Evaluating Ontology Development from the Extraction of Noun Phrases

Alexandra Moreira[1], **Alcione P. Oliveira**[1], **Jugurta Lisboa-Filho**[1]

[1]Departamento de Informática – Universidade Federal de Viçosa (UFV)
Viçosa –MG – Brazil

{alexandra.moreira,alcione,jugurta}@ufv.br

Abstract. *There are several methods for constructing an ontology. Among the automatic methods, one approach is the extraction of terms from domain documents and their subsequent extraction. In this case, the first step of the process is the extraction of noun phrases that are potential candidates to be components of the terminology of the area of interest. This article describes an automatic tool for the Brazilian Portuguese language that extracts noun phrases that can be adopted as terms for a certain domain. In addition, the system couples the extracted terms into a top-level ontology, which results in an initial ontology that can be further refined. To couple with the ontology an anchor term was used, and a statistic analysis showed that the use of the term anchor leads to an improvement in the performance of the system. The tool described in this article was used to select terms to be used in an ontology for the power sector domain. Also, the precision in the creation of the ontology was evaluated. The technique was able to generate the correct hierarchy for 70% of the terms.*

Keywords: Ontology Development; Noun Phrase Extraction; Brazilian Portuguese.

1. Introduction

Ontologies are important tools for operating information systems. They assign meaning to the terms of a domain and allow the exchange of information between systems and users [Moreira et al. 2004]. Nonetheless, the construction of these resources is complex and involves many hours of work by knowledge engineers and domain experts [Sanchez and Moreno 2004]. In order to mitigate this problem and speed up the development of ontologies, semi-automatic methods have been proposed. Among the automatic methods, one approach is the extraction of candidate terms from domain documents to be inserted in the ontology being built. In this case, the first step of the process is the extraction of noun phrases that have potential to be components of the terminology of the area of interest. This article describes an automatic tool for the Brazilian Portuguese language that extracts noun phrases that can be adopted as terms for a certain domain. In addition, the system couples the extracted terms into a top-level ontology, which results in an initial ontology that can be further refined. To couple the ontology an anchor term was used, and a statistic analysis showed that the use of the term anchor leads to an improvement in the performance of the system. The tool described in this article was used to select terms to be used in an ontology for the power sector domain. It showed better performance when compared to other tools developed for the Brazilian Portuguese. Also the precision in the creation of the ontology was evaluated. The technique was able to generate the correct hierarchy for 70% of the terms.

This paper is organized as follows: the next section presents researches previously developed that are related to this work; Section 2 provides an overview of the system; Section 3 describes the term extraction module; Section 4 describes the ontology building module; Section 5 presents the results obtained; and Section 6 presents the final conclusions.

2. Related Work

Term and concepts extraction from a textual database and automated ontology creation are active topics of research and there are several projects being developed.

Maynard et al. [Maynard et al. 2008] presented NLP techniques for ontology population, using a combination of rule-based approaches and statistical techniques for term recognition. They have also used contextual information to bootstrap learning. According to them, the experiments have shown promising results. The fundamental difference of our work is that they did not make use of a previously built and well-founded ontology to fit the extracted terms.

Carvalheira [da Cruz Carvalheira 2007] proposes a semi-automatic method for creating ontologies, which is our goal as well. The method uses linguistic and statistical resources to extract concepts and relations candidates to compose the ontology. However, the method does not eliminate the participation of an expert to determine which terms actually should be incorporated in the ontology to be built. In our case, we aim to reduce the manual workload, setting automatically, the concepts and relationships between concepts in a two-level hierarchy. Another distinguishing feature is that the work extracts terms using the Brown *corpus* containing American English texts. Our work focuses on Brazilian Portuguese, where there is a shortage of linguistic resources and a small number of proposals.

