Parallel corpus-based bilingual terminology extraction

Xavier Gómez Guinovart[1], Alberto Simões[2]

[1] Universidade de Vigo
xgg@uvigo.es

[2] Universidade do Minho
ambs@di.uminho.pt

Abstract : This paper presents a parallel corpora-based bilingual terminology extraction method based on the occurrence of bilingual morphosyntactic patterns in probabilistic translation dictionaries. We discuss an experiment focused on two language pairs – English-Galician and English-Portuguese, and show results which experimentally confirm the high degree of accuracy of the proposed extraction technique.

1 Introduction

This paper[1] presents a thorough analysis of a parallel corpora-based bilingual terminology extraction method based on the occurrence of bilingual morphosyntactic patterns in probabilistic translation dictionaries generated by *NATools* (`http://natools.sourceforge.net/`).

For the purpose of filtering and evaluating the results of term extraction we carried out an experiment in which both the level of lexical cohesion of the term candidates and their specificity with respect to a non-terminological corpus of the target language were taken into account. Testing was conducted for the language pairs English-Galician and English-Portuguese using the *Unesco Corpus* – which is part of the *CLUVI Parallel Corpus* (`http://sli.uvigo.es/CLUVI/`) (Guinovart & Sacau, 2004) – and the *JRC-Acquis* (Steinberger *et al.*, 2006), respectively. The evaluation results show a high degree of accuracy of the terminology extraction based on probabilistic translation dictionaries complemented by the bilingual syntactic patterns.

[1] This work has been funded by the Ministerio de Educación y Ciencia and the Fondo Europeo de Desenvolvemento Rexional (FEDER) within the project "Diseño e implementación de un servidor de recursos integrados para el desarrollo de tecnologías de la lengua gallega (RILG)" (HUM2006-11125-C02-01/FILO), and by the Consellaría de Innovación e Industria da Xunta de Galicia within the project "Desenvolvemento e aplicación de recursos integrados da lingua galega" (ref. INCITE08PXIB302185PR).

As for the evaluation of the terminological quality of the extracted terms, and given the lack of a comprehensive terminological database for Portuguese, a comparison with a hand-crafted normalised list of terms was performed only for Galician, for which we have the *bUSCatermos* (`http://www.usc.es/buscatermos/Caracteristicas.htm`) – a database with 126,338 Galician terms from all the fields collected by the Servizo de Normalización Lingüística at the University of Santiago de Compostela from a wide collection of dictionaries and glossaries, and the *Termoteca* (`http://sli.uvigo.es/termoteca/`) (Crespo *et al.*, 2008) – a corpus-based terminological databank with 6,621 Galician terms gathered by the TALG research group of the University of Vigo from the *Galician Technical Corpus* (`http://sli.uvigo.es/CTG/`) and the *CLUVI Corpus*.

The results of the system are used by a terminologists team at the University of Vigo as the basis for selecting English-Galician bilingual terms from the *CLUVI Corpus* in order to extend *Termoteca*.

