A Corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a Minority Language

Fco. Mario Barcala Rodríguez*, Eva Domínguez Noya*, Pablo Gamallo Otero†, Marisol López Martínez†, Eduardo Miguel Moscoso Mato†, Guillermo Rojo†, María Paula Santalla del Río†, Susana Sotelo Docío†

*Centro Ramón Piñeiro para a Investigación en Humanidades
Estrada Santiago-Noia, Km. 1 - A Barcia
E-15896, Santiago de Compostela, A Coruña, Spain
{fbarcala, edomin}@cirp.es
†Universidade de Santiago de Compostela
Facultade de Filoloxía, Burgo das Nacions, s/n
E-15782, Santiago de Compostela, A Coruña, Spain
{pablogam, fgmarsol, fgmato, guillermo.rojo, fempsr, fesdocio}@usc.es

Abstract

In this paper, we describe the compilation and structure of two linguistic resources, a corpus and a dictionary of terms in the field of economy, developed for Galician. In addition to this, we describe the use of these resources for the automatic extraction of multi-word terms by means of a combination of linguistic and statistical techniques. While doing this, special attention will be paid to the problems posed by minority languages such as Galician for the achievement of these tasks.

1. Introduction

The work described in this paper is part of a general project, RICOTERM-2[1], aimed at the development of a multi-lingual system for the re-formulation of queries posed by users of Internet interested in the search of information about some specialized communicative field, in our case, economy. The system is currently being developed for English, Spanish, Catalan, Basque and Galician. Its general design can be found in (Lorente, 2005). However, for the purposes of this document it is enough to point out that, to improve the result of the information retrieval task involved, the system will make use of techniques of query expansion in which a combination of methods of both term-only and full-text expansion are employed. For term-only expansion, the project plans to make use of a domain specific ontology. For full-text expansion, we try to prove the benefits of using a domain-specific corpus, structurally and linguistically annotated, in order to detect, by means of various integrated tools such as a terminology extractor, collocations or recurrent contextual phraseology of the terms included in the query or obtained after consulting the ontology.

A specific part of this general project, GARI-COTER[2], is mainly devoted to the development of the resources needed by such a system for one of the languages involved: Galician. In Sections 2. and 3. of this article we briefly describe the current stage of these resources: a galician corpus in the field of economy and a lexical collection of terms compiled from previously existing resources.

Besides this, and in line with the general approach underlying the RICOTERM project, we describe here the exploitation of the resources themselves to improve them in what can be seen as a bootstrapping process. Specifically, in Section 4. we describe the use of the corpus and the lexical resources to automatically extract multi-word terms of the field of economy in Galician.

To do this, we make use of a method that combines specific linguistic and statistical techniques in a way that can be compared to the widely-used approaches in the research community to deal with the task of terminology extraction.

Finally, as regards the development of resources, as well as for their application in terminology extraction strategies, we found that the situation of minority languages such as Galician constitutes a non-negligible difficulty. All along this document we have wanted to highlight this fact, very frequently not taken into account when designing terminology extraction techniques and, more generally, information retrieval systems.

2. The corpus

The first problem to be solved when trying to do automatic terminological extraction in a minority language like Galician is to get domain-specific documents. As already mentioned above, our research focus is in the field of economy.

The task of the development of a domain-specific corpus was divided in the following way: the development of a more general one, containing economy journal news, and of a specific one, with specialized texts of economy. This decision was taken on the basis of two reasons: the first type of corpus was much easier to obtain, but the second one was expected to be much richer from the point of view of terminology.

On the one hand, we had no problems to obtain documents for the first corpus, given that, thanks to a special agreement with the Center for Humanities Research Ramón Piñeiro[3], we could include in our corpus news collected in the CORGA (Reference Corpus of

Present-day Galician Language) corpus[4]. These news were already available both in electronic format and very carefully XML (eXtensible Markup Language)[5] structured. On the other hand, nevertheless, we had great difficulties to obtain documents for the collection of the specialized corpus, given that there were indeed very few economy specialized texts in Galician. In electronic format, only several texts, whose appropriatenes can in certain cases be arguable, could be found. We also had to encode them according to the above-mentioned XML structure.

