Merley S. Conrado, Thiago A. S. Pardo, and Solange O. Rezende

The Main Challenge of Semi-Automatic Term Extraction Methods

Abstract: Term extraction is the basis for many tasks such as building of taxonomies, ontologies and dictionaries, for translation, organization and retrieval of textual data. This paper studies the main challenge of semi-automatic term extraction methods, which is the difficulty to analyze the rank of candidates created by these methods. With the experimental evaluation performed in this work, it is possible to fairly compare a wide set of semi-automatic term extraction methods, which allows other future investigations. Additionally, we discovered which level of knowledge and threshold should be adopted for these methods in order to obtain good precision or F-measure. The results show there is not a unique method that is the best one for the three used corpora.

1 Introduction

Term extraction aims to identify a set of terminological units that best represent a specific domain corpus. Terms are fundamental in tasks for the building of (i) traditional lexicographical resources (such as glossaries and dictionaries) and (ii) computational resources (such as taxonomies and ontologies). Terms are also the basis for tasks such as information retrieval, summarisation, and text classification.

Traditionally, semi-automatic term extraction methods select candidate terms based on some linguistic knowledge [1]. After that, they apply measures or some combinations of measures (and/or heuristics) to form a rank of candidates [1–5]. Then, domain experts and/or terminologists analyze the rank in order to choose a threshold at which the candidates that have values above this threshold are selected as true terms. This analysis is subjective because it depends on personal human interpretation and domain knowledge, and it requires time to perform it.

Merley S. Conrado, Thiago A. S. Pardo, Solange O. Rezende: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, e-mail: {merleyc,taspardo,solange}@icmc.usp.br
Grants 2009/16142-3 Sao Paulo Research Foundation (FAPESP).

This subjectivity in analyzing the rank of candidates is the main challenge of semi-automatic term extraction methods.

Despite this challenge, the comparison of different extractors is a gap identified in the literature [6] since each research uses different corpora, preprocessing tools, and evaluation measures.

This paper aims to demonstrate how difficult it is to choose which candidates in a rank should be considered terms. For that, we perform and compare a wide set of term extraction methods that cover, separately, the three levels of knowledge used for term extraction: statistical, linguistic, and hybrid knowledge. We consider the same scenario when realizing the comparison of the extraction, i.e., we use the same corpora, the same textual preprocessing, and the same way of assessing results.

Our main contribution remains on demonstrate how difficult it is to analyze the rank of candidates created by semi-automatic term extraction methods. Additionally, in some cases, we discover which level of knowledge and threshold should be adopted for semi-automatic term extraction. Finally, with the experimental evaluation performed in this work, it is possible to fairly compare a wide set of semi-automatic term extraction methods, which allows other future investigations.

Next section describes the traditional term extraction methods and related work. Section 3 presents the measures used in the literature to extract terms, our experiments, results, and discussions. Finally, Section 4 presents conclusions and future work.

2 Related Work

Traditional term extraction select candidate terms based on some linguistic knowledge [1], e.g., to maintain only candidates that are nouns. Then, each candidate receives a value calculated by some statistical or hybrid measure or some combination of measures (and/or heuristics) [2–5]. These measures may be the candidate frequency or the accounting of distribution (e.g., the *weirdness* measure [7]) or occurrence probability (e.g., *glossEx* [3]) of candidates in a domain corpus and in a general language corpus. The candidates are ranked according to their values and those that have a minimum value of threshold in this rank are considered as potential terms of a specific domain. Domain experts and/or terminologists decide a threshold, which may be a fixed percentage or number of candidates to be considered. Manually choosing a threshold is not the best option since it has a high human cost. Semi-automatically choosing a threshold (e.g., an

expert decision considers a fixed number of candidates) is not the best option as well, however it requires less human presence.

Luhn [8] and LuhnDF [9] are semi-automatic methods that plot histograms from candidate terms based on, respectively, candidate frequencies (*tf*) and document frequencies (*df*). These histograms facilitate the visualization of any possible pattern that candidates may follow and, then, the histograms help to determine a threshold. Salton et al. [10] propose another method that suggests to consider candidates that have *df* between 1% and 10% of the total number of documents in a corpus. The TRUCKS approach [11] suggests to consider only the 30% first candidates in the rank created according to the *nc-value* measure [2]. There are also studies that consider different fixed values of candidates [2, 12–14]. Other studies explore a variation of result combinations (precision and recall, usually) [5, 15]. There are studies that compare some measures used to extract terms [14, 16, 17].

