IT: Moving Towards Real Multilingualism

Antoni Oliver and Cristina Borrell

Universitat Oberta de Catalunya, Spain

Abstract. The Linguamón-UOC Chair in Multilingualism of the Universitat Oberta de Catalunya has developed a project consisting of the automatic elaboration of linguistic resources, those including Catalan. For this, we have created an automatic extractor of terminology, which is freely distributed, multi-platform and adaptable to users' needs. One of its most important useful applications is the elaboration of glossaries, both monolingual and multilingual, based on a set of documents. The system automatically extracts the term candidates from the texts introduced by the user. Later, a specialist reviews the list of lexical units to verify the result. In the Chair in Multilingualism we have applied this tool to the expansion of the Eurovoc glossary with its Catalan version. We have extracted the equivalents for this language from the entries in Spanish and a bilingual Spanish – Catalan parallel corpus. From the application of the extractor, we obtained a multilingual glossary of 2 531 terms involving EU policies. In a second stage, we have used different techniques. Firstly, starting from the terms that were not yet translated by the first step, we have generated hypothesis of the Catalan translation of the Spanish Eurovoc terms, thus by the compositionality method. Secondly, we have applied statistically automatic translation, and finally we have checked the results in a monolingual corpus or by means of a restricted Web search. Our aim is to translate automatically as much terms as possible. With this project we try to demonstrate that having the suitable resources can remarkably reduce the translation costs.

1 Introduction

Multilingualism is one of the most visible aspects of the change society is living at present. A new challenge is growing, namely to cope with this change in any sphere of our lives. In this framework, the major aim of the Linguamón-UOC Chair in Multilingualism of the Universitat Oberta de Catalunya is promoting a concept of linguistic diversity that is sustainable, equitable and functional. For that reason, our team has developed several projects aiming to highlight the importance of linguistic richness and thus give a chance to all languages, even minor ones.

In this regard, the Linguamón-UOC Chair in Multilingualism has developed a set of open source tools aiming to automatically extract terminology. Our team has applied these tools to develop Catalan terminology resources related to the European Union. Currently, Catalan has an ambiguous status within the EU and very few documents are translated to this language. One of the reasons that are adduced not to translate them is the assumption of high costs and the consideration of Catalan as a minority language. With this project we try to demonstrate that having the suitable resources and using assisted and automatic translation tools, translation costs are

remarkably reduced. Moreover, with a modest budget, the right of Catalan citizens to have the documentation in their own language can be respected.

As we all know, the European Union produces official documents in 23 official languages[1]. The primary law documents must be written in all EU official languages, in order to respect the transparency, democracy and legitimacy that the Community promotes. Despite this, many Europeans encounter limitations when it comes to participating in the European institutions, as much of the multilingual and multicultural richness that exists within the borders of the EU is not taken into account.

In order to manage the volume of multilingual documents and assist both citizens and community workers, the European Union has set up a big amount of specialized language services[2]. For instance, the translation service in the European Commission is the most representative EU linguistic service, and one of the largest translation services in the world.

The European Commission's Directorate-General for Translation (DGT) has about 2,500 workers (thus apart from freelance translators) and produces about 1.3 million pages per year. The translation units are organized into departments, one for each EU language. In recent years, the translation process has been refined, especially with the incorporation of the technology offered by translation memories, so that translators can optimize their effort. Roughly, each translator uses to develop his or her work:

- Terminology (dictionaries, glossaries, terminological databases, etc.).
- Parallel and reference documents;
- Access to previously translated texts (especially through translation memories);
- Administration staff responsible for fulfilling the tasks of pre-edition and post-edition.

Another interesting aspect of the DGT is known as "field offices for multilingualism", which in some way, are responsible for promoting the richness of the linguistic diversity of Europe. The function of these branches is to adapt the Commission's communications to the languages of the region and act as a bridge between the local citizens and European institutions.

Thanks to the bilateral administrative arrangements between the EU institutions and the Spanish government[3], since June 13th 2005, Catalan citizens may apply to the institutions in their own language. They may do it in the written communications to the Council of the European Union and the European Commission, through an intermediate body designed by the state government in charge of translating the

[1] Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish.

[2] All information concerning the languages of Europe and the different language services, both internal and 'open', is available at the website of the EU "Languages and Europe".

[3] Council conclusions on the official use of other languages (2005 / C 148/01).

messages (from Catalan to Spanish), and send the translation to the institution concerned.

2 Tools

Eurovoc is a multilingual thesaurus covering the areas of activity of the European Union. Currently, it is available in Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish. It is the only resource that can be downloaded from its website.

Following this information, our team is able to develop a multilingual glossary with the tools described above. First of all, the team has selected the "available" languages for the project. English is the most popular language in the EU, followed by relatively far behind by French and German. The most logical thing then is to choose English (due to the amount of original documents available) and French, because a romance language will always be closer to Catalan, our target language, rather than German.

We will also take Spanish into consideration. This is one of the closest languages to Catalan, and above all, it is quite easy to find parallel documents in this linguistic combination (Spanish-Catalan). For instance, and even within the institutional field, the Official Journal of the Government of Catalonia[4] is published both in Catalan and Spanish.