Teline et al. [Teline et al. 2003] developed a term extractor called Exporter (Evaluation of Terminology Authomatic Extraction Methods for Portuguese Texts). The system was used in the BLOC-Eco project [Zavaglia et al. 2007], whose goal was to create a knowledge base with the ontological information about ecology terms in Brazilian Portuguese. The system is based on POS annotation and syntactic sequence patterns for unigrams, bigrams and trigrams. For example, one of the patterns for trigram would be `<Noun Preposition Adjective>`. Our study differs from the Exporter system as it does not have a limit on the size of the composed terms and for not having a fixed number of syntactic class sequence patterns.

Macken et al. [Macken et al. 2013] have created a bilingual terminology extraction system, called TExSIS, that uses a chunk based alignment method for the generation of candidate terms. The technique proposed requires multilingual *corpus* to perform the alignment of text segments. This fact distinguishes this research from the one proposed in this paper. Furthermore, the technique has not been tested in Portuguese.

Maia and Souza [Maia and Souza 2010] developed the software tool, named Ogma, to extract noun phrases from texts written in Portuguese. The aim of the authors was to check use the noun phrases as indexers for classifying documents. The Ogma tool is an extractor based on rules and is available on the Web[1]. The current research results

[1] http://www.luizmaia.com.br/ogma/

will be compared with the results obtained with the Ogma tool.

Kozareva [Kozareva 2014] proposed the creation of taxonomies automatically from a seed, a root term and from the mining of co-occurrence patterns of words. The proposed algorithm uses syntactic patterns to obtain hypernym candidates. The algorithm produces interesting results and it is able to propose new terms to preconstructed hierarchies, such as Wordnet. The main difference from our proposal is the fact that it does not use ontologies already defined, and therefore, it needs a later formalization.

Rani et al. [Rani et al. 2017] proposed the use of a text corpus of various topics to form an ontology using machine learning techniques. Two topic modeling algorithms were applied, namely LSI (Latent Semantic Indexing) & SVD (Singular Value Decomposition) and Mr.LDA (MapReduce Latent Dirichlet Allocation) for learning topic ontology. They were used to determine the statistical relationship between document and terms to build a topic ontology and ontology graph with minimum human intervention. The results obtained evidenced the effectiveness of using Mr.LDA topic modeling for Ontology Learning. It is an interesting approach, but it differs from our proposal to use machine learning techniques.

3. System Overview

This paper describes an automated tool created by the union of two modules developed by our research group, named EχTerm and AutOnGen (AUTomatic ONtology GENnerator). The EχTerm is used for extracting noun phrases that can be adopted as terms for a certain domain. The AutOnGen (AUTomatic ONtology GENnerator), receives as input those terms and assign meaning to them, building up an ontology. The system generates the ontology automatically, however, the ontology generated should be later examined by an expert to carry out adjustments. Some tests were also carried out, and it was shown that EχTerm has a better performance when compared to other tools with a similar goal that have been developed for the Brazilian Portuguese Language. Also, the ontology generated by AutOnGen had 64 to 70 percent of its terms correctly classified. This System tool is currently being used in the process of creating an ontology for the electrical power industry. The Fig. 1 shows the connection between the modules presented in this article and the flow of information processing.

4. The term extraction Module

First of all, it is important to emphasize the difference between noun phrases (NPs) and terms. NPs are syntactic structures whose semantic component indicate that they refer to entities in a discourse. On the other hand, terms are words or noun phrase that have a specific meaning in a particular language in a particular area. That is, they are used to define a concept in a specific domain and they have the syntactic and semantic aspects better defined. Having said that, a noun phrase is a candidate domain term that must pass the scrutiny of an expert.

