A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set*

Merley da Silva Conrado, Thiago A. Salgueiro Pardo, and Solange Oliveira Rezende
Laboratory of Computational Intelligence,
An Interinstitutional Center for Research and Development in Computational Linguistic,
Institute of Mathematical and Computer Sciences,
University of Sao Paulo (USP),
P.O. Box 668, 13561-970, Sao Carlos-SP, Brazil
{merleyc,taspardo,solange}@icmc.usp.br

Abstract

In this paper we propose an automatic term extraction approach that uses machine learning incorporating varied and rich features of candidate terms. In our preliminary experiments, we also tested different attribute selection methods to verify which features are more relevant for automatic term extraction. We achieved state of the art results for unigram extraction in Brazilian Portuguese.

1 Introduction

Terms are terminological units from specialised texts (Castellví et al., 2001). A term may be: (i) simple[1] (a single element), such as "*biodiversity*", or (ii) complex (more than one element), such as "*aquatic ecosystem*" and "*natural resource management*".

Automatic term extraction (ATE) methods aim to identify terminological units in specific domain corpora (Castellví et al., 2001). Such information is extremely useful for several tasks, from the linguistic perspective of building dictionaries, taxonomies and ontologies, to computational applications as information retrieval, extraction, and summarisation.

Although ATE has been researched for more than 20 years, there is still room for improvement. There are four major ATE problems. The first one is that the ATE approaches may extract terms that are not actual terms ("noise") or do not extract actual terms ("silence"). Considering the ecology domain, an example of silence is when a term (e.g., *pollination*), with low frequency, is not considered a candidate term (CT), and, therefore, it will not appear in the extracted term list if we consider its frequency. Regarding noise, if we consider that nouns may be terms and that adjectives may not, if an adjective (e.g., *ecological*) is mistakenly tagged as a noun, it will be wrongly extracted as a term. The second problem is the difficulty in dealing with extremely high number of candidates (called the high dimensionality of candidate representation) that requires time to process them. Since the ATE approaches generate large lists of TCs, we have the third problem that is the time and human effort spent for validating the TCs, which usually is manually performed. The fourth problem is that the results are still not satisfactory and there is a natural ATE challenge since the difficulty in obtaining a consensus among the experts about which words are terms of a specific domain (Vivaldi and Rodríguez, 2007).

Our proposed ATE approach uses machine learning (ML), since it has been achieving high precision values (Zhang et al., 2008; Foo and Merkel, 2010; Zhang et al., 2010; Loukachevitch, 2012). Although ML may also generate noise and silence, it facilitates the use of a large number of TCs and their features, since ML techniques learn by themselves how to recognize a term and then they save time extracting them.

Our approach differs from others because we adopt a rich feature set using varied knowledge levels. With this, it is possible to decrease the silence and noise and, consequently, to improve the ATE results. Our features range from simple statistical (e.g., term frequency) and linguistic (e.g., part of

*This research was supported by FAPESP (Proc. No. 2009/16142-3 and 2012/09375-4), Brazil.

[1]When we refer to *unigrams*, we mean *simple terms*.

Proceedings of the NAACL HLT 2013 Student Research Workshop, pages 16–23,
Atlanta, Georgia, 13 June 2013.

speech - POS) knowledge to more sophisticated hybrid knowledge, such as the analysis of the term context. As far as we know, the combined use of this specific knowledge has not been applied before. Another difference is that we apply 3 statistical features (Term Variance (Liu et al., 2005), Term Variance Quality (Dhillon et al., 2003), and Term Contribution (Liu et al., 2003)) that to date have only been used for attribute selection and not for term extraction. As far as we know, the combined use of this specific knowledge and feature feedback has not been applied before. We also propose 4 new linguistic features for ATE. All these features are detailed in Section 4. Finally, for the first time, ML is being applied in the task of ATE in Brazilian Portuguese (BP) corpora. Our approach may also be easily adapted to other languages.

We focus on extracting only unigram terms, since this is already a complex task. We run our experiments on 3 different corpora. Our main contribution is the improvement of precision (in the best case, we improve the results 11 times) and F-measure (in the best case, we improve 2 times).

Section 2 presents the main related work. Section 3 describes our ATE approach. Section 4 details the experiments, and Section 5 reports the results. Conclusions and future work are presented in Section 6.

