Exploring the inference role in automatic information extraction from texts

Denis A. de Araujo, Sandro J. Rigo
PIPCA, UNISINOS
denis.andrei.araujo@gmail.com,
rigo@unisinos.br

Carolina Muller, Rove Chishman
PPGLA, UNISINOS
São Leopoldo - Brazil
muller.carolina@ymail.com,
rove@unisinos.br

Abstract

In this paper we present a novel methodology for automatic information extraction from natural language texts, based on the integration of linguistic rules, multiple ontologies and inference resources, integrated with an abstraction layer for linguistic annotation and data representation. The SAURON system was developed to implement and integrate the methodology phases. The knowledge domain of legal realm has been used for the case study scenario through a corpus collected from the State Superior Court website in Brazil. The main contribution presented is related to the exploration of the flexibility of linguistic rules and domain knowledge representation, through their manipulation and integration by a reasoning system. Therefore, it is possible to the system to continuously interact with linguistic and domain experts in order to improve the set of linguistic rules or the ontology components. The results from the case study indicate that the proposed approach is effective for the legal domain.

1 Introduction

The aim of Information Extraction (IE) field is to develop tools and methodologies to identify, annotate and extract specific information from natural language text documents. Although the efforts in this field are not recent (Rillof, 2009), its growing importance and necessity certainly are related to the large volume of natural language text documents and relevant textual information currently stored in databases. Due to this context, the manual analysis of these resources becomes unfeasible. Therefore, text documents automatic processing stands as a necessity and the achievement of better results in IE systems allow improvements in the effectiveness of other related systems, such as, for instance, the Information Retrieval systems.

The initial systems of IE generally were related to specific domains as a way to achieve better results in their operation. Some of the main aspects considered by these IE systems are the frequency and position of domain-related terms (Wimalasurya, 2009). Related to these approach, some well-known works in the area (Bruninghaus, 2001; Jijkoun, 2004) do not consider certain important relation descriptions concerning linguistic and domain knowledge aspects. Recently, aiming to improve the flexibility and precision in IE, the use of domain knowledge, expressed by ontologies, is observed in several approaches (Saravanan, 2009; Daya 2010). Some other initiatives incorporate linguistic aspects in their design (Wyner, 2011; Moens, 1999; Amardeihl, 2005) in order to better treat natural language complex structures.

This paper presents a novel methodology for Information Extraction from natural language texts that combine domain knowledge with linguistic knowledge. The linguistics information is represented in form of ontologies and allows the application of automated reasoning algorithms. Therefore some improvements over related work are achieved. The first one is the wide use of semantic information described in domain ontologies, allowing reuse and the integration of multiple ontologies. The second is the incorporation of linguistic information, which is obtained from studies of the domain documents, composing flexible and precise rules for information extraction. The last one is the extensive use of an inference system, in order to integrate and process textual, domain and linguistic information. Moreover, an abstraction layer for linguistic annotation and data representation (Chiarcos, 2012a) is adopted as a key component of the methodology. The main benefit from this choice is the greater flexibility in the integration and processing of different parser originated annotations, as well as some facilities in the use of corpora originated from different sources.

Proceedings of the Joint Workshop on NLP&LOD and SWAIE, pages 33–40,
Hissar, Bulgaria, 12 September 2013.

The work has been developed in the context of a research group within the scope of the project "Semantic technologies and legal information retrieval systems"[1]. The group involved in this project aims to develop a conceptual-semantic model of the Brazilian legal domain, in order to integrate it into Information Retrieval systems targeted at legal documentation. The group has an interdisciplinary composition, comprising lawyers, linguists and computer science researchers.

2 Related Work

In this section we present two aspects of related works. The first one is more general and not related to IE techniques, but illustrates the increasing availability of data collections and data repositories, some of them fully integrated with several databases. The second aspect is related to the technical differences of the proposed approach from previous works.

There is a trend in providing facilitated access to documents in several specific domains and in the adoption of some standards to describe document collections. Therefore, these initiatives foster the generation of document patterns and repositories for annotation and automatic processing.

Since the case study adopted in this work is dedicated to the legal realm, some relevant examples of this situation are mentioned here. Currently, there are several initiatives underway to achieve a standard representation of legal documents, aimed at facilitating their automatic processing. In Brazil, LEXML project[2] is concerned with information representation and information retrieval, as well as some other projects in Italy also care about those issues (Brighi, 2009; Biagioli, 2005; Palmirani 2011). Some other examples of projects in this area can be cited, as the Institute of Legal Information Theory and Techniques[3], the results of Estrella Project[4] and Metalex standard proposal[5]. In general, standards and schemes are used in these initiatives, implemented in flexible formats, such as XML[6], fostering the generation of patterns for annotation and access. The existence of this trend in providing affordable computational formats shows the correct positioning of efforts to create automatic tools for legal text treatment.