The term extraction process takes as its starting point a *corpus* annotated with lexemes according to their syntactic classes. To carry out the annotation, we used the annotator Unigran Tagger from the NLTK package (Natural Language Toolkit)[2] trained

[2] http://www.nltk.org/

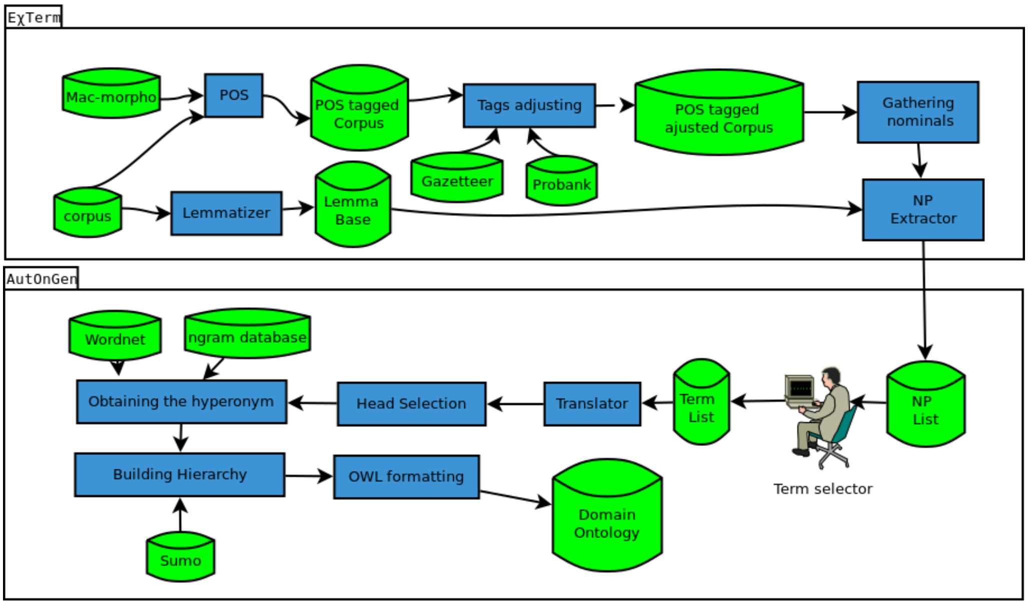

Figura 1. The interaction between the system modules. The system operates in pipeline fashion, where each module receives as input the output of the previous stage. The first phases belong to the system called E$_{\chi}$Term system. The phrases extraction module receives a *corpus* from a domain and issues a list of noun phrases. The noun phrases are examined by a specialist who extracts a list of terms that serve as input to the module that generates the ontology, called AutOnGen.

with the Mac-Morpho *corpus* [Aluísio et al. 2003]. After this step the resulting *corpus* undergoes an annotation adjustment step to reduce mistakes in the annotation process. For this purpose, the adjusting module uses a word database extracted from Probank.BR [Duran and Aluísio 2011], and a Brazilian proper names and Locations gazetteer.

The next step is the module most important one. It is the one executed by the submodule that performs the junction of the lexemes in order to create compound words or NPs. The joining is performed by a set of rules which basically combine separate nominal that may be united by prepositions and co-occurring adjectives. The rules adopted to detect NPs were the follows:

$$SN \rightarrow N(N \mid ADJ \mid < SPREP >)^*$$
$$SPREP \rightarrow PREP < SN >$$

where N=Noun, ADJ=Adjective, $SPREP$= Prepositional phrase, SN=Noun phrase and $PREP$= Preposition. Thus, the set of annotated terms for the noun phrase *“nível dos reservatórios das usinas hidrelétricas” (*level of hydroelectric power plant reservoirs)

```
('nível', 'N'), ('dos', 'PREP'), ('reservatórios', 'N'),
('das', 'PREP'), ('usinas', 'N'), ('hidrelétricas', 'N')
```

becomes

```
('nível dos reservatórios das usinas hidrelétricas', 'N').
```

```
mercado      mercado de Capitais
             mercado de energia
             mercado de capitais
             mercado do Grupo
             mercados
meta         meta de treinamento
             metas de crescimento
             metas de indicadores
             metas
metodologia  metodologia de cálculo para definição valores
             metodologia de cálculo das tarifas de fornecimento
```

Figura 2. A small segment of the initial grouping of extracted terms issued by the EχTerm system.