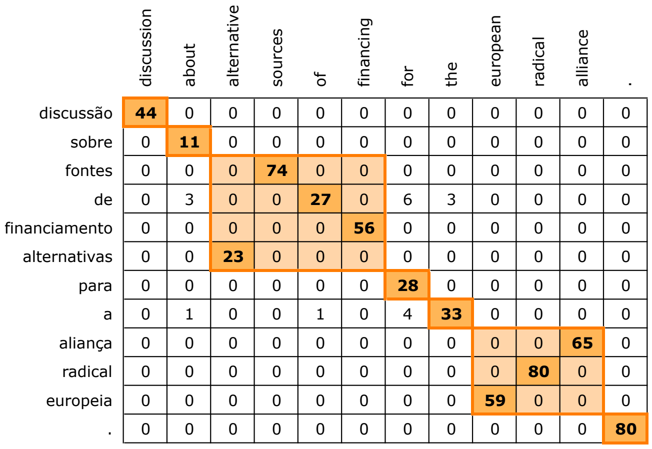
2 Extraction algorithm and metrics

The terminology extraction algorithm used in this study is based on *NATools* probabilistic translation dictionaries (Simões & Almeida, 2003) and was explained in detail in Simões & Guinovart (2009). *NATools* dictionaries, automatically extracted from sentence aligned parallel corpora, map words from a source language to a set of probable translations in a target language. Each of these translations have a probabilistic measure of translatability. This information enables the creation of an alignment matrix for any translation unit (figure 1) that includes in each cell the mutual translation probability for each word combination (from the source/target language). These matrixes can be used to extract bilingual terminology using translation patterns (Simões & Almeida, 2008) that specify how word order in the source language changes after translation takes place. Translation patterns may include morphological restrictions (for one or the both languages) defining the morphological categories allowed for the words matching the pattern. *NATools* relies on external morphological analyzers to validate the morphological restrictions. We used *jSpell* (Almeida & Pinto, 1994) for Portuguese and *FreeLing* (Atserias *et al.*, 2006) for Galician.

Moreover, following many other works on term extraction based on Dunning (1993), the system scores each term candidate with the log-likelihood measure, using the `Text::NSP` Perl module (`http://ngram.sourceforge.net/`). The minimum value for the partial trigrams is used for terms with more than tree constituents (Patry & Langlais, 2005).

3 Experiments and results

Our experiments focused on two language pairs, English–Galician and English–Portuguese and used two parallel corpora of very different sizes (table 1): the *Unesco Corpus* – a collection of 30 issues of the *Unesco Courier* (`http://www.`



	discussion	about	alternative	sources	of	financing	for	the	european	radical	alliance	.
discussão	**44**	0	0	0	0	0	0	0	0	0	0	0
sobre	0	**11**	0	0	0	0	0	0	0	0	0	0
fontes	0	0	0	**74**	0	0	0	0	0	0	0	0
de	0	3	0	0	**27**	0	6	3	0	0	0	0
financiamento	0	0	0	0	0	**56**	0	0	0	0	0	0
alternativas	0	0	**23**	0	0	0	0	0	0	0	0	0
para	0	0	0	0	0	0	**28**	0	0	0	0	0
a	0	1	0	0	1	0	4	**33**	0	0	0	0
aliança	0	0	0	0	0	0	0	0	0	0	**65**	0
radical	0	0	0	0	0	0	0	0	0	**80**	0	0
europeia	0	0	0	0	0	0	0	0	**59**	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	**80**

Figure 1: Alignment matrix with marked patterns (Portuguese–English)

`unesco.org/courier/`) in four languages, and the *JRC-Acquis* – a collection of parallel texts in 22 languages with the total body of Europea Union law applicable in the EU Member States.

Corpus	**Unesco**		**JRC-Acquis**	
Translation Units	47 903		1 315 907	
Tokens (source/target)	1 057 556	1 019 886	37 605 596	51 075 535
Forms (source/target)	50 866	66 515	283 061	295 923

Table 1: Parallel corpora

In addition, two literary corpora were used in the evaluation process for bigrams and trigrams exclusion (table 2): the BiVir Corpus – a Galician literary corpus containing works from the *Virtual Library of Universal Literature in Galician* (`http://www.bivir.com/`), and the *Compara* (`http://www.linguateca.pt/COMPARA/`) – a human-edited parallel corpus whose sentence alignment, lemmatization and POS tagging have been revised by human annotators (Frankenberg-Garcia & Santos, 2003).

Corpus	**Token**	**Bigrams**	**Trigrams**
BiVir	1 008 125	361 547	641 349
Compara	1 714 523	544 274	1 243 356

Table 2: Exclusion corpora

In order to evaluate the precision of the *NATools*-based term extraction algorithm, four translation patterns were defined, as shown in figure 2[2].