2 Related Work

There are several recent and interesting studies that are not focused on extracting unigrams (Estopà et al., 2000; Almeida and Vale, 2008; Zhang et al., 2008; Zhang et al., 2010; Nazar, 2011; Vivaldi et al., 2012; Lopes, 2012). Normally, ATE studies use corpora of different domain and language and, in some cases, the authors use different evaluation measures. Regardless of variation (e.g., the size of the test corpora), we mention studies that have highlighted results for **unigrams**[2]. When possible, we show the best precision (P) of the related work and its recall (R).

(Ventura and Silva, 2008) extracted terms using statistical measures that consider the predecessors and successors of TCs. They achieved, for English, P=81.5% and R=55.4% and, for Spanish, P=78.2% and R=60.8%. For Spanish, the Greek forms of a candidate and their prefix may help to extract terms (e.g., the Greek formant *laring* that belongs to the term ***laring****oespasm* in the medical domain) (Vivaldi and Rodríguez, 2007), achieving about P=55.4% and R=58.1%. For Spanish, (Gelbukh et al., 2010) compared TCs of a domain with words of a general corpus using Likelihood ratio based distance. They achieved P=92.5%. For Brazilian Portuguese, the ExPorTer methods are the only previous work that uniquely extract unigrams (Zavaglia et al., 2007). Therefore, they are the state of the art for unigrams extraction for BP. The linguistic ExPorTer considers terms that belong to some POS patterns and uses indicative phrases (such as *is defined as*) that may identify where terms are. It achieved P=2.74% and R=89.18%. The hybrid ExPorTer used these linguistic features with frequency and Likelihood ratio. The latter one obtained P=12.76% and R=23.25%.

3 Term Extraction Approach based on Machine Learning

In order to model the ATE task as a machine learning solution, we consider each word in the input texts[3] of a specific domain (except the stopwords) as a learning instance (candidate term). For each instance, we identify a set of features over which the classification is performed. The classification predicts which words are terms (unigrams) of a specific domain. We test different attribute selection methods in order to verify which features are more relevant to classify a term.

We start by preprocessing the input texts, as shown in Figure 1. This step consists of POS tagging the corpora and normalizing[4] the words of the texts. The normalization minimizes the second ATE problem because it allows working with a lower CT representation dimensionality. When working with a lower dimensionality, the words that do not help identify terms are eliminated. Consequently, fewer candidates should to be validated or refuted as terms (it would minimize the third ATE problem). When working with fewer candidates it also may improve the result quality (it handles the fourth ATE prob-

[2] It is not specified if (Zhang et al., 2010) extracted simple or complex terms.

[3] When we refer to *texts*, we mean *documents*.

[4] Normalization consists of standardizing the words by reducing their variations.

lem), and, definitely, it spends less time and fewer resources to carry out the experiments. By improving the results, consequently, we minimize silence and noise, which handles the first ATE problem. Afterwards, we remove stopwords.

In order to identify a set of features over which the classification is performed, we studied and tested several measures. The feature identification is the most important step of our approach. We divide the features into two types: (i) the features that obtain statistical, linguistic, and hybrid knowledge from the input corpus, such as TFIDF and POS, and (ii) the features that obtain these knowledge from measures that use other corpora besides the input corpus. The corpora belong to another domain that is different of the input corpus domain (called contrastive corpora) or not belong to any specific domain (called general corpora). Our hypothesis is that, with the joining of features of different levels of knowledge, it is possible to improve the ATE.

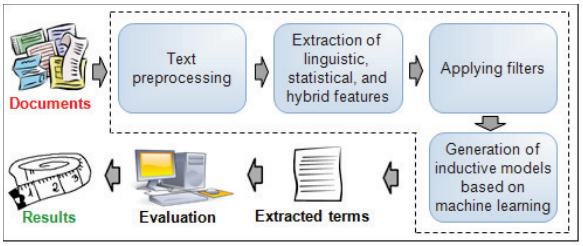

Figure 1: Term extraction approach proposed.

4 Experimental Setup

At this point, for obtaining the knowledge in order to extract terms, we tested 17 features that do not depend on general or contrastive corpora and 2 features that depend on these corpora. We intend to explore more features (and we will possibly propose new measures) that use contrastive or general corpora or any taxonomic structure. The experiments that expand the number of features are ongoing now.

