J Braz Comput Soc (2010) 16: 247–259
DOI 10.1007/s13173-010-0020-4

ORIGINAL PAPER

Extracting compound terms from domain corpora

Lucelene Lopes · Renata Vieira · Maria José Finatto · Daniel Martins

Received: 10 February 2010 / Accepted: 1 August 2010 / Published online: 20 August 2010

Abstract The need for domain ontologies motivates the research on structured information extraction from texts. A foundational part of this process is the identification of domain relevant compound terms. This paper presents an evaluation of compound terms extraction from a *corpus* of the domain of Pediatrics. Bigrams and trigrams were automatically extracted from a *corpus* composed by 283 texts from a Portuguese journal, Jornal de Pediatria, using three different extraction methods. Considering that these methods generate an elevated number of candidates, we analyzed the quality of the resulting terms according to different methods and cut-off points. The evaluation is reported by metrics such as precision, recall and f-measure, which are computed on the basis of a hand-made reference list of domain relevant compounds.

Keywords Term extraction · Statistical and linguistic methods · Ontology automatic construction · Extraction from corpora

L. Lopes (✉) · R. Vieira · D. Martins
PPGCC, FACIN, PUCRS, Av. Ipiranga 6681, Porto Alegre, Brazil
e-mail: lucelene.lopes@pucrs.br

R. Vieira
e-mail: renata.vieira@pucrs.br

D. Martins
e-mail: daniel.martins@pucrs.br

M.J. Finatto
DECLAVE, IL, UFRGS, Av. Bento Gonçalves, Porto Alegre, Brazil
e-mail: mfinatto@terra.com.br

1 Introduction

Concept hierarchies are fundamental to represent and manipulate specific domain knowledge. The first step when constructing a knowledge base is to determine the relevant terms of the domain. Besides being one of the steps for constructing ontologies [9], the extraction of relevant terms is important for the elaboration of a thesaurus (dictionary of terms) that can be used for automatic translation, information retrieval, indexing systems, *etc*.

Among ways to determine these terms, the options go from consulting a group of specialists of the domain of interest to an automatic process that requires little human interaction. In between we may consider that the manual process can be aided by computational tools that help the user to determine a list of candidate terms. Usually, term extraction is based on the analysis of a group of texts (*corpus*) of a domain of interest [9]. Our proposal of automatic extraction follows this line of work.

In this paper, for the extraction process we take into account both statistical analysis and more sophisticated approaches, employing linguistic data—such as morphosyntactic patterns and noun phrases. Therefore, this study is based on natural language statistical processing [19]. Specifically, this article presents experiments of compound terms extraction from a *corpus* in the pediatrics area. The texts of this *corpus* are linguistically annotated by the parser PALAVRAS [5], and the OntoLP tool [23] extracts a list of terms that are concept candidates. Each term has an associated relative frequency that is used as a criteria for its degree of relevance, since this is the standard output of the used tool.

The main contribution of this article is to show an analysis of absolute and relative cut-off points concerning the quality of the extraction of conceptually relevant terms in a *corpus*. Several options for cut-off points were compared with a reference list, previously produced to create a glossary for translators. We use quantitative metrics usually employed in the area of information retrieval: (*recall*, *precision* and *f-measure*). Since data obtained from real domain *corpora* are usually large, automatic extractions, like the one proposed here, are frequently the only practicable options for the treatment of a collection of texts.

It is important to keep in mind that the goal of this paper is not to propose a different form of term extraction itself, since the actual extraction process of the *corpus* is made through automatic tools that are treated as black boxes with an output in the form extracted term lists with their relative frequencies. In such way, our goal is to propose an automatic process to cut-off the extracted term lists in order to obtain a more precise list of relevant terms without losing too much relevant terms in this process.

This article is organized as follows. Section 2 presents related works regarding term extraction. Section 3 describes the *corpus* and the process for obtaining the list of terms considered as a reference list. Section 4 presents the proposed automatic extraction process, detailing the tools used in each stage and the metrics which allows a comparative analysis of the obtained results with the reference list. Section 5 shows the results for the experiments done with the pediatrics *corpus*. Finally, the conclusion summarizes the contribution of this work and suggests future topics for research.

2 Related work

The extraction of terms from *corpus* is a well known process, that is fundamental for several applications in Natural Language Processing. There is a general agreement in this area dividing extraction of terms in three main approaches:

- Statistics—the documents contained in the *corpus* are seen as a set of terms and their frequency of occurrence, proximity and other purely numeric information are measured;
- Linguistics—the texts are annotated with morphological, syntactic and semantic linguistic information;
- Hybrid—considers the union of the two previous approaches.

An example of purely statistical approach is the work of Baroni and Bernadini [3], in which algorithms are randomly used to automatically extract compound terms from the Web. One particularity of this work is that the construction of the *corpus* is part of the process. From a initial set of simple terms Google search is used to produce the *corpus* for term extraction. The process is simple: terms that frequently appears followed by connectors (for example: *de*, *do*, *da*, etc.) are searched.

A list of *stop words* is then created, as well as a list of irrelevant words that have a large frequency in the text but that are not connectors. The search for compound terms is based on heuristics, as for example, only considering terms that are above a frequency threshold and not considering terms that begin or end with connectors. Similarly, in our paper frequency thresholds to eliminate some of the extracted terms are used. The use of thresholds in the work of Baroni and Bernardini was not rigorously discussed, while in this article the main contribution is the analysis of different techniques for extraction of terms and thresholds of cut-off points.

The work of Navigli and Velardi [20] proposes a terminology extraction algorithm mainly based on statistical analysis of extraction of different *corpora* from different domains. The extracted terms of each domain are compared in order to perform a balanced analysis of each term relevance and consensus according to each domain. These relevance and consensus measures are deeply based on information theory, since they consider entropy of terms appearance and distribution among *corpora* and texts.

Therefore, our work differs from Navigli and Velardi's because we assume the existence of only one *corpus* over a specific domain and a single extraction procedure treated as a black box that outputs a list of extracted terms and their relative frequencies within the *corpus*.

Bourigault and Lame [7] uses a linguistic approach presenting a tool named SYNTEX [8] for the extraction of terms in a parsed *corpus* of the French language. The extraction considers noun phrases, morphosyntactic categories and the main syntactic relations (such as subject, direct object and prepositional complement, to the noun, to the verb and to the adjective). A specific lexicon of the domain is constructed at the same time that the syntactic analysis of the annotated *corpus* is being done. According to Bourigault and Lame, this technique allows a better adequacy to the *corpora* because each *corpus* has particularities of the domain that are specific, and, therefore, not predictable [7].

Even though the syntactic analysis of the experiments of our paper is different from the one of Bourigault and Lame, both works are similar for starting from a *corpus* annotated by a parser, as well as searching for compound terms that are noun phrases at least in one of the options of methods for extraction.

As a sequence of the linguistic approach in SYNTEX, Bourigault does a hybrid approach in UPERY [6], a tool that performs a distributional analysis of the terms linguistically extracted by SYNTEX. The simple or compound terms are analyzed in a syntactic context, and a network of words is constructed from each one of the phrases of the *corpus*.