All that linguistic potential can not be exploited, combined and adapted if we do not have any tool that helps us handle. That is why we have programmed and used an automatic extractor of terminology[5] that allows us to obtain the equivalent in Catalan from the Spanish Eurovoc terms. Basically, we have given the Spanish Eurovoc terms to the official parallel corpus in its Spanish version, thus receiving in return the official, Catalan version of the term.

Thanks to this collection of tools we can automatically select, manipulate and retrieve terminology, which is of great help for the translator. It is distributed under a free software license (GNU/GPL), so the user can freely download it, use it, distribute it or modify it following his needs. The tools are programmed in Perl, in a completely modular way, in order to provide, if necessary, its modification and adaptation. The package works for different platforms, for both Linux and Windows.

This tool allows users to:

- extract term candidates directly from the text;
- get help in finding the best translation for a word;
- automatically build glossaries, both monolingual and multilingual;
- create terminological databases for specific fields;
- automatically build lists of lexical units of a specific area.

If the input of the system is a monolingual corpus, the resulting output is a monolingual list of term candidates. This list must be reviewed manually by a

[4] DOGC `http://www.gencat.net/dogc/`

[5] Downloads and more information `http://ww.linguoc.cat`

linguist, who must confirm or reject the proposals automatically selected by the extractor. Otherwise, if the input consists of a set of documents in multiple languages, apart from the list we obtain a multilingual terminological resource, because the system automatically detects any possible equivalent translation of a candidate.

This tool kit is essentially a linguistic terminology extractor and a device for automatic translation equivalents search, programmed essentially with statistical methods.

1. The automatic terminological extractor takes a document, a set of documents or a parallel corpus and stop−words lists (lists of functional words) of the languages involved. The statistical method identifies term candidates based on word frequency in a corpus of expertise. The process, broadly speaking, is this: the system calculates statistically n-grams (sequence of n elements, usually from n = 2 to an n determined by the user, i.e. all combinations of two words, three words etc.). Logically, some of these combinations will not be relevant, that is why it is essential to filter the results with the list of empty words. The output of this module will already be the list of term candidates.
2. In the second case, the system picks a term candidate in a language A and selects a subset of the parallel corpus with all segments containing this term. If, then, the system makes a statistical terminological extraction in these segments, the result should be the term candidate in the language B. It is expected that the terminological unit more common in the segment in the language B would be the term that we searched in language A.

3 Procedure

By using the Eurovoc entries in Spanish (6.797 items) and the official, bilingual Spanish – Catalan parallel corpus we have extracted the equivalents in this language. The input corpus selected, the Official Journal of the Government of Catalonia (Diari Oficial de la Generalitat de Catalunya, from now on DOGC), is published daily from Monday to Friday (except holidays) in two equivalent editions, one in Catalan and another in Spanish. Basically, the laws, rules and regulations, general dispositions, agreements, resolutions, edicts, notices, ads and other acts of Government and its Administration are published in it, both in Catalan and Spanish.

Thanks to it, we have a parallel Catalan-Spanish corpus of institutional, political and legal areas, that it is why it is one of the Catalan official texts which shares more characteristics with the EU institutional documents. At the same time, as it offers a Spanish version, it optimizes the work of the extractor: it simply searches the Spanish Eurovoc terms in the Spanish version of the DOGC and returns the Catalan translation for it.

Although the statistical method resolves successfully the search in many cases, it is not always right. So, the program offers more than one candidate, leaving the user the final choice. The first result is always the one that has appeared more frequently.

Here is where the work really starts for the linguist. Now he or she has to manually review all the proposals made by the extractor, correct them when they are incorrect and even suggest translations if the extractor has been unable to offer any. This will happen in the cases where the Spanish word we are searching for does not appear in the DOGC. Either way, thanks to the extractor, the linguist saves time and research.

If the extractor did not act satisfactorily, the expert reviewing the outcome will have to deal with the types of problems mentioned here[6]:

1. Often, the first proposal for the Catalan translation is not complete. This happens especially in cases of terminological units consisting of more than one word, which somehow mislead statistical systems.
2. Sometimes, Catalan and Spanish versions are not equivalent. This happens not often, but it can be that the extractor mistakes and confuses the selection of terms that do not belong to the same field of expertise.
3. Sometimes the extractor proposes n-grams belonging to the same segment as the key term, but that do not correspond to the possible translation.
4. It can be that versions in other languages other than Spanish do not correspond to that of Catalan. It is recalled that the extractor takes the word in Spanish to make the choice of Catalan, so it never takes into account other languages (in this case, English and French)[7].
5. It can also happen that the version in Catalan is the development of an abbreviation or vice versa. The expert will have to take that into account and harmonize criteria.

In a second stage, we will translate the terms use different techniques. Firstly, starting from the terms that are not yet translated, we will generate hypothesis of the Catalan translation of the Spanish Eurovoc terms, thus by the compositionality method. Secondly, we will apply statistically automatic translation, and finally we will check the results in a monolingual corpus or by means of a restricted Web search. Our aim is to translate automatically as much terms as possible.

