Lopes and Vieira *Journal of the Brazilian Computer Society* [redacted]
DOI 10.1186/s13173-015-0025-0

Journal of the Brazilian Computer Society
a SpringerOpen Journal

RESEARCH Open Access

Evaluation of cutoff policies for term extraction

Lucelene Lopes* and Renata Vieira

Abstract

Background: This paper presents a policy to choose cutoff points to identify potentially relevant terms in a given domain. Term extraction methods usually generate term lists ordered according to a relevance criteria, and the literature is abundant to offer different relevance indices. However, very few studies turn their attention to how many terms should be kept, i.e., to a cutoff policy.

Methods: Our proposed policy provides an estimation of the portion of this list which preserves a good balance between recall and precision, adopting a refined term extraction and *tf-dcf* relevance index.

Results: A practical study was conducted based on terms extracted from a Brazilian Portuguese corpus, and the results were quantitatively analyzed according to a previously defined reference list.

Conclusions: Even thou different extraction procedures and different relevance indices could brought a different outcome, our policy seems to deliver a good balance for the method adopted in our experiments and it is likely to be able to be generalized to other methods.

Background

Automatic identification of relevant terms for a given domain is an extremely important task for a myriad of natural language processing applications. For instance, any ontology learning effort is doomed to fail if the concept identification step has a poor performance. In fact, any other steps to automatically build an ontology rely on the concept identification [1–4].

Also, text categorization applications can be much more effective if a good relevant term identification is available.

An important part in the process of identifying relevant terms is the extraction of terms and the computation of their frequencies of use as term relevance index. For the term extraction itself, many software tools are available [5, 6]. These tools usually offer high-quality extraction, regardless being implemented on the basis of linguistic or statistical approaches. As for relevance indices, many theoretical formulations are available [7–11] and it is safe to assume that a reliable relevance-based rank of extracted terms is not difficult to obtain.

Unfortunately, even assuming a nearly perfect ranked list of terms, it is still difficult to decide how many relevant terms are sufficiently relevant to cover the domain essential terms. Even the high-quality extraction procedures deliver some terms that maybe present in a corpus, but are not really relevant to the domain. The computation of a relevance index, and the ranking according to this index, is likely to push such irrelevant terms to the last positions of the ranked list.

Once the extracted terms are ranked according to the relevance, one needs to decide how many extracted terms are relevant enough. Such decision is made by adopting a cutoff point, i.e., choosing a point in the ranked list to discard all extracted terms below this point and to keep the terms from this point up. A cutoff point is present in virtually all term extraction endeavors, but almost all referred scientific works [4, 12–22] apply cutoff points based solely on guess and trial.

That being said, there are three traditional approaches to guess a cutoff point: absolute, threshold-based, and relative cutoff points. However, none of the referred work try to establish a generic policy to actually choose a point to cutoff the ranked term list. The major contribution of our paper is, therefore, to propose a policy to choose an actual cutoff point to a given term list, i.e., to offer researchers and practitioners an algorithmic solution to choose the discard point, instead of leaving this choice to a guess.

*Correspondence: lucelene.lopes@pucrs.br
PPGCC-FACIN-PUCRS, Av. Ipiranga, 6681, Porto Alegre, Brazil

It is important to call the reader's attention about the difference between an approach to choose a cutoff point and a cutoff point policy. An approach to choose a cutoff point is a method to determine how to cut off a term list, e.g., to keep the n top ranked terms of a list. A cutoff point policy is a more ambitious task, since it consists in the definition of a way to determine a cutoff point and the actual point where the list should be cut off, e.g., to keep the n top ranked terms of a list and a formula to estimate the value of n.

The application scope of the term selection policy proposed in this paper implies the following assumptions:

- Given a domain corpus, a list of terms is extracted, and, for the purpose of this paper, we will focus on the relevant terms identification on a list of terms extracted from a domain corpus, even thou most of the definitions made here can be generalized to lists of terms built from virtually any other sources than domain corpora;
- A relevance index is computed (usually based on some sort of frequency in the corpus) assigning to each term a numeric value proportional to its domain relevance; hence, the term list can be ranked according to this index;
- Some of the extracted terms are sufficiently relevant and should be considered representative of the domain, those terms will be referred as valid terms;
- Some of the extracted terms, on the contrary, are not sufficiently relevant and should be discarded; those terms will be referred as invalid terms;
- Each extracted term is either valid or invalid.

In consequence, our policy will represent a contribution to practical cases where these assumptions are true. However, our proposed policy can still be beneficial even to practical cases where not all these assumptions can be verified. For a start, the definition of valid and invalid terms is frequently arguable. While some terms are clearly classified as valid or invalid, others terms remain in a grey area and they need to be arbitrarily classified into one group or another.

Another issue is the accuracy of the relevance index, since not necessarily all valid terms will be ranked in higher positions than all invalid terms. That being said, giving the efficacy of the relevance indices [23], it is only reasonable to assume that the relevance index-based ranking is strongly correlated to the terms' actual relevance. Note that this assumption is probably as secure as to assume that terms that are not present in the list of extracted terms can never be considered relevant.

As a final consequence of these assumptions, the identification of the relevant terms of a domain consists in choosing a cutoff point to the ranked list aiming to have the valid terms above this point and the invalid terms below. In other words, it is necessary to find out the point that separates the valid and invalid terms.

By applying precision and recall metrics, which are common in information retrieval domain [24], it becomes clear that to maximize the precision, one only needs to consider very few or possibly just the most relevant terms and to discard all others. Similarly, it is easy to maximize the recall by considering all extracted terms, i.e., not discarding any. Besides those two trivial approaches, the challenge is to balance precision and recall keeping a set of extracted terms with most, ideally all, relevant terms and very few, or ideally none, irrelevant terms.

In this context, the challenge is to find out a trade-off between maximizing the precision (cutting off many terms) and maximizing the recall (cutting off few terms). Therefore, this paper's goal is to propose a term selection policy as the definition of a cutoff point to the list of extracted terms in order to balance precision and recall with respect to a reference list indicating valid terms, i.e., terms judge as relevant to the domain. To fulfill this objective, a practical study is developed through the application of three traditional cutoff point approaches (absolute, threshold, and relative) to a term list extracted from a corpus.

The result of these approaches for several parameters is compared to a reference list (gold standard) assumed to be the full set of valid terms for the corpus. Such experiments allow the quantitative comparison of traditional cutoff policies. Consequently, a new and more effective policy is proposed.

It is important to stress that in all approaches, it is possible to achieve better and worse precision-recall balances, but our claim is that using traditional approaches, there is no simple way to determine whereas a given threshold would result in a reasonably balanced cutoff point. Therefore, it is not our goal to show that our policy offers a better reduction of extracted term list sizes than traditional approaches. Instead, we want to propose a way to estimate a cutoff point that takes into account more than any traditional approach alone.

This section presents the corpora employed in this paper's experiments, the reference list corresponding to the relevant terms of the corpora, a brief description of the extraction tool and relevance index computation and, finally, the formulation of traditional information retrieval metrics. Next, the three traditional cutoff policy approaches: "Absolute", "Threshold-based" and "Relative" and their results over a practical examples of different arbitrarily chosen cutoff points are presented.