The work of Park et al. [22] also proposes automatic term extraction based on linguistic analysis. Unlike the extraction procedure used here, Park's work proposes a deeper analysis of syntactic structure of terms in order to locate not only relevant terms, which is called terminology identification, but to chose among relevant terms the terms that are suitable to automatic glossary extraction. Once again, the objective of our work differs from such linguistic approach since it is not our goal to propose a change in the way the extraction procedure is actually done, but how to deal with a list of extracted terms and their relative frequencies.

Other examples of term extraction with hybrid approach are the recent works of Aubin and Hamon [1] and Fortuna, Lavrac and Velardi [11]. Both these works describe experiments with term extraction from text *corpora* in order to produce a topic ontology (a concept hierarchy). While Aubin and Hamon [1] use a dedicated tool (YATEA) to extract the terms, Fortuna, Lavrac and Velardi [11] use an integrated environment, called, OntoGen to perfom the extraction of terms and determination of concepts hierarchy. Due to the hybrid approach, these three works are close to ours, even though neither Bourigault, Aubin and Hamon, nor Fortuna et al. exploits more than one extraction method.

3 Corpus and reference list

The *corpus* used in the experiments comprises 283 Portuguese texts extracted from the bilingual journal *Jornal de Pediatria* (http://www.jped.com.br/). The number of words is 785,448. This *corpus* was organized by Coulthard to study translation patterns [10]. This *corpus* can be used for research purposes freely and it is available at http://www.pget.ufsc.br.

Although the experiments included in our paper were performed with this specific *corpus*, it is important to highlight that the same process can be used for any other Portuguese annotated *corpus* using the parser PALAVRAS. However, a list of reference terms is necessary in order to analyze the efficiency of the process. In fact, the main restriction to reproduce the process described in this paper to other *corpora* is to find not only the *corpus*, but also a reference list.

The list of reference terms was built by the TEXTQUIM-TEXTECC group at the Federal University of Rio Grande do Sul (http://www.ufrgs.br/textecc). The objective of selecting and listing these terms from the pediatrics *corpus* was to elaborate a glossary to support translation students. The first glossary is a dictionary per se, and the second is a catalogue of recurrent and relevant expressions for translation students without specialized knowledge in the domain of pediatrics. The group performed an extraction of n-grams from raw texts (without linguistic annotation) from the *corpus* in order to identify items for inclusion in these glossaries.

In this process, only those n-grams with more than 5 occurrences in the *corpus* were used. From a list of 36,741 n-grams, a filtering process was applied based on heuristics that resulted in a list with 3,645 n-grams, the relevant candidates to be part of the glossaries. These heuristics were developed aiming to exclude groups of words that were not appropriate for the generation of entries. For example, terms that begin or end with prepositions, such as in *para aleitamento materno* (to maternal breastfeeding), were changed by the exclusion of these prepositions. Terms that started with verbs were also excluded, such as in *promover o aleitamento* (to promote breastfeeding).

A further step in this heuristics was an assessment of the relevance of the expressions, performed by 5 translation students with some knowledge of the domain of pediatrics. These results were again refined by a manual verification aiming at making the reference more adequate for the purpose of creating an ontology (concept definition), due to the fact that the initial aim was to create glossaries for translation students. Finally, a list of 2,151 terms was provided, being 1,421 bigrams and 730 trigrams. Terms that had a composition of more than 3 words were not considered for this work.

4 Automatic extraction

The automatic extraction of terms is performed in three stages: the linguistic annotation of the *corpus* (4.1); the extraction of terms using OntoLP (4.2); and the selection of terms by cut-off point (4.3).

4.1 Annotation of the corpus

The linguistic annotation the *corpus* is performed by the parser PALAVRAS [5]. PALAVRAS performs a syntactic analysis by constructing a tree in which terminal nodes (leaves of trees) are words from the text and the terminals represent categories from the phrase structure. The input texts are in ASCII format (txt) and the output is represented in XML files. The XML files contain all words and their morphological characteristics. The phrases are annotated according to their syntactic functions.

For example, the phrase "*Muitos avanços ocorreram.*" is annotated as presented in Fig. 1.

The encoding follows the principles stated by *Corpus Encoding Standart* (CES) [13]. It is based in two main elements:

- *struct* is related to the lexical structure to be represented through an attribute *type* (*token*, *pos* and *phases*);
- *feat* is a sub-element of *struct* that describes the characteristics if this structure, has as attribute *name* and *value*, by which the type of information and its value are respectively explicitly set.

```
TOKEN
- <struct to="6" type="token" from="0">
  <feat name="id" value="t1" />
  <feat name="base" value="Muitos" />
  </struct>
- <struct to="14" type="token" from="7">
  <feat name="id" value="t2" />
  <feat name="base" value="avanços" />
  </struct>
- <struct to="24" type="token" from="15">
  <feat name="id" value="t3" />
  <feat name="base" value="ocorreram" />
  </struct>

PHRASE
- <struct to="t2" type="phrase" from="t1">
  <feat name="id" value="phr1" />
  <feat name="cat" value="np" />
  <feat name="function" value="S" />
  <feat name="head" value="t2" />
  </struct>
- <struct to="t9" type="phrase" from="t3">
  <feat name="id" value="phr5" />
  <feat name="cat" value="x" />
  <feat name="function" value="CJT" />
  </struct>

POS
- <struct type="pos">
  <feat name="id" value="pos1" />
  <feat name="class" value="pron-indef" />
  <feat name="tokenref" value="t1" />
  <feat name="canon" value="muito" />
  <feat name="gender" value="DET" />
  <feat name="number" value="M" />
  <feat name="complement" value="quant" />
  </struct>
- <struct type="pos">
  <feat name="id" value="pos2" />
  <feat name="class" value="n" />
  <feat name="tokenref" value="t2" />
  <feat name="canon" value="avan\c{c}o" />
  <feat name="gender" value="M" />
  <feat name="number" value="P" />
  <feat name="semantic" value="act" />
  </struct>
- <struct type="pos">
  <feat name="id" value="pos3" />
  <feat name="class" value="v-fin" />
  <feat name="tokenref" value="t3" />
  <feat name="canon" value="ocorrer" />
  <feat name="complement" value="predco" />
  <feat name="complement" value="fmc" />
  <feat name="complement" value="mv" />
  <feat name="tense" value="PS/MQP" />
  <feat name="person" value="3P" />
  <feat name="n_form" value="VFIN" />
  <feat name="mode" value="IND" />
  </struct>
```

Fig. 1 Example of XCES representation

The information in XCES files is separated in three files considering the linguistic levels of annotation. File *token.xml* has all lexical units of texts (terminals). File *pos.xml* has specific information from each lexical unit. File *phases.xml* has syntagmatic information of the phrases (non-terminals).

4.2 Extraction of candidate terms

In this stage, the XML files are set as input for the OntoLP tool, where the extraction of relevant compound terms is made. For the extraction of terms, it is necessary to follow two basic steps of the plugin OntoLP: extraction of simple terms and extraction of compound terms.

The extraction of simple terms is necessary to determine the compound terms. The OntoLP has two methods for the extraction of simple terms: Grammatical Class and Head of the Noun Phrase, detailed in [23]. In the experiments presented in this article, the Grammatical Class method was used. Altough this method allows the user to specify certain grammatical classes for the extraction, for this paper, simple terms of all grammatical classes were extracted.