The last part of the process is to create a two-level hierarchy for the purpose of grouping the terms and facilitating the creation of an ontology. The terms have been grouped into classes named by the lemma form of the first lexeme of the term. The leftmost nominal in a noun phrase usually is the head of the phrase in Portuguese. In order to obtain the lemma form we used the lemmatizer available in NILC website (Interinstitutional Center for Computational Linguistics USP)[3]. The steps are summarized in Fig.1 and make up the system known as EχTerm.

An excerpt from the final file issued by the process can be seen in Fig.2. It is the result of applying the system to an annual report of a company for generating and transmitting electrical energy. As it can be seen, the system was able to extract noun phrases from the company's annual report that, even for a non-specialist, appear to be related to the domain.

After executing this module, the terms undergo a manual selection process by a specialist, and after that, serve as input to the module that generates the ontology based on a top-level ontology. The ontology building module is described in the next section.

5. The Ontology Building Module

Once the noun phrases have been selected, we can now call them terms. These terms feed the AutOnGen module that generates a preliminary ontology that can later be edited. This section describes how the AutOnGen module works. The module was partially described in [Moreira et al. 2016], but was not analyzed whether the technique used was statistically relevant. In the results section of this article this analysis is presented.

The module uses a top level ontology (SUMO [Pease et al. 2002]) and WordNet 3.0[Miller 1995] to assign meaning to the terms. We chose to use Princeton's Wordnet in place of Wordnet.br [Dias-da Silva 2006] due to the broader lexical coverage of the first and the emphasis mostly on verbs by the latter. The terms selected by the previous module were written in Brazilian Portuguese. As the next module is based on the use of lexical bases in English, the first task of the module AutOnGen is to translate the terms

[3] http://www.nilc.icmc.usp.br/nilc/index.php

into the English. For translating the terms the Google©translator was used through the libtranslate[4] library. The translation done by Google is based on the calculated probability on an immense base of translations and this translation is best accomplished if some contextual information, such as co-occurring terms, is provided. Therefore, providing the term in isolation may lead to erroneous translation. Therefore, to minimize the occurrence of spurious translation, a context indicator term, named *domain anchor*, was added to the previous term during the translation. The term anchor must be carefully chosen, and some tests should be performed to determine the best choice. As stated by [Moreira et al. 2016], the *domain anchor* may ensure, for instance, that the term "acordo" be translated as "agreement" and not "wake up" by simply adding the term "negócios" (business) as *domain anchor*.

After the translation phase, the terms were syntactically annotated by the POS tagger in order to identify the core lexeme of the noun phrase. This is done in order to obtain the term head that will later be applied in a query to WordNet to obtain the hypernym of the term head. This is particularly important in the case of composite terms. In this case, when WordNet does not return the hypernym of the compound term, the term head is used in a new query for the hypernym search.

The term head of a compound term is selected according to the Right-hand Head Rule (RHHR), proposed by [Williams 1981] that states that if the composite term does not have prepositions or conjunctions, it will be selected the last nominal of the compound term, otherwise it will be selected the last nominal occurring before the preposition or the conjunction.

The Wordnet query may return more than one hypernym candidate term, and in this case, some form of disambiguation is required. The disambiguation is done by verifying which term is the most likely to be the hypernym term via the ngrams database provided by Google[5]. The probability is obtained by performing a query to the base using as an argument of the query the concatenation of the original term with the candidate hypernym. Thus, for the term "business" and the hypernym candidate term "group"is formed the compound term "group business". According to [Moreira et al. 2016], the underlying hypothesis of this technique is that the construction <*noun noun*>, where the second term modifies or qualifies the first, it is a common linguistic construction in the English language. This type of compound was called subsumptive compound by Marchand [Marchand 1969] *apud* [Lieber and Stekauer 2009]. In the results section, it is checked if such an approach leads to better results than a random selection.

Fig. 3 shows the probability of occurrence of a bigram involving the "business"term and hypernym candidates returned by WordNet. The graph of the Fig. 3 was obtained from the Google Ngram viewer. It's important to highlight that the probabilities returned by Google may vary a great deal depending on the probability calculation period (it was used from 1980 to 2012), the *corpus* used (it was used American English version 2012) and the smoothing factor applied (3).