4 Results

To perform the experiments we used a parallel corpus Spanish-Catalan (the DOGC) with all issues from 3.544 (dated 2/01/2002) to 5.118 (dated 24/04/2008). The corpus has been segmented by sentences and automatically aligned. It consists of a total of 9 492 333 segments and 121 145 821 words in both languages.

[6] These errors were found after the implementation of the first version of the extractor, reason why most of which disappear due to the improvements in later versions. This analysis helped us discover what changes were important to improve the performance of the extractor.

[7] Ultimately, this would depend on the quality of the Eurovoc equivalents, which is outside the scope of this study.

The results presented were first obtained for an increment value of n ± 1 and the search algorithm of translation equivalents was not optimized. Thus the first candidate to appear was the one that showed the greatest frequency of occurrence in the parallel subcorpus of all the segments containing the original term. The table shows the percentage of equivalents appearing in first position, in the first three and in the top ten positions.

Position	**Percentage**
1	43,4
1 – 3	73,4
1 – 10	86,2

Table 1. Results obtained with the algorithm without optimization.

Analysing the results we reached the following conclusions:

- When candidates get into the top positions with the same frequency, the candidate who comes in first place can be any of them. This is because hashes are used to store the candidates, and these data structures do not have a certain order.
- Often the first candidate has a higher frequency than the original word. This means that mono−word terms will appear in the first place even if the original term was multi−word.

To minimize the consequences of the first phenomenon we programmed an optimization of the translations' equivalent search algorithm which consisted in gathering all the proposals that had the highest frequency and apply the algorithm on the edit distance (the closer the candidate is to the original, the more likely to be the correct one it is), the results improved as follows:

Position	**Percentage**
1	62,2
1 - 3	82,1
1 - 10	88,2

Table 2. Results obtained grouping the more frequent candidates and using the edit distance.

To avoid the problem produced by candidates with a fake higher frequency we have applied an optimization similar to the one above, but that including all three candidates with the highest frequencies. Thus we obtain the following results:

Position	Percentage
1	74,7
1 - 3	80,4
1 - 10	88,2

Table 3. Results obtained by pooling the three candidates with the higher frequencies and using the edit distance.

As we can see, the results have improved markedly, especially in the top positions.

Results of the second stage will appear during the second half of the year.

5 Conclusions and future work

In the development of the project we have obtained a multilingual glossary of 2 364 terms used in EU official documents, available in four languages, among which Catalan. For the moment, above all the official language resources it has only been possible to process the Eurovoc, which is the only tool that offers the possibility to be directly downloaded.

During the development we could improve the algorithm three times, in order to obtain better results. The final result is that the extractor has a 75% of success in the very first position. The success percentage raises to 88.2% when we look at the first ten positions: that means that within the first ten choices the translator can find the correct translation candidate. From the application of the extractor, we obtained a multilingual glossary of 2 364 terms involving EU policies.

In a second stage, we have used different techniques. Firstly, starting from the terms that are not yet translated (6 797 – 2 364 = 4 433 terms left), we have generated hypothesis of the Catalan translation of the Spanish Eurovoc terms, thus by the compositionality method. Secondly, we have applied statistically automatic translation, and finally we have checked the results in a monolingual corpus or by means of a restricted Web search. Our aim is to translate automatically as many terms as possible.

In sum, we have shown that some basic and simple tools can help us create multilingual resources that can expedite the process of translating documents. With the right tools, the process of generation of these resources is very flexible and, therefore, economic. In addition, all tools used are freely distributed and therefore available for the whole community[8].

This methodology can be applied to a large number of languages. The only requirements are namely to have a parallel corpus of the desired areas of expertise and a short list of empty words of the pairs of languages involved. One should keep in mind that the statistical methodology for the extraction of terminology does not work well for agglutinative languages.

[8] Download and more information about the tools: http://multilingualismchair.uoc.edu.

Future work is divided into two lines:

- On the one hand, to apply the methodology described to other European terminology resources, such as IATE.
- On the other hand, to continue working on improving and optimizing the extraction tools.

References

[1] European Comission. (2008). Un reto provechoso. *Cómo la multiplicidad de lenguas podría contribuír a la consolidación de Europa* [on line]. Brussels. Internet: http://ec.europa.eu/education/languages/archive/doc/maalouf/report_es.pdf

[2] Oliver, A., Vazquez, M. & More, J. (2007). Linguoc LexTerm: una herramienta de extracción automática de terminología gratuita (Tomo 11) [on line]. *Poughkeepsie: Translation Journal*, number 4. ISSN 1536-7207. Internet: http://accurapid.com/journal/42linguoc.htm

[3] European Comission. (2004). Translation and drafting aids in the European Union languages [on line]. Brussels. Internet: http://ec.europa.eu/translation/index_en.htm

[4] Diari Oficial de la Generalitat de Catalunya (DOGC) [on line]. Barcelona: Generalitat de Catalunya, 2008. Internet: http://www.gencat.net/dogc/

[5] Eurovoc. Multilingual Thesaurus [on line]. Brussels: European Union, 2005. Internet: http://europa.eu/eurovoc/