Three methods for the extraction of compound terms, implemented in OntoLP, were used: n-grams, morphosyntactic patterns, and noun phrase.

The n-gram method (NG) extracts sequences of n words from the text and uses statistical measurements to evaluate the probability that each of the sequences has to be classified as a term, that is, the more frequently two words appear together, higher is the chance they can be considered bigrams [24]. In this sense, the NG method is not based on linguistic information, its analysis is purely statistical.

In the morphosyntactic patterns method (MP) the identification of n-grams is done by combining words that follow a pattern of grammatical categories, such as [noun]_[adjective], or [noun]_[pronoun]_[noun]. The MP method is linguistic based, since the grammatical composition of a term determines if this term will be considered an n-gram to be extracted [18]. The morphosyntactic patterns used in OntoLP are specific for Portuguese language and were initially defined in [4].

The noun phrase method (NP) tries to identify n-grams annotated as a noun phrase by the parser PALAVRAS, that is, a set of n words organized around the head of a noun phrase. So, the NP method has more linguistic complexity, since it is based on full syntactic analysis of the terms.

The OntoLP tool applies, in the extraction process, some heuristics that restrict the selected terms according to the characteristics of each method. The heuristics are[1]:

- canonical form: noun, number and gender flexions are removed;
- article: terms beginning with an article have this article removed;
- preposition+article: terms containing a pair of preposition followed by article have this pair considered as a single word;
- preposition: terms beginning or ending with a preposition are discarded;

[1] More details about these heuristics can be found in [23].

- word composition: terms containing words with special characters or punctuation marks are discarded.

4.3 Evaluation of cut-off points

In general, the output of OntoLP generates a large list of terms containing both relevant and irrelevant terms [15]. In this sense, it is interesting to look for a way to reduce the size of these lists, excluding a minimum number of relevant terms. In order to reduce this number of irrelevant terms, the first step must be to order the terms according to their relevance. It is necessary, then, to define a criteria that better represents the relevance of each term. By elimination, it seems for us that, among other given information, the relative frequency[2] is the most accurate way to represent this relevance. So, the first step to find relevant terms in a *corpus* is to classify these terms according to their decrease in the relative frequency.

In this paper we want to define, once the terms are ordered, the point where we should discard the less frequent terms. However, it is not the goal of this paper to discover the best cut-off point to the specific Journal of Pediatric *corpus*, which is used in this paper experiments. In fact, the best cut-off point to this specific *corpus* could be obtained by a simple binary search, but our intention is to analyze the Pediatrics *corps* as an example to be generalized to other *corpora*.

It is important to point out, that only the relative frequency within the whole *corpus* is used as criteria because the relevance of terms is not affected by the distribution of occurrences among the *corpus* texts. Therefore, other indexes as *tf-idf* [14] or rank [21] were not considered.

Among several options it is possible to assume an absolute arbitrary cut-off point, for example, to discard all the terms that present a relative frequency under 10^{-5} (1E-5). Another option is to adopt relative cut-off points. An example of it would be maintaining the first 20% terms from the ordered list. Independently from the chosen approach, it is important to establish the more adequate threshold and the best cut-off percentage to be applied. In the following section, experiments will be shown concerning these questions.

5 Experiments with the corpus

The experiments were done over the pediatrics *corpus* from which three lists of bigrams and three lists of trigrams were extracted by applying different methods: n-gram analysis (NG), morphosyntactic patterns (MP) and noun phrases (NP). Thereby, the extraction process described has produced six lists of terms.

For these lists all the phases in the process presented in Sect. 4 were applied. The calculation of the cut-off point was generated for each one of the lists of candidate terms: 10 lists with terms extracted by absolute cut-off points and 10 lists with terms extracted by relative cut-off points.

For the 10 lists generated by absolute cut-off points, the thresholds used were: 1E-5, 9E-6, 8E-6, 7E-6, 6E-6, 5E-6, 4E-6, 3E-6, 2E-6 and 1E-6. For the 10 lists generated by relative cut-off points, the following percentages from the list of candidate terms were chosen: 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 80% and 100%. The 100% case means that no term in the candidate list was discarded.

In order to evaluate each one of the options of cut-off points, the lists of extracted terms were compared with the lists of bigrams and trigrams that were obtained manually (Sect. 3). The metrics used for the evaluation of the extracted terms lists are precision (P), recall (R) and f-measure (F).

5.1 Metrics

Precision (P) indicates the capacity that the method has to identify the correct terms, considering the reference list, and it is calculated with the formula (1), which is the ratio between the number of terms found in the reference list (*RL*) and the total number of extracted terms (*EL*), i.e., the cardinality of the intersection of the sets *RL* and *EL* by the cardinality of set *EL*.

$$P = \frac{|RL \cap EL|}{|EL|}. \tag{1}$$

Recall (R) indicates the quantity of correct terms extracted by the method and it is calculated through the formula (2).

$$R = \frac{|RL \cap EL|}{|RL|}. \tag{2}$$

F-measure (F) is considered a harmonic measure between precision and recall, and it is given by the formula (3).

$$F = \frac{2 \times P \times R}{P + R}. \tag{3}$$

There is a lack of similar numerical results for Portuguese term extraction, but even for extraction over languages with more abundant material the results usually found for precision, recall and f-measure vary from very low to reasonably high values. For example, the work of Hulth [12] which performs keyword extraction over English written texts obtain precision, recall and f-measure values from 29%, 37% and 33% to 41%, 55% and 45%, respectively. However, an example of anthroponyms and proper names extraction in Portuguese found in the work of Baptista et al. [2] delivers

[2] The relative frequency of a term is computed as its number of occurrences divided by the total number of terms extracted, including repeated terms.

Table 1 Extracted terms and intersection with the reference—absolute cut-off points

Extraction	Number of	Cut-off Points									
Methods	Terms	1E-5	9E-6	8E-6	7E-6	6E-6	5E-6	4E-6	3E-6	2E-6	1E-6
bi-NG	$\|EL\|$	540	646	772	955	1255	1255	1722	2537	4610	15607
	$\|RL \cap EL\|$	335	404	482	579	733	733	804	838	872	935
bi-MP	$\|EL\|$	600	703	852	1044	1373	1373	1856	2705	4782	14565
	$\|RL \cap EL\|$	404	466	561	664	794	794	841	866	878	906
bi-NP	$\|EL\|$	328	403	494	626	820	820	1121	1729	3192	10512
	$\|RL \cap EL\|$	199	245	295	375	457	457	564	679	769	861
tri-NG	$\|EL\|$	317	402	513	682	882	882	1262	2059	4228	20930
	$\|RL \cap EL\|$	180	221	279	348	420	420	445	468	487	526
tri-MP	$\|EL\|$	365	469	591	776	1035	1035	1498	2425	4874	21448
	$\|RL \cap EL\|$	185	228	283	336	407	407	426	433	440	450
tri-NP	$\|EL\|$	94	109	129	188	285	285	387	621	1336	7069
	$\|RL \cap EL\|$	57	68	78	109	164	164	205	248	296	359

Table 2 Extracted terms and intersection with the reference—relative cut-off points

