Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches

Lucelene Lopes[1], Leandro Henrique M. de Oliveira[2], and Renata Vieira[1]

[1] Faculdade de Informática – FACIN – PUCRS
Porto Alegre – RS – Brazil
{lucelene.lopes, renata.vieira}@pucrs.br
[2] Embrapa Informática Agropecuária – CNPTIA
Campinas – SP – Brazil
leandro@cnptia.embrapa.br

Abstract. This paper discuss different extraction methods, linguistic informed and pure statistical, through a comparative analysis using usual metrics: Precision, Recall and F-measure. The experiments were made over a *corpus* from Pediatrics in portuguese, and the extracted terms were compared with a reference list.

1 Introduction

The process of building ontologies is difficult and its relevance for the management, organization and dissemination of a specific domain of knowledge is well known. Extraction from text has been explored as a method for automatic ontology building. Such text-based approaches start with term extraction, which is the fundamental step, *i.e.*, the success of further steps depends on it, since the extracted terms give the basic conceptual representation of a given domain.

Usually, automatic term extraction processes are based on the analysis of a group of texts (*corpus*) of a domain of interest [5]. This paper draws a comparative analysis of two approaches for term extraction (linguistic and statistical) from a specific *corpus* from the Pediatrics area.

The linguistic based method receives a syntactically annotated *corpus* and extracts terms using an analysis based on the most frequent noun phrases. In this sense, this method is similar to the one used by Bourigault *et al.* [4].

The second method clearly follows a statistical approach, in which the terms are extracted through an analysis of their frequency in the *corpus*, discarding terms from a *stoplist*. Thus, this method is similar to the ones proposed by Aubin and Hamon [1] and Fortuna *et al.* [9].

Section 2 describes the *corpus* and the reference list of terms; Sections 3 and 4 present the linguistic and statistical approaches; Section 5 discusses the results; finally we summarize this paper contribution and suggest future work.

2 *Corpus* and Reference List

The *corpus* used in the experiments is composed of 283 texts in Portuguese extracted from *Jornal de Pediatria* (http://www.jped.com.br), with a total of

785,448 words. In order to verify the efficiency of the process, we use a reference list of terms. The reference list was build up by TEXTQUIM-TEXTECC project (`http://www.ufrgs.br/textecc`). The primary goal of this list was to create glossaries for translation support. To identify items for these glossaries, terms were extracted from plain texts (without any linguistic annotation). In this process, terms with less than 4 occurrences in the *corpus* were discarded. Based on an initial list of 36,741 terms, an automatic filtering process, based on heuristics, a list with 3,645 candidate terms was generated to be integrated in the glossaries.

After the filtering process based on heuristics, a manual assessment of the relevance of the terms was performed manually. Finally, a list of 2,150 terms was produced containing 1,420 bigrams and 730 trigrams. The complete reference list is available in the OntoLp portal (`http://www.inf.pucrs.br/~ontolp`). The same corpus has been used in previous experiments, such as [11].

3 Linguistic Approach – EχATO$_{\mathcal{LP}}$

In this approach, the process of extracting terms begins with the linguistic annotation of the *corpus*, which was performed by the PALAVRAS parser [3]. Each word from each phrase is annotated according to its syntactic function, its morphological characteristics and a semantic tag.

From the annotated texts, Noun Phrases (NPs) were extracted. Unlike isolated words, NPs meaning tends to be more stable [7]. In this paper, the analysis of the extraction is focused only in NPs with 2 (bigrams) and 3 (trigrams) words. The linguistic extraction method is performed by a software tool called EχATO$_{\mathcal{LP}}$ [12].