All these studies use assorted ways to select a threshold in the candidate rank using different corpora and extracting, sometimes, simple terms and, other times, complex terms. This paper evaluates a wide set of simple term extraction methods comparing different thresholds and considering the same scenario of extraction. To the best of our knowledge, there is no reseaarch that evaluates this set of methods in a same scenario.

3 Evaluation of Traditional Term Extraction Methods

We performed and compared different semi-automatic extraction methods of simple terms. For the experiments, we use three corpora of different domains in the Portuguese language. The DE corpus [18] has 347 texts about distance education; the ECO corpus [13] contains 390 texts of the ecology domain; and the Nanoscience and Nanotechnology (N&N) corpus [19] has 1,057 texts.

In order to minimize the interference of the preprocessing in the term extraction, we carefully preprocessed all the texts as follows:

1. We identify the encoding of each document in order to correctly read the words. Without this identification, we would incorrectly find "*p*" and "*s-teste*" instead of "*pós-teste*" (*post-test*). We also transformed all letters to lowercase.

2. We remove stopwords[1] and special characters, such as \, |, , and ^.
3. We clean the texts, such as to convert "*humanos1*" to "*humanos*" (*humans*), "*que_a*" to "*que*" (*that*) and "*a*" (*an*), and "*tam-be'm*" to "*também*" (*too*). In these examples, *humans* would be a candidate term and the other words might be removed (because they are stopwords).
4. We identify part-of-speech (*pos*) tags using the Palavras parser [20].
5. We normalize the words using stemming[2].
6. As domain experts, we also consider compound terms (e.g., "*bate-papo*" – *chat*) as simple terms.

At the end of preprocessing, we obtained 9,997, 16,013, and 41,335 stemmed candidates, respectively, for the ECO, DE, and N&N corpora.

3.1 Term Extraction Methods

Each preprocessed unigram of each corpus was considered a candidate term. For all candidates of each corpus, we evaluated, separately, 21 simple term extraction methods. These methods are divided into three levels of knowledge: statistical, linguistic, and hybrid knowledge.

Each statistical method applies some statistical measure (Table 1) in order to quantify termhood, i.e., to express how much a candidate is related to the corpus domain. In Table 1, D is the number of documents in a corpus (c); f_{d_x,t_j} is the frequency of t_j (j^{th} candidate term) in the d_x (x^{th} document); $1 - p(0;\lambda_j)$ is the Poisson probability of a document with at least one occurrence; and W is the amount of corpus words.

We used the *tv, tvq*, and *tc* measures – normally applied to the attribute selection tasks (identified by *) – because they were considered relevant measures for extracting terms in the work of [26]. We used the *n-gram length* measure to verify if terms of a specific domain have different length (in characters) of words in general language or of terms in another domain. E.g., the longest term of the ecology domain ("*territorialidade*" – *territoriality*) contains 16 characters, while the longest term of the N&N domain ("*hidroxipropilmetilcelulose*" – *hydroxypropylmethylcellulose*) has 26 characters.

1 Stoplist and Indicative Phrase list are available at http://sites.labic.icmc.usp.br/merleyc/ThesisData/

2 PTStemmer: a stemming toolkit for the Portuguese language – http://code.google.com/p/ptstemmer/

Tab. 1: The statistical measures.