The corpora and reference lists

This paper's experiments were conducted over five different Portuguese domain corpora. The first one is

the Pediatrics corpus (Ped) [25] composed of 281 texts extracted from the *Jornal de Pediatria*, a Brazilian scholar journal on Pediatrics.

The more important particularity of this corpus is the availability of a reference list of relevant bigrams and trigrams. This reference list was manually built in the context of TEXTCC research project, conducted by Maria José Finatto group at the Federal University of Rio Grande do Sul (UFRGS).

This list is composed of 1,534 bigrams and 2,660 trigrams chosen among the extracted terms that were found at least four times in the corpus. Terminology students and domain specialists (pediatricians) issued their opinion whereas a term was representative of the Pediatrics domain, and only terms with all participant agreement were kept. This list is publicly available at Finatto's group web page: http://www6.ufrgs.br/textecc/.

The other four domain corpora employed in this paper's experiments were developed in Lucelene Lopes Ph.D. thesis [3]. These corpora cover the topics of stochastic modeling (SM), data mining (DM), parallel processing (PP), and geology (Geo). All these corpora are composed not only by scientific articles but also by larger documents as Ph.D. thesis and M.Sc. dissertations. Table 1 describes numerical characteristics for the five mentioned corpora.

An important information about these corpora is the fact that they were carefully revised manually to prevent frequent textual problems such as char encoding and inclusion of non-phrasal information. Unfortunately, the four additional corpora (SM, DM, PP, and Geo) do not have previously established reference lists. Therefore, no cutoff numerical comparison with a gold standard was possible to these corpora. However, they were needed as contrastive corpora to compute the relevance index for the pediatrics corpus extracted terms (see the "The relevance index" section). Additionally, we also apply the traditional approaches and our proposed policy over the terms extracted from these contrastive corpora in order to observe the sheer impact on the term list size.

The reader interested in further information about these corpora construction and its characteristics can find detailed information in a previous publication [26]. As for the corpora data itself, the reader may freely download all these corpora from the PUCRS Natural Language Processing group at http://www.inf.pucrs.br/peg/lucelenelopes/ll_crp.html; therefore, all experiments conducted here could be replicated or compared with other approaches.

Table 1 Corpora characteristics

Corpus	Texts	Sentences	Words	Terms
Ped	281	27,724	835,412	180,120
SM	88	44,222	1,173,401	252,168
DM	53	42,932	1,127,816	244,439
PP	62	40,928	1,086,771	241,145
Geo	234	69,461	2,010,527	436,401

The extraction tool

All experiments with cutoff points conducted in this paper were based on the list of extracted terms from the five aforementioned corpora. Each of these corpora were submitted to the EχATO$_{lp}$ software tool [6], a term extractor based on linguistic knowledge [20].

The full extraction procedure includes a previous annotation of part-of-speech (POS) and grammatical information performed by the PALAVRAS parser [27]. The annotated texts are submitted to EχATO$_{lp}$ extraction software tool that basically identifies all noun phrases and submits these noun phrases to a set of linguistic heuristics to either avoid unappropriated terms [28].

As result, a rather large set of terms is extracted from each corpora. Figure 1 depicts how many terms classified according to their number of words (unigrams, bigrams, etc. —the terms indicated as M-grams have 10 or more words) were extracted from each corpora. In this picture, we present the number of noun phrases identified originally by PALAVRAS (left hand side bars—indicated with an asterisk) and the number of actually extracted terms after applying EχATO$_{lp}$ heuristics (right hand side bars).

The use of EχATO$_{lp}$ heuristics provides a considerable re-arrangement of extracted terms, including a large number of unigrams and a significant reduction of terms with more than one word (bigrams, trigrams, and so on). As the reader may verify in a specific publication about the employed extraction [28], such re-arrangement is highly beneficial to quality of the extracted terms.

Table 2 presents numerically the number of extracted terms produced by EχATO$_{lp}$ for each corpora. Note that this number of terms reflects the number of terms after applying all EχATO$_{lp}$ linguistic heuristics, i.e., it corresponds to the number depicted in the right hand side columns of Fig. 1.

While the total number of extracted terms represented in Table 2 stands for every term extracted from the corpora, a more important number of terms is expressed in Table 3. In this table, the number of terms express how many distinct terms were extracted. As expected, there are much more repetition in the smaller n-grams.

The relevance index

Among the myriad of options to use as relevance index, the experiments conducted in this paper consider the term frequency, disjoint corpora frequency (*tf-dcf*) index [23] as estimation of each term relevance. This index is based on the use of contrasting corpora, i.e., it considers that each term relevance is directly proportional to the number of occurrences of the term in the target corpus

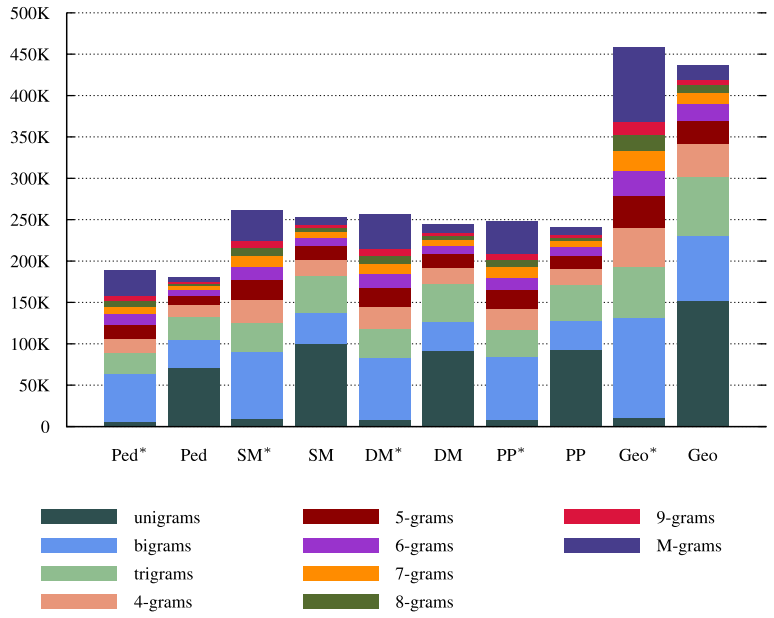

Fig. 1 Total number of extracted terms

and inversely proportional to the number of occurrences in other corpora.

Numerically, the *tf-dcf* index of a term t belonging to a corpus c, considering a set of contrasting corpora $\mathcal{G}$ is expressed as:

$$tf\text{-}dcf_t^{(c)} = \frac{tf_t^{(c)}}{\prod\limits_{\forall g \in \mathcal{G}} 1 + \log\left(1 + tf_t^{(g)}\right)} \tag{1}$$

where $tf_t^{(c)}$ express the absolute frequency, i.e., the number of occurrences, of term t in corpus c.

The main characteristic of the *tf-dcf* index is to express the relevance of a term as its absolute frequency in the corpus, if this term does not appear in the contrasting corpora. The terms of the target corpus that also appears in the contrasting corpora are penalized geometrically according to its absolute frequency in each contrasting corpora. Consequently, a term that appears in many contrasting corpora tends to suffer a great penalization.