EχATO$_{\mathcal{LP}}$ – Automatic Extractor of Terms for Ontologies in Portuguese Language – is a software tool that receives an annotated *corpus* and extracts automatically all NPs, classifying them according to its number of words. The tool uses a group of heuristics to refine the process of extraction. These heuristics are based on linguistic knowledge and it aims to discard or to refine the terms that were identified by parser as NPs, such as eventual parser errors or terms presenting lack of terminological relevance. Specifically, the heuristics applied to improve or to discard terms identified as NPs by PALAVRAS are:

- **removal of conjunctions at the end of a noun phrase**: when the input text has two implicit noun phrases connected by a conjunction, the parser keeps the conjunction in the first noun phrase. For example, “doença cardíaca e pulmonar”(in English: cardiac and pulmonary disease) is annotated just as “doença cardíaca e”, *i.e.*, it inserts an error in the explicit term “doença cardíaca” and misses the implicit term “doença pulmonar”;
- **removal of pronouns at the beginning of a noun phrase**: the reference established by the pronoun is too difficult to be investigated during the term extraction, hence, the removal of the pronoun preserves the noun and its complements. For example, “aquelas crianças recém-nascidas”(in English: those new born children), becomes just “crianças recém-nascidas”;
- **removal of articles**: since articles do not carry conceptual information about the term, they are always removed, regardless if they appear at the beginning,

the middle, or even if they were contracted with a preposition (which is quite common in Portuguese). For example, "a vida do bêbe" (in English: the life of the baby) is represented "a vida de o bêbe" (decontraction of the preposition "de" and article "o"); after article removal it becomes "vida de bêbe";

- **removal of terms with numerals**: when a NP has numerals in their numeric (24) or written form ("vinte e quatro", in English: twenty four) inside it, the term is discarded;
- **removal of terms with symbolic characters**: when a NP has characters other than upper and lowercase letters with or without valid portuguese accentuations[3], plus hyphen ("-") and underscore ("_") characters, it is also unlikely to be a relevant term;
- **removal of terms with parser errors**:
  - NPs ending with a preposition (syntactic structure recognition error);
  - NPs in which the head is neither a noun, proper name, adjective, nor participle past verb (syntactic structure recognition error);
  - NPs with circular references in the annotated file (tree representation error);
  - NPs with words with more than 128 characters (agglutination into a single term of a full clause, or even a quite large proper name);
  - NPs composed by more than 256 characters (a similar agglutination error).

The NPs extracted can be composed by any number of words, including those with just one word. This is typically the case of terms that were composed by an article plus a noun and that had the article removed, but it can also be produced by nouns originally composed by a single word in a role of subjects or objects of a clause. In practice, the EχATO$\mathcal{LP}$ groups NPs in ten lists that have NPs with 1 to 9 words, and the last list contains NPs with 10 or more words. EχATO$\mathcal{LP}$ generates each one of these ten lists of terms in decrescent order of frequency in the *corpus*. Thus, these lists can be easily submitted to cut-off points.

In the extraction of terms performed in this paper we considered NPs that have the absolute frequency greater or equal to 4 occurrences in the *corpus*, *i.e.*, NPs that appear 3, 2 or 1 single time were not included in the final list of terms. This selection of terms inserts a statistical component in the process of extraction.

4 Statistical Approach – NSP

NSP Tool – *Ngrams Statistic Package* [2] – is a set of programs developed to identify and extract *ngrams* from *corpus*. In this paper, we use `count.pl`, which is the main program of the NSP package. This program needs the following parameters:

- **the size of *ngram***: in this paper only bigrams and trigrams were considered;
- **the *stoplist***: can be specified in the NSP syntax including functional words with a high frequency, such as prepositions, articles, conjunctions, and a significative quantity of adverbs that do not present any terminological value. Moreover, usual demarcation words in technical texts as "Introdução", "Referências", "Bibliografia";

[3] The valid accentuations in Portuguese are: "á", "é", "í", "ó", "ú", "â", "ê", "ô", "à", "ü", "ã", "õ" and "ç". These accentuations appear in both upper and lowercase letters.

- **a cut-off threshold**: informs NSP which values of absolute frequency of *ngrams* must be eliminated during the processing. In this paper a cut-off threshold was equal to 4; and
- **a word formation rule**: allows the definition and specification of which pattern of words must be selected by the program in given execution. For the experiments in this paper, the rule for formation of tokens was used in order to accept composed words with upper or lowercase letters and other common accents found in Portuguese, as well as the hyphen.

