Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação

Building Domain Specific *Corpora* in Portuguese Language

Lucelene Lopes, Renata Vieira

Relatório Técnico N^0 062

Porto Alegre, Dezembro de 2010

Abstract

This report presents the effort to build five domain specific *corpora* to glossary construction, information extraction and ontology construction. The whole building effort is detailed by the explanation on how the texts were chosen, validated, converted to a common format and, particularly, how they have been subject to a careful refinement in order to keep only relevant and well-formed sentences. The builded *corpora* are described by its numerical characteristics and practical applications are suggested.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Many works in Natural Language Processing (NLP) area are based on domain specific *corpus*, *i.e.*, a set of texts of a given domain considered to be representative of this domain. Formally, *corpora* (the plural of *corpus*) are sets of linguistic data belonging to the oral or written use of a language duly systematized according to criteria wide and deep enough to be considered representative of the linguistic use [15].

The use of *corpora* in the scientific process becomes relevant because of its impartiality and reliable indication of frequencies of the forms, since it represents the language reality without any theoretical preconceptions [14]. Despite being long and laborious, the *corpus* building process is justifiable, since once builded it can be used to different applications, *e.g.*, automatic term extraction. Once the relevant terms are extracted, a glossary construction or even an ontology learning process can be started.

The work of [11] presents the automatic extraction of relevant terms from a Pediatrics *corpus* builded by Coulthard [4]. This work describes with a relative success that the automatic extraction of terms is comparable to a list of terms manually generated by a group of Linguistic and Pediatric specialists over a couple of years [17].

Such example of application demonstrates that the *corpus* availability allow a considerable save in time and expensive specialist resources. In fact, the availability of *corpora* from different scientific domains represents an important asset in order to identify the relevant terms with a considerable lower cost than the use of specialist of the domain, and probably with more reliable results, since it will furnish the terms that are actually used in the area avoiding possible preconceptions from the specialists themselves.

The use of domain specific *corpora* is quite common in many languages. One example is the work of [3] which uses a Law *corpus* in French language to extract noun phrases in order to build an ontology.

Another examples of *corpus*-based term extraction are the works of [7] which presents an ontology extraction from a German language *corpus* composed by texts from the intranet of an insurance company, and [8] which describes the creation of an huge 55 million words bilingual (Irish and English) *corpus*.

The construction of automatic tools to extract information from *corpora* also is very popular as the works of [6] which describes a term extractor for Dutch language *corpora*, and [13] which presents an automatic tool for extraction of terms applied to *corpora* in English and Chinese languages.

This report presents the building process of five original domain specific *corpora* in Portuguese (brazilian, actually) language. Additionally, the result, *i.e.*, the *corpora* main characteristics are presented, and some initial application are described. The domain of the five *corpora* are:

- Petroleum Geology (PG);
- General Geology (GG);
- Databases and Datamining (DD);
- Stochastic Modeling (SM);
- Parallel Processing (PP).

Besides the importance of the choice of which texts to include, the most important issue in the *corpus* building process proposed in this report is the careful refinement of the texts in order to produce a reliable language resource. It is vital to keep in mind that the generated *corpus* will probably be the input of subsequent software tools, and at least some of them may perform a linguistic annotation. Therefore, it is quite interesting to generate texts with well-formed sentences, *i.e.*, texts as free as possible from pitfalls to the next NLP tools to be employed.

For this reason, in this report the previously existing *corpus* on Pediatrics [4] is submitted to the same careful refinements as the other five original *corpora*. To illustrate the benefit of such procedure, some traditional metrics (precision, recall and f-measure) are taken from a term extraction procedure applied to the *corpus* as proposed by Coulthard and after the application

of the refinements. It is shown that the refinements improve the quality of the *corpus* since the precision of the extracted terms clearly increases as the texts are subject to the refinements.

This report is organized as follows. Chapter 2 describes the steps in the construction of the five original *corpora*. Chapter 3 presents the application of the refinements to the previously existing *corpus* on Pediatrics and the gains achieved in term extraction according to numerical metrics. Chapter 4 presents the main characteristics of the five original *corpora* and some results obtained with automatic term extraction. Finally, the conclusion suggest some future work to be developed using the present five *corpora* and summarizes the contribution of the presented *corpus* building technique.

Chapter 2

The building process

In order to build the five original *corpora* a four step process was made (Figure 2.1). Although quite intuitive, this process was organized to minimize the involvement of specialists in the domain, even though it represents an increase of the work of less specialized people, *i.e.*, in our experiments, computer science and linguistic students.