Different methods are used for filtering the results of term extraction: identification of unlikely term candidates because of their similarity with a lexical

[2] The EN–GL patterns are similar but with FreeLing specific tag names. Some examples of the extracted terms can be found in Simões & Guinovart (2009).

```
[R1] A B = B[CAT<-/nc/] A[CAT<-/(a_nc|adj)/];
[R2] A B = B[CAT<-/nc/] "de"|"do"|"da"|"dos"|"das" A[CAT<-/(a_nc|nc)/];
[R3] A "of"|"in"|"for" B = A[CAT<-/nc/] "de"|"do"|"da"|"dos"|"das" B[CAT<-/nc/];
[R4] A B C = C[CAT<-/nc/] A[CAT<-/(adj|a_nc)/] B[CAT<-/(adj|a_nc)/];
```

Figure 2: EN–PT bilingual syntactic patterns

pattern, ranking of candidates by virtue of some score of lexical association, and assessment of term specificity with respect to some kind of non-terminological corpus of the language, among others (Hong *et al.*, 2001).

With the first filtering method, term candidates beginning or ending with any of the words of a list of stop words are removed from the list. This method, however, does not apply to the results of *NATools* complemented with bilingual syntactic patterns, since term candidates obtained by *NATools* respect the defined morphologic restrictions.

Another well-known method for filtering the results of extraction consists of calculating the lexical association of candidates in the corpus using one of the possible scores to test the strength of this association. The extractor in *NATools* calculates the log-likelihood ratio score (Dunning, 1993). However, this score does not carry any significance as a discriminatory factor when assessing the outcome of our terminology extraction method, presumably because the quality of selection based on a probabilistic translation dictionary derived from the parallel corpus and filtered with patterns ensures a fairly high minimum cohesion between the components of the candidate terms (Simões & Guinovart, 2009).

Therefore, we decided to check the accuracy of the term extraction of *NATools* with bilingual syntactic patterns using a non-terminological corpus of exclusion as a filter. The exclusion corpus will determine the identification (and exclusion) of unlikely term candidates. Literary corpora, unlike corpora of news articles, for instance, usually contain very few terminological units. A literary corpus, as a corpus of exclusion for term extraction, represents a very safe filter. When using a literary corpus as a filter, there are more false candidates identified as such than correct candidates wrongly identified as false ones. We created lists of word n-grams from the exclusion corpora *BiVir* and *Compara*, and applied these lists as criteria for filtering and evaluation of *NATools*-based terminology extraction.

The evaluation results (table 3) point to a high precision of the *NATools*-based extraction algorithm. As shown in the first column of the table, the 12,689 translation equivalences (TE) identified in the *Unesco Corpus* using *NATools* with the EN-GL bilingual syntactic patterns depicted in figure 2 represent 7,250 candidate bilingual term pairs (term candidates or TC) (57% of TE) after eliminating repeated TE. When filtering that list of TC with the list of word bi- and trigrams from the *BiVir Corpus*, we obtain a list of 6,949 Galician terms from TC (corresponding to 96% of TC) which are not present in the exclusion corpus, and a complementary list of 301 Galician term candidates (only 4% of TC) identified as erroneous term candidates due to their presence in the exclusion corpus. Thus, these scores show a precision of 96% in the *NATools*-based term extraction from the *Unesco Corpus*.

As for the experiments with the *JRC-Acquis*, the 717,293 TE identified with the EN-PT bilingual syntactic patterns shown in figure 2 represent 72,952 TC (only 10.2% of TE) after eliminating repeated TE. Differences between the TE/TC ratio of the *Unesco Corpus* and and that of the *JRC-Acquis* (57% vs. 10.2%) lie in the lexical density (percentage of different words in a text) of the two corpora. When filtering that list of TC with the list of n-grams from the *Compara*, we get a list of 63,744 Portuguese terms from TC (corresponding to 87.4% of TC) which are not present in the exclusion corpus, and a complementary list of 6,949 Portuguese term candidates (12.6% of TC) identified as unlikely term candidates because of their presence in the exclusion corpus. Differences in the precision scores of term extraction between the *Unesco Corpus* and the *JRC-Acquis* (96% vs. 87.4%) lie in the different size of the corpora (and of the exclusion corpora) and also in their level of lexical density and terminological specificity.