NSP also offers a post-processing, which for this paper experiments was used to remove proper names. Hence, terms like "São Paulo" and "Sociedade Brasileira" were excluded. This task did not use any sophisticated linguistic knowledge because it simply excludes terms in which words begin with uppercase letters.

5 Experiments

The lists extracted by both approaches were compared to the reference list composed by 1,420 bigrams and 730 trigrams. The linguistic approach (EχATO$\mathcal{LP}$) has generated 1,248 bigrams and 608 trigrams, while the statistical approach (NSP) produced 3,709 bigrams and 2,550 trigrams.

Table 1 presents the number of terms found in each one of the approaches for diverse cut-off thresholds according to the frequency of the terms. The last column (Full) indicates the number of terms presented for the complete lists generated by EχATO$\mathcal{LP}$ and NSP Tools. The other columns indicate the cardinality of lists reduced by the application of absolute cut-off criteria. In each of these columns just the 100, 200, 300, 400 and 500 first terms of the extracted terms list were considered.

Table 1. Number of terms found in several cut-off lists

Methods of	Number of	Size of the List					
Extraction	Terms	100	200	300	400	500	Full
bigrams	$\|EL\|$	100	200	300	400	500	1248
EχATO$\mathcal{LP}$	$\|RL \cap EL\|$	77	147	213	275	331	686
bigrams	$\|EL\|$	100	200	300	400	500	3709
NSP	$\|RL \cap EL\|$	66	117	175	223	269	1230
trigrams	$\|EL\|$	100	200	300	400	500	608
EχATO$\mathcal{LP}$	$\|RL \cap EL\|$	48	97	151	206	236	276
trigrams	$\|EL\|$	100	200	300	400	500	2550
NSP	$\|RL \cap EL\|$	39	71	110	147	186	556

A first superficial analysis of these numbers appears to indicate an advantage towards the statistical approach because the quantity of extracted terms is clearly higher. However, it is necessary to take into account not only the number of terms found ($|RL \cap EL|$), but also the size of each one of the extracted lists ($|EL|$) and the size of the reference list ($|RL|$).

The following quantitative metrics express Precision (P) and Recall (R), as well as the equilibrium between these two indexes (F-measure - F):

$$P = \frac{|RL \cap EL|}{|EL|} \qquad R = \frac{|RL \cap EL|}{|RL|} \qquad F = \frac{2 \times P \times R}{P + R}$$

Computing those metrics, Table 2 shows that statistical approach Recall results were higher, but other indexes were favorable to the linguistic approach. However, once again this analysis of results can be considered shallow, since we are not taking into account the distribution of the correct terms in the extracted list (extracted terms present in RL) for each approach.

Table 2. Metrics found in several cut-off lists

Methods of Extraction	Metrics	Size of the List					
		100	200	300	400	500	Full
bigrams EχATO$\mathcal{LP}$	P	77.00%	73.50%	71.00%	68.75%	66.20%	54.97%
	R	5.42%	10.35%	15.00%	19.37%	23.31%	48.31%
	F	10.13%	18.15%	24.77%	30.22%	34.48%	51.42%
bigrams NSP	P	66.00%	58.50%	58.33%	55.75%	53.80%	36.16%
	R	4.65%	8.24%	12.32%	15.70%	18.94%	86.62%
	F	8.68%	14.44%	20.35%	24.51%	28.02%	47.96%
trigrams EχATO$\mathcal{LP}$	P	48.00%	48.50%	50.33%	51.50%	47.20%	45.39%
	R	6.58%	13.29%	20.68%	28.22%	32.33%	37.81%
	F	11.57%	20.86%	29.32%	36.46%	38.37%	41.26%
trigrams NSP	P	39.00%	35.50%	36.67%	36.75%	37.20%	21.80%
	R	5.34%	9.73%	15.07%	20.14%	25.48%	76.16%
	F	9.40%	15.27%	21.36%	26.02%	30.24%	33.90%

6 Conclusion

Despite of the difficulties found in the linguistic annotation of the *corpus*, we can conclude that the linguistic approach used by EχATO$\mathcal{LP}$ offers better results and, therefore, it is more appropriated for the purpose of identification of concepts for automatic ontologies extraction. It is worth to mention that the difficulty in the syntactic annotation consists in finding a reliable tool (*parser*) and the conversion from the parser output to the linguistic extraction tool. Actually, some problems such as the lower Recall in EχATO$\mathcal{LP}$ for bigrams and trigrams, may be explained by errors in annotation inherited from the parser that were not be corrected.