We used 3 corpora of different domains in the Portuguese language. The EaD corpus (Souza and Di Felippo, 2010) has 347 texts about distance education and has a gold standard with 118 terms[5] (Gianoti and Di Felippo, 2011). The second one is the ECO[6] corpus (Zavaglia et al., 2007). It contains 390 texts of ecology domain and its gold standard has 322 unigrams. The latter is the Nanoscience and Nanotechnology (N&N) corpus (Coleti et al., 2008) that contains 1,057 texts. Its gold standard has 1,794 unigrams (Coleti et al., 2008; Coleti et al., 2009).

In order to preprocess these corpora, we POS tagged them using the PALAVRAS parser (Bick, 2000) and normalized their words using a stemming[7] technique. Stemming was chosen because of its capacity to group similar word meanings, and its use decreases representation dimensionality of candidate terms, which minimizes the second and third ATE problems. Afterwards, we removed the stopwords[8], the conjugation of the verb "to be", punctuation, numbers, accents, and the words composed of only one character are removed.

We identify and calculate 19 features in which 11 features are used for ATE in the literature, 3 features are normally applied to the attribute selection tasks (identified by *), 1 normally used for Named Entity Recognition (identified by **), and we created 4 new features (identified by $^{\Delta}$). These features are shown in Table 1, accompanied by the hypotheses that underlie their use. They are also divided into 3 levels of knowledge: statistical, linguistic, and hybrid.

For the ***S*** feature, we removed stopwords at the beginning and at the end of these phrases. For ***POS***, we assumed that terms may also be adjectives (Almeida and Vale, 2008), besides nouns and verbs. For ***GC*** and ***Freq_GC***, we used the NILC Corpus[9] as a general corpus, which contains 40 million words. We created and used 40 indicative phrases (***NPs***). For example, considering *are composed of* as an IP in *All organisms are composed of one or more cells*, we would consider *organisms* and *cells* as TCs. For features related to CT stem, we analyzed, e.g., the words *educative, educators, education* and *educate* that came from the stem *educ*. Therefore, *educ* may

[5](Gianoti and Di Felippo, 2011) stated that the EaD unigram gold standard has 59 terms, but in this paper we used 118 unigrams that the authors provided us prior to their work.

[6]ECO corpus - http://www.nilc.icmc.usp.br/nilc/projects/bloc-eco.htm

[7]PTStemmer: A Stemming toolkit for the Portuguese language - http://code.google.com/p/ptstemmer/

[8]Stoplist and Indicative Phrase list are avaiable in http://www2.icmc.usp.br/ merleyc/

[9]NILC Corpus - http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm

Table 1: Features of candidate terms.

Feature	*Description*	*Hypothesis*
The eight linguistic features		
S	noun and prepositional phrases	terms are noun phrases and, sometimes, prepositional phrases
N_S	head of phrases	heads of noun and prepositional phrases
POS	noun, proper noun, and adjective	terms follow some patterns
IP	indicative phrases	IPs may identify definitions/descriptions that may be terms
N_noun $^{\Delta}$	number of nouns	stemmed terms come from higher number of nouns than adjectives or verbs
N_adj $^{\Delta}$	number of adjectives	
N_verb $^{\Delta}$	number of verbs	
N_PO $^{\Delta}$	total of words from which stemmed TCs come from	
The seven statistical features		
SG**	n-gram length	each domain has a term pattern
TF	Term Frequency	terms have neither low nor very high frequencies
DF	Document Frequency	terms appear in at least certain number of documents
TFIDF	Term Frequency Inverse Document Frequency (Salton and Buckley, 1987)	terms are very common in the corpus but they occur in few documents in this corpus
TCo*	Term Contribution (Liu et al., 2003)	terms help to distinguish the different documents
TV*	Term Variance (Liu et al., 2005)	terms do not have low frequency in documents and maintain a non-uniform distribution throughout corpus (higher variance)
TVQ*	Term Variance Quality (Dhillon et al., 2003)	
The four hybrid features		
GC	CT occurrence in general corpus	terms do not occur with high frequency in a general corpus
Freq_GC	CT frequency in GC	
C-value	the potential of a CT to be a term (Frantzi et al., 1998)	the C-value helps to extract terms
NC-value	CT context (Frantzi et al., 1998)	candidate context helps to extract terms

have as features ***N_Noun*** = 2 (*educators* and *education*), ***N_Adj*** = 1 *educative*, ***N_Verb*** = 1 (*educate*), and ***N_PO*** = 4 (total number of words). Our hypothesis is that stemmed candidates that were originated from a higher number of nouns than adjectives or verbs will be terms. Finally, we used NC-Value adapted to unigrams (Barrón-Cedeño et al., 2009).