As a result of this work, we could compile a general corpus constituted by 609 newspaper news which include 206510 words in 7892 sencences, and a specific one constituted by 14 books and 2 specialized journals which include 801702 words in 34588 sentences.

Apart from being collected, every document in the specialized corpus (each book or article from a specialized journal) has been classified by an expert according to two different taxonomies of the field. As a result of this classification, we can at least ensure that, with respect to the documents taken from the specialized journals, the corpus is reasonably representative of the field. The same, however, cannot be ensured for the book texts, for reasons that, when dealing with minority languages as Galician, are obvious: as very few texts of this type are available, only in extremely particular circumstances, one can decide not to include an available electronic text in a specialized corpus of a minority language.

2.1. Corpus encoding

As we have already pointed out, documents are structured according to the XML standard. Each document has a header which includes bibliographical details, as well as the argument or arguments of the document, this being followed by the text of the document itself, structured up to the sentence level. For example the XML structure of a single news item is:

```
... preambles of XML standard ...
<noticia> (single news item)
  <cabeceira_noticia> (header)
    <nome_publicación>
      name of the publication
    </nome_publicación>
    <editorial>publisher</editorial>
... more bibliographic information ...
    <identificador>
      single news item identifier
    </identificador>
    <autor>author</autor>
    </área_temática>
      argument
    </área_temática>
  </cabeceira_noticia>
  <contido_noticia> (content)
    <titular> (title)
      <parágrafo> (paragraph)
        <oración>sentence</oración>
      </parágrafo>
    </titular>
    <resumen> (summary)
      <parágrafo>
        <oración>sentence</oración>
... more sentences ...
      </parágrafo>
    </resumen>
    <corpo> (content)
      <parágrafo>
        <oración>sentence</oración>
... more sentences ...
      </parágrafo>
... more paragraphs ...
  </contido_noticia>
</noticia>
```

2.2. Corpus annotation

In order to use morphosyntactic information to perform automatic terminological extraction in the way we describe below, Section 4., the corpus was annotated with *POS* (Part-of-Speech) information. The tagset used is based on the one developed by the XIADA (Tagger/Lemmatizer of Present-day Galician Language) project[6]. It consists of approximately 370 tags and is designed according to the guidelines EAGLES (Expert Advisory Group on Language Engineering Standards (EAGLES), 1996).

In the first step, this tagset identifies the morphological category, and in the second one, it identifies the grammatical attributes considered relevant for the corresponding category. In the development of this tagset, the completeness of morphological descriptions was given preference over the introduction of any syntactic information in its widest sense. The latter was, in fact, reduced to the specification for only certain elements of certain categories of their functional capabilities in terms of nucleus and modifiers.

To annotate the general corpus, we have made use of the Galician default trained tagger developed by the XIADA project (Barcala et al., 2006; Graña and M. A. Alonso, 2002; Graña et al., 2002). As this tagger can manage XML information, the result was a set of documents encoded in an intermediate XML format which integrates *POS* information.

After the automatic annotation of the corpus, we performed a manual revision of its results. To do this, we have used a simple generic XML editor (XMLMind Editor[7]) adapted with Cascade Stylesheets[8]. In this stage, we took a great advantage of the tagger's intermediate XML format, which allowed us to do this task much less cumbersome.

Once the manual check of the general corpus was accomplished, the tagger was first trained again with the data of the general corpus, and then used to tag the specific one. The result of this second automatic annotation process was not manually revised.

Finally, the tagger's XML intermediate format is automatically simplified. The final format is similar to the one previosly shown, but includes *POS* information within the sentence structure:

```
...
```

```
<oración> (sentence)
  <expresión>
    full text of the sentence
  </expresión>
  <análise> (analysis)
    <análise_unidade> (analysis unit)
      <unidade>
        lexical unit to be analysed
      </unidade>
      <constituínte> (constituent)
        <forma>word</forma>
        <etiqueta>POS tag</etiqueta>
        <lema>lemma</lema>
      </constituínte>
... more constiuents if necessary ...
    </análise_unidade>
... more analysis units
  </análise>
</oración>
...
```

Let's notice especially the presence of *constituents* in the format. Although in the great majority of cases lexical units have only one constituent, this element is needed, and mainly used, to handle verb forms with enclitic pronouns, which may, in fact, have a very complex compound structure in Galician. By using constituents, however, those compounds can be efficiently accounted for, on the basis of their segmentation into a verb part and as many additional parts as enclitic pronouns attached to the verb, each one, as the verb part, analyzed separately. This phenomenon is correctly managed by the tagger, so we could get rid of it (Barcala et al., 2006) before further processing of the corpus for terminology extraction itself, see Section 4.