Extraction	Number of	Cut-off Points									
Methods	Terms	1%	5%	10%	20%	30%	40%	50%	60%	80%	100%
bi-NG	$\|EL\|$	156	780	1560	3121	4682	6242	7803	9364	12485	15607
	$\|RL \cap EL\|$	109	486	780	850	872	878	880	889	914	935
bi-MP	$\|EL\|$	145	728	1456	2913	4369	5826	7282	8739	11652	14565
	$\|RL \cap EL\|$	110	480	798	867	877	878	879	883	899	906
bi-NP	$\|EL\|$	105	525	1055	2102	3153	4204	5256	6307	8409	10512
	$\|RL \cap EL\|$	64	316	543	702	769	782	794	811	841	861
tri-NG	$\|EL\|$	209	1046	2093	4186	6279	8372	10465	12558	16744	20930
	$\|RL \cap EL\|$	117	429	468	487	494	498	503	507	520	526
tri-MP	$\|EL\|$	214	1072	2144	4289	6434	8579	10724	12868	17158	21448
	$\|RL \cap EL\|$	108	409	431	440	442	443	443	444	447	450
tri-NP	$\|EL\|$	70	353	706	1413	2120	2827	3534	4241	5655	7069
	$\|RL \cap EL\|$	41	189	252	298	308	318	326	330	347	359

precision, recall and f-measure varying from 30%, 20% and 33% to 97%, 87% and 73%, respectively.

5.2 Results—absolute cut-off points

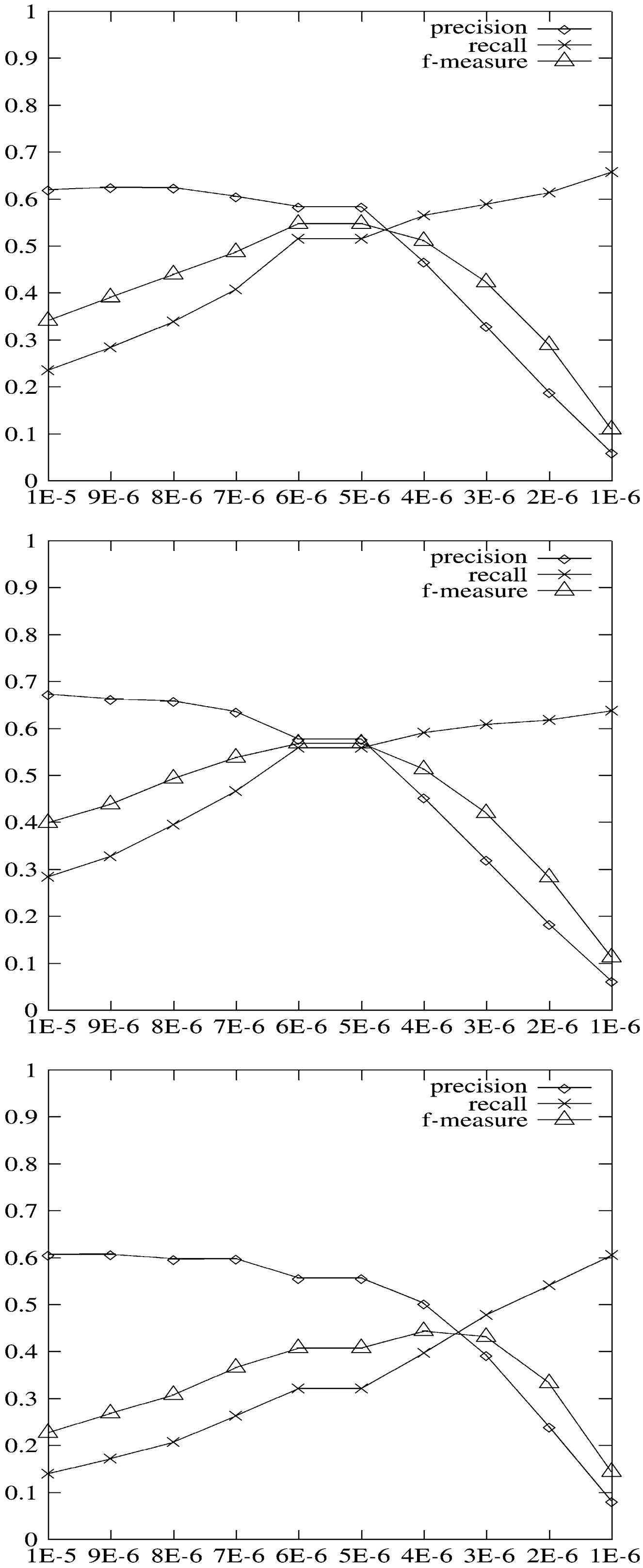
Table 1 presents the number of extracted terms ($|EL|$) and the portion of these terms that were found in the reference list ($|RL \cap EL|$), containing 1,421 bigrams and 730 trigrams. The graphics of Figs. 2 and 3 present precision, recall and f-measure metrics, according to the results of the absolute cut-off point.

We can see that the absolute cut-off points with a high threshold (1E-5) are too restrictive, since recall is always under 0.3, even if precision becomes quite good (always above 0.5). On the other hand, low thresholds (1E-6) do not reduce the number of extracted terms, i.e., all candidate terms are kept (as in 100%).

A balance between precision and recall is frequently found with intermediate cut-off points (5E-6 and 6E-6) that present the highest values of f-measure for all the cases except for Noun Phrases (bi-NP and tri-NP), which had a better balance with the cut-off point 4E-6. Another interesting remark is that there were no terms with relative frequency between 5E-6 and 6E-6 in none of the extracted lists, so the results for these two cut-off points are equal.

5.3 Results—relative cut-off points

Table 2 presents the total number of extracted terms ($|EL|$) and the portion of these terms that were found in the reference list ($|RL \cap EL|$). The graphics of Figs. 4 and 5 present



n-gram (NG)

threshold	precision	recall	f-measure
1E-5	0.6204	0.2357	0.3417
9E-6	0.6254	0.2843	0.3909
8E-6	0.6244	0.3392	0.4396
7E-6	0.6063	0.4075	0.4874
6E-6	0.5841	0.5158	0.5478
5E-6	0.5841	0.5158	0.5478
4E-6	0.4669	0.5658	0.5116
3E-6	0.3303	0.5897	0.4234
2E-6	0.1892	0.6137	0.2892
1E-6	0.0599	0.6580	0.1098

Morphosyntatic Patterns (MP)

threshold	precision	recall	f-measure
1E-5	0.6733	0.2843	0.3998
9E-6	0.6629	0.3279	0.4388
8E-6	0.6585	0.3948	0.4936
7E-6	0.6360	0.4673	0.5387
6E-6	0.5783	0.5588	0.5684
5E-6	0.5783	0.5588	0.5684
4E-6	0.4531	0.5918	0.5133
3E-6	0.3201	0.6094	0.4198
2E-6	0.1836	0.6179	0.2831
1E-6	0.0622	0.6376	0.1133