After calculating the features for each unigram (candidate term), the CT representation has high dimensionality (it is the second ATE problem) and, hence, the experiments may take a considerable amount of time to be executed. To decrease this dimensionality and, consequently, the number of TCs (which corresponds to the second and third ATE problems, respectively), we tested two different cut-offs, which preserve only TCs that occur in at least two documents in the corpus. The first cut-off is called *C1*. In the second one (called *C2*), the candidates must be noun and prepositional phrases and also follow some of these POS: nouns, proper nouns, verbs, and adjectives. The number of obtained candidates (stems) was 10,524, 14,385, and 46,203, for the ECO, EaD, and N&N corpora, respectively. When using the *C1* cut-off, we decreased to 55,15%, 45,82%, and 57,04%, and *C2* decreased 63.10%, 63.18%, 66.94% in relation to the number of all the obtained candidates (without cutt-offs).

5 Experimental Evaluation and Results

The first evaluation aimed to identify which features must be used for ATE (see Section 3). For that, we applied 2 methods that select attributes by evaluating the attribute subsets. Their evaluation is based on consistency (CBF) and correlation (CFS). We also tested search methods. The combination of these methods, available in WEKA (Hall et al., 2009), is: CFS_SubsetEval using the RankSearch Filter as search method (*CFS_R*), CFS_SubsetEval using the BestFirst as search method (*CFS_BF*), CBF_SubsetEval using the Ranking Filter (*C_R*), and CBF_SubsetEval using the Greedy Stepwise (*C_G*). These methods return feature sets that are considered the most representative for the term classification (Table 2). For the EaD corpus, the *CG* attribute selection method did not select any feature. For our experiments, we also considered all the features (referred by *All*). Additionally, we compared the use of two cut-off types for each feature set, *C1* and *C2*, detailed in Section 4.

For both evaluations[8], we chose largely known inductors in the machine learning area. They represent different learning paradigms: JRip (Rule Induction), Naïve Bayes (Probabilistic), J48 (Decision Tree) with confidence factor of 25%, and SMO (Statistical Learning). All of these algorithms are avail-

Table 2: Features chosen by the attribute selection methods.

Methods	*Corpora*		
	EaD	**ECO**	**N&N**
CFS_R	TFIDF, TV, TVQ, IP, N_Noun, N_Adj	TFIDF, TV, TVQ, POS, N_Noun	Freq, TFIDF, TVQ, IP, Cvalue, N_Noun, POS, N_Adj, N_PO
CFS_BF	Same as in the CFS_R method.	TFIDF, TVQ, TCo, POS	Freq, TFIDF, TV, IP, Cvalue, N_Noun, POS, N_Adj, N_PO
C_R	Freq, DF, TFIDF, TV, TVQ, TCo, IP, GC, POS, FreqGC, NCvalue, Cvalue, N_Adj, N_Noun, N_Verb, N_PO	Freq, DF, TFIDF, TV, TVQ, TCo, GC, Cvalue, NCvalue, IP, S, N_S, POS, N_Noun, N_Adj, N_Verb, N_PO	Freq, DF, TFIDF, TV, TVQ, TCo, GC, IP, S, Cvalue, POS, NCvalue, N_S, N_Noun, N_Adj, N_Verb, N_PO
C_G	Method did not select any feature.	Freq, DF, TFIDF, TV, TVQ, GC, IP, N_S, NCvalue, S, N_Noun, POS, N_Adj, N_PO	Freq, DF, TFIDF, S, TV, TVQ, TCo, IP, NCvalue, N_S, POS, GC, N_Noun, N_PO, N_Verb, N_Adj

able in WEKA and described in (Witten and Frank, 2005). We run the experiments on a 10 fold cross-validation and calculated the precision, recall, and F-measure scores of term classification according to the gold standard of unigrams of each corpus. Using default parameter values for SMO, the results were lower than the other inductors. Due to this fact and the lack of space in the paper, we do not present the SMO results here.