3. Lexical resources

One of the needs, and a goal on itself, for terminology extraction as described below is the compilation of a database of terms in the field of economy[9]. Two techniques were used to obtain this database of terms: the automatic extraction from the domain corpus, as described in Section 4., and the manual compilation of terms from a wide range of sources which include electronic glossaries and dictionaries. In this section we are going to describe the lexical resources developed using the second technique, as well as the sources from which they could be obtained. Although we will not go into detail with respect to this for each of the sources examined, we want to remark here that Galician is a language which has recently undergone –it still undergoes– a process of normalisation, which means that in the collection of terms from different sources we had to handle the different forms in which same words can be transcribed.

The sources[10] considered were quite heterogeneous, as can be deduced from Table 1: two dictionaries (*Eiras* and *Formoso*, one of them trilingual), two electronic glossaries freely available from the web, and the section of economy of the terminological database built by the Linguistic Normalization Service of the University of Santiago de Compostela (a very large terminological database which tries to cover the terminology of several scientific fields).

The last one is the most reliable and accurate, since it was carefully collected from 26 different sources and includes very rich and varied information, such as the equivalence of terms in other languages, information about semantic relations such as synonymy or hiperonymy, and definitions. Dictionaries also must be considered good and reliable sources: they include definitions and translations, as well as a not too exhaustive information about synonyms and antonyms.

Not only with respect to quality (volume of information for each term), but also to quantity (number of terms supplied), these three resources are more important than the others: in addition, in effect, to the fact that more terms are indeed gathered in them, the percentage of unique terms in these resources is also higher (see Table 1).

The internet glossaries, then, must be considered minor resources, both in number of terms and with respect to the information included for each term.

Each source was encoded using XML and a common structure defined by a DTD, the one that is used for the Gari-Coter term database.

Source	**Terms**	**Type**	**Unique lemmas**
Eiras	3232	dictionary	2291 (70,88%)
SNL	2894	terminological DB	1746 (60,33%)
Formoso	1346	multiling. dictionary	839 (62,33%)
Panlatino	273	multiling. glossary	20 (7,32%)
galego.org	153	glossary	46 (30%)
Gari-Coterm	6046		

Table 1: Sources of the lexical resources

3.1. Dictionary encoding

The Gari-Coter list of terms was encoded according to the XML standard as a result of merging the different sources described above. Each term is enclosed within the tag `<term>`, and includes exhaustive information about lemma, part-of-speech and definition, and in most cases it includes also the equivalence in other languages, as well as some semantic information about synonyms or hyperonyms.

In future work, we plan to convert this XML-based resource into a relational database with a web interface. This will quite easily allow us to generate subsets of the list in accordance with specific restrictions, something which we expect that will be very useful to perform sub-domain terminology extraction.

4. Terminology extraction

Terms are seen here as useful indexing units in IR applications. So, they must be good from a semantic point of view, that is, they must capture as much as possible the meaning of a domain-specific corpus. Moreover, it has been recognized that single words are not always useful for the terminological representation of domain-specific texts.

terms	similar multi-words	Dice
forza de traballo (*labour force*)	man de obra (*labour force*)	0.15
	medio de produción (*production means*)	0.08
gasto público (*public spending*)	diñeiro en circulación (*money supply*)	0.12
	déficit comercial (*trade deficit*)	0.10
tecido industrial (*business network*)	Baixa Idade Media* (*Late Middle Ages**)	0.12
	explotación agraria (*land cultivation*)	0.11
taxa de crecemento (*growth rate*)	ritmo de crecemento (*rhythms of growth*)	0.11
	maior crecemento* (*bigger growth*)	0.11
	taxa de paro (*rate of unemployment*)	0.11
enerxía renovable (*renewable energy*)	enerxía solar (*solar energy*)	0.13

Table 2: 5 terms and their similar multi-words

For this purpose, multi-word expressions seem to be more appropriate. In this section, we describe an approach to automatically extract multi-word terms.