Acronym	Measure	Equation
1. *n-gram length*	Number of characters in a *n*-gram	–
2. *tf*	Term frequency	$\sum_{x=1}^{D} f_{d_x,t_j}$
3. *rf*	Relative frequency	tf_{t_j}/W
4. *atf*	Average term frequency	tf_{t_j}/df_{t_j}
5. *ridf*	Residual inverse document frequency [21]	$\left(\log_2\left(\frac{D}{df_{t_j}}\right)\right) - \log_2\left(\frac{1}{1-p(0;\lambda_j)}\right)$
6. *df*	Document frequency	$\sum_{x=1}^{D}(1 \mid f_{d_x,t_j} \neq 0)$
7. *tf-idf*	Term frequency – inverse document frequency [22]	$tf_{d_x,t_j} \times \log\left(\frac{D}{df_{t_j}}\right)$
8. *tv**	Term variance [23]	$\sum_{x=1}^{D}[f_{d_x,t_j} - \bar{f}_{t_j}]^2$
9. *tvq**	Term variance quality [24]	$\sum_{x=1}^{D} f_{d_x,t_j}^2 - \frac{1}{D}\left[\sum_{x=1}^{D} f_{d_x,t_j}\right]^2$
10. *tc**	Term contribution [25]	$\sum_{x=1}^{D}\sum_{y=1}^{D} f_{d_x,t_j} \times idf_{t_j} \times f_{d_x,t_j} \times idf_{t_j}$

We expect that the statistical measures *tf, rf, atf*[3]*, ridf*[3], and *tf-idf* help to identify frequent domain terms. The *df* measure counts in how many documents in the corpus the candidate terms occur. Then, we expect that *df* identifies candidates that represent the corpus by assuming they occur in at least a minimal amount of documents.

There are frequent terms in the corpus, but there are also rare terms or those that have the same frequency of non-terms. The statistical measures are not able to identify these differences. For this reason, we also evaluated four linguistic methods of extracting terms that follow different ways for obtaining linguistic knowledge aiming to identify terms. We used the annotation provided by the

3 We used the implementation available at https://code.google.com/p/jatetoolkit/

Palavras parser [20]. The first linguistic method of extracting terms considers terms are *noun phrases*. The second linguistic method (*pos*) assumes terms are nouns. The third linguistic method (*k_noun_phrase*) considers terms are kernels of noun phrases, since they represent the meaningful kernels of terms, as discussed in [27]. For instance, the noun phrase "*Os autótrofos terrestres*" (*The terrestrial autotrophics*) belongs to the ecology domain and the experts of this domain consider only the kernel of this phrase as a term, i.e., *autotrophics*. It is also expected that, if the texts' authors define or describe some word, the latter is important for the text domain and, for this reason, this word is possibly a term. For example, in "*A segunda possibilidade de coalescência é descrita por...*" (*The second possibility of coalescing is described by*), the term *coalescence* from the nanoscience and nanotechnology domain may be identified because it is near the indicative phrase *is described by*. Thus, the fourth linguist extraction method (*ip*) considers terms are those that occur near some indicative phrase.

When considering only statistical measures, it is not possible to identify, e.g., terms with similar frequencies to non-terms. Similarly, when considering only linguistic measures, it is not possible to identify terms that follow the same patterns of non-terms. Then, we expect that to consider the statistical and linguistic knowledge together may optimize the term identification. For example, the verb *to propose* can be quite frequent in technical texts from a specific domain. Although, if we assume a term should follow both a linguistic pattern (e.g., being a noun) and a statistical pattern (e.g., being frequent in the corpus), this verb – even if it is frequent – will be correctly identified as non-term.

For this reason, we also evaluated six hybrid methods of extracting terms found in the literature. Two of these methods use statistical and linguistic knowledge to identify terms by applying, separately, the *c-value* and *nc-value* measures (Table 2). The other hybrid methods statistically analyze information of general language corpus by applying, separately, the *gc_freq., weirdness³, thd, tds*, and *glossEx³* measures (Table 2). The latter methods assume, in general, that terms have very low frequencies or that do not appear in general language corpus. We used the NILC[4] corpus of general language with 40 million words. For the identification of phrases used in the *c-value* and *nc-value* measures and to find kernels of noun phrases and part-of-speech tags, we used the annotation provided by the Palavras parser [20].

In Table 2, $r_{t_j}^{(c)}$ is the ordination value of the candidate t_j in a specific domain corpus c; g is a general language corpus; $td(t_j)$ is domain specificity of t_j; and $tc(t_j)$ is the cohesion of the t_j candidate. For *c-value*, T_{t_j} is the candidate set with length

4 NILC Corpus – http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm

Tab. 2: The hybrid measures.