Noun Phrases (NP)

threshold	precision	recall	f-measure
1E-5	0.6067	0.1400	0.2276
9E-6	0.6079	0.1724	0.2686
8E-6	0.5972	0.2076	0.3081
7E-6	0.5990	0.2639	0.3664
6E-6	0.5573	0.3216	0.4079
5E-6	0.5573	0.3216	0.4079
4E-6	0.5031	0.3969	0.4437
3E-6	0.3927	0.4778	0.4311
2E-6	0.2409	0.5412	0.3334
1E-6	0.0819	0.6059	0.1443

Fig. 2 Metrics for absolute cut-off points—bigrams

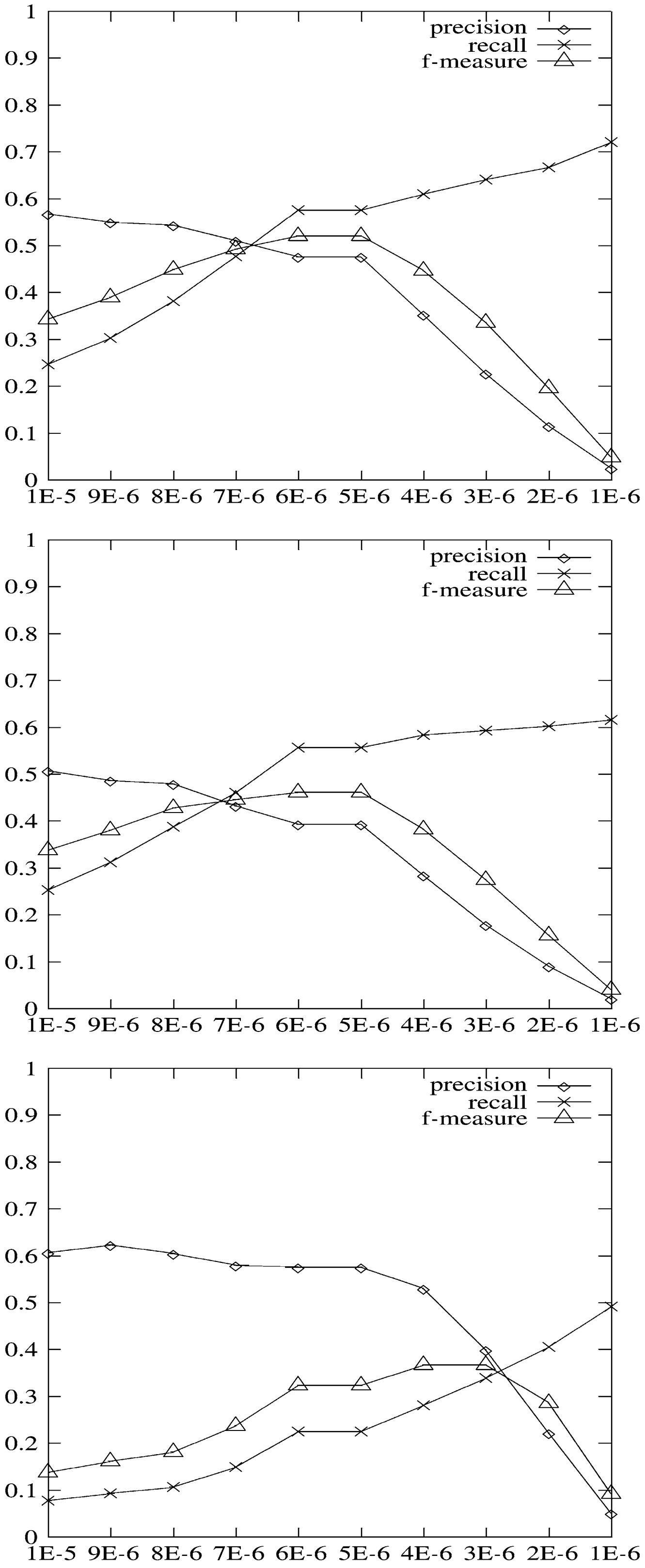


Fig. 3 Metrics for absolute cut-off points—trigrams

n-gram (NG)

threshold	precision	recall	f-measure
1E-5	0.5678	0.2466	0.3438
9E-6	0.5498	0.3027	0.3905
8E-6	0.5439	0.3822	0.4489
7E-6	0.5103	0.4767	0.4929
6E-6	0.4762	0.5753	0.5211
5E-6	0.4762	0.5753	0.5211
4E-6	0.3526	0.6096	0.4468
3E-6	0.2273	0.6411	0.3356
2E-6	0.1152	0.6671	0.1965
1E-6	0.0251	0.7205	0.0486

Morphosyntactic Patterns (MP)

threshold	precision	recall	f-measure
1E-5	0.5068	0.2534	0.3379
9E-6	0.4861	0.3123	0.3803
8E-6	0.4788	0.3877	0.4285
7E-6	0.4330	0.4603	0.4462
6E-6	0.3932	0.5575	0.4612
5E-6	0.3932	0.5575	0.4612
4E-6	0.2844	0.5836	0.3824
3E-6	0.1786	0.5932	0.2745
2E-6	0.0903	0.6027	0.1570
1E-6	0.0210	0.6164	0.0406

Noun Phrases (NP)

threshold	precision	recall	f-measure
1E-5	0.6064	0.0781	0.1383
9E-6	0.6239	0.0932	0.1621
8E-6	0.6047	0.1068	0.1816
7E-6	0.5798	0.1493	0.2375
6E-6	0.5754	0.2247	0.3232
5E-6	0.5754	0.2247	0.3232
4E-6	0.5297	0.2808	0.3671
3E-6	0.3994	0.3397	0.3671
2E-6	0.2216	0.4055	0.2865
1E-6	0.0508	0.4918	0.0921