The first step consists in collect a considerable number of scientific texts from the Internet in various electronic formats, *e.g.*, `.pdf`, `.ps`, `.doc`, `.tex`, *etc.*. This step is done by non-specialized students searching public databases of thesis, dissertations, technical reports and conference and journal papers with keywords or titles or abstracts containing the words of the domain name. Specifically, for the construction of the five *corpus* the Brazilian Digital Library of Thesis and Dissertations (BDTD) [1] and the Google Scholar [5] were the basic sources for the search performed.

The second step was the only step in which a domain specialist was involved. In this step a very shallow analysis of the texts was made only to select which texts were actually relevant to the domain. It is important to notice that the specialist was not required to fully read the texts, but only to consider according to its experience if the text could indeed be considered as belonging to the domain or not. Nevertheless, a considerable number of texts were discarded during the selection step (see Table 2.1 at the end of this chapter).

The third step consists in transform the electronic format of the selected texts into a common textual format using extended `ascii` representation[1].

[1]The use of an extended `ascii` was necessary, since Portuguese texts always have non-

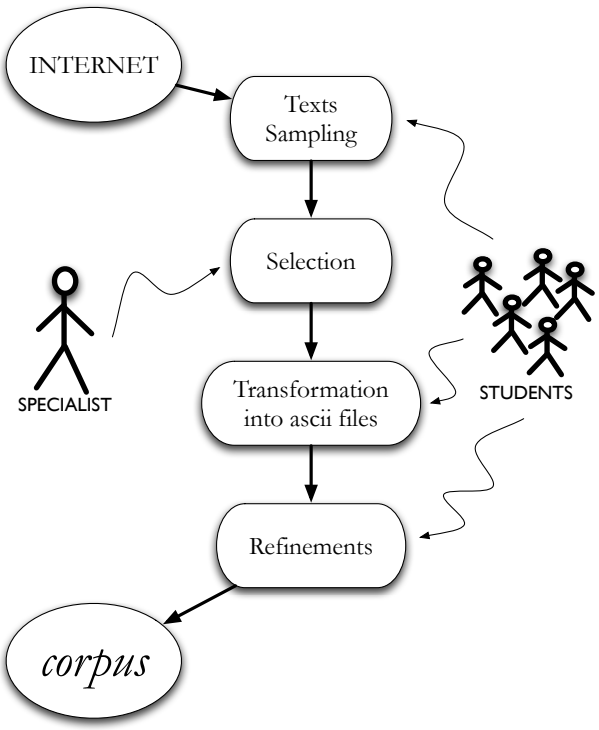

Figure 2.1: Four steps building process.

For most of the selected texts, an automatic converter called Entrelinhas [16] was used, but some already textual formats like `.tex` (LaTeX files) were only transformed by the exclusion of specific LaTeX commands.

Probably the most important and more laborious step of the process was the fourth step in which the texts were subject to a semi-automatic application of a set of refinements in order to keep in the text only valid and coherent Portuguese language sentences. In this step titles, keywords, abstracts in other languages, figures, tables, captions and acknowledgments were removed, not by having low relevance, but by not being valid sentences. The future *corpus* application was imagined as a linguistic procedure, so only valid sentences can be correctly recognized. This removal is very laborious since its use is far from uniform in the scientific texts. However, in order to reduce the manual effort, some automatic refinements were applied through

standard `ascii` characters for the accentuated letters (á, é, í, ó, ú, â, ê, ô, ü, à, ã, õ, ç) in their lower and upper case versions.

the use simple scripts and regular expressions search and replace option of Notepad++ word processor [12].

One important point in this methodology is that the more tedious steps were performed by non-specialized students and only the Selection step was performed by a specialist. Applying the process to the building of the five original *corpora*, a considerable amount of information was processed. Table 2.1 summarizes the number of texts considered in the Text Sampling step (Before Selection) and the actual number of texts considered in the following steps (After Selection). The texts were divided in three groups: the Ph.D. thesis (**T**), the M.Sc. dissertations (**D**) and the technical reports and conference and journal papers (**O**).

Corpus	Before Selection			After Selection			Number of Kept Texts
	T	**D**	**O**	**T**	**D**	**O**	
PG	23	46	136	6	22	67	95
GG	32	30	203	11	10	118	139
DD	30	97	51	8	32	13	53
SM	31	70	90	6	33	49	88
PP	43	114	78	9	27	26	62

Table 2.1: Number of texts processed during the *corpora* construction.

Chapter 3

The refinement of texts

To illustrate the benefits of the refinements of the texts, we conduct an experiment with the Pediatrics *corpus* [4]. This *corpus* is composed by 283 texts from papers of the Brazilian Journal of Pediatrics, it has around 750 thousand words and it has been created without any particular concern with refinements of the sentences. For this *corpus* a reference list with the more relevant terms with two and three words (bigrams and trigrams, respectively) was generated [17].