Regarding the technical aspects, it is important to identify the main differences presented by the proposed methodology from previous works, which resides mainly in the representation of linguistic information in form of ontologies and in the extensive use of inference mechanisms and linguistic rules to identify relevant events. This novel approach was not found in the reviewed material and it brings to the implemented system the possibilities of achieving better precision and flexibility, as described in the results analysis.

Some initial efforts in Information Extraction for legal realm were conceived with the same syntactic pattern approach observed in other fields (Jijkoun, 2004; Oard, 2010). These works presents no capability to cope with some important linguistic relations and also lack flexibility to maintain the sets of syntactic patterns used.

To overcome such aspects, some works apply knowledge representation as a resource to improve the domain information possibilities, since the IE is dependent of specific vocabulary and related to proper concepts. The use of ontologies is adopted in several works (Saravanan, 2009; Soysal, 2010) and allow improvements in domain concepts representation. In such works, in general, the ontologies are mainly used as concepts repositories, dedicated to help in search operations, therefore with little exploration of inference and reasoning possibilities.

The linguistics information is also applied in several works (Moens, 1999; Mazzei, 2009; Cederberg, 2003) and these approaches contribute to the understanding of the great importance of using linguistic structures in IE, since they allow a more precise analysis of the texts. Extending these initiatives, some proposals suggest the use of ontologies combined with linguistic analysis (Amardeilh, 2005; Palmirani, 2011; Lenci, 2007). The main argument in these cases is the possible improvements integrating the linguistic and domain knowledge, providing a better basis to the text analysis. In these approaches, however, there is not an integrated representation of the domain knowledge and the linguistic analysis, as provided by the proposed methodology in this work.

3 Proposed methodology

The proposed methodology has two phases, called linguistic phase and computational phase. In the first one the focus of attention is the cor-

[1] Legal Research Support Program No. 020/2010/CAPES/CNJ

[2] http://projeto.lexml.gov.br

[3] http://www.ittig.cnr.it/

[4] http://www.estrellaproject.org/

[5] http://www.metalex.eu/

[6] http://www.w3.org/XML/

pus study, which is necessary to build the necessary domain ontology and linguistic rules. The second phase objective is to integrate linguistic rules with domain ontologies through the use of an inference system and the abstraction layer for linguistic annotation and data representation. This phase is therefore based on the use of Natural Language Processing techniques (LIDDY, 2003), ontology and inference resources. The outcome of this phase is a knowledge base composed by the relevant information identified.

To illustrate the overall methodology integration aspects, the Figure 1 shows the main elements of each phase. As indicated in the Figure 1, in the linguistic phase the desired corpus is studied and then are generated two ontologies: the ontology with the linguistic rules and the domain ontology. The linguistic rules, depending on their complexity, are formalized in OWL (McGuinness, 2004), through logical axioms in Description Logic (Baader, 2003) or SWRL[7]. The domain ontology is formalized in OWL language.

The computational phase aims to provide the text documents processing and the integration of the domain ontology with the ontology containing the linguistic rules. We propose in our methodology that the natural language text documents submitted to the IE process should be first treated by a deep linguistic parser and then represented in OWL with the POWLA data model (Chiarcos, 2012). This data model represents corpora structures through linguistic concepts in OWL, therefore allowing the use of the linguistic rules and the domain ontology concepts in an integrated and flexible manner. When necessary, some optimizations can be performed in order to ensure that the represented text do not generate excessive and not useful information.

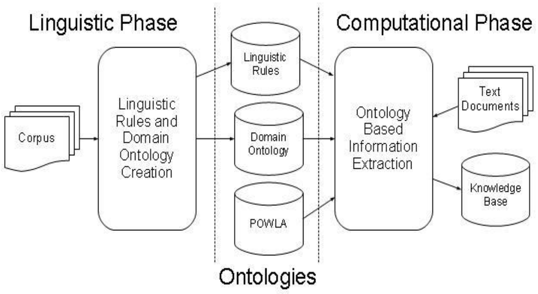


Figure 1. General view of the proposed methodology phases

As an outcome of these methodological choices, we can indicated the following positive aspects: (a) the IE process can be done with data originated from different linguistic parsers; (b) the linguistic rules can be formulated considering more than one annotation layer; (c) the reasoning system integrates the description of the domain, the linguistic rules and the documents linguistic annotation; (d) the knowledge representation of linguistic rules, domain and linguistic annotation can be manipulated in a flexible way. Some details in the proposed methodology are described below.

