2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)

Automatic Extraction of Domain Specific Non-Taxonomic Relations from Portuguese Corpora

Vinicius H. Ferreira
Computer Science Dept.
IFRS Technology Institute
Feliz, Brazil
vinicius.ferreira@feliz.ifrs.edu.br

Lucelene Lopes Renata Vieira
Computer Science Dept.
PUCRS University
Porto Alegre, Brazil
{lucelene.lopes, renata.vieira}@pucrs.br

Maria José Finatto
Linguistic, Philology and Literature Dept.
UFRGS University
Porto Alegre, Brazil
mfinatto@terra.com.br

***Abstract*—This paper presents an automatic method to extract domain specific non-taxonomic relations from previously processed Brazilian Portuguese corpora. The proposed method is detailed and exemplified through a five corpora experiment. The obtained relations can be visualized and handled through an intuitive web interface and the results were evaluated by an human made analysis. The results show the positive performance of the extraction method and their perspectives for different kind of linguistic applications.**

I. Introduction

Ontology Learning (OL) is based on Natural Language Processing, Machine Learning and Data Mining techniques to help automatic and semi-automatic ontology construction [1]. The source for OL can be structured data (data bases), semi-structured data (dictionaries), or non structured data (natural language texts), which is the more abundant form of information available [2].

Information extraction from texts may occur in a myriad of forms, but one of the most popular task is the search for concepts in domain corpora. Such kind of task starts from basic term extraction, then it performs term relevance estimation to identify the concepts, *e.g.*, [3], [4]. Among these initiatives, the work of Lopes [4] is the only one applied to Brazilian Portuguese language.

One of the possible outputs of Lopes' extraction process is a linguistic resource with the relevant terms, assumed to be concepts, and its context information. Such context information includes[1]: the canonical form and pos-tag (noun, adjective, *etc.*) of each concept, its grammatical role in the sentence (subject, direct object, *etc.*) and the predicate (in its canonical form) to whom the concept plays its grammatical role.

In the process of OL, the non-taxonomic extraction phase has been acknowledged as the more complex and also the more neglected one [5], [6]. According to these authors, and also in the present paper, a non-taxonomic relation is the relation between concepts, usually expressed by a verb, or a verbal phrase, that does not establish an hierarchy among concepts. An example of such kind of relation in the Law domain is the relation "Represents" that occurs between the concept "Lawyer" and the concept 'Client" [2].

[1]The process defined in Lopes' work [4] has various linguistic resource outputs, each providing contextual information. However, in this paper we will be limited to the mentioned list of concepts and the cited contextual information, since only this data is necessary to the method proposed here.

In the literature one may encounter methods of non-taxonomic relations extraction as those of Sanchez and Moreno [5], and Villaverde *et al.* [6]. However, none of these can be applied to Brazilian Portuguese language texts, nor they use contextual information related to the concepts.

The EχATO$_{lp}$ tool [7] allows a sophisticated concept extraction process based on linguistic approach, since it identifies noun phrases tagged by PALAVRAS parser [8] and it applies linguistic extraction heuristics [9]. EχATO$_{lp}$ also uses a statistic-based relevance index called *tf-dcf* [10] to classify the terms candidate to concepts. To our proposed method it is relevant the EχATO$_{lp}$ output that delivers a list of terms and its contextual information, as well as a list of concept candidates and its *tf-dcf* index. These two lists provide the information needed to identify non-taxonomic relations of the domain.

The contribution brought by our work, besides being applied to Brazilian Portuguese texts, resides in the source of information to identify the relations. The proposed process starts with the list of more relevant terms and its contextual information as delivered by EχATO$_{lp}$ tool. From this information, sets of triples *<Concept 1, Verb, Concept 2>* are generated if in a same sentence *Concept 1* is found as subject of *Verb* and *Concept 2* is found as object of this same *Verb*. Additionaly, statistical analysis are made to estimate the relevance of each relation with respect to the domain corpus where it was found. Therefore, an important contribution of our work is the fact that we discover domain specific relations.

The proposed process was applied to five domain corpora, generating non-taxonomic relations to each corpora. This result was examined in three ways: *(i)* analysis of extracted relations specificity through comparison with the verbs encountered in a generic corpus with texts from a brazilian newspaper called "*Diário Gaúcho*"; *(ii)* analysis of the domain concepts; and *(iii)* analysis of the extracted non-taxonomic relations from the point of view of human specialists in the context of semantic role analysis.