Our strategy consists of two steps. First, a list of seed terms is semi-automatically selected from the annotated corpus making use of available glossaries and resources. Then, we use that list as a set of positive examples to identify multi-word units with similar contextual distribution in the corpus. Similar multi-word units will be considered as new term candidates.

4.1. Term samples

The first objective is to build a starting list of positive term examples. For this purpose, we follow a very basic strategy. First, some morpho-syntactic patterns are used as endogenous constraints to select a generic list of multi-word units from the annotated corpus. Four nominal patterns are used:

$noun - adj$
$adj - noun$
$noun - noun$
$noun - prep - noun$

Then, a statistical filter is applied to identify those multi-word units in the generic list with a high degree of cohesion. The glue measure employed in the filtering process is *SCP*, defined in (Silva et al., 1999). Finally, the filtered list is revised by hand using as gold standard the available terminological resources described above, in Section 3.

4.2. Corpus-based similarity

The second objective is to learn new candidate terms by making use of both the annotated corpus and the list of positive examples selected in the previous step. For this purpose, we follow a method based on exogenous (i.e. contextual) information (Basili et al., 2001; Maynard and Ananiadou, 1999; Cimiano and Völker, 2005). The assumption the method is based on is the following: a multi-word unit that appears in the same local contexts as a given multi-word term should also be considered as a term. So, we implemented an algorithm calculating the similarity between terms and multi-word units on the basis of contextual features extracted from the corpus. The multi-word units compared to the list of term samples are all those selected using the 4 nominal patterns described above.

Lexico-syntactic contexts of multi-word units are extracted from the POS tagged corpus using pattern matching techniques (articles and pronouns are previously removed). For instance, given the expressions:

"loss of *labour force*"
"*labour force* of a country"

containing the compound noun "labour force", two contexts are extracted:

< loss of [NOUN] >
< [NOUN] of country >

where NOUN stands for the head category of the multi-word unit. To extract lexico-syntactic contexts, we follow the notion of *co-requirements* introduced in (Gamallo et al., 2005). According to this notion, two words (*head* and *dependent* words) related by a syntactic dependency are mutually constrained. They impose linguistic requirements on each other. A pre-fixed "Predicate-Argument" organization does not exist. The head imposes syntactic and semantic constraints on the words that fill the dependent position, as well as the dependent word imposes specific restrictions on the kind of head it depends on. Experimental tests showed that co-requirements permits a finer-grained characterization of "meaningful" syntactic contexts.

Once lexico-syntactic contexts have been extracted, they are associated to their co-occurring multi-word units in order to build a collocation database. Each multi-word unit (term or not) is defined as a vector where each lexico-syntactic context corresponds to a feature. Before starting to compute similarity between vectors, sparse contexts are filtered out. A context is sparse if it has high word dispersion. Dispersion is defined as the number of different multi-word units occurring with a lexico-syntactic context divided by the total number of different multi-word units in the training corpus. So, the vector space is only constituted by those lexico-syntactic contexts whose multi-word unit dispersion is lower than an empirically set threshold.

Each multi-word term of the starting list is compared to the rest of multi-word units in the corpus using Dice coefficient as similarity measure. Similarity between a

	Accuracy
Test list 1	.74
Test list 2	.70
Test size	160

Table 3: Evaluation of candidate terms

multi-word term, t, and a multi-word unit, mw, which is not in the starting list of term samples, is computed as follows:

$$Dice(t, mw) = \frac{2 * \sum_i min(f(t, c_i), f(mw, c_i))}{f(t) + f(mw)}$$

where $f(t, c_i)$ represents the number of times t co-occurs with the context c_i. Likewise, $f(mw, c_i)$ represents the number of times mw co-occurs with the context c_i. For each term, we select the k most similar multi-word units (where $k = 5$) with a Dice score $>= 0.05$. Table 2 shows the most similar multi-word units associated to five terms of the starting list. Similar multi-word units are considered to be candidate terms. Those extracted multi-word units with asterisk are odd terms.