After obtaining the hypernym, the rest of the hierarchy is obtained through successive queries to the Sumo ontology, where each query searches for the hypernym of the

[4] http://www.nongnu.org/libtranslate/

[5] https://books.google.com/ngrams

current term. The final output of the system is an owl file containing the entire hierarchy. The Fig. 4 displays a small segment of a generated ontology for the electrical energy sector.

6. Results

Ideally, to assess whether the proposed process is a viable alternative for term extraction, it would be necessary to compare their performance against other tools when applied to a pair of "corpus × terms" established in advance. However, as already mentioned, there is a lack of both tools and *corpus* of tests for the Brazilian Portuguese.

In order to carry out a test to show the true potential of the tool, it was made a comparison with another tool, the Ogma tool [Maia and Souza 2010], using a more appropriate *corpus*. To verify the precision and recall the results were compared with a list of terms extracted from the *corpus* manually by an expert. Is worth mentioning that we compared the output of tools that emit a list of noun phrases, with the list of terms

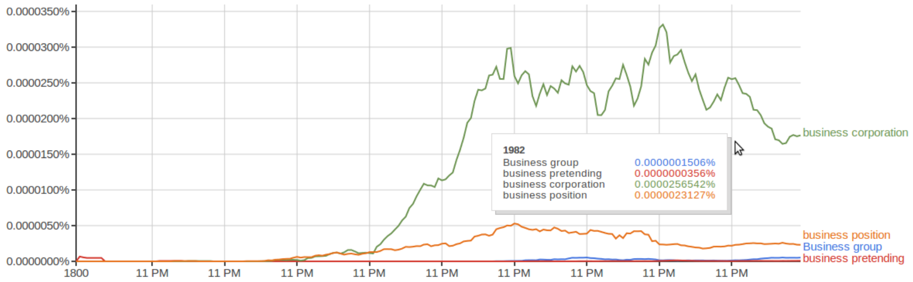

Figura 3. A screen capture of the Google Ngram viewer. The y-axis denotes the probability of occurrence of n-gram, and the x-axis denotes the time.

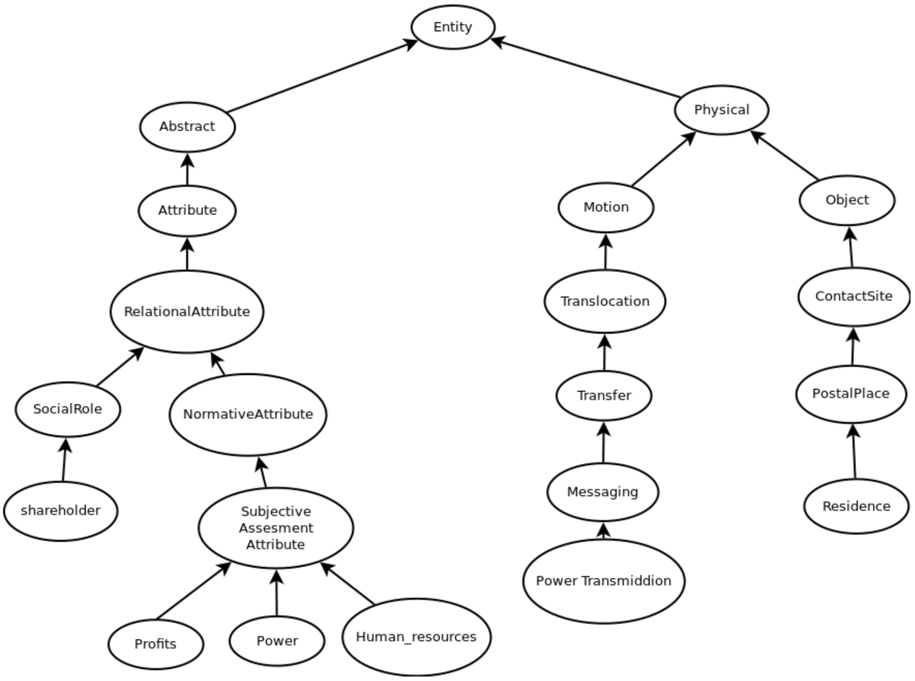

Figura 4. A generated ontology sample. The terms in the sheets show the terms extracted from the corpus and translated into the English.

prepared by the expert. The expert has prepared a list of 142 terms.