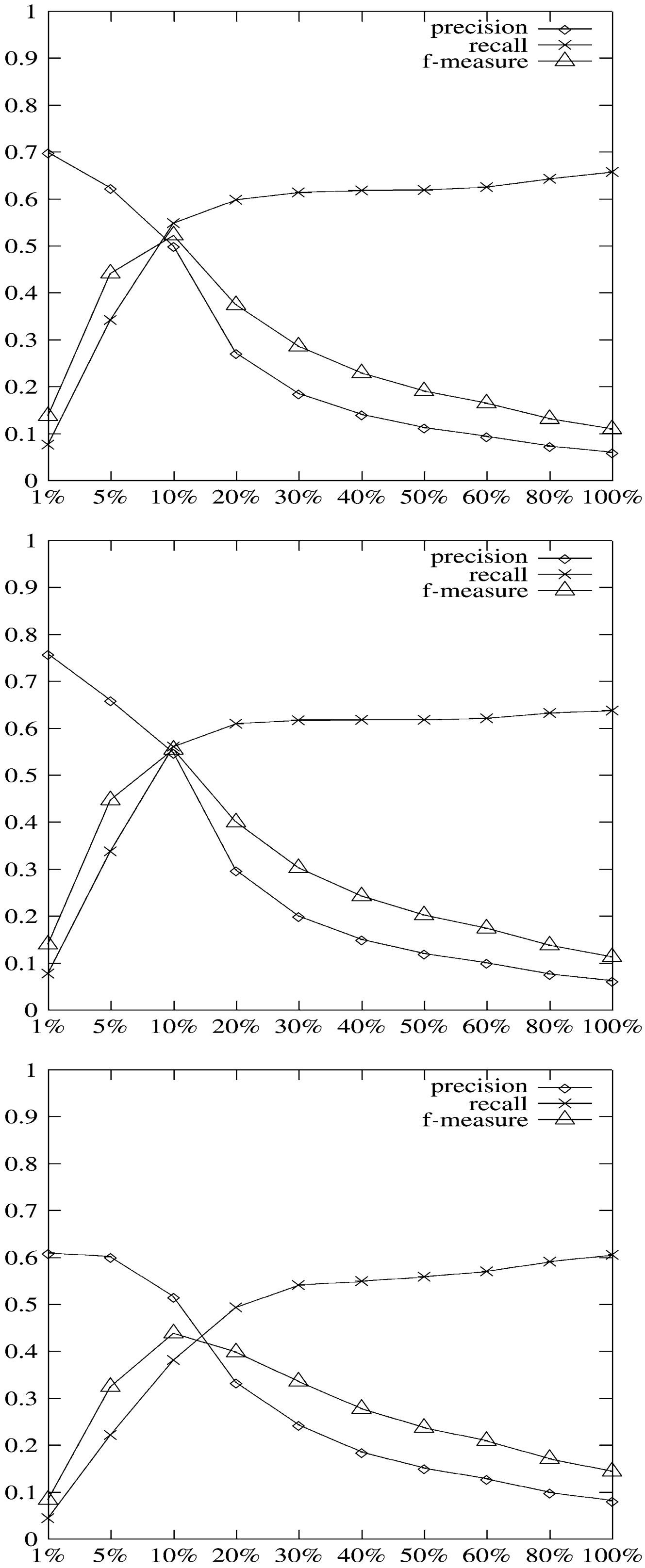


Fig. 4 Metrics for relative cut-off points—bigrams

n-gram (NG)

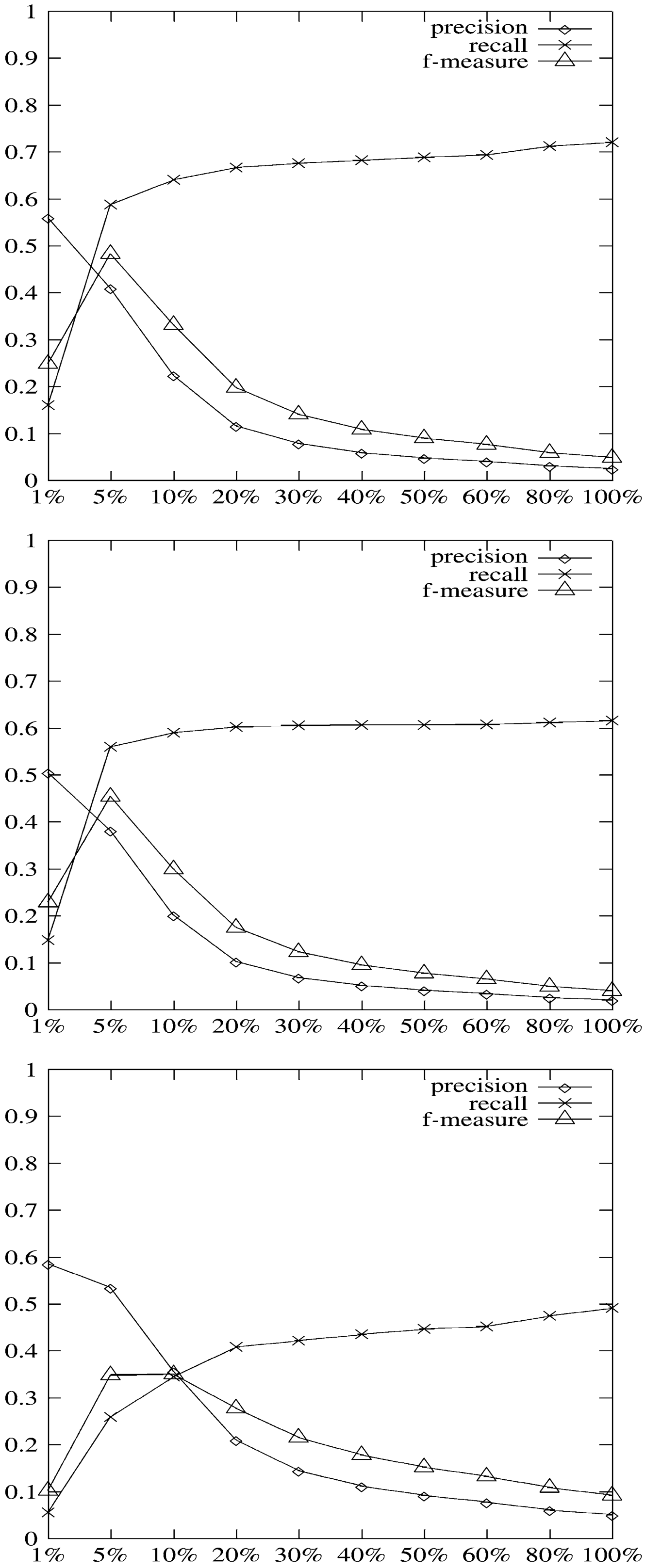
threshold	precision	recall	f-measure
1%	0.6987	0.0767	0.1382
5%	0.6231	0.3420	0.4416
10%	0.5000	0.5489	0.5233
20%	0.2723	0.5982	0.3743
30%	0.1862	0.6137	0.2858
40%	0.1407	0.6179	0.2292
50%	0.1128	0.6193	0.1908
60%	0.0949	0.6256	0.1649
80%	0.0732	0.6432	0.1315
100%	0.0599	0.6580	0.1098

Morphosyntatic Patterns (MP)

threshold	precision	recall	f-measure
1%	0.7586	0.0774	0.1405
5%	0.6593	0.3378	0.4467
10%	0.5481	0.5616	0.5547
20%	0.2976	0.6101	0.4001
30%	0.2007	0.6172	0.3029
40%	0.1507	0.6179	0.2423
50%	0.1207	0.6186	0.2020
60%	0.1010	0.6214	0.1738
80%	0.0772	0.6327	0.1375
100%	0.0622	0.6376	0.1133

Noun Phrases (NP)

threshold	precision	recall	f-measure
1%	0.6095	0.0450	0.0839
5%	0.6019	0.2224	0.3248
10%	0.5167	0.3821	0.4393
20%	0.3340	0.4940	0.3985
30%	0.2439	0.5412	0.3362
40%	0.1860	0.5503	0.2780
50%	0.1511	0.5588	0.2378
60%	0.1286	0.5707	0.2099
80%	0.1000	0.5918	0.1711
100%	0.0819	0.6059	0.1443



n-gram (NG)

threshold	precision	recall	f-measure
1%	0.5598	0.1603	0.2492
5%	0.4101	0.5877	0.4831
10%	0.2236	0.6411	0.3316
20%	0.1163	0.6671	0.1981
30%	0.0787	0.6767	0.1410
40%	0.0595	0.6822	0.1094
50%	0.0481	0.6890	0.0899
60%	0.0404	0.6945	0.0763
80%	0.0311	0.7123	0.0595
100%	0.0251	0.7205	0.0486