The best precision obtained for the EaD corpus using the term classification, 66.66%, was achieved by the *C_R* attribute selection method with the *C2* cut-off (*C_R-C2*) using the JRIP inductor. The best recall score, 20.96%, was obtained using Naïve Bayes with the *CFS_R-C1* method. The best F-measure was 17.58% using the J48 inductor with *C_R-C2*. For the ECO corpus, the best precision was 60% obtained with the J48 inductor with confidence factor of 25% and the *C_R-C1* method. The best recall was 21.40% with JRIP and the *C_G-C1* method. Our best F-measure was 24.26% obtained with Naïve Bayes using the *CFS_R-C1* method. For the N&N corpus, the best precision score was 61.03% using JRIP. The best recall was 52.53% and the best F-measure score was 54.04%, both using J48 inductor with confidence factor of 25%. The three results used the *All-C2* method.

Table 3 shows the comparison of our best results with 2 baselines, which are the well-known term frequency and TFIDF, using our stoplist. We also considered all the stemmed words of these corpora as CT, except the stopwords, and we calculated the precision, recall, and F-measure scores for these words as well. Finally, we compared our results with the third baseline, which is the only previous work that uniquely extracts unigrams (Zavaglia et al., 2007), described in Section 2. Therefore, this is the state of the art for unigrams extraction for Portuguese. In order to compare this work with our results of the EaD and N&N corpora, we implemented the ATE method of Zavaglia et al. We have to mention that this method uses the normalization technique called lemmatization instead of stemming, which we used in our method. The only difference between our implementation descriptions and the original method is that we POS tagged and lemmatizated the texts using the same parser (PALAVRAS[10] (Bick, 2000)) used in our experiments instead of the MXPOST tagger (Ratnaparkhi, 1996).

For all used corpora, we obtained better results of precision and F-measure comparing with the baselines. In general, we improve the ATE precision scores, for the EaD corpus, eleven times (from 6.1% to 66.66%) and, for the N&N corpus, one and a half times (from 35.4% to 61.03%), both comparing our results with the use of TFIDF. For the ECO corpus, we improve four and a half times (from 12.9% to 60%), by comparing with the use of frequency. We improve the ATE F-measure scores, for the EaD corpus, one and a half times (from 10.93% to 17.58%); for the ECO corpus, we slightly improve the results (from 20.64% to 24.26%); and, for the N&N corpus, two times (from 28.12% to 54.04%). The last three cases are based on the best F-measure values obtained using TFIDF. Regarding recall, on the one hand, the linguistic ExPorTer method (detailed in Section 2), to which we also compare our results, achieved better recall for all used corpora, about 89%. On the other hand, its precision (about 2%) and F-measure (about 4%) were significantly lower than our results.

Finally, if we compare our results with the results of all stemmed words, with the exception of the stopwords, the recall values of the latter are high (about 76%) for all used corpora. However, the precision scores are extremely low (about 1.26%), because it used almost all words of the texts.

[10] As all NLP tools for general domains, PALAVRAS is not excellent for specific domains. However, as it would be expensive (time and manual work) to customize it for each specific domain that we presented in this paper, we decided use it, even though there are error tagging.

Table 3: Comparison with baselines.

Method	*Precision (%)*	*Recall (%)*	*F-Measure (%)*
The EaD corpus			
JRIP with C_R-C2	**66.66**	8.06	14.38
Naïve Bayes with CFS_R-C1	13.19	20.96	16.19
J48 with F.C. of 0.25 with C_R-C2	27.58	12.9	**17.58**
Ling. ExPorTer	0.33	**89.70**	0.66
Hyb. ExPorTer	0.07	17.64	0.15
Frequency	5.9	50.86	10.57
TFIDF	6.1	52.58	10.93
All the corpus	0.52	62.9	1.04
The ECO corpus			
J48 with F.C. of 0.25 with C_R-C1	**60.00**	6.02	10.94
JRIP with C_G-C1	23.44	21.40	22.38
Naïve Bayes with CFS_R-C1	33.33	19.06	**24.26**
Ling. ExPorTer	2.74	**89.18**	5.32
Hyb. ExPorTer	12.76	23.25	16.48
Frequency	12.9	43.28	19.87
TFIDF	13.4	44.96	20.64
All the corpus	1.48	99.07	2.92
The N&N corpus			
JRIP with All-C2	**61.03**	27.73	38.14
J48 with F.C. of 0.25 with All-C2	55.64	52.53	**54.04**
Ling. ExPorTer	3.75	**89.40**	7.20
Hyb. ExPorTer	1.68	35.35	3.22
Frequency	31.6	20.83	25.1
TFIDF	35.4	23.33	28.12
All the corpus	1.83	66.99	3.57

6 Conclusions and Future Work

This paper described ongoing experiments about unigrams extraction using ML. Our first contribution regarding the experiments was to create 4 features and to test 4 features that normally are applied to other tasks and not for automatic term extraction.

