Approaches and Strategies to Extract Relevant Terms: How are they being applied?

J. Valaski, S. Reinehr, and A. Malucelli

joselaine.valaski@pucpr.br,sheila.reinehr@pucpr.br, malu@ppgia.pucpr.br

PPGIa, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil

Abstract - *One of the goals of the term extraction is to identify and structure relevant information form texts. Despite advances in recent years, there are still several challenges for the development of efficient tools and methods related to this activity. Term extraction in specific domain knowledge is even a more complex issue, because sometimes you cannot rely on an initial vocabulary to support the extraction. Motivated by this question, an exploratory research to identify the approaches and strategies that are being applied to term extraction has been conducted. The main goal was to identify the most significant one. The research began with 174 proposals and finished with 25 that achieved the filter criteria. This research presents the main data related to these selected proposals.*

Keywords: Relevant Term, Term Extraction, Exploratory Research, Domain Knowledge.

1 INTRODUCTION

The recent years have witnessed a proliferation in unstructured text data [1]. In a world where the amount of digital data grows over more than 50% per year, any means to structure this data becomes increasingly relevant [2]. Difficulties arise when knowledge is contained in a textual format and no support is available. In these cases, techniques for automatic processing of a textual content are required [3].

In Natural Language Processing (NLP), several techniques exist for extracting terms from large collections of general and specific texts [1].

In general, these techniques are applied to extract relevant terms from texts. With these results it is possible to generate structured and formalized information. The goals of these techniques are diverse, such as, creating concept maps to assist in the learning process [4], building domain ontology [5], creating semantic model domain [6-7], extracting skills in job advertisements [8], improving search on the Web [9] and extracting important topics in a blog [10].

For the identification of relevant terms, distinct approaches and strategies can be applied, such as, linguistic, statistics and hybrid. Despite several proposals of tools and methods for the relevant term extractions, most extant term extraction algorithms are inadequate to address the challenges posed by domain-specific texts. A major challenge is the sparse nature of these texts, which do not offer reliable statistical evidence, and severely compromise the algorithms' performance [1].

The term extraction can be supported by an initial vocabulary or not. However, when you need to identify terms from a specific domain, in most cases, you cannot count on the support of this initial vocabulary. This restriction makes the development of algorithms with a good performance even more challenging.

In this context, it is considered relevant to obtain an overview of the proposals presented in this area. It is believed that with these results, it is possible to identify approaches and strategies that have been applied with greater relevance. These parameters can be used as criteria on the selection of tools and/or methods. Moreover, its parameters may support the improvement of existing tools and/or methods.

As it has not been found on scientific bases a work that presented these results in a systematic way, this research aimed to conduct this exploratory type of research method. The research was conducted through the analysis of scientific papers available at Scholar search site.

This research is structured as follows: Section 2 presents a brief theoretical framework about the most common approaches and strategies applied to the relevant terms' extraction. Section 3 describes in detail how the exploratory research was conducted. Section 4 presents and discusses the main results of this research. Section 5 concludes the research and presents the future work.

2 THEORICAL BACKGROUND

Several methods are available to extract terms from a set of documents. These methods can be broadly categorized into three different approaches: linguistic approaches, statistical approaches, and hybrid approaches [1][11].

2.1 Linguistic approach

Linguistic approach uses NLP for term extraction. Linguistic methods are often implemented as Part-of-Speech

(POS) filters, which accepts, as terms, any noun sequences containing optional adjectives and/or prepositions. A POS-tagger labels the part-of-speech of terms (e.g., adjective, noun, verb, etc.) appearing in a text. In general, there are three main techniques on this approach: syntactic, morphological and semantic analysis.

- Syntactic analysis: identifying the syntactic function of the word, such as noun, adjective and verb. POS-tagger tools are applied on this identification;
- Morphological analysis: derivation a term's form, e.g., whether a terms used in singular or plural form. The procedure lemmatization is used on this analysis. The lemmatization allows to group together in a single attribute the multiple morphological forms of words which have a common semantics; and
- Semantic analysis: identifying the meaning of words, normally obtained by means of an external base, for example, WordNet.