In fact, compared to other relevance indices [9–11], *tf-dcf* is often more effective because it penalizes terms according not only to the number of occurrences in contrasting corpora but also to the number of contrasting corpora in which it appears.

Nevertheless, the *tf-dcf* index mathematical structure keeps a semantic associated to the absolute frequency, since the numeric value of *tf-dcf* of a given term is

Table 2 Total number of extracted terms

Corpus	Ped	SM	DM	PP	Geo
Unigrams	71,327	100,425	91,370	93,433	151,755
Bigrams	33,340	37,608	35,727	35,233	78,490
Trigrams	27,587	43,905	45,450	43,303	71,377
4-grams	15,555	19,905	19,212	19,354	39,625
5-grams	10,067	16,388	17,199	15,897	28,785
6-grams	6973	9893	9683	9612	19,877
7-grams	4659	7159	7440	6901	13,597
8-grams	3186	4700	5013	4756	9493
9-grams	2218	3402	3628	3424	6547
M-grams	5208	8783	9717	9232	16,855
Total	180,120	252,168	244,439	241,145	436,401

Table 3 Number of distinct extracted terms

Corpus	Ped	SM	DM	PP	Geo
Unigrams	5949	4323	4199	4361	7679
Bigrams	15,485	14,107	14,804	14,301	30,775
Trigrams	18,172	18,875	19,976	19,976	37,210
4-grams	13,104	14,506	14,024	14,997	30,295
5-grams	9223	12,239	12,349	12,809	23,621
6-grams	6676	8410	8236	8484	17,190
7-grams	4516	6187	6348	6305	12,045
8-grams	3095	4210	4450	4404	8523
9-grams	2161	3061	3232	3216	5905
M-grams	5078	8077	8906	8726	15,383
Total	83,459	85,926	96,524	97,579	176,581

bounded by its absolute frequency. As a consequence, the numeric value of the *tf-dcf* remains quite intuitive.

Information retrieval metrics

All comparisons of term lists made in this paper compute three information theory metrics that are frequently employed in a myriad of scientific works dealing with information retrieval. The metrics are precision (P), recall (R), and F-measure (F) [24].

These metrics consider the existence of two sets of different origin, e.g., two term lists. One of these sets is considered reliable, e.g., a reference list, denoted $\mathcal{RL}$. The other set is the one to be compared to the reliable one, e.g., a extracted term list, denoted $\mathcal{EL}$.

The precision of the set $\mathcal{EL}$ with respect to the reference set $\mathcal{RL}$ is expressed by the ratio between the number of identified terms that are also present in the reference list, i.e., the cardinality of the intersection between $\mathcal{EL}$ and $\mathcal{RL}$ and the number of identified terms, i.e., the cardinality of $\mathcal{EL}$. Identified terms, in this context, mean terms that were extracted from the corpus and considered relevant to the target domain. Tuus, precision expresses the percentage of terms correctly considered as relevant to the domain, i.e., the percentage of true positives with respect to the all positive terms. Numerically:

$$P = \frac{|\ \mathcal{RL} \cap \mathcal{EL}\ |}{|\ \mathcal{EL}\ |} \tag{2}$$

Recall index expresses the ratio between the number of identified terms that are also present in the reference list and the number of terms in the reference list, i.e., the cardinality of $\mathcal{RL}$. In such way, recall expresses the percentage of reference list terms that were identified, i.e., the percentage of true positives with respect to all true terms. Numerically:

$$R = \frac{|\ \mathcal{RL} \cap \mathcal{EL}\ |}{|\ \mathcal{RL}\ |} \tag{3}$$

As mentioned, considering as relevant very few extracted terms tends to improve the precision at the expense of recall reduction. On the contrary, considering as relevant a large number of extracted terms tends to increase the recall and reduce the precision. Consequently, it is necessary to find a way to evaluate the balance between precision and recall. F-measure numerically express such balance as the harmonic average between precision and recall, i.e., F-measure value is always a value between P and R, and as the difference between P and R grows, the value of F tends to be closer to the smaller one.

$$F = \frac{2 \times P \times R}{P + R} \tag{4}$$

The use of those metrics is largely spread in several domains [29–31]. The Natural Language Processing area and, specially in term extraction, many works numerically support their contribution using such metrics [7, 32–35].

Absolute approach

The simplest and more popular way to apply cutoff points is to choose an arbitrary number of terms to consider as relevant. This approach is called absolute cutoff point, and its difficulty resides in estimating how many terms should be considered, since there is not a reasonable way to estimate which number will be adequate.

Despite all that, several works in the literature use this approach and arbitrarily choosing cutoff points with empirical basis [12–16, 18–22]. It is worth noting that often this approach is accompanied by the manipulation of terms in separated lists according to the number of words, i.e., lists of unigrams, lists of bigrams, and so on.

Such kind of manipulation seems to make sense since the relevance indices tend to behave quite differently according to the number of words in each term. For instance, the absolute frequency of unigrams tends to be much higher than the absolute frequency of bigrams. In consequence, a bigram that appears, for example, 20 times, tends to be much more relevant than a unigram that also appears 20 times. For instance, considering the Geology corpus previously mentioned, the more frequent unigram was *bacia* ("basin" in English) with 2390 occurrences, while the more frequent bigram was *matéria orgânica* ("organic matter" in English) with only 430 occurrences.

Experiments with absolute cutoff points to terms extracted from pediatrics corpus

By applying the absolute cutoff points method to the bigrams and trigrams of the previously mentioned pediatrics corpus, we obtain precision, recall, and F-measure as shown in Fig. 2 for the considered reference lists mentioned in the "The corpora and reference lists" section (corresponding numeric values are given in Table 4). For these experiments, we arbitrarily choose absolute cutoff points from 100 to 3500 terms with an increment of 100.

The first observation from Fig. 2 is the fact that the precision and recall curves do cross as expected, since the precision drops and recall increases as the absolute cutoff point increases, i.e., the considered extracted terms lists becomes less restrictive. Such crossing occurs nearly at 1500 bigrams and 2700 trigrams, which is not a surprise since the reference lists have, respectively, 1534 bigrams and 2660 trigrams. Obviously, lists with less terms than the reference list will never reach a 100 % recall.

The numeric values presented in Fig. 2 indicate that the best balanced choice was found for 2000 bigrams, with 81 %, and 3300 trigrams, with 84 %. It is interesting to notice that the best F-measure value are not found when the precision and recall curves cross (1500 bigrams