Acronym	Measure	Equation
1. *gc_freq.*	Term frequency in a general language corpus	$\sum_{x=1}^{D^{(g)}} f_{d_x,t_j}$
2. *weirdness*	Term distribution in a domain corpus and general language corpus [7]	$\left(tf_{t_j}^{(c)}/W^{(c)}\right) / \left(tf_{t_j}^{(g)}/W^{(g)}\right)$
3. *thd*	Termhood index: weighted term frequency in a domain corpus and general language corpus [4]	$\left(r_{t_j}^{(c)}/W^{(c)}\right) - \left(r_{t_j}^{(g)}/W^{(g)}\right)$
4. *tds*	Term domain specificity [28]	$\left(P(t_j^{(c)})/P(t_j^{(g)})\right) = \frac{\textit{prob. in domain } c}{\textit{prob. in corpus } g}$
5. *glossEx*	Occurrence probability of a term in a domain corpus and general language corpus [3]	$a * td(t_j) + b * tc(t_j)$, default a=0.9, b = 0.1.
6. *c-value*	Frequency of a andidate with certain *pos* in the domain corpus	$(1 + \log_2 \lvert t_j \rvert) \times \log_2 \lvert t_j \rvert \times tf(t_j)$, if $t_j \notin$ a V; otherwise
	and its frequency inside other longer candidate terms [2, 29]	$(1 + \log_2 \lvert t_j \rvert) \times \log_2 \lvert t_j \rvert \left(tf(t_j) - \frac{1}{P(T_{t_j})} \sum_{b \in T} f(b)\right)$.
7. *nc-value*	The context in which the candidate occurs is relevant [2]	$0.8\ \textit{c-value}_{t_j} + 0.2 \sum_{b \in C_{t_j}} f_{t_j}(b)(t(w)/nc)$

in grams larger than t_j and that contains t_j; $P(T_{t_j})$ is the number of such candidates (types) including the type of t_j; and V is the set of neighbours of t_j. For *nc-value*, C_{t_j} is the set of words in the context of the candidate t_j; b is a context word for the candidate t_j; $f_{t_j}(b)$ is the occurrence frequency of b as a context word for t_j; w is the calculated weight for b as a context word; and nc is the total number of candidates considered in the corpus.

3.2 Results and Discussion

For the term extractors that use statistical and hybrid measures (Tables 1 and 2), which result in continuous values, the candidates are decreasingly ordered by

Tab. 3: Extraction Method that uses *tf*: The ECO corpus.

#Cand.	#ET	P(%)	R(%)	FM(%)
50	21	42,00	7,00	12,00
100	32	32,00	10,67	16,00
150	45	30,00	15,00	20,00
200	52	26,00	17,33	20,80
250	61	24,40	20,33	22,18
300	69	23,00	23,00	23,00
350	77	22,00	25,67	23,69
400	85	21,25	28,33	24,29
		•••		
9800	298	3,04	99,33	5,90
9850	298	3,03	99,33	5,87
9900	298	3,01	99,33	5,84
9950	299	3,01	99,67	5,83

their chance of being terms, considering the value obtained in the calculation of each measure. In this way, those candidates that have the best values in accordance with a certain measure are at the top of the rank. Then, we calculated precision, recall, and F-measure considering different cutoff points in this rank, starting with the first 50 ordered candidates, the first 100 candidates, 150, and so on until the total number of candidates of each corpus. To calculate precision, recall, and f-measure, we used gold standards of the ECO, DE, and N&N corpora, which contain, respectively, 322, 118, and 1,794 simple terms, and, after stemming them, we have 300, 112, and 1,543 stemmed terms. The authors of the ECO corpus built its gold standard with the unigrams that occur, at the same time, in 2 books, 2 specialized glossaries, 1 online dictionary, all related to the ecology domain. Differently, an expert of the distance education domain decided which noun phrases, that satisfied certain conditions, should be considered terms. For the elaboration of the N&N gold standard, its authors applied statistical methods to selected candidate terms, then manually a linguist removed some of them and, finally, an expert decided which of these candidates are terms.

