EχATO$_{\mathcal{LP}}$ – a tool for domain relevant terms extraction

Lucelene Lopes, Paulo Fernandes, Renata Vieira, Guilherme Fedrizzi, and Daniel Martins

Faculdade de Informática – FACIN – PUCRS
Porto Alegre – RS – Brazil
{lucelene.lopes, paulo.fernandes, renata.vieira,
guilherme.fedrizzi}@pucrs.br

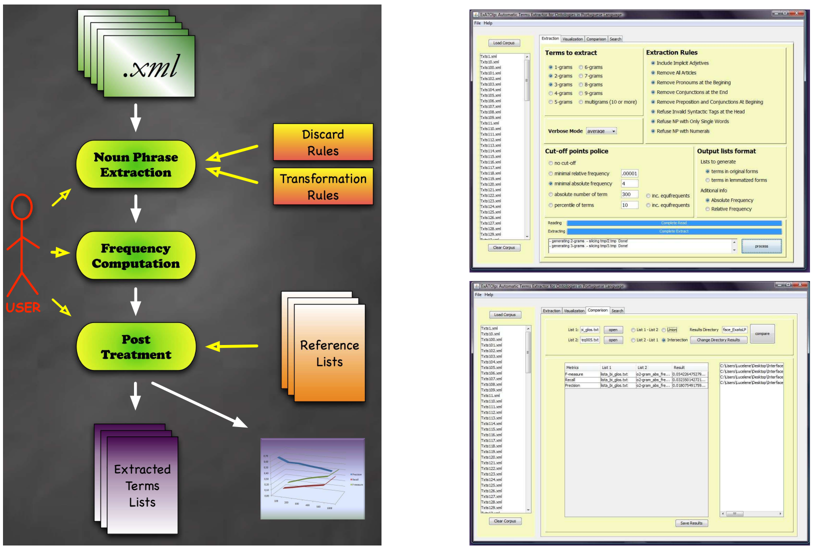
Abstract. This paper presents a software tool to extract relevant terms from Portuguese texts. EχATO$_{\mathcal{LP}}$ extracts the most frequent noun phrases in an annotated *corpus*. The annotation is provided by the PALAVRAS parser. The software tool offers different options to improve the quality of extraction that goes from post-treatment of the parser annotation to application of linguistic and statistical criteria. EχATO$_{\mathcal{LP}}$ also offers some additional features to compare extracted terms with reference lists, to compute efficiency numerical indexes and to search for terms in the *corpus*.

Term extraction from *corpora* is usually the basis of many Natural Language Processing (NLP) task such as automatic glossary construction [7], text categorization [4] and even ontology learning [3]. Term extraction, as many other NLP applications, can benefit from both linguistic and statistical approaches, as the combination of these two approaches often offers better results than each of the approaches separately.

EχATO$_{\mathcal{LP}}$ software tool [6] thus uses both linguistic and statistical approaches to extract and select domain significant terms from a an annotated domain *corpus*. From a linguistic point of view, the extraction is based on the syntactic annotation performed by the parser PALAVRAS [2]. The candidate terms are terms annotated as noun phrases by the parser according to an extra set of discard and transformation rules. From a statistical point of view, those candidate terms are subject to frequency analysis, *i.e.*, in order to select the more frequent ones.

Figure 1(a) graphically presents the software architecture. The basic input is a set of `.xml` files which are the files with the annotated texts of the *corpus*. The extraction process consider a set of discard and transformation rules to, respectively, discard some noun phrases that may be unwanted, *e.g.*, noun phrases with numerals, or to adapt some noun phrases to the purpose of extraction, *e.g.*, remove articles. The user can chose by the tool options which rules of these two sets are to be applied. Figure 1(b) upper screenshot presents the interface where the user can choose from all these options.