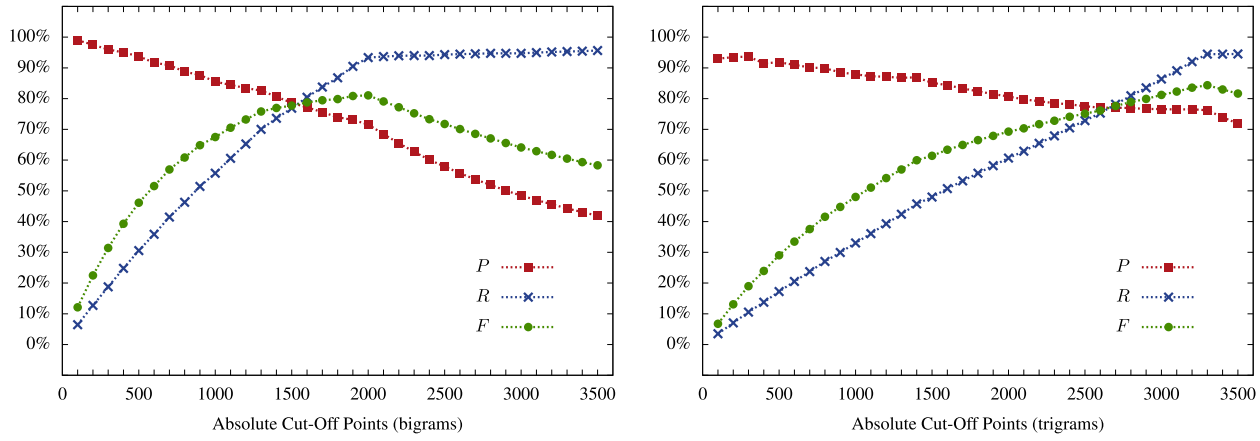

Fig. 2 Precision (*P*), recall (*R*), and F-measure (*F*) using absolute cutoff points

and 2700 trigrams). The optimal values of F-measure are found in the situation where the precision and recall values present an inflection point. This is particularly clear observing Fig. 2 bigrams curve for recall and its flexion point at 2000.

It is observable that as the list sizes grow, the precision values drop slower than the recall values increase. In fact, the optimal F-measure values seem to be associated to the near stagnation of the recall values. For bigrams, it occurs with a list of 2000 terms, when the recall is around 93 %. This is a curious behavior, since it means that nearly 500 bigrams more than the reference list must be considered. This represents that approximately 25 % of the considered bigrams are invalid (out of reference list). Hence, the precision value stays around 72 %.

For trigrams, an analogous behavior is observed, since approximately 600 additional trigrams must be considered (3300 trigrams for 2660 in the reference list) in order to deliver the higher F-measure value. However, both precision and recall values are slightly higher (76 and 94 %) than for bigrams; thus, the F-measure is a little higher for trigrams.

Threshold-based approach

Another popular approach to establish cutoff points is to choose a threshold for the relevance index. Once again, it is an arbitrary choice of which index value is considered high enough to represent a valid term.

A reasonable number of works in the area adopt the threshold method using the simplest relevance index, the absolute frequency. This is the case of the Bourigault and Lame work [13] that assumes that terms occurring at least 10 times in the corpus are relevant enough. Obviously, such choice is greatly dependent on the size of corpus since occurring nine times maybe not relevant to Borrigault and Lame experiment, but it is quite significative for a small corpus 100,000 words.

Additionally, as mentioned before, the number of occurrences also varies considerably according to the number of words. Consequently, a same threshold for unigrams and for bigrams may represent very different levels of restriction.

Some authors [15] claim that cutoff points by threshold values of absolute frequency have a direct relation with the corpus size (number of words). Such intuitive assumption may be useful for many practical cases, but it is a known fact that the number of occurrences of terms in a corpus decreases exponentially [36].

The exact format of exponential decrease may vary according to the extraction method. For instance, terms extracted according to a pure statistical method tend to follow a Zipf law. According to Zipf law formulation, the frequency of a word in a corpus is inversely proportional to its rank. Therefore, the second more frequent word tends to appear the half of times than the more frequent one, and the third more frequent word tends to occur one third of times as the more frequent one [37].

However, linguistic-based extraction process tend to behave differently. For instance, the four more frequent unigrams in the pediatrics corpus are *criança* (“child”) with 2055 occurrences, *estudo* (“study”) with 1529 occurrences, *paciente* (“patient”) with 1505 occurrences, and *doença* (“disease”) with 784 occurrences.

Therefore, it is impractical to consider a generic formula to estimate a reasonable threshold considering solely the corpus size. Because of that, our experiments were conducted considering some arbitrary values of threshold.

Experimenting threshold-based cutoff points to terms extracted from pediatrics corpus

To conduct the experiments using the threshold-based cutoff points over a list os ranked terms, it is necessary to choose threshold values in accordance with the relevance index. It is important to recall that the *tf-dcf*

Table 4 Precision (*P*), recall (*R*), F-measure (*F*), and list sizes using absolute cutoff points

Absolute cutoff points	Bigrams				Trigrams			
	P (%)	*R* (%)	*F* (%)	$\lvert \mathcal{EL} \cap \mathcal{RL} \rvert$	*P* (%)	*R* (%)	*F* (%)	$\lvert \mathcal{EL} \cap \mathcal{RL} \rvert$
100	99	6	12	99	93	3	7	93
200	98	13	22	195	94	7	13	187
300	96	19	31	288	94	11	19	281
400	95	25	39	380	92	14	24	366
500	94	31	46	469	92	17	29	459
600	92	36	52	550	91	21	33	546
700	91	41	57	636	90	24	38	631
800	89	46	61	710	90	27	42	719
900	88	51	65	789	89	30	45	797
1000	86	56	67	855	88	33	48	879
1100	84	61	71	929	87	36	51	960
1200	83	65	73	1001	87	39	54	1045
1300	83	70	76	1074	87	42	57	1128
1400	81	74	77	1129	87	46	60	1217
1500	79	77	78	1179	85	48	61	1277
1600	77	80	79	1234	84	51	63	1350
1700	76	84	79	1285	83	53	65	1416
1800	74	87	80	1331	82	56	66	1483
1900	73	90	81	1388	81	58	68	1548
2000	72	93	*81*	1432	81	61	69	1614
2100	68	94	79	1437	80	63	70	1674
2200	66	94	77	1441	79	65	72	1742
2300	63	94	75	1442	79	68	73	1806
2400	60	94	73	1442	78	70	74	1875
2500	58	94	72	1447	78	73	75	1938
2600	56	94	70	1449	77	75	76	2003
2700	54	95	69	1451	77	78	78	2080
2800	52	95	67	1453	77	81	79	2154
2900	50	95	66	1453	77	84	80	2222
3000	48	95	64	1453	77	86	81	2298
3100	47	95	63	1457	76	89	82	2370
3200	46	95	62	1460	77	92	84	2449
3300	44	95	60	1462	76	94	*84*	2514
3400	43	95	59	1464	74	94	83	2514
3500	42	96	58	1467	72	95	82	2515

Best F-measure values are indicated in italics

index has a similar semantic to the absolute frequency; therefore, we can choose threshold values having in mind the same semantics, i.e., for threshold purposes, the *tf-dcf* index can be understood as the plain number of occurrences.

Figure 3 shows precision, recall, and F-measure for different threshold values from 0 to 15 for bigrams and trigrams of the pediatrics corpus (numeric values given in Table 5). In these experiments, threshold 0 means to consider all extracted terms, i.e., do not discard extracted terms. As the threshold value increases, the list becomes more restrictive.

Analogously to the analysis of the results for absolute cutoff points, Fig. 3 starts showing that once again, the precision and recall curves cross each other. However, the threshold values seem insufficient in terms of granularity to locate the inflection points as were observed in Fig. 2.

The optimal F-measure values were found for threshold 3 (bigrams) and for threshold 2 (trigrams) with values as good as the best absolute cutoff points (2000 bigrams and 3300 trigrams). The adjacent threshold values show significant drops in the F-measure value, illustrating the poor granularity of the threshold-based cutoff points.