Table 3 shows the results[5] of extraction method that uses the *tf* measure with the ECO corpus. This table also highlights the highest precision (P (%) = 42.00), recall (R (%) = 99.67), and F-measure (FM (%) = 24.29) achieved when using the *tf*

5 All the term extraction results for the three corpora using each measure are available at http://sites.labic.icmc.usp.br/merleyc/ThesisData/.

measure with, respectively, the first 50, 400, and 9,950 candidates of the rank (# Cand.).

The linguistic methods result a unique and fixed total number of extracted candidates with which we calculated the evaluation measures. For example, all the 5,030 noun phrases identified in the ECO corpus are considered extracted candidates. Considering there are 279 terms (#ET) in these candidates, 5.55% of the extracted candidates are terms (P(%)), identified 93% of the total of terms in this domain (R(%)), and achieved a balance between precision and recall equal to 10.47% (FM(%)).

Table 4 shows the best and worst results[5] for precision, recall, and F-measure of the term extraction methods using different measures considering various cut-off points in the candidate rank.

We observe that the best precision (52% – line 1 in Table 4) of the ECO corpus was achieved when using the top 50 candidates ordered by the *tc* measure. The best recall (100% – line 7) was reached with *gc_freq.*, *weirdness*, and *thd* when using more than 88% (8,850 candidates) of the corpus. The best F-measure (29.43% – line 13) used the top 400 candidates ordered by *tvq*. Regarding the DE corpus, the best precision (36% – line 2) was performed when using the first 50 candidates better ranked by the *tds* measure. The best recall (100% – line 8) was reached using *tf, rf, atf, df, tf-idf, tv, tvq, tc, gc_freq.*, and *c-value* considering more than 11,305 candidates (> 70% of the DE corpus). On the other hand, the best F-measure (22.22% – line 14) was achieved using the top 50 candidates ordered by the *tds* measure. Finally, for the N&N corpus, the best precision (66% – line 3) was obtained using the top 50 candidates ranked by the *tvq* measure. The best recall (94.10% – line 9) was achieved with *tf* and *rf* using more than 14,900 candidates (> 36% of the corpus). The best F-measure (36.22% – line 15) was reached with the first 3,150 candidates ordered by *tf-idf*.

Figure 1 shows the relation between the number of extracted candidates and values of precision, recall, and F-measure obtained with these candidates considering the ECO corpus. Due to the limited page number of this paper, we only show the graphic of the ECO corpus, however, we discuss the results of the three used corpora. For these three corpora, the highest precisions are achieved using the same amount of candidates, which is 50. The recall values reach around 100%, however, most of these cases use almost the entire corpus. Accordingly, the recall values are not considered good results since, if the entire corpus is used, the results would be equal or similar (see also lines 19–21 in Table 4). There are methods that achieve around 20% to 30% of F-measure for the ECO and DE corpora when using from 0.31% to 4,5% of these corpora (from 50 to 450 candidates). Meanwhile, for the N&N corpus, it is necessary to use 7.62% of the corpus (3150 candidates) to obtain the best F-measure (36.22%).

Tab. 4: Summary of term extraction method results.

	Line	Corpora	Measures	#Cand.	#ET	P (%)	R (%)	FM (%)
Best	1	ECO	tc	50	26	52,00	8,67	14,86
precision	2	DE	tds, glossEx	50	18	36,00	16,07	22,22
	3	N&N	tvq	50	33	66,00	2,14	4,14
Worst	4	ECO	n-gram length, tf, rf, atf, df, tf-idf, tv, tvq, tc, c-value	9950	299	3,01	99,67	5,83
precision	5	DE	ridf	15000	64	0,43	57,14	0,85
	6	N&N	ridf	14950	16	0,11	1,04	0,19
Best	7	ECO	gc_freq., weirdness, thd, nc-value	> 8850	300	3,02 - 3,39	100,00	5,85 - 6,56
recall	8	DE	tf, rf, atf, df, tc, tf-idf, tv, tvq, gc_freq., c-value	> 11100	112	0,75 - 1,01	100,00	1,48 - 2,00
	9	N&N	tf, rf	> 14900	1452	9,68 - 9,74	94,10	17,55 - 17,66
	10	ECO	tds, glossEx	50	2 - 4	4,00	0,67 - 1,33	1,14 - 2,00
Best	11	DE	thd	50	6	12,00	5,36	7,41
recall	12	N&N	weirdness ridf	50 14950	16	0,11	1,04	0,19
	13	ECO	tvq	400	103	25,75	34,33	29,43
Best	14	DE	tds, glossEx	50	18	36,00	16,07	22,22
F-measure	15	N&N	tf-idf	3150	850	26,98	55,09	36,22
	16	ECO	tds, glossEx	50	2 - 4	4,00	0,67 - 1,33	1,14 - 2,00
Worst	17	DE	ridf	15000	64	0,43	57,14	0,85
F-measure	18	N&N	ridf	14950	16	0,11	1,04	0,19
	19	ECO	–	9950	300	3,02	100,00	5,85
Entire	20	DE	–	16000	112	0,70	100,00	1,39
corpus	21	N&N	–	41335	1543	0,04	100,00	0,07