Morphosyntatic Patterns (MP)

threshold	precision	recall	f-measure
1%	0.5047	0.1479	0.2288
5%	0.3815	0.5603	0.4539
10%	0.2010	0.5904	0.2999
20%	0.1026	0.6027	0.1753
30%	0.0687	0.6055	0.1234
40%	0.0516	0.6068	0.0952
50%	0.0413	0.6068	0.0774
60%	0.0345	0.6082	0.0653
80%	0.0261	0.6123	0.0500
100%	0.0210	0.6164	0.0406

Noun Phrases (NP)

threshold	precision	recall	f-measure
1%	0.5857	0.0562	0.1025
5%	0.5354	0.2589	0.3490
10%	0.3569	0.3452	0.3510
20%	0.2109	0.4082	0.2781
30%	0.1453	0.4219	0.2161
40%	0.1125	0.4356	0.1788
50%	0.0922	0.4466	0.1529
60%	0.0778	0.4521	0.1328
80%	0.0614	0.4753	0.1087
100%	0.0508	0.4918	0.0921

Fig. 5 Metrics for relative cut-off points—trigrams

precision, recall and f-measure according to relative cut-off points.

The first observation about the graphics in 4 and 5 is the low precision (always below 9%) when no cut-off point is applied (100%). On the other hand, while more restrictive cut-off points were applied, the precision has grown substantially, reaching, in cut-off points of 1%, at least 0.6 in bigrams and 0.5 in trigrams. However, for all of these cases, there is a significative decrease in recall.

Then, in all cases of bigrams tested, a larger value of balance between recall and precision (f-measure) with relative cut-off points of 10% is seen. For the cases of trigrams, a larger value of f-measure for absolute cut-off points in 5E-6 and 4E-6 is seen, but the relative cut-off points around 5% and 10% were practically as good as the absolute cut-off points. For both absolute and relative cut-off points the graphics presented have shown an inversion in precision and recall values according to the cut-off points adopted. This fact makes the results obtained reliable, since it confirms the cut-off points to be applied.

5.4 Comparison of methods

The extraction methods consider different complexity of linguistic information, which grows in this order: NG, MP and NP. It is expected that methods with a higher linguistic complexity present a better overall result than those relying mainly on statistical analysis [17].

In the results without cut-off points we can see that larger values of f-measure for bigrams are organized consistently with the complexity of linguistic information, i.e.: NG(0.1098) < MP(0.1133) < NP(0.1443). For trigrams, likewise, there is a slight variation that inverts the order between NG and MP methods, resulting in: MP(0.0406) < NG(0.0486) < NP(0.0921). Hence, for the chosen *corpus* (283 texts in Portuguese from the journal *Jornal de Pediatria*) and with the extraction tools used (PALAVRAS, OntoLP), the Noun Phrase method, which has the highest linguistic complexity, presented higher results before the application of cut-off points.

Nevertheless, observing the results obtained with the application of cut-off points, the hierarchy between methods differs. Table 3 presents the hierarchy for each relative and absolute cut-off points for bigrams and trigrams according with the higher f-measure. The last line presents the hierarchy of the highest values of f-measure found in all cut-off points.

For absolute cut-off points, either for bigrams or trigrams, we can perceive that the noun phrase method (NP) presents a higher f-measure for values 3E-6, 2E-6 and 1E-6. Observing the results with relative cut-off points we note that it is similar to those found with absolute cut-off points, in which the NP method presents the lower value of f-measure, considering cut-off points of 10% for bigrams and 5% for trigrams.

Observing the absolute cut-off points (Fig. 2) for each method, the best f-measure for bigrams is: NP(0.4437 in 4E-6) < NG(0.5478 in 6E-6) < MP(0.5684 in 6E-6). For trigrams (Fig. 3), the best values of f-measure found were: NP(0.3671 in 4E-6) < MP(0.4612 in 6E-6) < NG(0.5211 in 6E-6). For the relative cut-off points, the higher values of f-measure for bigrams (Fig. 4) were: NP(0.4393 in 10%) < NG(0.5233 in 10%) < MP(0.5547 in 10%), while for trigrams (Fig. 5) we found: NP(0.3510 in 10%) < MP(0.4539 in 5%) < NG(0.4831 in 5%). In this way, the MP method had the highest value of f-measure for bigrams and NG the highest value for f-measure for trigrams.

In general, considering the f-measure, the NP method, despite its larger complexity in linguistic information, achieved lower values than the other methods (NG and MP). However, we could observe a higher precision in the NP method for the extraction of trigrams, in all of the cut-off points. For bigrams, the NP method is more precise in the most inclusive cut-off points, from 4E-6 absolute and from 20% for the relative cut-off points.

We could note that for all the methods the best results are situated in 10% (bigrams) and 5% (trigrams). This option for cut-off point represented a reduction of approximately 92% from the total of the list of terms extracted by OntoLP. Nevertheless, even with this great reduction, most part of the discarded terms were not in the reference list (85% of the terms considered relevant were kept). For the best absolute cut-off points (6E-6 for bigrams and trigrams), the reduction in the size of the list of terms extracted was around 93%. Similarly, 85% of the terms actually in the reference list were kept.

We could also verify that for the methods applied in bigrams, a relative cut-off point of 10% is very similar to an absolute cut-off point of 6E-6. But, for trigrams, the same absolute cut-off point corresponds to a relative cut-off point of around 5%.

6 Conclusion

The automatic extraction of conceptually relevant terms in a given domain is not a simple process. The process can be accelerated with automatic extraction methods. It is worth pointing that the technique for the extraction of terms available in OntoLP already offers an in depth solution, which is, however, of low precision if performed without any human intervention or selection technique, such as the one evaluated in this article.

Starting from a specific domain *corpus*, such as the one from pediatrics, we extracted lists of bigrams and trigrams that were evaluated in cut-off points. An important aspect in the extraction of relevant terms is to balance precision and recall, looking for a high f-measure. Through the experiments in this article, we could observe where the highest values of f-measure were obtained, identifying the best cut-off

Table 3 Hierarchy of the cut-off point methods

Absolute Cut-off Points	Hierarchy for bigrams	Hierarchy for trigrams	Relative Cut-off Points	Hierarchy for bigrams	Hierarchy for trigrams
1E-5	NP<NG<MP	NP<MP<NG	1%	NP<NG<MP	NP<MP<NG
9E-6	NP<NG<MP	NP<MP<NG	5%	NP<NG<MP	NP<MP<NG
8E-6	NP<NG<MP	NP<MP<NG	10%	NP<NG<MP	MP<NG<NP
7E-6	NP<NG<MP	NP<MP<NG	20%	NG<NP<MP	MP<NG<NP
6E-6	NP<NG<MP	NP<MP<NG	30%	NG<MP<NP	MP<NG<NP
5E-6	NP<NG<MP	NP<MP<NG	40%	NG<MP<NP	MP<NG<NP
4E-6	NP<NG<MP	NP<MP<NG	50%	NG<MP<NP	MP<NG<NP
3E-6	MP<NG<NP	MP<NG<NP	60%	NG<MP<NP	MP<NG<NP
2E-6	MP<NG<NP	MP<NG<NP	80%	NG<MP<NP	MP<NG<NP
1E-6	NG<MP<NP	MP<NG<NP	100%	NG<MP<NP	MP<NG<NP
better f-measure	NP<NG<MP	NP<MP<NG	better f-measure	NP<NG<MP	NP<MP<NG

points for all of the methods (NG, MP and NP). The maximum recall obtained was always lower than 73%, which means that at least 27% of the terms in the reference list were not automatically extracted.