The reference list was originally manually created with a deep involvement of domain specialists and the ultimate goal of this list was to build a list of compound terms to help human translation. However, the resulting list of 1,420 bigrams and 730 trigrams can be considered the relevant terms for the Pediatrics domain, at least according to this *corpus*.

A previous work [9] using a semantic annotation tool, the parser PALAVRAS [2], and a noun phrase extractor, the EχATO$_{LP}$ tool [10], extracted 1,248 bigrams and 608 trigrams. The intersection between the terms extracted manually and the terms extracted by PALAVRAS and EχATO$_{LP}$ tools was 686 bigrams and 276 trigrams. The quality of such automatic extraction can be computed using the traditional precision (P), recall (R) and f-measure (F) metrics, *i.e.*:

$$P = \frac{|RL \cap EL|}{|EL|}$$

$$R = \frac{|RL \cap EL|}{|RL|}$$

$$F = \frac{2 \times P \times R}{P + R}$$

where $|RL|$ is the cardinality of the reference list (the list extracted manually), $|EL|$ is the cardinality of the automatically extracted list (the list extracted by PALAVRAS and EχATO$_{LP}$) and $|RL \cap EL|$ is the cardinality of the intersection between the two lists.

Computing these metrics for the experiment with the Pediatrics *corpus* without any refinements in the texts the result for bigrams and trigrams are:

$$P = 54.97\% \quad R = 48.31\% \quad F = 51.42\% \quad \text{(bigrams)}$$

$$P = 45.39\% \quad R = 37.81\% \quad F = 41.26\% \quad \text{(trigrams)}$$

Applying the refinements to remove from the texts all information that result in invalid sentences, as it was applied in the construction of the five original *corpora*, the number of automatically extracted bigrams and trigrams reduce to 1124 and 555, respectively. The intersection between the reference lists and the new automatically extracted lists stayed the same as before the refinements, *i.e.*, 686 bigrams and 276 trigrams present in the reference lists were extracted. Such result represents a clear improvement in the quality of the extraction, since the refinements applied to the *corpus* avoid the extraction of terms that are not relevant. In fact, computing the metrics for the experiment with the refined Pediatric *corpus* the following results were obtained:

$$P = 61.03\% \quad R = 48.31\% \quad F = 53.93\% \quad \text{(bigrams)}$$

$$P = 49.73\% \quad R = 37.81\% \quad F = 42.96\% \quad \text{(trigrams)}$$

Such results demonstrate a significant increase in the precision without any loss in the recall, *i.e.*, the quality of the extraction is bigger with a refined *corpus*. The 283 original texts of the Pediatrics *corpus* transformation with the refinements was quite important for some texts. Actually, 2 of the 283 texts were completely removed from the *corpus*, since these two texts did not had any complete sentence.

As said before, the experiment with the existing Pediatrics *corpus* illustrate the benefits of a careful refinement of the texts.

Chapter 4

Corpora characteristics

The five original *corpora* built were chosen according to the fields of expertise of a multidisciplinary research group in order to make easier the communication between researchers that did not share a common formation. The specific areas of the *corpora* are grouped into two domain from Earth Sciences (Petroleum Geology - PG and General Geology - GG) and three domains from Computer Science (Databases and Datamining - DD, Stochastic Modeling - SM and Parallel Processing - PP). Table 4.1 summarizes the characteristics of the five *corpora* constructed.

corpus	Number of Texts	Number of Sentences	Number of Words
PG	95	29,318	843,809
GG	139	39,648	1,165,220
DD	53	42,653	1,127,190
SM	88	44,111	1,176,016
PP	62	40,774	1,085,842

Table 4.1: *Corpora* characteristics.

To exemplify the use of the *corpora*, we use them as input to the PALAVRAS parser [2] to linguistic annotation and to EχATO$_{LP}$ term extractor in order to extract their relevant terms. The number of extracted terms to each *corpus* is presented in Table 4.2 were the number of extracted terms is classified according to the number of words in the terms, *i.e.*, unigrams, bigrams, trigrams, quadrigrams and multigrams (ngrams with five or

more words).

corpus	unigrams	bigrams	trigrams	quadrigrams	multigrams	total
PG	4,512	16,700	18,678	15,715	55,605	107,994
GG	84,765	51,715	43,336	27,705	207,521	295,185
DD	3,857	16,784	19,191	17,163	56,995	119,136
SM	3,733	15,656	19,069	17,249	55,707	112,185
PP	3,952	15,989	20,285	18,051	58,277	119,078

Table 4.2: Number of extracted terms to each *corpus*.