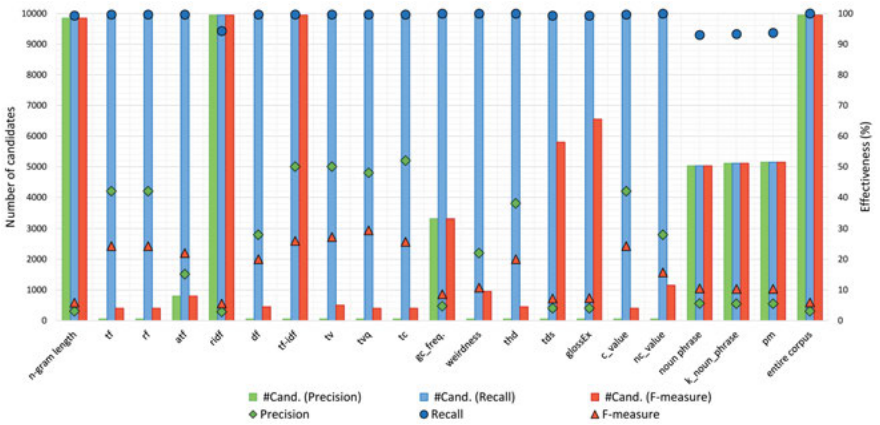

Fig. 1: Precision, recall, and F-measure *vs*. amount of extracted candidates – The ECO corpus.

Regarding the knowledge level (linguistic, statistical, or hybrid) of the methods, in general, the highest precision results were obtained using the statistical methods or some hybrid methods (such as *tds* or *c-value*). Interestingly, those methods that use measures do not commonly used to extract terms (*tc* or *tvq*), were a good option considering their precision values. The recall results of the statistical and hybrid methods were very similar because their best recall results use almost 100% of the candidates; this fact makes the recall independent of the used measures. An exception of the statement about the independence of recall compared to the measures is the use of the linguistic measures, since the extractors that use them have the highest recall values: 93.67%, 96.43%, and 88.32% using 51%, 49%, and 42% of the candidates, respectively, for the ECO, DE, and N&N corpora.

When using statistical or hybrid extractors, F-measure generally maintains the same pattern. It is noteworthy that the linguistic extractors obtain low precision and F-measure results (between 1.35 to 8.09%). This fact give us evidence the linguistic extractors should use measures of other knowledge levels as well. The linguistic extractors achieve high recall results (between 85 and 96%), which means they are able to remove part of the non-terms without excluding many terms. Therefore, we proved that better results are obtained when the statistical measures are applied on a candidate list that was previously filtered based on some linguistic measure, as stated by [1].

4 Conclusions and Future Work

With the experiments performed in this work, we demonstrated how difficult it is to analyze the rank of candidates created by semi-automatic term extraction methods.

Based on our experiments, we list below some suggestions to be followed by the traditional semi-automatic methods of simple term extraction. In order to achieve good precisions, we suggest to consider the first 50 candidates ordered by some of the statistical or hybrid methods. It was not possible, however, to identify which method is the best one to reach good recall, since the highest recall values were only achieved using (almost) the entire corpus, which is not recommended. Linguistic methods showed to be promising for that. It was not possible to identify a unique method that is the best one for the three used corpora. Nevertheless, regarding the threshold used in the candidate ranks, the statistical (except *n-gram length* and *ridf*) and hybrid (except *gc_freq.*) methods are the most desirable when aiming to achieve high precision and F-measure.