When extracting terms for a certain domain, they however do not consider the relevance of a term for that domain. Since linguistic methods rely on the syntactic structure, they identify terms according to the unithood property [5].

2.2 Statistical approach

Differently than the linguistic approaches, statistical methods do not use the linguistic characteristics of terms, but rely solely on statistical measures to extract terms [5]. These statistical methods are applied to acquire the relevance of a term for a domain.

Statistical approach concerning the termhood, it is statistically determined based on the observation that the highly frequent expressions in a domain specific corpus are likely to denote relevant terms. Another termhood estimation technique is that of corpus comparison, in which a domain-specific corpus is compared against a collection of general texts. Expressions that are more likely in the domain-specific corpus are then treated as domain-specific terms.

In general, the strategies on statistical approach are based on frequency with some variation: absolute frequency, frequency with comparison, frequency with weight and co-occurrence:

- Absolute frequency: count the absolute frequency of a term in a document.
- Frequency with comparison: count the frequency of a term in a document considering the frequency of a term in another document. An example is to use a domain-specific corpus to compare against a collection of general texts;
- Frequency with weight: count the frequency of a term in a document and apply distinct weight. As an example, a term that appears on the document title can have a bigger weight than a term that appears on the document body.
- Co-occurrence: count the frequency of two or more terms together. In this strategy the compound terms are considered more relevant than simple terms.

3 METHOD

An exploratory research was conducted to identify general and specific information applied to relevant terms extraction. This identification aimed to map the most relevant approaches and strategies.

The search was performed using the search engine Scholar. The Scholar was used because it enables the identification of diversified publications, such as, dissertations, theses, technical reports, articles, among others. The search on only renowned scientific bases, such as, Springer Link, Science Direct and IEEE, could limit the number of publications. This conclusion was obtained by a preliminary test.

The period between 2004 and 2014 has been defined for the selection of the publications. This criterion was established considering that this research does not aim to explore the evolution of all proposals for the extraction of relevant terms. The main goal was to identify the latest tools and methods related to this activity. It is believed that the period of 10 years is enough for this scenario.

The search string used was: "term extraction" and "tool" and "relevant term." The survey was conducted on 05/31/2014 and 174 publications were returned. The research was also performed without a date filter. In this scenario 212 publications were returned. This showed that the largest number of publications related to this topic, actually happened in the last 10 years. Thus, the filter was kept in date (2004-2014) and 174 publications were evaluated.

Among these 174 publications, patents, citations, books or files without access were identified. Thus, the publications without access were excluded, resulting in 96 publications with documents available for analysis. Among the 96 publications available, 52 were selected by reading the abstract. By reading these abstracts, it was feasible to identify the possibility of these proposals to be related to some tool or method applied to the extraction of relevant terms.

These 52 publications were read in full to ensure the application of some tool or method for the extraction of relevant terms. Furthermore, studies that depended on an initial vocabulary for term extraction were excluded. The goal of this research was to identify proposals that did not depend on an initial vocabulary to extract relevant terms. This restriction is important especially in specific domains where normally there is not an initial vocabulary to support this activity.

As a final result, 25 proposals were obtained, and the following information was extracted: year of publication, language of the processed text and tools applied on extraction approach. The results are shown below.

4 RESULTS AND DISCUSSION

The results are presented in two perspectives: general data and tools and the applied approaches and strategies. In the first perspective, information related to proposals origin and tools used are presented. In the second perspective, the approaches and strategies applied to relevant term identification are presented. The symbol UI (Unidentified Information) was used to indicate situations where information could be identified.

4.1 General data and tools

Table 1 presents the general data, such as: publication year, origin country of publication and idiom of the processed text.