4.3. Experiments and evaluation

Experiments have been carried out over the annotated corpus described in Section 2. The starting glossary of terms contains 150 entries, while the final list of candidate terms we have extracted contains 740 multi-word units. To evaluate the accuracy of the system, we randomly selected 2 test lists of 160 multi-word units from the final list. A human evaluator decided if they are correct or incorrect terms. Table 3 depicts the accuracy scores, where *accuracy* is defined as the number of correct terms divided by the total number of test words.

The main problem of our strategy is that co-occurrences of multi-word units are still more sparse than those of simple words. Indeed, corpus-based algorithms to extract any information on multi-word units (for instance, information on *termhood*) require larger domain-specific corpus. This is a challenge for minority languages.

Notes

[1] *Terminological and Discursive Control for Information Retrieval in Specialized Communicative Fields, by means of Specific Linguistic Resources and a Re-Elaborator of Queries*, financed between 2004 and 2007 by the Ministry of Science and Technology of the Spanish Government.

[2] *Development and Multilingual Integration of Linguistic Resources in Galician for Information Retrieval by means of Strategies of Terminological and Discursive Control in Specialized Communicative Fields*, financed between 2004 and 2007 by the Ministry of Science and Technology of the Spanish Government.

[3] http://www.cirp.es. [Consulted: june, 2, 2007].

[4] http://corpus.cirp.es/corgaxml. [Consulted: june, 2, 2007].

[5] http://www.w3.org/XML/

[6] http://corpus.cirp.es/xiada, 0.1.0 version. [Consulted: june, 2, 2007].

[7] http://www.xmlmind.com

[8] http://www.w3.org/Style/CSS/

[9] In the course of the Gari-Coter project, this database is going to be integrated in an ontology of the field of economy.

[10] **Eiras**: Eiras Rey, A.: *Dicionario de economía*, to be published.

Formoso: Formoso Gosende, V. (coord.) (1997): *Diccionario de termos económicos e empresariais galego-castelán-inglés*. Santiago de Compostela: Confederación de Empresarios de Galicia.

Panlatin Electronic Commerce Glossary: http://fon.gs/panlatino

Glossary about commerce from galego.org: http://galego.org/vocabularios/ccomercial.html

SNL: http://www.usc.es/en/servizos/portadas/snl.jsp

5. References

Barcala, F. M., M. A. Molinero, and E. Domínguez, 2006. XML rules for enclitic segmentation. In A. Quesada-Arencibia, J. C. Rodríguez-Rodríguez, R. Moreno-Díaz (jr.), and R. Moreno-Díaz (eds.), *Computer Aided Systems Theory (Extendeds Abstracts)*. Las Palmas de Gran Canaria.

Basili, R., M. Pazienza, and F. M. Zanzotto, 2001. Modelling syntactic context in automatic term extraction. In *3th Conference on Recent Advances in Natural Language Processing, RANLP2001*.

Cimiano, P. and J. Völker, 2005. Text2Onto - A framework for ontology learning and data-driven change discovery. In *10th International Conference on Applications of Natural Language to Information Systems (NLDB'2005)*.

Expert Advisory Group on Language Engineering Standards (EAGLES), 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages. EAGLES Document EAG-CLWG-MORPHSYN/R. Technical report.

Gamallo, P., A. Agustini, and G. Lopes, 2005. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–146.

Graña, J., F. M. Barcala, and J. Vilares, 2002. Formal methods of tokenization for part-of-speech tagging. In *Computational Linguistics and Intelligent Text Processing*. LNAI, Springer-Verlag, pages 240–249.

Graña, J. and M. Vilares M. A. Alonso, 2002. A common solution for tokenization and part-of-speech tagging: One-pass Viterbi algorithm vs. iterative approaches. In *Text, Speech and Dialogue*. LNAI, Springer-Verlag, pages 3–10.

Lorente, M., 2005. Ontología sobre economía y recuperación de información [on line]. *Hipertext.net*, (3). http://www.hipertext.net. [Consulted: january, 30, 2007].

Maynard, D. and S. Ananiadou, 1999. Identifying contextual information for term extraction. In *5th International Congress on Terminology and Knowledge Engineering (TKE'90)*.

Silva, J. F., G. Dias, S. Guilloré, and G. P. Lopes, 1999. Using LocalMaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Progress in Artificial Intelligence*. LNAI, Springer-Verlag, pages 113–132.