Corpora	**Unesco**	**JRC-Acquis**
Language	GL	PT
Trans. Equiv.	12 689	717 293
Term Cand.	7 250 (57%)	72 952 (10.2%)
Excluded TC	301 (4%)	9 208 (12.6%)
Not-excl. TC	6 949 (96%)	63 744 (87.4%)

Table 3: Extraction results

Finally, regarding the terminological quality of the extracted terms, the comparison of the 6,949 Galician terms identified by this method and filtered by the *BiVir* literary corpus, by one side, with the gold standard list formed by the 129,269 unique terms Galician terms found in the *bUSCatermos* and the *Termoteca*, by the other side, shows that only the 7.5% of the terms (521 terms) selected in the corpus by our method are part of the gold standard list. In some cases, the reason of this mismatch lies in the lack of lemmatisation in extraction. For instance, the extractor identifies “alimentos naturais”, but the gold standard list contains the lemmatised version of the term, namely, “alimento natural.” But more frequently the reason lies in the obvious fact that no term listing contains all the terms in a language. So we have found in our results a lot of genuine terms like “acceso directo”, “acción cidadá”, “acción humanitaria”, “acordo de paz”, “aeroporto internacional”, “ministerio de defensa” or “abusos sexuais”, which are not included in the list of 129,269 terms of our gold standard.

4 Conclusions

Bilingual terminology extraction from parallel corpora based on probabilistic translation dictionaries and complemented with bilingual syntactic patterns shows high rates of accuracy. At the present stage of development of the term extractor included in the *NATools* package, any word which is not recognized by the morphological analyzer cannot be part of a term candidate and some feasible candidates may be ignored. To avoid this the easiest solution would consist of considering any non-recognized word as a noun (obviously, a decision with risks). As for the evaluation of term quality, we must point the difficulty both in acquiring an undisputed gold standard for a language, as in interpreting the evaluation results due to the fact that no term listing contains all the terms in a language.

References

Almeida J. J. & Pinto U. (1994). Jspell – um módulo para análise léxica genérica de linguagem natural. In *Actas do X Encontro da Associação Portuguesa de Linguística*, p. 1–15.

Atserias J., Casas B., Comelles E., González M., Padró L. & Padró M. (2006). FreeLing 1.3: syntactic and semantic services in an open-source NLP library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, p. 48–55.

Crespo A., Clemente X. M. G., Guinovart X. G. & López S. (2008). XML-based extraction of terminological information from corpora. In J. C. Ramalho, J. C. Lopes & S. Abreu, Eds., *XATA 2008 — 6^a Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas*, p. 28–39.

Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

Frankenberg-Garcia A. & Santos D. (2003). Introducing COMPARA, the Portuguese-English parallel translation corpus. In S. B. Federico Zanettin & D. Stewart, Eds., *Corpora in Translation Education*, p. 71–87. Manchester: St. Jerome Publishing.

Guinovart X. G. & Sacau E. (2004). Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo). In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, p. 1179–1182.

Hong M., Fissaha S. & Haller J. (2001). Hybrid filtering for extraction of term candidates from German technical texts. In *Proceedings of Terminologie et Intelligence Artificielle*.

Patry A. & Langlais P. (2005). Corpus-based terminology extraction. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*, p. 313–321.

Simões A. & Almeida J. J. (2008). Bilingual terminology extraction based on translation patterns. *Procesamiento del Lenguaje Natural*, **41**, 281–288.

Simões A. & Guinovart X. G. (2009). Terminology extraction from English-Portuguese and English-Galician parallel corpora based on probabilistic translation dictionaries and bilingual syntactic patterns. In A. Teixeira, M. S. Dias & D. Braga, Eds., *I Iberian SLTech 2009*, p. 13–16, Porto Salvo, Portugal.

Simões A. M. & Almeida J. J. (2003). NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, **31**, 217–224.

Steinberger R., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufiş D. & Varga D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation.*