To illustrate some of these terms (terms of the *RL* that were not extracted by any of the methods), we have bigrams: *cordão umbilical* (umbilical cord), *leites artificiais* (artificial milks), *região cervical* (cervical region); and trigrams: *recém nascidos internados* (newborns hospitalized), *aleitamento materno complementado* (complemented breastfeeding, *bronquiolite viral aguda* (acute viral bronchiolitis). On the other hand, some of the terms extracted that seem to be important were not found in the reference list. This was, for example, the case of bigrams: *criança pequena* (small child), *baixo peso* (low weight), *nível plasmático* (plasmatic level); and trigrams: *profissional de saúde* (healthcare professional), *mês de vida* (month of life), *terapia intensiva pediátrica* (intensive pediatric therapy).

Although, these missing terms deserve deeper studies, it is not the goal of the paper to evaluate the particular tools that were used (PALAVRAS, OntoLP, etc.). Indeed, the focus is to verify which were the best thresholds to maximize precision and recall.

In the best cases, 93% of the terms extracted were discarded, with a loss of only 15% of the terms in the reference. This represented high increasing of precision (for example, PM bigrams from 6% to 76%), with a sensible decreasing of recall (for example, NG trigrams from 72% to 59%). That means the use of cut-off points allows the selection of the terms that are really relevant for the domain.

Considering that the cut-off point tends to repeat, no matter what the method of extraction is, it is reasonable to imagine that the estimative of the best cut-off points is valid for other *corpus* and perhaps also for other languages, because the analyses are based on the relative frequency of terms, a criteria that is language independent.

The results are useful to highlight the importance of verifying a threshold in the process of extraction, concerning the *corpus* and the objective of the vocabulary being extracted. An incremental analysis of the terms that are candidate to be concepts seems to be more adequate in tools of semi-automatic extraction. Such tools can warn the user about the decreasing in cost-benefit as the extraction of terms proceeds.

A natural future work is the investigation of the same experiment to other corpora. However, to repeat this experiment to another corpus is quite difficult, since there is not many Portuguese corpora available, and it is even more rare to find a corpus with a reliable reference list to compare the precision and recall indexes.

An interesting future work is the comparison of these experiments with other term extraction tools, as the ExATO$_{lp}$ software tool [16]. This tool presented better results than statistical approaches [17], but unlike OntoLP, it provides only noun phrase based extraction.

After further experiments with other corpora and other extraction tools, a more consistent future is work is to investigate the composition of hierarchies concerning extracted terms. Such investigation will allow a more reliable qualitative analysis of the extracted terms, but also it will be the natural next step in order to develop a topic ontology.

References

1. Aubin S, Hamon T (2006) Improving term extraction with terminological resources. In: FinTAL 2006. LNAI, vol 4139, pp 380–387
2. Baptista J, Batista F, Mamede N (2006) Building a dictionary of anthroponyms. In: Proceedings of 7th PROPOR

3. Baroni M, Bernadini S (2004) BootCaT: Bootstrapping Corpora and Terms from the Web. In: Proceedings of the 4th LREC, pp 1313–1316
4. Baségio T (2006) Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua Portuguesa do Brasil. Dissertation (MSc), PUCRS
5. Bick E (2000) The parsing system "Palavras": automatic grammatical analysis of Portuguese in a constraint grammar framework. PhD thesis, Arhus University
6. Bourigault D (2002) UPERY: un outil d'analyse distributionnelle étendue pour la construction d'ontologies a partir de corpus. In: TALN, Nancy
7. Bourigault D, Lame G (2002) Analyse distributionnelle et structuration de terminologie—application a la construction d'une ontologie documentaire du Droit. In: TAL, vol 43(1), pp 1–22
8. Bourigault D, Fabre C, Frérot C, Jacques M, Ozdowska S (2005) SYNTEX, analyseur syntaxique de corpus. In: TALN, Dourdan
9. Buitelaar P, Cimiano P, Magnini B (2005) Ontology learning from text: An overview. In: Buitelaar P, Cimiano P, Magnini B (eds) Ontology learning from text: methods, evaluation and applications. Frontiers in artificial intelligence and applications, vol 123. IOS Press, Amsterdam
10. Coulthard RJ (2005) The application of corpus methodology to translation: the JPED parallel corpus and the pediatrics comparable corpus. Dissertation (MSc), UFSC
11. Fortuna B, Lavrac N, Velardi P (2008) Advancing topic ontology learning through term extraction. In: PRICAI 2008. LNAI, vol 5351, pp 626–635
12. Hulth A (2004) Enhancing linguistically oriented automatic keyword extraction. In: HLT-NAACL, ACL
13. Ide N, Bonhomme P, Romary L (2000) Xces: An xml-based encoding standart for linguistic corpora. In: Proceedings of the second LREC
14. Lavelli A, Sebastiani F, Zanoli R (2004) Distributional term representations: an experimental comparison. In: Proceedings of the 13th ACM CIKM, pp 615–624
15. Lopes L, Vieira R, Finatto MJ, Zanette A, Martins D, Ribeiro LC Jr (2009) Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. RECIIS—Electron J Commun Inf Innov Health 3(1):72–84
16. Lopes L, Fernandes P, Vieira R, Fedrizzi G (2009) ExATOlp: An automatic tool for term extraction from Portuguese language corpora. In: Proceedings of the fourth language & technology conference: human language technologies as a challenge for computer science and linguistics, LTC'09, Faculty of Mathematics and Computer Science of Adam Mickiewicz University, November, 2009, pp 427–431
17. Lopes L, Oliveira LH, Vieira R. (2010) Portuguese term extraction methods: comparing linguistic and statistical approaches. In: Proceedings of the 9th PROPOR
18. Maedche A, Staab S (2000) Semi-automatic engineering of ontologies from text. In: Proceedings of the 12th SEKE
19. Manning CD Schutze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge
20. Navigli R, Velardi P (2002) Semantic interpretation of terminological strings. In: Proceedings of the 6th TKE, INIST-CNRS, Vandoeuvre-lès-Nancy, France
21. Pazienza MT, Pennacchiotti M, Zanzotto FM (2005) Terminology extraction: an analysis of linguistic and statistical approaches. In: Sirmakessis S (ed) Knowlodge mining. Studies in fuzziness and soft computing, vol 185. Springer, Berlin
22. Park Y, Bird R, Bougarev B (2002) Automatic glossary extraction: Beyond terminology identification. In: Proceedings of the 19th COLING, Taipei, Taiwan
23. Ribeiro LC (2008) OntoLP: Construção semi-automática de ontologias a partir de textos da língua portuguesa. Dissertation (MSc), UNISINOS
24. Suchanek FM, Ifrim G, Andweikum G (2006) Leila: Learning to extract information by linguistic analysis. In: Proceedings of the 2nd workshop on ontology learning and population. Association for computational linguistics