Our second contribution is related to the first and fourth ATE problems, which are the existence of silence and noise and low ATE results, respectively. We achieved state of art results for unigrams in Brazilian Portuguese. We improved, for all used corpora, precision (in the best case, we improve the results 11 times using the EaD corpus) and F-measure (in the best case, 2 times using the N&N corpus) and, consequently, we minimized silence and noise.

The third contribution is about the features that are better for extracting domain terms. All the tested attribute selection methods indicated the TFIDF as an essential feature for ATE. 90.9% of the methods selected N_Noun and TVQ, and 81.81% selected TV, IP, N_adj, and POS as relevant features. However, only one of these methods chose Freq_GC, and none of them chose the SG feature. Regarding the levels of knowledge - statistical, linguistic, and hybrid - in which each feature was classified, at least 45.45% of the methods chose 6 statistical, 5 linguistic, and 3 hybrid features. We also observed that the best F-measures (see Tables 2 and 3) were obtained when using at least linguistic and statistical features together. This fact proves that our main hypothesis is true, because we improved the ATE results by joining features of different levels of knowledge. Additionally, we allow the user to choose the features that are better for term extraction.

As the fourth contribution, we minimized the problem of high dimensionality (as mentioned, the second ATE problem) by means of the use of two different cut-offs (*C1* and *C2*). By reducing the number of TCs, fewer candidates were validated or refuted as terms and, consequently, we minimized the third ATE problem, which is the time and human effort for validating the TCs. However, we still perceived the need to reduce more the number of candidates. Therefore, for future work, we intend to use instance selection techniques to reduce the term representation.

We believe to have achieved significant results for the experiments realized to date. Experiments using more features that dependent on general corpus are ongoing. We will also possibly propose new features and will use taxonomic structure in order to improve more the results. For using the taxonomic structure, we intend to create a conventional taxonomy (Miiller and Dorre, 1999) is created using the input corpus. Therefore, we may identify more features for the instances considering this taxonomy. For example, normally in a taxonomy's leaf specific words of a domain happen, consequently, terms should appear there. Additionally, we are encouraged to adapt these features for bigram and trigram terms as well.

References

G. M. B. Almeida and O. A. Vale. 2008. Do texto ao termo: interação entre terminologia, morfologia e

linguística de corpus na extração semi-automática de termos. In A. N. Isquerdo and M. J. B. Finatto, editors, *As Ciências do Léxico: Lexicologia, Lexicografia e Terminologia*, volume IV, pages 483–499. UFMS, MS, Brazil, 1 edition.

A. Barrón-Cedeño, G. Sierra, P. Drouin, and S. Ananiadou. 2009. An improved automatic term recognition method for spanish. In *Proc of the 10th Int. CNF on Computational Linguistics and Intelligent Text Processing*, pages 125–136, Berlin, Heidelberg. Springer-Verlag.

E. Bick. 2000. *The Parsing System "PALAVRAS". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Arhus, Arhus.

M. T. Cabré Castellví, R. Estopà Bagot, and Jordi Vivaldi Palatresi. 2001. Automatic term detection: a review of current systems. In D. Bourigault, C. Jacquemin, and M-C. L'Homme, editors, *Recent Advances in Computational Terminology*, pages 53–88, Amsterdam/Philadelphia. John Benjamins.

J. S. Coleti, D. F. Mattos, L. C. Genoves Junior, A. Candido Junior, A. Di Felippo, G. M. B. Almeida, S. M. Aluísio, and O. N. Oliveira Junior. 2008. *Compilação de Corpus em Língua Portuguesa na área de Nanociência/Nanotecnologia: Problemas e soluções*, volume 1. Tagnin and Vale., SP, Brazil, 192 edition.

J. S. Coleti, D. F. Mattos, and G. M. B. Almeida. 2009. Primeiro dicionário de nanociência e nanotecnologia em língua portuguesa. In Marcelo Fila Pecenin, Valdemir Miotello, and Talita Aparecida Oliveira, editors, *II Encontro Acadêmico de Letras (EALE)*, pages 1–10. Caderno de Resumos do II EALE.

I. Dhillon, J. Kogan, and C. Nicholas. 2003. Feature selection and document clustering. In M. W. Berry, editor, *Survey of Text Mining*, pages 73–100. Springer.