This paper is organized as follows: Section 2 presents related works to extract non-taxonomic relations. Section 3 describes this paper main contribution, *i.e.*, our proposed method to extract relations from processed corpora. Section 4 exemplifies the proposed method to five corpora, and analyze the obtained results via an human made systematic observation. Finally, the conclusion summarizes this paper contribution and points out future works.

978-1-4799-2902-3/13 $31.00 © 2013 IEEE
DOI 10.1109/WI-IAT.2013.167

II. Overview of Related Work

Table I presents a comparison of works on non-taxonomic relation extraction that are similar to the one presented in this paper. In this table it is shown: the name, year of publication and reference to each work (*work*), the source of input data (*data source*), the indication (yes or no) if this work looks for the verbs to identify the relations (*verbs*), and the language(s) to whom the work was tested (*language*).

TABLE I. Comparison with related work.

work	*data source*	*verbs*	*language*
Maedche and Staab (2000) [11]	domain corpus	no	German
Schutz and Buitellar (2005) [12]	domain corpus	yes	English German
Sanchez and Moreno (2008) [5]	web	yes	English
Villaverde *et al.* (2009) [6]	domain corpus, candidate concepts and concept hierarchy	yes	English
Weichselbraun *et al.* (2009) [13]	ontology and domain corpus	yes	English
Serra and Girardi (2011) [2]	domain corpus	no	English

III. The Proposed Method

Unlike the presented related work, the method proposed in this paper starts from previously identified domain concepts. Consequently, it is not necessary to directly process the domain corpus, allowing the exclusive focus on the non-taxonomic relation extraction. Figure 1 presents a representation of the proposed method split in five steps detailed in this section.

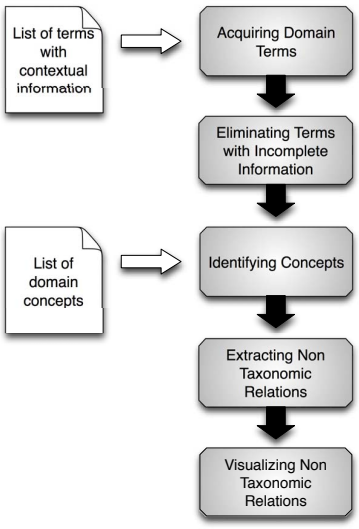

Fig. 1. Proposed method overview.

A. Acquiring Domain Terms

This step represents the capture of domain terms previously extracted by EχATO$_{lp}$ software tool with all its contextual information. As mentioned before, the relevant information for the proposed method are the canonical form of each term, its pos-tag, its grammatical function, the predicate (in the canonical form) to whom the term plays its grammatical function and its "address", *i.e.*, the sentence and position of the predicate inside the sentence.

B. Eliminating Terms with Incomplete Information

The second step consists in eliminating terms with missing essential contextual information to discover non-taxonomic relations. Specifically, the only terms kept are those with all information previously mentioned. In fact, are discarded terms where the grammatical function is different from subject or object (in all possible forms, *i.e.*, indirect, direct, *etc.*); and terms without predicate associated.

C. Identifying Concepts

The third step intends to discard terms that are not sufficiently relevant to be considered concepts of the domain. For that, all terms remaining from the second step are compared to the concept candidates list established by EχATO$_{lp}$ according to *tf-dcf* computation and the choice of cut-off points to keep around 15% from all terms extracted from the corpus.

It is important to call the reader attention that similar works, *e.g.*, [2], [5], [6], perform this concept identification at the beginning of their methods. Since our method relies on the choice made by EχATO$_{lp}$, this decision can be made later on. Nevertheless, this is a very important step in our method, and all other methods alike, since the goal is to extract non-taxonomic relations among concepts, and consequently only concepts must be taken into account.

D. Extracting Non-Taxonomic Relations

The fourth step is probably the most important one in the whole method. As in almost all similar works [2], [5], [6], [12], [13], this step extracts non-taxonomic relations between concepts inside a same sentence. Specifically, it discovers a relation between subject and object of the same predicate.

In the extraction methods proposed by Sanchéz and Moreno [5], Villaverde *et al.* [6], and Serra and Girardi [2], to each relation discovered a numeric value, *i.e.*, an index, is assigned to indicate the relation relevance to the target domain. According to such index, the discovered relations are ranked to later selection by domain specialists. Similarly, in our method, after identifying each relation two relevance indices are computed. Those indices are: *(i)* the accumulated frequency, and *(ii)* the shared frequency. To both indices, the basic information considered is the *tf-dcf* [10] index of each concept that belongs to the relation.

