win98, Win2k or WinXP and it is necessary to have installed on your computer IE 5.0 or greater, XML 4.0, ActivePerl 5.8.0. We found that even with a poor quality of the corpus preprocessing, our system is very robust. Any improvement in data preparation would boost the performance of this system.

Acknowledgements

The work reported here was carried with support from the European project FF-Poirot, no. IST-2001-38248. and

from the Romanian Ministry of Education and Research under the CORINT programme.

The evaluation of the term translation equivalents was performed by Koen Kerremans of CVC Brussels.

References

Gale, W. A., Church, K. (1993). A Program for Aligning Sentences in Bilingual Corpora. In Computational Linguistics, 19 (1), 75—102.

Melamed, D. (2001). Empirical Methods for Exploiting Parallel Texts. Cambridge, MA: MIT Press.

Tufiş, D. (2002). A cheap and fast way to build useful translation lexicons. In Proceedings of the 19th International Conference on Computational Linguistics (pp. 1030--1036), Taipei.

Tufiş, D., Barbu, A.-M. (2002). Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing. In International Journal of Speech Technology, Kluwer Academic Publishers, 5(3), 199--209.

Tufiş, D., Barbu, A-M, Ion, R. (2004). Extracting multilingual lexicons from parallel corpora. In Computers and the Humanities, Kluwer Academic Publishers (to appear, 38 p).

Appendix
The graphical interface to TREQ(-ALL)