Tool	N. extracted NPs	match w/ human	Precision P	recall R	$F1 = \frac{2PR}{P+R}$
Ogma	211	19	9,0%	13,38%	10,76
EχTerm	132	49	37,12%	34,5%	35,76

Tabela 1. Comparison of results with Ogma tool. Although the tools have used similar techniques, the proposed system is superior in all performance measures. The comparison was made with a list 142 of terms produced by a specialist from the same text used by automatic extractors.

In Table 1 one can see that the proposed tool exhibited a higher performance than the Ogma tool. It was higher in both precision and recall and the F-measure strongly reflects this superiority. We believe the reason for this is that the rules of Ogma tool produce NPs that include some elements not belonging to a term. For example, the tool issued the noun phrase "os dados financeiros" (the financial data) and, in this case, the more appropriate term candidate would be "dados financeiros" (financial data). Furthermore, the tool produces some spurious noun phrases, such as "às suas" (to their) and does not join some nominal to form a compound term, such as "empresa" and "controladora" to form "empresa controladora" (controlling company).

In order to test the AutOnGen Module, we re-execute the EχTerm module with a larger *corpus* and extracted 4114 terms. From these extracted terms we randomly sampled 100 terms for testing. The output of the system was analyzed to verify the system performance. Table 2 shows the result of running the system. The *Related* column displays the number of items that were framed in a wrong sense, but somehow related. The *spurious* column displays the number of items that were framed in a totally wrong sense. The *Not in WordNet* column displays the number of items that are not included in the WordNet database. A *related* framing is, for instance, to state that *Relationship agent* is a type of *Relation* when the correct would be to state that it is a type of *Agent*. A *spurious* framing is, for instance, to state that *electricity* is a type of *EmotionalState* when the correct would be to state that it is a type of *Energy*.

Tabela 2. Execution result. *Related*: **number of items assigned wrong but somehow related sense.** *spurious*: **number of items assigned a totally wrong sense.** *Not in WordNet*: **number of items not included in the WordNet database.**

Correct	*Related*	*spurious*	*Not in WordNet*
64	16	17	3

Examining the results presented in Table 2 certain conclusions can be drawn. Only three items in a hundred were not found in WordNet, which attests to its wide lexical coverage. Other lexical bases were attempted, such as DBpedia and FrameNet, but failed to obtain the same performance. 64 items were understood as classified correctly, 16 classified erroneously, but related and 17 totally wrong. All the terms have been translated correctly. The reasons for the wrong framing fall into two categories: 1) the term has multiple senses, and the system chose the incorrect sense to the domain; or 2) there is no appropriate option to frame the term.

To try to improve the selection of hypernym when WordNet returns more than a concept, a small change in the system was performed: instead of using a bigram formed by the concatenation of the hypernym term with the hyponym term, it was used a bigram formed by the concatenation of an anchor term with the candidate hypernym. The anchor term must be composed of only a Word to form a bigram and benefit from extensive coverage of bigram from the google base. The Table 3 shows the result of running the system using this modification and the anchor term "enterprise".

Tabela 3. Execution result with the additional module

Correct	*Related*	*spurious*	*Not in WordNet*
70	17	10	3

The results in the Table 3 show that there was an improvement in the number of correctly classified terms. Also, there has been a substantial improvement in the reduction of spurious cases. This may be the clue that this is a good way to be followed, but the result can vary greatly depending on the term anchor adopted.