Regarding the four extractors that use measures (*tv, tvq, tc*, and *n-gram length*) normally applied to other tasks instead of term extraction, we observe that *tv, tvq*, and *tc* were responsible for at least one of the highest results of each corpus. Therefore, as expected, these three measures are good options for extracting terms. However, *n-gram length* reached lower results than the other measures used in this research. Then, we conclude that, contrary to expectations, there is no difference in the length (in characters) between terms and non-terms of these corpora. Finally, this experiments demonstrate how difficult and subjective it is to determine a threshold in the candidate term ranking. Our future work remains on combining the measures and exploring new ones.

Bibliography

[1] Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M.: Terminology extraction: An analysis of linguistic and statistical approaches. In Sirmakessis, S., ed.: Knowledge Mining Series: Studies in Fuzziness and Soft Computing. Springer Verlag (2005) 255–279

[2] Frantzi, K.T., Ananiadou, S., Tsujii, J.I.: The C-value/NC-value method of automatic recognition for multi-word terms. In: PROC of the 2nd European CNF on Research and Advanced Technology for Digital Libraries (ECDL), London, UK, Springer-Verlag (1998) 585–604

[3] Kozakov, L., Park, Y., Fin, T.H., Drissi, Y., Doganata, Y.N., , Confino, T.: Glossary extraction and knowledge in large organisations via semantic Web technologies. In: PROC of the 6th INT Semantic Web CNF (ISWC) and the 2nd Asian Semantic Web CNF. (2004)

[4] Kit, C., Liu, X.: Measuring mono-word termhood by rank difference via corpus comparison. Terminology **14**(2) (2008) 204–229
[5] Vivaldi, J., Cabrera-Diego, L.A., Sierra, G., Pozzi, M.: Using wikipedia to validate the terminology found in a corpus of basic textbooks. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Tapias, D., eds.: PROC of the 8th INT CNF on Language Resources and Evaluation (LREC), Istanbul, Turkey, ELRA (2012) 3820–3827
[6] Conrado, M.S., Di Felippo, A., Pardo, T.S., Rezende, S.O.: A survey of automatic term extraction for Brazilian Portuguese. Journal of the Brazilian Computer Society (JBCS) **20**(1) (2014) 12
[7] Ahmad, K., Gillam, L., Tostevin, L.: University of surrey participation in TREC8: Weirdness indexing for logical document extrapolation and retrieval (WILDER). In: PROC of the Text REtrieval CNF (TREC). (1999) 1–8
[8] Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal os Research and Development **2**(2) (1958) 159–165
[9] Nogueira, B.M.: Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos. Master's thesis, Institute of Mathematics and Computer Science (ICMC) – University of Sao Paulo (USP), São Carlos, SP (2009)
[10] Salton, G., Yang, C.S., Yu, C.T.: A theory of term importance in automatic text analysis. Journal of the American Association Science **1**(26) (1975) 33–44
[11] Maynard, D., Ananiadou, S.: Identifying terms by their family and friends. In: PROC of 18th INT CNF on Computational Linguistics (COLING), Saarbrucken, Germany (2000) 530–536
[12] Pantel, P., Lin, D.: A statistical corpus-based term extractor. In: PROC of the 14th Biennial CNF of the Canadian Society on COMP Studies of Intelligence (AI), London, UK, Springer-Verlag (2001) 36–46
[13] Zavaglia, C., Aluísio, S.M., Nunes, M.G.V., Oliveira, L.H.M.: Estrutura ontológica e unidades lexicais: uma aplicação computacional no domínio da ecologia. PROC of the 5th Wkp em Tecnologia da Informação e da Linguagem Humana (TIL) – Anais do XXVII Congresso da Sociedade Brasileira da Computação (SBC), Rio de Janeiro (2007) 1575–1584
[14] Lopes, L., Vieira, R.: Aplicando pontos de corte para listas de termos extraídos. In: PROC of the 9th Brazilian Symposium in Information and Human Language Technology (STIL), Fortaleza, Brasil, Sociedade Brasileira de Computação (2013) 79–87
[15] Vivaldi, J., Rodríguez, H.: Evaluation of terms and term extraction systems: A practical approach. Terminology: INT Journal of Theoretical and Applied Issues in Specialized Communication **13**(2) (2007) 225–248
[16] Knoth, P., Schmidt, M., Smrz, P., Zdráhal, Z.: Towards a framework for comparing automatic term recognition methods. In: CNF Znalosti. (2009)
[17] Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.: A comparative evaluation of term recognition algorithms. In Calzolari et al., ed.: PROC of the 6th INT CNF on Language Resources and Evaluation (LREC), Marrakech, Morocco, ELRA (2008) 2108–2113
[18] Souza, J.W.C., Felippo, A.D.: Um exercício em linguística de corpus no âmbito do projeto TermiNet. Technical Report NILC-TR-10-08, Institute of Mathematics and Computer Science (ICMC) – University of Sao Paulo (USP), Sao Carlos, SP (2010)
[19] Coleti, J.S., Mattos, D.F., Genoves Junior, L.C., Candido Junior, A., Di Felippo, A., Almeida, G.M.B., Aluísio, S.M., Oliveira Junior, O.N.: Compilação de corpus em Língua Portuguesa na área de nanociência/nanotecnologia: Problemas e soluções. In of Sao Paulo (USP), U., ed.: Avanços da linguistica de Corpus no Brasil. Volume 1. 192 edn. Stella E. O.Tagnin; Oto Araújo Vale. (Org.), Sao Paulo, SP (2008) 167–191