On the other hand, a statistical approach has the advantage of being easier to adapt. Actually, regardless the *corpus* specific domain and language, the use of NSP requires only the construction of a *stoplist* and a set of rules for the construction of valid words. The simplicity of the statistical approach contributes to identify a great number of terms. This fact explains the higher Recall indexes for this approach experiments that occurs when extracting both bigrams and trigrams. However, the same simplicity that contributes for the increasing in Recall takes its toll by reducing Precision index.

As general conclusion, it is possible to assert that if there is reliance in the linguistic annotation tool and its adaptation to the term extractor, it may compensate to use a linguistic approach. A future work in the study of these approaches it to consider other *corpus*, in order to strength these claims.

Another natural sequence to the work presented in this paper would be the identification of hierarchies of terms, in order to proceed with the ontology construction. This task is considerably more complex and tasks like detection of synonyms and verb analysis should be considered. One interesting topic for research is to adapt the extraction tools (EχATO$_{\mathcal{LP}}$ and NSP) to deal other measures of relevance for extracted terms. This is the case of traditional approaches [6, 8], but also more recent techniques based on the well-known *tf-idf* [10] or other measures based on "perplexity" of extracted terms [13].

This paper has shown a quantitative assessment where most refined process of extraction based on linguistic information overcome simpler statistical approaches.

References

1. AUBIN, S.; HAMON, T. Improving term extraction with terminological resources. FinTAL 2006, *LNAI 4139*, pp. 380-387, 2006.
2. BANBERJEE, S.; PEDERSEN, T. The Design, Implementation, and Use of the Ngram Statistics Package. In: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Feb., 2003, Mexico City.
3. BICK, E. *The parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* PhD thesis, Arhus University, 2000.
4. BOURIGAULT, D.; FABRE, C.; FRROT,C.; JACQUES, M.; OZDOWSKA, S. SYNTEX, analyseur syntaxique de corpus, *TALN*, Dourdan 2005.
5. BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. In: P-Buitelaar, Cimiano, P.; and Magnini, B. (Ed.). *Ontology Learning from Text: Methods, Evaluation and Applications*, v. 123 of Frontiers in Artificial Intelligence and Apllications. IOS Press, 2005.
6. CHANG, J.-S.; SU, K.-Y. *A multivariate gaussian mixture model for automatic compound word extraction.* Techinical Report, Department of Electrical Engineering, National Tsing-Hua University, 1997.
7. COLLINGS, M. J. *Head-driven statistical models for natural language parsing.* PhD thesis, UPenn, 1999.
8. DUNNING, T. Accurate methods for statistics of surprise and coincidence. *Computational Linguistic*, 19(1):61–74, 1993.
9. FORTUNA, B; LAVRAC, N.; VELARDI, P. Advancing topic ontology learning through term extraction. PRICAI 2008, *LNAI 5351*, pp. 626-635, 2008.
10. LAVELLI, A.; SEBASTIANI, F.; ZANOLI, R. Distributional term representations: an experimental comparison. *CIKM'04*, Washington, USA, pp. 615–624, 2004.
11. LOPES, L.; VIEIRA, R.; FINATTO, M. J.; ZANETTE, A.; MARTINS, D.; RIBEIRO JR, L. C. Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. *RECIIS*, 3(1): 72-84, 2009.
12. LOPES, L.; FERNANDES, P.; VIEIRA, R.; FEDRIZZI, G. ExATOlp - An automatic tool for term extraction from portuguese language corpora. *LTC'09*, Poznam, Poland, 2009.
13. YOSHIDA, M.; NAKAGAWA, H. Automatic term extraction based on perplexity of compound words. *IJCNLP*, pp. 269–279, 2005