Table 1. General data

Reference	Year	Country	Text idiom
[1]	2013	Netherlands	English
[3]	2011	Italy	Italian
[4]	2009	Brazil	Portuguese
[5]	2012	Italy	English
[6]	2010	Netherlands	English
[9]	2008	China	English
[10]	2012	South Korea	English
[11]	2014	Netherlands	English
[12]	2013	Iran	Farsi
[13]	2013	Brazil	Portuguese
[14]	2010	China	English
[15]	2008	Germany	English
[16]	2012	India	English
[17]	2012	Brazil	Portuguese
[18]	2005	Germany	English
[19]	2012	Spain	English Spanish
[20]	2013	Slovak Republic	**UI**
[21]	2004	China	Chinese
[22]	2013	Sweden	Swedish
[23]	2009	Netherlands Spain Italy	English Dutch Spanish
[24]	2012	Mexico	**UI**
[25]	2010	Brazil	Portuguese
[26]	2011	Tunisia	English
[27]	2013	Germany	English
[28]	2007	Austria	English

In order to obtain a better analysis, the data presented on Table 1 was summarized and is presented in Table 2. The data summarization allows some important information, which are discussed as follow.

From 2012 a bigger number of proposals related to the relevant term identification was observed. This could support the idea that the importance of this activity has increased in the last few years. It is important to emphasize that 2014 is a special period, because this research was performed in the first semester. Thus, it is expected that the number of publications is lower this year.

There is a big concentration of researches on European countries (Netherlands, Germany, Italy, Austria, Spain and Sweden). Among the analyzed proposals, 14 had European origin. In the American continent, Brazil is the highlight country and in the Asian continent, China is the highlight country.

Table 2. General data summary

Year	Qty	Country	Qty	Idiom	Qty
2004	1	Netherlands	4	English	15
2005	1	Brazil	4	Portuguese	4
2007	1	China	3	Spanish	2
2008	2	Germany	3	Chinese	1
2009	2	Italy	3	Swedish	1
2010	3	Spain	2	Farsi	1
2011	2	Austria	1	Italian	1
2012	6	South Korea	1	Dutch	1
2013	6	India	1	**UI**	2
2014	1	Iran	1		
		Tunisia	1		
		Mexico	1		
		Slovak Republic	1		
		Sweden	1		

English is the main language used to process texts. Among 25 analyzed proposals, 15 used a text written in English in order to evaluate the extraction instruments of relevant terms. The bigger the amount of applied works in English language is, it results in more significant advances in tools and methods applied to this language, than to any other.

Table 3 presents data related to applied tools. In this table, for each proposal the tools applied are presented in order to: perform term extraction, linguistic annotation (e.g., parser, tagger, dependency relationship) and other support (e.g., documentation indexing, n-gram extraction, summarization and semantic indexing).

Also in order to make a better analysis possible, the data showed on Table 3 was summarized and are presented in Table 4. The data summarization allows some important information, which is discussed as follows.

Nine different proposals of tools applied to relevant term extraction were found. However, only "EXATO LP" was found in more than one distinct proposal. However, it is important to emphasize that both proposals that used "EXATO LP," belong to the same research group. This result shows several tools but each one of them with its own approach. This can be seen more clearly in the following results.

It is important to emphasize that 15 proposals did not present a tool in order to perform relevant term extraction. Only an algorithm was applied, as it can be observed in the next results as well.

Table 3. Applied tools

Reference	Tools		
	Term extraction	**Linguistic Annotation**	**Other**
[1]	ExtTerm	Stanford	**UI**
[3]	**UI**		
[4]	**UI**		
[5]	Extractor (SAOD)	**UI**	Lucene
[6]	**UI**	FreeLing2 Alpino	**UI**
[9]	**UI**	Qtag	**UI**
[10]	**UI**		
[11]	ATCT	Stanford	**UI**
[12]	**UI**	Bijankhan	**UI**
[13]	**UI**		
[14]	**UI**	Survey parser	**UI**
[15]	SProUT	MINIPAR	Lucene
[16]	**UI**		
[17]	EXATO LP	PALAVRAS LX-center	**UI**
[18]	**UI**	Genia YamCha-Chunker	**UI**
[19]	**UI**	**UI**	Ngram statistics package
[20]	**UI**		
[21]	**UI**		
[22]	IPhraxtor	Connexor machinese syntax	**UI**
[23]	Tybot	**UI**	**UI**
[24]	**UI**		
[25]	EXATO LP	PALAVRAS	NSP tool
[26]	**UI**	**UI**	TextTiling LSI
[27]	ATExTA	**UI**	**UI**
[28]	Protégé plugin	**UI**	**UI**

These results may confirm the idea of the recent evolution of relevant term extraction. Many algorithms have been proposed but there not enough tools able to integrate the best algorithms results.