In such way, the use of a threshold-based cutoff point alone seems to be too gross to find a balance between precision and recall. Even though, the definition of an arbitrary threshold is often found in many term extraction works [13, 15].

Relative approach

A rarely found alternative [17] to the absolute and threshold approaches is the use of relative cutoff points. In this approach, the number of terms to be considered is a fixed proportion of the extracted terms. As in the other approaches, it is still necessary to arbitrarily choose the percentage of terms to consider, but the biggest advantage of the relative cutoff approach is to offer an option that is independent of the extraction process or corpora size, even thought it is still dependent on the size of the extracted term list.

Specifically, this approach consists in the selection of a percentage of the top ranked extracted terms. Consequently, this approach only needs to choose a numeric value between 0 and 100 %, being those close to 0 % very restrictive cutoff points (delivering high precision) and those close to 100 % (delivering high recall).

Experimenting relative cutoff points to terms extracted from pediatrics corpus

As for the other traditional approaches, we conduct our experiments adopting different relative values for cutoff points. Specifically, we investigate the metrics for cutoff

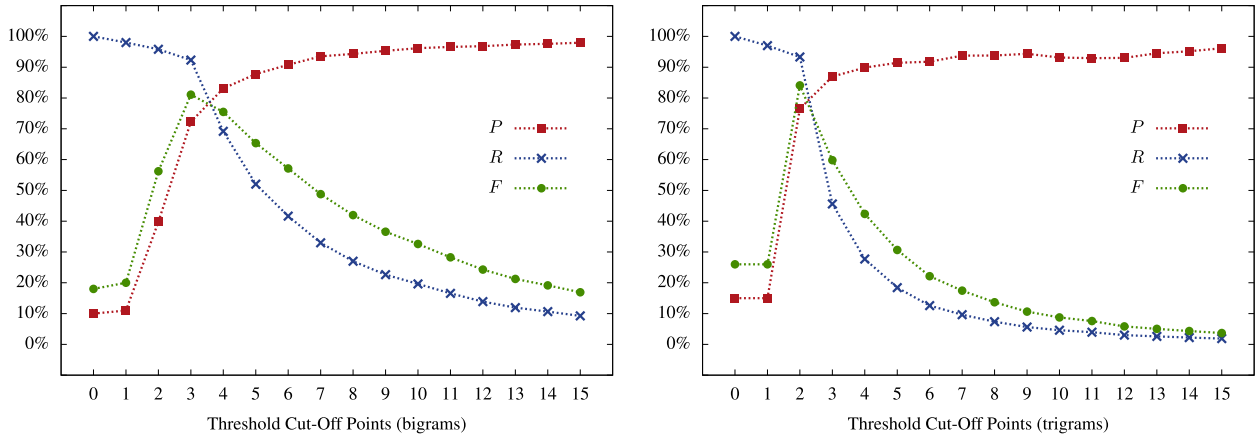

Fig. 3 Precision (*P*), recall (*R*), and F-measure (*F*) using threshold-based cutoff points

points considering from 1 % up to 30 % of bigrams and trigrams extracted from the pediatrics corpus.

Figure 4 depicts the precision, recall, and F-measure values obtained for these relative cutoff points (numeric values at Table 6). The first observation from the metrics in Fig. 4 is that the granularity of results is, like for the absolute cutoff point results, good enough to observe the inflections of precision and recall curves.

The more balanced values were obtained for 13 % of the bigrams and 18 % of the trigrams. In fact, we notice that for relative cutoff points from 8 to 15 % (for bigrams) and from 14 to 22 % (for trigrams), a F-measure value of at

Table 5 Precision (*P*), recall (*R*), F-measure (*F*), and list sizes using threshold-based cutoff points

Threshold cutoff points	Bigrams					Trigrams				
	P (%)	*R* (%)	*F* (%)	$\mid \mathcal{EL} \mid$	$\mid \mathcal{EL} \cap \mathcal{RL} \mid$	*P* (%)	*R* (%)	*F* (%)	$\mid \mathcal{EL} \mid$	$\mid \mathcal{EL} \cap \mathcal{RL} \mid$
0	10	100	18	15,487	1534	15	100	26	18,174	2660
1	11	98	20	13,589	1502	15	97	26	17,227	2577
2	39	96	56	3698	1470	77	93	*84*	3245	2483
3	72	92	*81*	1959	1416	87	46	60	1395	1213
4	83	69	75	1277	1061	90	28	42	820	737
5	88	52	65	909	798	91	18	31	536	490
6	91	42	57	704	639	92	13	22	365	335
7	94	33	49	541	506	94	10	17	273	256
8	94	27	42	439	414	94	7	14	209	196
9	95	23	37	364	347	94	6	11	159	150
10	96	20	33	313	301	93	5	9	131	122
11	97	17	28	263	254	93	4	8	113	105
12	97	14	24	220	213	93	3	6	86	80
13	97	12	21	188	183	95	3	5	73	69
14	98	11	19	167	163	95	2	4	62	59
15	98	9	17	143	142	96	2	4	52	50

Best F-measure values are indicated in italics

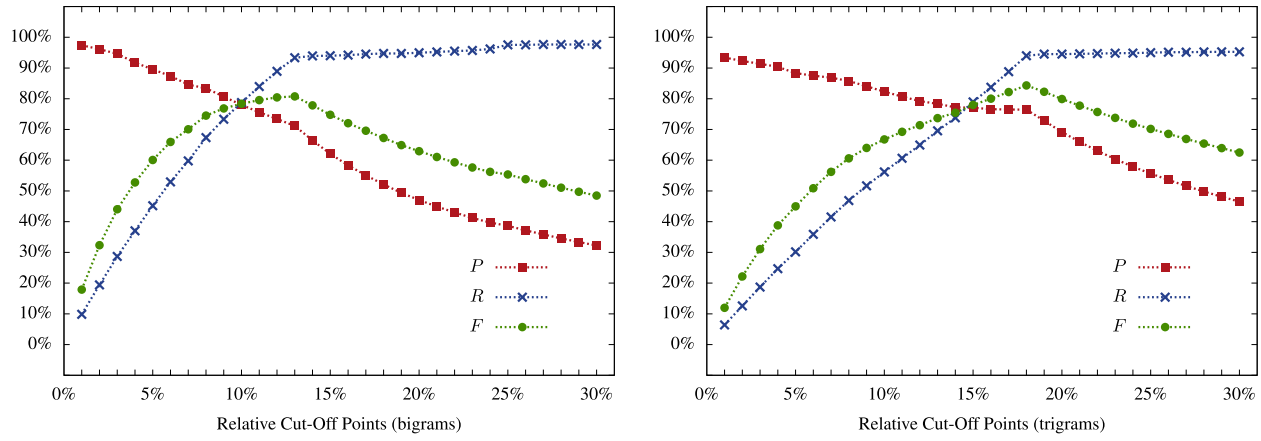

Fig. 4 Precision (*P*), recall (*R*), and F-measure (*F*) using relative cutoff points

least 75 % is achieved. Thus, it is possible to consider that a relative cutoff point around 14 and 15 % was a good trade-off for bigrams and trigrams.