It is interesting to notice that the General Geology (GG) *corpus* is very different from the others, since from nearly the same number of words and sentences, it produces a much larger number of relevant terms (almost three times the number of terms of the others). This peculiar behavior may be due to the domain intrinsic writing style, or maybe by the fact that in Geology a large number of proper names is employed.

The other interesting remark that can be made from the observation of Table 4.2 is that, except from *corpus* GG, the number of terms distributed according to the number of words is fairly the same for quite different domains. This observation let us believe that well-formed *corpora* have similar characteristics, and only the difference in the terms extracted actually distinguish them.

Obviously, this analysis of the extracted terms of each *corpus* deserves a deeper investigation, but such work is out of this report scope.

Chapter 5

Conclusion

This report presents the effort to build five original *corpora* since the choice of texts, selection by a specialist, conversion of electronic format and text refinement. In particular, the importance of text refinement step was illustrated with an existing Pediatrics *corpus*. Therefore, it is reasonable to believe that the five builded *corpora* must be quite good to further linguistic analysis.

The scientific contribution of this report is two-fold, since not only the five *corpora* are good language resources to be used by the NLP community, but also the process of *corpus* construction is a valid framework to develop new valid and reliable *corpora*.

As said before, this building effort is inserted in a broader research initiative that congregates researchers from different domain areas. These new *corpora* are already being used to extract relevant terms in each domain in order to build glossaries to help the scientific exchanges among researchers from different domains.

Besides this on going application, these *corpora* can also be employed to other application, *e.g.*, as concepts extraction for ontology learning or even more sophisticated tasks as relation extraction. In fact, there is a myriad of potential applications for the new *corpora*.

5.1 Acknowledgements

The authors would like to express their gratitude to the students that helped in the *corpora* building: Daniel Martins, Kamila Ail da Costa, Guilhereme

Rodegheri and Eduardo Schwingel Diederichsen. We thank as well the researchers of the PALEOPROSPEC project at the Computer Science Department (FACIN) of the PUCRS University that contribute in the *corpora* building process as specialist of the five domains.

Bibliography

[1] Biblioteca digital brasileira de teses e dissertações, February 2010. [online] `http://bdtd.ibict.br`.

[2] E. Bick. *The parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.d. thesis, Arhus University, 2008.

[3] D. Bourigault, C. Fabre, C. Frérot, M. Jacques, and S. Ozdowska. Syntex: analyseur syntaxique de corpus. In *Actes de la 12ème TALN*, Dourdan, 2005. ATALA.

[4] R. J. Coulthard. The application of corpus methodology to translation: the jped parallel corpus and the pediatrics comparable corpus. M.sc. dissertation, UFSC, Florianópolis, Brazil, 2005.

[5] Google scholar, February 2010. [online].

[6] N. Gregoire. Duelme: a dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, Open Access(doi: 10.1007/s10579-009-9094-z), 2009.

[7] J. Kietz, R. Volz, and A. Maedche. Extracting a domain-specific ontology from a corporate intranet. In *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning*, volume 7, pages 167–175, Morristown, NJ, 2000. Association for Computational Linguistics.

[8] A. Kilgarriff, M. Rundell, and E. U. Dhonnchadha. Efficient corpus development for lexicography: building the new corpus for ireland. *Language Resources and Evaluation*, 40(7):127–152, 2006.

[9] L. Lopes, L. H. M. de Oliveira, and R. Vieira. Portuguese term extraction methods: Comparing linguistic and statistical approaches. In *PROPOR 2010 – International Conference on Computational Processing of Portuguese Language*, 2010.

[10] L. Lopes, P. Fernandes, R. Vieira, and G. Fedrizzi. Exatolp - an automatic tool for term extraction from portuguese language corpora. In *LTCÕ09 - 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 167–175, Poznan, Poland, 2009. Adam Mickiewicz University.

[11] L. Lopes, R. Vieira, M. J. Finatto, A. Zanette, D. Martins, and L. C. Ribeiro Jr. Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. *RECIIS*, 3(1):72–84, 2009.

[12] Notepad++. The web site name, February 2010. [online].

[13] P. Pantel and D. Lin. A statistical corpus-based term extractor. In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 36–46, New York, USA, 2001. ACM Press.

[14] M. A. Perini. *Princípios de linguística descritiva: introdução ao pensamento gramatical*. Parábola, São Paulo, Brazil, 2007.

[15] A. Sanchez and P. Cantos. *CUMBRE – Corpus Linguistico del Español Contemporaneo – Fundamentos, Metodologia, y Aplicaciones*. SEGL, Madri, Spain, 1996.

[16] F. P. Silveira. Entrelinhas - uma ferramenta para processamento e análise de corpus. M.sc. dissertation, PUCRS, Porto Alegre, Brazil, 2008.

[17] Textcc – textos técnicos e científicos, February 2010. [online].