The accumulated frequency index of a relation r (*facum*$_r$) is computed as the sum of the *tf-dcf* index of both concepts related by each instance of relation r (Eq. 1) This index aims to estimate the relevance of relation r to the domain as a direct function of the concepts related by r relevance, regardless if the concepts were subjects or objects in the sentence.

$$facum_r = \sum_{\forall(t_1,r,t_2)} tf\text{-}dcf_{t_1} + tf\text{-}dcf_{t_2} \quad (1)$$

Observing equation 1, it is important to notice that $\forall(t_1, r, t_2)$ represents all instances of relation r. Therefore, the *tf-dcf* value of a given term t must be add as many times as term t appear as subject or object of instances of r. It is usual to have a same term being repeatedly accounted, since the more frequent relations often repeatedly relate the same terms.

The shared frequency of a relation r ($fcompart_r$) is computed as the ratio between the accumulated frequency of a relation and the number of instances that this relation has (Eq. 2). This index can be seen as the average relevance of instances of a given relation.

$$fcompart_r = \frac{facum_r}{\sum_{\forall(t_1,r,t_2)} 1} \quad (2)$$

Equation 2 can be seen as a normalization of the accumulated frequency index. In contrast with the *facum* index, the *fcompart* index indicates as relevant a relation that relates relevant concepts. However, the *fcompart* index is not affected by number of instances, but solely to the relevance of its related concepts. It is important to call the reader attention that both presented indices have their interest to rank relations, and our method does not privilege one over the other. Both indices are available, and according to the application of the extracted non-taxonomic relations one may decide to rank the relations following any possible weighting of each index.

E. Visualizing Non-Taxonomic Relations

The fifth and last step of the proposed method is the visualization of the extracted non-taxonomic relations. This step is a simple one, but it is, nonetheless, essential to make the relations available to human observation and handling. Although none of the similar works presents a visualization step, Schutz and Buitelaar [12], and also Villaverde *et al.* [6], mention the use of a web interface to show extracted relation available to domain specialists.

In our method, the visualization plays an important role and, therefore, the extracted relations were made available through a public software tool where non-taxonomic relations extracted from five different corpora can be browsed by each of its components, *i.e.*: *(i)* by the subject concept, *(ii)* by the predicate, and *(iii)* by the object concept. Within each of these options, the non-taxonomic relations can be ranked according to both presented index, *i.e.*: $facum_r$, and $fcompart_r$. This software tool is written in Brazilian Portuguese, and it is available at the address: http://www.vhflabs.com.br/nontax.

IV. Application and Analysis

To validate the proposed method, a set of experiments were performed over five domain corpora. After that, the extracted non-taxonomic relations were made available to linguistic students and an evaluation was made. The next subsections describe this two validation steps.

A. Application to Five Corpora

The case study performed took five previously available corpora on the following domains: Pediatrics (Ped - 281 texts), Geology (Geo - 234 texts), Data Mining (DM - 53 texts), Stochastic Modeling (SM - 88 texts), and Parallel Processing (PP - 62 texts). The Ped corpus was originally developed by Coulthard [14], and it contains texts from the Brazilian Journal of Pediatrics. The other corpora were assembled by Lopes [4] and they are composed by conference and journal papers, dissertation and thesis published by researchers or available at the Brazilian Repository of Thesis and Dissertations.

Table II presents some information about these corpora, according to their processing by EχATO$_{lp}$ software tool, and the application of our proposed method, specifically, how many instances of relations were discovered to each corpus and how many distinct relations (last column).

TABLE II. Corpora characteristics.

Corpus	Original Characteristics			EχATO$_{lp}$	Extracted	
	sentences	words	terms	concepts	instances	relations
Ped	27,724	835,412	180,120	8,270	159	93
Geo	69,461	2,010,527	436,401	25,173	306	129
DM	42,932	1,127,816	244,439	12,816	128	65
SM	44,222	1,173,401	252,178	12,582	228	94
PP	40,928	1,086,771	241,145	11,591	374	160

B. Human Made Analysis

This result was examined in three ways: *(1)* analysis of extracted relations specificity through comparison with the verbs encountered in a generic corpus from a brazilian newspaper called Diário Gaúcho (compiled by PorPopular Project http://www.ufrgs.br/textecc/porlexbras/porpopular/; *(2)* analysis of the domain concepts; and *(3)* analysis of the extracted non-taxonomic relations from the point of view of human specialists in the context of semantic role analysis (35 Terminology undergraduate students and 1 Ph.D. student at the Translation Program of UFRGS University).