It has been also observed that several proposals used linguistic annotation to support the relevant term extraction. Twelve distinct tools have been identified. Only "PALAVRAS" and "Stanford" tools have been applied in more than two proposals. Moreover, documentation indexing, summarization and n-gram extraction tools to support the relevant term extraction have also been identified.

The main results obtained by these data were as follows: there is relevant application of linguistic annotation tools to identify features that enable the relevant terms identification; there are several proposals without a supportive tool, and among the proposals that use a tool, the approaches are very diversified.

Table 4. Tools summary

Term extraction	Qty	Linguistic Annotation	Qty	Other	Qty
EXATO LP	2	PALAVRAS	2	Lucene	2
ATCT	1	Stanford	2	Ngram statistics package	1
ATExTA	1	YamCha-Chunker	1	NSP tool	1
SProUT	1	Connexor machinese syntax	1	TextTiling	1
Protégé plugin	1	FreeLing	1	LSI	1
IPhraxtor	1	Genia	1		
Tybot	1	LX-center	1		
ExtTerm	1	MINIPAR	1		
Extractor	1	Survey parser	1		
UI	15	Bijankhan	1		
		Alpino	1		
		Qtag	1		

Due to the approach diversification, it was considered important to identify the ones that are being applied with more emphasis. In order to obtain this result, for each proposal, approaches and strategies applied were identified. The results are presented as follows.

4.2 Applied approaches and strategies

For each analyzed proposal, with the use of tools or not, applied approaches and strategies to relevant term extraction were identified. Table 5 presents these results. The approach and strategies presented in Section 2 to classify the proposals were used.

Linguistic approach

Among the 25 analyzed proposals, 17 distinct proposals that applied linguistic approach to relevant term extraction were identified. Considering the linguistic approach, the syntactic analysis was the most used one. Using the syntactic analysis strategies, 11 proposals [1][9][11-12][14-15][17-18][22-23][25] suggest the noun term identification as a main feature to relevant term extraction. Some of these proposals also suggest the adjective [1][12][15], verb [1] and preposition [1][12] term identification.

The proposals that used morphological analysis, in a general manner, had the goal to prepare the term to syntactic analysis. Applying the lemmatization process to reduce the word to its root, can be cited as an example. Only two proposals were identified considering the semantic analysis [6][23].

Statistical approach

All the 25 proposals applied strategies from statistical approach. The co-occurrence [1][6][9][11-12][16][18-21][24-25] and frequency with comparison [1][3][5][9-11][13][15][17][20][28] were the most used strategies from the statistical approach. The most highlighted co-occurrence strategy were: techniques to identify n-grams [9][19][24-25], MI (Mutual Information) metric [1][6][21] and C-Value metric [12].

In the frequency with comparison strategy, the highlighted metrics were as follows: TF-IDF (Term Frequency - Inverse Document Frequency) [3][9-10][20][28], TF-DCF (Term Frequency - Disjoint Corpora Frequency) [13][17], KF-IDF [15] and IG (Information Gain) [5].

Despite being a simple strategy, the absolute frequency was also applied in some proposals [4][22-23][26-27]. The frequency with weight was applied as follow: type of structure where the term appears, such as title [11]; typed of syntactic structure where the term appears [14], type of annotation performed by the user [20], and the position of a candidate term in the hierarch, the hypernym relations between candidate terms [23].

Hybrid approach

Among the 25 analyzed proposals, 17 applied strategies combined with linguistic and statistical approach. These results show that the relevant term extraction is more efficient if distinct strategies are applied together.

Considering the results obtained from the exploratory research, the following strategies were identified as the most relevant:

- Using of syntactic analysis to consider a noun;
- Applying the TF-IDF metric; and
- Identification of co-occurrence to consider a compound term.