R. Estopà, J. Vivaldi, and M. T. Cabré. 2000. Use of greek and latin forms for term detection. In *Proc of the 2nd on LREC*, pages 855–861, Greece. ELRA.

J. Foo and M. Merkel. 2010. Using machine learning to perform automatic term recognition. In N. Bel, B. Daille, and A. Vasiljevs, editors, *Proc of the 7th LREC - Wksp on Methods for automatic acquisition of Language Resources and their Evaluation Methods*, pages 49–54.

K. T. Frantzi, S. Ananiadou, and J. I. Tsujii. 1998. The C-value/NC-value method of automatic recognition for multi-word terms. In *Proc of the 2nd ECDL*, pages 585–604, London, UK. Springer-Verlag.

A. F. Gelbukh, G. Sidorov, E. Lavin-Villa, and L. Chanona-Hernández. 2010. Automatic term extraction using log-likelihood based comparison with general reference corpus. In *NLDB*, pages 248–255.

A. C. Gianoti and A. Di Felippo. 2011. Extração de conhecimento terminológico no projeto TermiNet. Technical Report NILC-TR-11-01, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, SP, Brazil.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. In *SIGKDD-ACM*, volume 11, pages 10–18.

T. Liu, S. Liu, and Z. Chen. 2003. An evaluation on feature selection for text clustering. In *Proceedings of the 10th Int. CNF on Machine Learning*, pages 488–495, San Francisco, CA, USA. Morgan Kaufmann.

L. Liu, J. Kang, J. Yu, and Z. Wang. 2005. A comparative study on unsupervised feature selection methods for text clustering. In *Proc of IEEE NLP-KE*, pages 597–601.

L. Lopes. 2012. *Extração automática de conceitos a partir de textos em língua portugesa*. Ph.D. thesis, Porto Alegre, RS. Pontifícia Universidade do Rio Grande do Sul (PUCRS).

N. Loukachevitch. 2012. Automatic term recognition needs multiple evidence. In N. Calzolari, K. Choukri, T. Declerck, M. Dogan, B. Maegaard, J. Mariani, Odijk, and S. Piperidis, editors, *Proc of the 8th on LREC*, pages 2401–2407, Istanbul, Turkey. ELRA.

A. Miiller and J. Dorre. 1999. The taxgen framework: Automating the generation of a taxonomy for a large document collection. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences (HICSS)*, volume 2, pages 2034–2042, Washington, DC, USA. IEEE Computer Society.

R. Nazar. 2011. A statistical approach to term extraction. *Int. Journal of English Studies*, 11(2).

A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. *Proc of the CNF on EMNLP*, pages 491–497.

G. Salton and C. Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA.

J. W. C. Souza and A. Di Felippo. 2010. Um exercício em lingüistica de corpus no âmbito do projeto TermiNet. Technical Report NILC-TR-10-08, ICMC - USP, SP, Brazil.

J. Ventura and J. F. Silva. 2008. Ranking and extraction of relevant single words in text. In Cesare Rossi, editor, *Brain, Vision and AI*, pages 265–284. InTech, Education and Publishing.

J. Vivaldi and H. Rodríguez. 2007. Evaluation of terms and term extraction systems: A practical approach. *Terminology*, 13(2):225–248.

J. Vivaldi, L. A. Cabrera-Diego, G. Sierra, and M. Pozzi. 2012. Using wikipedia to validate the terminology found in a corpus of basic textbooks. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proc of the 8th Int. CNF on LREC*, Istanbul, Turkey. ELRA.

I. H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

C. Zavaglia, L. H. M. Oliveira, M. G. V. Nunes, and S. M. Aluísio. 2007. Estrutura ontológica e unidades lexicais: uma aplicação computacional no domínio da ecologia. In *Proc. of the 5th Wksp em Tecnologia da Informação e da Linguagem Humana*, pages 1575–1584, RJ, Brazil. SBC.

Z. Zhang, J. Iria, C. Brewster, and F. Ciravegna. 2008. A comparative evaluation of term recognition algorithms. In N. Calzolari (CNF Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, *Proc of the 6th on LREC*, pages 2108–2113, Marrakech, Morocco. ELRA.

X. Zhang, Y. Song, and A. Fang. 2010. Term recognition using conditional random fields. In *Proc of IEEE NLP-KE*, pages 333–336.