1) Relation specificity: The specificity of the extracted relations was established with the help of three linguistic students. These students receive the 20 more frequent predicates found in a generic corpus assembled with texts from a brazilian popular newspaper called "*Diário Gaúcho*".

Each student proceed by looking for predicates composed by only one word (a verb) that was adjacent to either the subject or the object of the sentence. Then the 20 more frequent verbs were considered, discarding Portuguese connection verbs[2], and the verb "*haver*" ("to exist" in English).

Next, the linguistic students searched the generic corpus frequent verbs among the non-taxonomic relations of each corpus. If found, the most relevant relation instance, *i.e.*, the instance with the higher *facum*, was searched in the WebCorp [15] tool[3]. Assuming that the more frequent verbs of the generic corpus represent generic relations, it was desirable to not found occurrences in the WebCorp of this verb with the same subject or the same object.

Fortunately, the analysis of specificity of extracted relations was considered satisfactory, since most of extracted relations were not found at all as frequent verbs of the generic corpus. Additionally, in the rare case of found the same verb, never a pair "subject+verb", nor a pair "verb+object" among the extracted relation triplets was also found in the WebCorp. This result indicates that the extracted non-taxonomic relations are indeed specific to the domain corpus.

[2]In Portuguese the connection verbs are: "*ser*" ("to be" in English), "*estar*" ("to be" in English), "*permanencer*" ("to stay" in English), "*ficar*" ("to stay" in English).

[3]WebCorp tool is a concordanciator for Internet textual material, *i.e.*, it is a software that looks for some word or expression in a large set of texts and shows how this word or expression is employed in sentences of these texts.

2) Concepts found as subjects and objects: The analysis of concepts found as subjects or objects was conducted by 35 linguistic students. In this analysis, the students were split in groups, and each group attempt to choose in decrescent order of relevance the ten more relevant subjects and the ten more relevant objects of each domain corpus. This analysis was made with the careful reading of all texts in each corpus, but they also had access to the WebCorp tool in order to help them choosing concept candidates by correlation between the concept and the domain name.

After this hand made selection of relevant subjects and objects, the linguistic students have accessed the visualization tool proposed in the previous section in order to verify how their manual choice of relevant subjects and objects matched (or not) the subjects and objects indicated as more relevant by the proposed method. Obviously, such analysis is more related to the quality of concepts identified by EχATO$_{lp}$. However, it also indicated that our proposed method kept the relevance initially indicated by EχATO$_{lp}$, after choosing the more relevant non-taxonomic relations.

There were draw curves with the average relevance estimated by the students and curves with the relevance estimated by the *tf-dcf* index provided by EχATO$_{lp}$ original output. Once again, the analysis has shown a good correlation (over 80%) between the relevance estimated by linguistic students and automatically computed with *tf-dcf*.

3) Analysis of semantic roles: The third experiment was made by the subjective analysis of the non-taxonomic visualization tool performed by a Ph.D. student in linguistic studies that reported their use in the context of semantic roles research. According to the Ph.D. student evaluation, the visualization tool and the available relations are very useful to observe verbal placement. Additionally, the presented relations duly inserted in an ontology may provide important insights concerning the interdependency among concepts, as well as, a clear panorama of domain specific jargon.

However, the evaluation pointed out two negative points related to semantic roles of the extracted relations that deserve improvements in the future. First, the relation do not include verbal complements, since in some cases the sentence has more information than just the subject and object of the predicate. For instance, the evaluation mentions the relation found in the corpus of Geology between the concepts "Harris" (a proper noun) and "ausência de Biotita" ("absence of Biotite" in English) that happens with the verb "associar" ("to associate" in English). For this specific relation, Harris was the person who found out the association between the absence of Biotite and another concept that was not discovered by the proposed method. The second down point identified was the absence of prepositions with the predicates, since in many cases the prepositions are necessary to fully define the relations.

V. Final Considerations

This work presents a method to extract non-taxonomic relations from domain corpora in Brazilian Portuguese, which is by this only fact an original contribution in itself. It is also noticeable that by using state of the art relevance index (*tf-dcf*), the proposed method delivers relations that are specific to domain corpora.