Evidently, these results may not generalize to other extracted term lists, or even other reference lists. Nevertheless, we see no reason to consider that these precision-recall balanced region would not appear in other experiments. It is important to notice that unlike the other traditional approaches, relative cutoff points seem to deliver a larger reasonable balanced region. For instance, absolute and threshold-based cutoff points do not have cutoff values where both bigrams and trigrams have F-measure above 75. Observing Table 4, bigrams cutoff points deliver F-measure above 75 % for absolute values between 1300 and 2300, while trigrams F-measure above 75 % is found for values from 2500 through 3500. Observing Table 5, bigrams cutoff points deliver F-measure above 75 % for values 3 and 4, while for trigrams only threshold 2 delivers F-measure above 75 %.

Pushing the analysis a little bit further than just the numerical analysis of the metrics over bigrams and trigrams in comparison to the reference list, we observe that larger terms, i.e., terms with four or more words are likely to have low values of the relevance index. Specifically, we observe that even the 15 % top ranked 4-grams contain terms that were found one single time in the corpus. Such rare terms are even more likely to be found in the top ranked positions for lists as the number of words of the terms increase. For instance, among the 1567 extracted 8-grams, only 38 terms were found more than once in the corpus, and for the 1093 extracted 9-grams, the number of terms found at least twice in the corpus was only 30.

Consequently, it seems that even relative cutoff points cannot be considered alone as an effective way to estimate a precision-recall balanced point for all extracted term lists, despite the finding of reasonable values of F-measure (above 75 %) for values of 14 and 15 % in bigrams and trigrams lists.

Ultimately, these experiments were our motivation to look for an alternative way to estimate cutoff points without being limited to the traditional approaches of absolute, threshold-based, and relative values. Hence, we suggest the adoption of a policy to choose a cutoff point for virtually any extracted term list.

Methods

The results for the traditional approaches presented in the previous section have shown some cutoff point values delivering balanced percentages of precision and recall, i.e., higher F-measures values. However, the goal of this paper is more than finding out good parameters for the bigrams and trigrams of the pediatrics corpus. Our goal is to propose a policy that has potential to be applied to any extracted term lists delivering satisfactory results.

Finding a single policy, regardless of the approach, does not seem possible since even for the two lists submitted to numerical experiments, it was difficult to define one single situation delivering optimal F-measure values to bigrams and trigrams. Therefore, we propose a hybrid method to estimate a reasonable cutoff point, and considering the experiments conducted in this paper, we will assume that an F-measure above or equal to 75 % is sufficiently balanced. Specifically, we propose a policy combining threshold-based and relative approaches.

It is important to clarify that the use of absolute cutoff points is not feasible, since in a generalization context, it is not possible to suggest any arbitrary fixed number of terms to consider. The approaches based on relative and threshold cutoff points are naturally flexible to the problem, since the size of the extracted list is sufficiently dependent on the corpus size.

Following the same reasoning, it is not enough to consider a same threshold to all extracted term lists;

Table 6 Precision (*P*), Recall (*R*), F-measure (*F*) and List Sizes using relative cutoff points

Relative cutoff points (%)	Bigrams					Trigrams				
	P (%)	*R* (%)	*F* (%)	$\mid \mathcal{EL} \mid$	$\mid \mathcal{EL} \cap \mathcal{RL} \mid$	*P* (%)	*R* (%)	*F* (%)	$\mid \mathcal{EL} \mid$	$\mid \mathcal{EL} \cap \mathcal{RL} \mid$
1	97	10	18	155	151	93	6	12	182	170
2	96	19	32	310	298	92	13	22	363	335
3	95	29	44	465	440	91	19	31	545	498
4	92	37	53	619	568	90	25	39	727	657
5	90	45	60	774	693	88	30	45	909	803
6	87	53	66	929	812	88	36	51	1090	954
7	85	60	70	1084	917	87	42	56	1272	1105
8	83	67	75	1239	1033	86	47	61	1454	1247
9	81	73	77	1394	1125	84	52	64	1636	1375
10	78	79	78	1549	1208	82	56	67	1817	1495
11	76	84	80	1704	1288	81	61	69	1999	1613
12	73	89	80	1858	1364	79	65	71	2181	1728
13	71	93	*81*	2013	1432	78	70	74	2363	1851
14	66	94	78	2168	1441	77	74	75	2544	1964
15	62	94	75	2323	1442	77	79	78	2726	2101
16	58	94	72	2478	1445	77	84	80	2908	2228
17	55	95	70	2633	1450	76	89	82	3090	2363
18	52	95	67	2788	1453	76	94	*84*	3271	2501
19	49	95	65	2943	1453	73	95	82	3453	2515
20	47	95	63	3097	1457	69	95	80	3635	2515
21	45	95	61	3252	1461	66	95	78	3817	2518
22	43	96	59	3407	1465	63	95	76	3998	2520
23	41	96	58	3562	1468	60	95	74	4180	2523
24	40	96	56	3717	1476	58	95	72	4362	2524
25	39	98	55	3872	1496	56	95	70	4544	2528
26	37	98	54	4027	1496	54	95	69	4725	2532
27	36	98	52	4181	1498	52	95	67	4907	2532
28	35	98	51	4336	1498	50	95	65	5089	2534
29	33	98	50	4491	1498	48	95	64	5270	2535
30	32	98	48	4646	1498	46	95	62	5452	2535

Best F-measure values are indicated in italics

however, it seems reasonable to adopt a small threshold to avoid considering terms that would drop significantly the precision values. Therefore, we propose to adopt as the basic step of the proposed policy to consider only terms that have a *tf-dcf* value equal or superior to 2.

This choice of a threshold 2 for *tf-dcf* is conservative, since for bigrams of the pediatrics corpus, a threshold 3 would be a better choice. It is worthy to remember that for the trigrams of pediatrics corpus, the choice of threshold 3 would discard an important number of valid trigrams.

A more reliable part of the proposed policy is the adoption of a relative cutoff point of 15 % of the extracted terms. As seen before, a 15 % relative cutoff point would be a good trade-off for bigrams and trigrams of the pediatrics corpus. It is important to remind that the purpose of our proposed policy is to offer a practical way to estimate a good cutoff point to a given extracted term list. Such estimation was not possible using only traditional approaches as absolute or threshold-based approaches.

Even for relative cutoff points, where the experiments with values 14 and 15 % delivered reasonably high F-measure values (above 75 %) for bigrams and trigrams, we had observe that terms with more words (4-grams and up), therefore likely to be less relevant, were encountered in the top ranked 15 % extracted terms. Consequently, we suggest the policy considering, initially, the use of a 15 % relative cutoff point, followed by a restriction accepting only terms with a relevance index above or equal to 2, i.e., a composition with a threshold-based cutoff point.

Summarizing our cutoff policy, we propose to:

- consider the extracted terms as a set of lists with terms with the same number of words, duly ranked by *tf-dcf* relevance index;
- discard 85 % of the extracted terms of each list, keeping the top 15 % terms with the higher *tf-dcf* values;
- discard among the remaining 15 % of each list, the terms with a *tf-dcf* value inferior to 2.