Once the candidate terms are extracted, their frequencies in the *corpus* and in each text is computed. Then, by user choice, some of the candidate terms can be selected according to different criteria, *e.g.*, keeping only the 10% more frequent



Terms to extract

● 1-grams ○ 6-grams
● 2-grams ○ 7-grams
● 3-grams ○ 8-grams
○ 4-grams ○ 9-grams
○ 5-grams ○ multigrams (10 or more)

Verbose Mode: average

Extraction Rules

● Include Implicit Adjetives
● Remove All Articles
● Remove Pronoums at the Begining
● Remove Conjunctions at the End
● Remove Preposition and Conjunctions At Begining
● Refuse Invalid Syntactic Tags at the Head
● Refuse NP with Only Single Words
● Refuse NP with Numerals

Cut-off points police

○ no cut-off
○ minimal relative frequency .00001
● minimal absolute frequency 4
○ absolute number of term 300 ○ inc. equifrequents
○ percentile of terms 10 ○ inc. equifrequents

Output lists format

Lists to generate
● terms in original forms
○ terms in lemmatized forms

Aditional info
● Absolute Frequency
○ Relative Frequency

Reading: Complete Read
Extracting: Complete Extract

- generating 2-grams - slicing tmp/2.tmp Done!
- generating 3-grams - slicing tmp/3.tmp Done!

List 1: ɔi_glos.txt
List 2: 'eq005.txt
○ List 1 - List 2 ○ Union
○ List 2 - List 1 ● Intersection
Results Directory: face_ExatoLP

Metrics	List 1	List 2	Result
F-measure	lista_bi_glos.txt	o2-gram_abs_fre...	0.054226475279...
Recall	lista_bi_glos.txt	o2-gram_abs_fre...	0.032350142721...
Precision	lista_bi_glos.txt	o2-gram_abs_fre...	0.018075491759...

C:\Users\Lucelene\Desktop\Interface
C:\Users\Lucelene\Desktop\Interface
C:\Users\Lucelene\Desktop\Interface
C:\Users\Lucelene\Desktop\Interface

(a) EχATO$_{\mathcal{LP}}$ architecture (b) Screenshots

Fig. 1. Tool architecture and interface examples.

terms in all *corpus*. The lists of extracted terms can be compared with reference lists in order to compute usual evaluation metrics like precision, recall and f-measure. Figure 1(b) lower screenshot presents the interface for this comparison.

EχATO$_{\mathcal{LP}}$ is an ongoing project, therefore many future extensions are planned. One of the next extensions planned is to aggregate a bootstrap method to increase the quality of selected terms in a similar way as the work of Baroni and Bernardini [1]. Another planned extension is to compute other types of frequency, *e.g.*, the popular *tf-idf* [5].

References

1. BARONI, M.; BERNARDINI, S. BootCaT: Bootstrapping Corpora and Terms from the Web, *LREC 2004*,(1313–1316), 2004.
2. BICK, E. *The parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Arhus University, 2000.
3. BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. In: Buitelaar, P.; Cimiano, P.; Magnini, B. (Ed.). *Ontology Learning from Text: Methods, Evaluation and Applications*, v. 123 of Frontiers in Artificial Intelligence and Applications. IOS Press, 2005.
4. CAI, L.; HOFMANN, T. Hierarchical document categorization with support vector machines. *13th CIKM*, (78–87), ACM Press, 2004.
5. LAVELLI, A.; SEBASTIANI, F.; ZANOLI, R. Distributional term representations: an experimental comparison. *13th CIKM*, (615–624), ACM Press, 2004.
6. LOPES, L.; FERNANDES, P.; VIEIRA, R.; FEDRIZZI, G. ExATOlp - An automatic tool for term extraction from portuguese language corpora. *LTC'09*, (167–175), 2009.

7. NAVIGLI, R.; VELARDI, P. Glossextractor: A web application to automatically create a domain glossary. *AI*IA*, LNCS 4733, (339–349), Springer-Verlag, 2007.