In order to test whether the term anchor is a viable resource, an experiment was conducted. From a base of 724 terms that have more than one candidate for the hypernym term, were randomly selected 100 terms which were henceforth, divided into 10 groups of 10 terms. For each term of each group were taken three actions: 1) hypernym was randomly selected; 2) the hypernym was selected using a bigram formed by the concatenation of the hypernym term with hyponym term; and 3) the hypernym was selected using the concatenation of an anchor term with the candidate hypernym. After that the result was punctuated as follows: If the selected term was spurious received 0 points; if the selected term was somehow connected received 1 point; and if the term was well selected for the area received 2 points. Then the points were added for each approach and for each group. (Table 4)

Tabela 4. Score for each approach applied in groups of 10 terms.

	1	2	3	4	5	6	7	8	9	10
X = random	8	6	11	10	9	17	9	11	10	13
Y = hypo-hyper	13	9	9	10	13	12	10	11	15	15
Z = anchor	14	11	11	15	14	15	14	16	14	15

Apparently the term anchor presents an improvement, and only in the Group 6 random selection was superior. But does this superiority was a coincidence or is it statistically significant? To verify that a paired t-test[6] was performed on the results.

For this analysis the degrees of freedom is $d.o.f = n - 1 = 9$ and the critical value for t with degrees of freedom = 9 and $\alpha = 0.05$ (95% confidence) is 2.263. The null hypothesis is that the means of the two samples are equal. Therefore, comparing the random assignment with the assignment using the anchor term we have:

$$\bar{XD} \approx -3.5 \tag{1}$$

[6] http://www.encyclopediaofmath.org/index.php?title=Student_test

$$S_{X_D} = \sqrt{\frac{1}{n-1}\sum_{n}^{i=1}\left(X_{Di} - \bar{XD}\right)^2} \approx 2.6352 \tag{2}$$

$$t = \frac{\bar{XD}}{\frac{S_{X_D}}{\sqrt{n}}} \approx -4.2001 \tag{3}$$

where $\bar{XD}$ is the mean of differences between pairs and S_{X_D} is the standard deviation of differences between pairs. The absolute value of the calculated t exceeds the critical value (4.2001>2.263), hence the null hypothesis is rejected . So the means are significantly different and the use an anchor term improves the selection. On the other hand, the use a bigram formed by the concatenation of the hypernym term with hyponym term resulted in an absolute value of the calculated t smaller than the critical value (1.2851<2.263), so the means are not significantly different and, therefore, we cannot say that this approach offers enhancements over the random choice of hypernym.

7. Conclusions

Satisfactory results in the execution of tools as described in this work depend largely on the performance of more basic tools on which they rely upon. These more basic tools would be the lemmatizers, POS taggers, Stemmers and, parsers. Despite the efforts of some Brazilian research groups to provide these tools, there is still a lack of basic tools for the Brazilian Portuguese Language and there is room for improvement in this area. The results obtained in this study suggest that a bottom-up approach for extracting terms can increase the recall without a great loss of precision. However, more tests and improvements are needed in order to reduce the terms extraction process dependency of a human expert.

It was hard to find tools available to perform a comparison. Only one was found to download and the results showed that the tool proposed in this paper has a notably superior performance. Due to an inappropriate *corpus*, the tool had a low performance, but the proposed tool showed a better recall rate. The next step is to place a statistical filter trained with a larger *corpus* to filter out spurious terms and improve the accuracy of the tool. In the case of the AutOnGen module, the ontology generated can benefit from all the definitions and relationships designed for the SUMO ontology, but the system does not generate relationships between the terms of the domain. Thus, it's not captured relationships as *power company has shareholders.* This problem must be addressed in future versions.

The system was applied to a list of terms extracted from the electrical power domain. The generated ontology had 64 to 70 percent of its terms correctly classified and the ontology expressed in OWL language could be readily edited by various tools available. The critical point of the process is the proper selection of the hypernym for the term translated. Using the probability of google n-gram database over a bigram formed by an anchor term combined with the hypernym term candidate showed the best results. It is planned for inclusion in the next version of the system, a more suitable technique for selecting the best hypernym of a word among the options returned by WordNet. Probably, the use of a domain-oriented *corpus* would produce better results, but that will be tested in a next version.

Acknowledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and also by the funding agencies FAPEMIG and CNPq.

Referências

Aluísio, S., Pelizzoni, J., Marchi, A. R., de Oliveira, L., Manenti, R., and Marquiafável, V. (2003). *Computational Processing of the Portuguese Language*, chapter An account of the challenge of tagging a reference corpus for brazilian portuguese, pages 110–117. Springer.

da Cruz Carvalheira, L. C. (2007). *Método semi-automático de construção de ontologias parciais de domínio com base em textos.* PhD thesis, Universidade de São Paulo, São Paulo.

Dias-da Silva, B. C. (2006). Wordnet. br: An exercise of human language technology research. In *Proceedings of the Third International WordNet Conference-GWC*, pages 22–26.

Duran, M. S. and Aluísio, S. M. (2011). Propbank-Br: a brazilian portuguese corpus annotated with semantic role labels. In *8th Brazilian symposium in information and human language technology*, pages 164–168, Cuiaba, Brazil. Sociedade Brasileira de Computação.

Kozareva, Z. (2014). *Text Mining*, chapter Simple, fast and accurate taxonomy learning, pages 41–62. Springer.

Lieber, R. and Stekauer, P. (2009). *The Oxford handbook of compounding.* Oxford University Press, Oxford.

Macken, L., Lefever, E., and Hoste, V. (2013). TExSIS: bilingual terminology extraction from parallel corpora using Chunk-based Alignment. *Terminology*, 19(1):1–30.

Maia, L. C. G. and Souza, R. R. (2010). Uso de sintagmas nominais na classificação automática de documentos eletrônicos. *Perspectivas em Ciência da Informação*, 15(1):154–172.

Marchand, H. (1969). *The categories and types of present-day English word-formation: a synchronic-diachronic approach.* Verlag C. H. Beck, München.

Maynard, D., Li, Y., and Peters, W. (2008). *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, volume 167, chapter Nlp techniques for term extraction and ontology population, pages 107–127. Ios Press.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Moreira, A., Alvarenga, L., and Oliveira, A. P. (2004). O nível do conhecimento e os instrumentos de representação: tesauros e ontologias. *DataGramaZero-Revista de Ciência da Informação*, 5(6):1–25.

Moreira, A., Lisboa Filho, J., and Oliveira, A. (2016). Automatic creation of ontology using a lexical database: an application for the energy sector. In *International Conference on Applications of Natural Language to Information Systems*, pages 415–420. Springer.

Pease, A., Niles, I., and Li, J. (2002). The suggested upper merged ontology: a large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28, Edmonton, Alberta, Canada. AAAI Press.

Rani, M., Dhar, A. K., and Vyas, O. (2017). Semi-automatic terminology ontology learning based on topic modeling. *Engineering Applications of Artificial Intelligence*, 63:108–125.

Sanchez, D. and Moreno, A. (2004). Creating ontologies from Web documents. *Recent Advances in Artificial Intelligence Research and Development*, 113:11–18.

Teline, M. F., Almeida, G., and Aluisio, S. M. (2003). Extração manual e automática de terminologia: comparando abordagens e critérios. In *16th Brazilian Symposium on Computer Graphics and Image Processing-SIBGRAPI*, São Carlos, Brazil. IEEE Computer Society.

Williams, E. (1981). On the notions “lexically related” and “head of a word”. *Linguistic inquiry*, 12(2):245–274.

Zavaglia, C., Aluísio, S., Nunes, M. G. V., and Oliveira, L. (2007). Estrutura ontológica e unidades lexicais: uma aplicação computacional no domínio da ecologia. In *Proceedings of the 5th Workshop in Information and Human Language Technology*, pages 1575–84, Rio de Janeiro, Brazil.