All of these 3 strategies can be used together to rank the terms. The position that the term appears can be used to measure the relevance.

Among all proposals analyzed, only 3 [1][9][11] applied the 3 strategies identified as most relevant. In [1], a framework called ExtTerm oriented to term extraction is proposed. In [9] the strategies to improve the search on the Web are proposed. Finally in [11] a framework called ATCT to automatic build domain taxonomy from texts is proposed. In all proposals there is not availability of a tool to public use.

Table 5. Applied approaches and strategies

Reference	Approaches/Strategies						
	Linguistic			Statistical			
	Morphological	Syntactic	Semantic	Absolute frequency	Frequency with comparison	Frequency with weight	Co-occurrence
[1]		X			X		X
[3]	X				X		
[4]				X			
[5]					X		
[6]	X		X				X
[9]		X			X		X
[10]					X		
[11]		X			X	X	X
[12]		X					X
[13]					X		
[14]		X				X	
[15]		X			X		
[16]							X
[17]	X	X			X		
[18]		X					X
[19]							X
[20]	X				X	X	X
[21]							X
[22]		X		X			
[23]		X	X	X		X	
[24]							X
[25]		X					X
[26]	X			X			
[27]	X			X			
[28]	X				X		
Total	**7**	**11**	**2**	**5**	**11**	**4**	**12**

The obtained results give guide on the features to be considered to choose a tool to relevant term extraction. Through broader search, using the search engine Google, it was possible to identify free tools available to relevant term extraction. Most of these tools are not mentioned in the proposals analyzed in this exploratory research.

An evolution of this research could be a survey of the tools that meet the strategies indicated as the most relevant.

5 CONCLUSION AND FUTURE WORK

The application of techniques to relevant term extraction is a crucial activity to transform non-structured texts in structured and formalized information. In order to support this activity, there are several tool, methods and approaches.

Despite the diversity of proposals, we have not found a systematic map presenting how these instruments are being applied. We believe that this overview is important to support the decision on what tools and methods to choose. Moreover, this overview can suggest the improvement of existing tools and methods. This research had the goal of performing this systematic map.

The main obtained results were: the hybrid approach applications that bonds syntactic and statistical approach were the most relevant ones. The strategies that privilege the noun identification, compound term and use TF-IDF metrics are the most significant.

This survey was used to define the most efficient instruments to relevant term extraction. These results are being applied on an environment to support the semiautomatic build of ontological conceptual model. The next step is to integrate selected tools in order to obtain relevant term extraction on specific domain in this environment.

REFERENCES

[1] Ittoo, A., and Bouma, G. Term extraction from sparse, ungrammatical domain-specific documents. Expert Systems with Applications, v, 40, n. 7, p. 2530–2540, 2013.

[2] IDC, The digital universe decade — are you ready?, from http://www.emc.com/ collateral/demos/microsites/idc-digital-universe/iview.htm 2010.

[3] Amato, F., Mazzeo, A., and Scippacercol, S. A method for the evaluation of the peculiar lexicon significance. Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation, v. 2, n. 1, p. 54–64, 2011.

[4] Dalmolin, L. C. D., Nassar, S. M., Bastos, R. C., and Matheus G. P. A Concept Map Extractor Tool for Teaching and Learning. In *9th IEEE International Conference on Advanced Learning Technologies*, p.18–20, 2009.

[5] Eynard, D., Matteucci, M., and Marfia. A modular framework to learn seed ontologies from text. Semi-Automatic Ontology Development: Processes and Resources, 2012.

[6] Bosma, W., and Vossen, P. Bootstrapping Language Neutral Term Extraction. In 7th Language Resources and Evaluation Conference (LREC), 2010.

[7] Valaski, J., Reinehr S., and Malucelli, A. Environment for Requirements Elicitation Supported by Ontology-Based Conceptual Models: A Proposal. In Proceedings of the 2014 International Conference on Software Engineering Research and Practice (SERP'14), ISBN 1-60132-286-0, Las Vegas, USA, p. 144-150, 2014

[8] Reichhold, M., Kerschbaumer, J., Fliedl, G., and Winkler, C. Automatic Generation of User Role Profiles for Optimizing Enterprise Search. In ICSSEA, v. 43, p. 1-8, 2012.