Results and discussion

Applying this policy to the extracted terms of the five corpora described in the "The corpora and reference lists" section, we obtain the number of terms presented in Table 7. In order to have a graphic impression of the proposed cutoff point application over the five corpora, Fig. 5 depicts the reduction in terms of the number of remaining terms. The left hand side bars depict the number of terms as outputted by the extraction (Table 3), while right hand side bars (identified with an asterisk) depict the number of terms considered as relevant to each domain (Table 7).

Table 7 Number of identified relevant terms

Corpus	Ped	SM	DM	PP	Geo
Unigrams	892	648	630	654	1152
Bigrams	2323	2116	2221	2145	4616
Trigrams	2726	2831	2871	2996	5582
4-grams	1192	2176	2104	2072	4544
5-grams	560	1836	1852	1602	3281
6-grams	221	1000	949	739	1990
7-grams	124	690	728	458	1267
8-grams	70	407	450	295	855
9-grams	52	281	309	188	562
M-grams	113	599	705	442	1326
Total	8273	12,584	12,819	11,591	25,175

As mentioned before, due to a lack of reference lists, we can only observe the reduction of the sheer number of remaining terms for all the extracted lists. It is also noticeable that lists of large terms were subject to reductions below 15 %, since such lists have a large number of terms found only once in their corpora. Nevertheless, for the bigrams and trigrams of the pediatrics corpus, it is possible to compare the remaining terms with the reference lists. Such comparison delivers the following values of precision, recall, and F-measure:

$$P = 62\,\%\ \ R = 94\,\%\ \ F = 75\,\% \quad \text{for bigrams}$$

$$P = 77\,\%\ \ R = 79\,\%\ \ F = 78\,\% \quad \text{for trigrams}$$

Conclusions

The main purpose of this paper was to propose a policy to estimate a reasonably balanced cutoff point that may be applicable to virtually any list of extracted terms. The first difficulty in such kind of work was to find closely similar work, since much of the available literature dealing with the choice of cutoff points do not focus on alternatives to estimate a fairly balanced cutoff point. In fact, at the authors' best knowledge, most of authors limit themselves to briefly justify the adoption of a specific value for a traditional cutoff approach. Therefore, we believe that our work is an original contribution by focusing on the alternatives to estimate cutoff points and proposing a policy to be employed by practitioners and researchers that have to face the choice of a cutoff point.

In this paper, several experiments with cutoff points were conducted. For all these experiments, we started with a qualified list of extracted terms, duly ranked with an effective relevance index. Unfortunately, among all corpora employed, only the pediatrics corpus had a reliable and independently developed reference list to be used to the numerical analysis.

In such context, we were able to propose a hybrid method to identify the relevant terms to domain. The results were informally analyzed, and it seems that the considered extracted terms are rather relevant and specific to each domain. However, a formal analysis of these lists is an interesting future work. Such work was not yet carried out, due to the lack of domain specialists. Nevertheless, the practical application of the proposed policy to reduce the extracted terms was empirically approved by reducing with quality the term extracted in applied linguistic projects conducted by TEXTCC research group leaded by Maria José Finatto at the Federal University of Rio Grande do Sul (UFRGS).

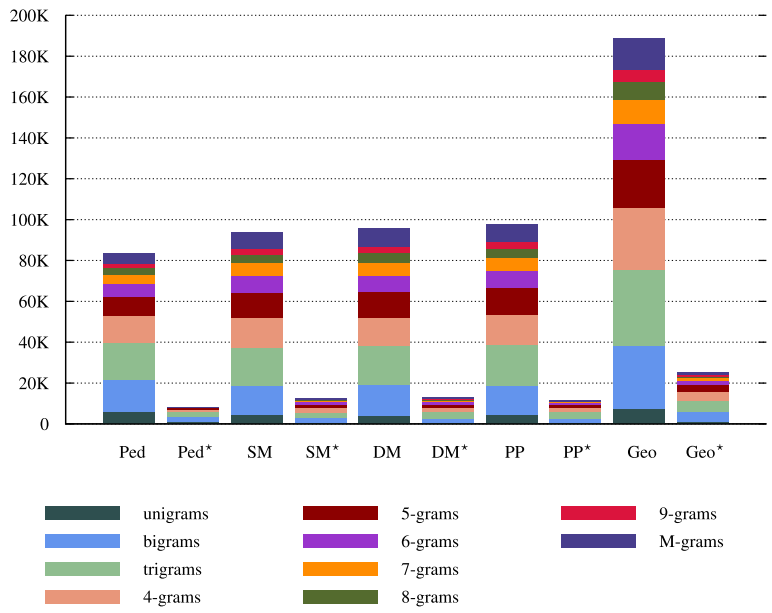

Fig. 5 Comparative analysis of the proposed policy application

Another interesting future work to consider is to look for more sophisticated analysis of the terms, adopting, besides the *tf-dcf* index, other indications of term relevance. We could gather term information regarding its grammatical role, for instance, terms that are subject of sentence may be more inclined to be relevant. Such kind of initiative would correspond to a multi-valued analysis that can be carried out even using machine learning tools [38]. Among such theoretical tools, it is natural to employ classification methods [39–41] to improve the quality of the relevant term selection.

Despite of such ambitious future work, the current state of the policy proposed in this paper is already being employed in several research initiatives for Brazilian Portuguese. However, new controlled experiments would improve the confidence in our policy. Other suggestions of future work include the use of other extraction tools, e.g., [5, 13, 20], or even other relevance indices, e.g., [10, 11, 42]. The main obstacle to these works is the availability of corpora accompanied with reference lists.

It is also important to mention that the experience with a larger set of corpora and reference lists may lead to a revision of the proposed policy. After all, the choice of a good cutoff point is dependent of many factors, as the quality of the extraction procedure, and the quality of relevance ranking of terms. Consequently, the results presented in our paper may not generalize to other extraction efforts, including similar extraction procedures applied to other languages.

At the authors' best knowledge, there are very few similar works to evaluate the options of cutoff points in general and none for term extraction from Portuguese corpora. Therefore, this paper offers some help to researchers and practitioners to define cutoff points to their own experiments.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
This work was part of LL PhD dissertation and RV was her advisor. LL carried out most of the investigation, implementation, experiments, and writing. RV gave important ideas to the project and also contributed in the writing and revision of this manuscript. Both authors read and approved the final manuscript.

Acknowledgements
Lucelene Lopes is funded by CAPES and FAPERGS DOCFIX grant. Renata Vieira is partially funded by CNPq.