The evaluation experiments conducted illustrated the specificity of the extracted relations (1), and they have shown a strong correlation between human evaluation and automatic choice of subject and object concepts (2). The third experiment was helpful to point out future work to be done as the need to include prepositions and not only predicates to define a relation. Nevertheless, even in its present state the proposed method was helpful to observe interdependency among concepts and language patterns, as mentioned.

References

[1] C. Biemann, "Ontology learning from text: A survey of methods," *LDV Forum*, vol. 20, no. 2, pp. 75–93, 2005.

[2] I. Serra and R. Girardi, "A process for extracting non-taxonomic relationships of ontologies from text," *Intelligent Information Management*, vol. 3, pp. 119–124, July 2011. [Online]. Available: http://www.scirp.org/journal/PaperInformation.aspx?paperID=5865

[3] P. Drouin, "Detection of domain specific terminology using corpora comparison," in *Proc. of the 4th Int. Conf. on Language Resources and Evaluation (LREC) 2004*, M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, Eds., ELRA. Lisbon, Portugal: European Language Resources Association, May 2004, pp. 79–82.

[4] L. Lopes, "Extração automática de conceitos a partir de textos em língua portuguesa," Ph.D. dissertation, PUCRS University - Computer Science Department, Porto Alegre, Brazil, 2012.

[5] D. Sánchez and A. Moreno, "Learning non-taxonomic relationships from web documents for domain ontology construction," *Data Knowl. Eng.*, vol. 64, no. 3, pp. 600–623, Mar. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2007.10.001

[6] J. Villaverde, A. Persson, D. Godoy, and A. Amandi, "Supporting the discovery and labeling of non-taxonomic relationships in ontology learning," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 288–10 294, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417409000943

[7] L. Lopes, P. Fernandes, R. Vieira, and G. Fedrizzi, "ExATO lp – An Automatic Tool for Term Extraction from Portuguese Language Corpora," in *Proc. of the 4th Language & Technology Conference (LTC '09)*, Faculty of Mathematics and Computer Science. Poznan, Poland: Adam Mickiewicz University, November 2009, pp. 427–431.

[8] E. Bick, "The parsing system PALAVRAS: automatic grammatical analysis of portuguese in constraint grammar framework," Ph.D. dissertation, Arhus University, Arhus, Danemark, 2000.

[9] L. Lopes and R. Vieira, "Heuristics to improve ontology term extraction," in *PROPOR 2012 – Int. Conf. on Computational Processing of Portuguese Language*, ser. LNCS 7243, 2012, pp. 85–92. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-28885-2_9

[10] L. Lopes, P. Fernandes, and R. Vieira, "Domain term relevance through tf-dcf," in *Proc. of the 2012 Int. Conf. on Artificial Intelligence (ICAI 2012)*. Las Vegas, USA: CSREA Press, 2012, pp. 1001–1007.

[11] A. Maedche and S. Staab, "Learning ontologies for the semantic web," in *SemWeb*, 2001.

[12] A. Schutz and P. Buitelaar, "Relext: A tool for relation extraction from text in ontology extension," in *The Semantic Web ISWC 2005*, ser. LNCS, Y. Gil, E. Motta, V. Benjamins, and M. Musen, Eds. Springer Berlin Heidelberg, 2005, vol. 3729, pp. 593–606. [Online]. Available: http://dx.doi.org/10.1007/11574620_43

[13] A. Weichselbraun, G. Wohlgenannt, A. Scharl, M. Granitzer, T. Neidhart, and A. Juffinger, "Discovery and evaluation of non-taxonomic relations in domain ontologies," *Int. J. Metadata Semant. Ontologies*, vol. 4, no. 3, pp. 212–222, Aug. 2009. [Online]. Available: http://dx.doi.org/10.1504/IJMSO.2009.027755

[14] R. J. Coulthard, "The application of Corpus Methodology to Translation: the JPED parallel corpus and the Pediatrics comparable corpus," Ph.D. dissertation, UFSC, Florianópolis, Brazil, 2005.

[15] P. Resnik and A. Elkiss, "The linguist's search engine: an overview," in *Proc. of the ACL 2005*, ser. ACLdemo '05. Stroudsburg, PA, USA: ACL, 2005, pp. 33–36. [Online]. Available: http://dx.doi.org/10.3115/1225753.1225762