[9] Chen, T., Chang, P., and Teng, W. Supporting Informational Web Search with Interactive Explorations. In IEEE International Conference on Signal Image Technology and Internet Based Systems, p.153–60, 2008.

[10] Park, J., Lee, S., Jung, H., and Lee, J. Topic word selection for blogs by topic richness using web search result clustering. In Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication - ICUIMC, 2012.

[11] Meijer, K., Frasincar, F., and Hogenboom, F. A semantic approach for extracting domain taxonomies from text. Decision Support Systems, v. 62, p. 78–93, 2014.

[12] Behin-Faraz, H., Passban, P., and Shokrollahi-Far, M. A reliable linguistic filter for Farsi term extraction. The 5th Conference on Information and Knowledge Technology, p. 328–31, 2013.

[13] Fernandes, P., Furquim, L. O. C, and Lopes, L. A Supervised Method to Enhance Vocabulary with the Creation of Domain Specific Lexica. In IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), p. 139–42, 2013.

[14] Zhang, X., and Fang A. C. An ATE system based on probabilistic relations between terms and syntactic functions. In 10th International Conference on Statistical Analysis, 2010.

[15] XU, F. Bootstrapping Relation Extraction from Semantic Seeds. Thesi, 2008.

[16] David, M. R., and Samuel, S. Clustering of PubMed abstracts using nearer terms of the domain. Bioinformation, v. 8, n 1, p. 20–25, 2012.

[17] Lopes, L., Vieira R., Fernandes, P., and Couto, G. Exatolp: extraction of language resources from Portuguese corpora. In International Conference on Computational Processing of the Portuguese Language – PROPOR, p. 45–47, 2012.

[18] Wermter, J., and Hahn, U. Finding new terminology in very large corpora. In K-CAP '05 Proceedings of the 3rd international conference on Knowledge capture, p. 137–44, 2005.

[19] Vàzquez, M., and Oliver, A. Improving Term Candidate Validation Using Ranking Metrics. In 3rd World Conference on Information Technology, v. 3, p. 1348–1359, 2013.

[20] Harinek, J., and Simkom M. Improving term extraction by utilizing user annotations. Proceedings of the 2013 ACM symposium on Document engineering - DocEng, 2013.

[21] Pu, H., and Chien, L. Integrating log-based and text-based methods towards automatic web thesaurus construction. In Proceedings of the American Society for Information Science and Technology, v. 41, n. 1, p. 463–471, 2005.

[22] Merkel, M., Foo, J., and Ahrenberg, L. IPhraxtor - A linguistically informed system for extraction of term

candidates. In Proceedings of the 19th Nordic Conference on Computational Linguistics, p. 121–132, 2013.

[23] Bosma, W., Vossen, P., and Soroa, A. KAF: a generic semantic annotation format. In Proceedings of the GL2009 Workshop on Semantic Annotation, 2009.

[24] Vieyra, S., Suárez-Figueroa, M. C., Estrada, H., and Martinez, A. knOWLearn: a reuse-based approach for building ontologies in a semi-automatic way. 2012.

[25] Lopes, L., Oliveira, L. H., and Vieira, R. Portuguese term extraction methods: Comparing linguistic and statistical approaches. In Proceedings of the 9th PROPOR, p. 1–6, 2010.

[26] Chabi, A. H., Kboubi, F., and Ahmed, M. B. Thematic analysis and visualization of textual corpus. International Journal of Computer Science & Engineering Survey, v. 2, n. 40, 2011.

[27] Houy, C., Sainbuyan, K., Fettke, P., and Loos, P. Towards automated analysis of fads and trends in information systems research: Concept, implementation and exemplary application in the context of business process management research." In IEEE 7th International Conference on Research Challenges in Information Science (RCIS), p. 1–11, 2013.

[28] Pammer, V. Two Protégé plug-ins for supporting document-based ontology engineering and ontological annotation at document-level. In 10th International Protégé, p. 1–3, 2007.