[20] Bick, E.: The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Arhus, Arhus (2000)
[21] Church, K.W.: One term or two? In: PROC of the 18th Annual INT CNF on Research and Development in Information Retrieval (SIGIR), New York, NY, USA, ACM (1995) 310–318
[22] Salton, G., Buckley, C.: Term weighting approaches in automatic text retrieval. Technical report, Cornell University, Ithaca, NY, USA (1987)
[23] Liu, L., Kang, J., Yu, J., Wang, Z.: A comparative study on unsupervised feature selection methods for text clustering. In: PROC of IEEE INT CNF on Natural Language Processing and Knowledge Engineering (NLP-KE). (2005) 597–601
[24] Dhillon, I., Kogan, J., Nicholas, C.: Feature selection and document clustering. In Berry, M.W., ed.: Survey of Text Mining. Springer (2003) 73–100
[25] Liu, T., Liu, S., Chen, Z.: An evaluation on feature selection for text clustering. In: PROC of the 10th INT CNF on Machine Learning (ICML), San Francisco, CA, Morgan Kaufmann (2003) 488–495
[26] Conrado, M.S., Pardo, T.A.S., Rezende, S.O.: Exploration of a rich feature set for automatic term extraction. In Castro, F., Gelbukh, A., González, M., eds.: Advances in Artificial Intelligence and its Applications (MICAI). LNCS. Springer (2013) 342–354
[27] Estopa, R., Martí, J., Burgos, D., Fernández, S., Jara, C., Monserrat, S., Montané, A., noz, P.M., Quispe, W., Rivadeneira, M., Rojas, E., Sabater, M., Salazar, H., Samara, A., Santis, R., Seghezzi, N., Souto., M.: La identificación de unidades terminológicas en contexto: de la teoría a la práctica. In: Terminología y Derecho: Complejidad de la Comunicación Multilingue, Cabré, T. and Bach, C. and Martí, J. (2005) 1–21
[28] Park, Y., Patwardhan, S., Visweswariah, K., Gates, S.C.: An empirical analysis of word error rate and keyword error rate. In: 9th Annual CNF of the INT Speech Communication Association (INTERSPEECH), INT Speech Communication Association (ISCA) (2008) 2070–2073
[29] Barrón-Cedeño, A., Sierra, G., Drouin, P., Ananiadou, S.: An improved automatic term recognition method for spanish. In: PROC of the 10th INT CNF on COMP Linguistics and Intelligent Text Processing (CICLing), Berlin, Heidelberg, Springer-Verlag (2009) 125–136