Received: 2 May 2014 Accepted: 9 June 2015
Published online: 14 July 2015

References
1. Maedche A, Staab S (2001) Ontology learning for the semantic web. IEEE Intell Syst 16(2):72–79. doi:10.1109/5254.920602
2. Cimiano P (2006) Ontology learning and population from text: algorithms, evaluation and applications. Springer
3. Lopes L (2012) Extração automática de conceitos a partir de textos em língua portuguesa. PhD thesis, PUCRS University - Computer Science Department, Porto Alegre, Brazil
4. Conrado M, Pardo T, Rezende S (2013) A machine learning approach to automatic term extraction using a rich feature set. In: Proceedings of the 2013 NAACL HLT Student Research Workshop. Association for Computational Linguistics. pp 16–23. http://aclweb.org/anthology/N13-2003
5. Banerjee S, Pedersen T (2003) The design, implementation and use of the Ngram statistics package. In: CICLing'03 Proceedings of the 4th international conference on Computational linguistics and intelligent text processing. Springer-Verlag, Berlin, Heildelberg. pp 370–381

6. Lopes L, Fernandes P, Vieira R, Fedrizzi G (2009) ExATO lp—an automatic tool for term extraction from Portuguese language corpora. In: Proceedings of the 4th Language & Technology Conference (LTC '09). Faculty of Mathematics and Computer Science of Adam Mickiewicz University. pp 427–431
7. Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge, USA
8. Chung TM (2003) A corpus comparison approach for terminology extraction. Terminology 9(2):221–246
9. Kit C, Liu X (2008) Measuring mono-word termhood by rank difference via corpus comparison. Terminology 14(2):204–229
10. Park Y, Patwardhan S, Visweswariah K, Gates SC (2008) An empirical analysis of word error rate and keyword error rate. In: INTERSPEECH. pp 2070–2073
11. Kim SN, Baldwin T, Kan MY (2009) Extracting domain-specific words—a statistical approach. In: Pizzato L, Schwitter R (eds). Proceedings of the 2009 Australasian Language Technology Association Workshop. Australasian Language Technology Association, Sydney, Australia. pp 94–98
12. Pao ML (1978) Automatic text analysis based on transition phenomena of word occurrences. J Am Soc Inform Sci 29(3):121–124. doi:10.1002/asi.4630290303
13. Bourigault D, Lame G (2002) Analyse distributionnelle et structuration de terminologie. application a la construction d'une ontologie documentaire du droit. Traitement automatique des langues 43(1):129–150
14. Milios E, Zhang Y, He B, Dong L (2003) Automatic term extraction and document similarity in special text corpora. In: 6th Conference of the Pacific Association for Computational Linguistics, Halifax, Nova Scotia, Canada. pp 275–284. http://users.cs.dal.ca/~eem/res/pubs/pubs/pacling2003.pdf
15. Wermter J, Hahn U (2005) Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In: Proc. of the Conf. on Human Language Technology. HLT '05. Assoc. for Comput. Ling., Stroudsburg, PA, USA. pp 843–850. doi:10.3115/1220575.1220681. http://dx.doi.org/10.3115/1220575.1220681
16. Yang H, Callan J (2008) Ontology generation for large email collections. In: Proceedings of the 2008 International Conference on Digital Government Research. dg.o '08. Digital Government Society of North America. pp 254–261. http://dl.acm.org/citation.cfm?id=1367832.1367875
17. Maynard D, Li Y, Peters W (2008) NLP techniques for term extraction and ontology population. In: Proceedings of the 2008 Conference on Ontology Learning and Population. IOS Press, Amsterdam, The Netherlands. pp 107–127. http://dl.acm.org/citation.cfm?id=1563823.1563834
18. Lopes L, Vieira R, Finatto MJ, Zanette A, Martins D, Ribeiro Jr LC (2009) Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. RECIIS 3(1):72–84
19. Evert S (2010) Google web 1T 5-grams made easy (but not for the computer). In: Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop. WAC-6 '10. Association for Computational Linguistics, Stroudsburg, PA, USA. pp 32–40. http://dl.acm.org/citation.cfm?id=1868765.1868770
20. Lopes L, Oliveira LH, Vieira R (2010) Portuguese term extraction methods: comparing linguistic and statistical approaches. In: PROPOR 2010 – International Conference on Computational Processing of Portuguese Language
21. Awawdeh R, Anderson T (2010) Improving search in tag-based systems with automatically extracted keywords. In: Bi Y, Williams M.-A. (eds). Knowledge Science, Engineering and Management. Lecture Notes in Computer Science. Springer Vol. 6291. pp 378–387. http://dx.doi.org/10.1007/978-3-642-15280-1_35
22. Ding J, Zhou S, Guan J (2011) miRFam: an effective automatic miRNA classification method based on n-grams and a multiclass SVM. BMC Bioinformatics 12(1):216. doi:10.1186/1471-2105-12-216
23. Lopes L, Fernandes P, Vieira R (2012) Domain term relevance through tf-dcf. In: Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012). CSREA Press, Las Vegas, USA. pp 1001–1007
24. van Rijsbergen CJ (1975) Information retrieval. Butterworths, London, UK
25. Coulthard RJ (2005) The application of corpus methodology to translation: the JPED parallel corpus and the Pediatrics comparable corpus. PhD thesis, UFSC, Florianópolis, Brazil
26. Lopes L, Vieira R (2013) Building domain specific parsed corpora in portuguese language. In: Proceedings of the X National Meeting on Artificial and Computational Intelligence (ENIAC). pp 1–12
27. Bick E (2000) The parsing system PALAVRAS: automatic grammatical analysis of portuguese in constraint grammar framework. PhD thesis, Arhus University
28. Lopes L, Vieira R (2012) Improving quality of portuguese term extraction. In: PROPOR 2012 – International Conference on Computational Processing of Portuguese Language
29. Boreczky JS, Rowe LA (1996) Comparison of video shot boundary detection techniques. J Electron Imaging 5(2):122–128
30. Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M (2000) Automatic extraction of protein interactions from scientific abstracts. In: Pacific Symposium on Biocomputing Vol. 5. pp 538–549
31. Fernandes P, Lopes L, Ruiz DDA (2010) The impact of random samples in ensemble classifiers. In: SAC'10: Proceedings of the 2010 ACM Symposium on Applied Computing. ACM, New York, USA. pp 1002–1009. doi:10.1145/1774088.1774300
32. Witten IH, Moffat A, Bell TC (1999) Managing gigabytes: compressing and indexing documents and images. Morgan Kaufmann, San Francisco
33. Hulth A (2004) Enhancing linguistically oriented automatic keyword extraction. In: Proceedings of HLT-NAACL 2004: Short Papers. HLT/NAACL. ACM, New York, USA. pp 17–20
34. Lopes L, Vieira R, Finatto MJ, Martins D (2010) Extracting compound terms from domain corpora. J Braz Comput Soc 16:247–259. doi:10.1007/s13173-010-0020-4
35. da Silva Conrado M, Felippo A, Salgueiro Pardo T, Rezende S (2014) A survey of automatic term extraction for brazilian portuguese. J Braz Comput Soc 20(1):12. doi:10.1186/1678-4804-20-12
36. Spärck-Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc 28(1):11–21. doi:10.1108/eb026526
37. Zipf GK (1935) The psycho-biology of language—an introduction to dynamic philology. Houghton-Mifflin Company, Boston, USA
38. Mitchell T (1997) Machine learning. McGraw-Hill
39. Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: bagging, boosting and variants. Mach Learn 36(1–2):105–139
40. Lopes L, Scalabrin EE, Fernandes P (2008) An empirical study of combined classifiers for knowledge discovery on medical data bases. In: APweb 2008 Workshops (LNCS 4977). pp 110–121
41. Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques. 3rd edn. Morgan Kaufmann
42. Wu HC, Luk RWP, Wong KF, Kwok KL (2008) Interpreting TF-IDF term weights as making relevance decisions. ACM Trans Inform Syst 26:13–11337. doi:10.1145/1361684.1361686