3.1 Domain Ontology construction

In general, IE is context dependent, since there are specific vocabulary and textual constructions more frequently observed in each knowledge field. Due to this situation, our methodology applies domain ontology to describe the important concepts of the targeted area. This domain ontology is created through a semantic analysis of terms and their relations, during the study of natural language texts describing the desired domain area.

We consider that using the domain ontology together with linguistic rules can improve precision and recall in the computational phase. One of the main aspects supporting this consideration is the integration of concepts and relations described in the domain ontology with elements of the linguistic rules ontology, thus allowing greater accuracy.

Also it is important to note that the reuse aspect of ontologies can be a very important element to foster the methodology application in different knowledge domains. In the case study related in this work, aiming at the legal realm, the domain ontology was created using the categories recommended by (Minghelli, 2011), that are: Legal Events, Legal Institutions, Legal Documents and Legal Participants.

With these categories, is possible to capture and describe specific contexts that assist in the interpretation of textual information found in the text documents. It also enables the identification of various important relations, such as dependency and composition. The integration established between the domain ontology and the linguistic rules foster the specification of references between elements described in the domain ontology and linguistic elements.

3.2 Description of Linguistic Rules

Linguistic experts define the linguistic rules applied in our methodology, in order to better represent the knowledge involved in textual analy-

[7] http://www.w3.org/Submission/SWRL/

sis. The experts in the knowledge field are also involved in this phase.

Therefore, the linguistic rules represent the reflections of the linguistic experts about the textual constructions, based on linguistic corpus analysis and interactions with other experts in the knowledge area.

In order to achieve flexibility and to maintain the greater amount of semantic information, in our methodology, the linguistic rules will be expressed in sets of constructions in Description Logic and SWRL rules. These rules also combine the concepts of the domain ontology, and therefore are able to correctly and precisely identify terms and excerpts of the text documents analyzed.

These documents are represented using the POWLA/OWL data model. One of the advantages of this approach is the flexibility for describing linguistic rules. Since the basic elements of the text are available, together with more complex components, such as sentences or phrasal structures, the linguistic rules can be expressed using all these aspects. This expands the possibilities of the linguists in the description of the rules. This context is possible through the use of multiple ontologies, which are specialized in different components, such as the annotation layer, the domain concepts, and the linguistic rules specification.

Despite the higher computational cost that this approach can present when compared with some other options, the results, as described in the result analysis section, presents a good precision and are not dependent of a large volume of documents to generate basic and reference models.

3.3 Computational Phase: the SAURON System

The computational phase of the methodology suggested is implemented in the SAURON system, developed in Java Language[8] and the OWL Api[9] support, integrating the Pellet reasoner[10].

This system is inspired in the unifying logic layer of the standard technology stack for semantic web[11], since one of the objectives of this system is to unify the use of several semantic technologies applied. The system provides the necessary support to the tasks involving Natural Language Processing, such as the text preprocessing, the syntactic parser access and some format conversions tasks.

The first computational process performed on the text documents is to convert them to OWL representation. To do this, we first apply the widely adopted Palavras parser (Bick, 2000), which is a morph-syntactic parser for Portuguese. The result produced by the parser after the text document analysis is a file in TIGER-XML format (König, 2003). This file contains a hierarchical structure of the sentences from the original document and the linguistic annotations (Bick, 2005) about terms that compose them. TIGER-XML file has many linguistic annotations that represent a rich source of data to carry out the identification of information in automated systems.

The large amount of linguistic information generated by linguistic parser will be used to make ontological inferences. To accomplish this we used the POWLA data model to convert the TIGER-XML format to OWL. For this task we adapted a script developed originally to convert documents from TIGER-XML to POWLA (Chiarcos 2012a). After this initial processing of the texts, the SAURON System integrates the textual information, the ontologies containing the linguistic rules and the domain knowledge produced at linguistic phase. This is done through a process of ontologies integration and the use of the inference engine, responsible for identifying the concepts in the text documents processed.

4 Experiment description

To obtain and evaluate results with the use of the developed methodology, we conducted an experiment in the legal realm. To better demonstrate the methodology aspects, the next sections describe the domain ontology created, then some of the linguistic rules construction and, finally, the obtained results. The experiment was conducted with a corpus of 200 documents, composed of 39.895 sentences, that was obtained from RS State Superior Court (in Portuguese, *Tribunal Superior do Rio Grande do Sul - TJRS*). The results of the automatic extraction of events were manually reviewed by experts for the identification of its correction and to latter recall and precision metrics application.

4.1 Domain ontology creation

To implement the tasks of the linguistic phase in this case study, we adopted the following methodology: selection of corpus, relevant term ex-

[8] http://docs.oracle.com/javase/

[9] http://owlapi.sourceforge.net/

[10] http://clarkparsia.com/pellet/

[11] http://www.w3.org/2001/12/semweb-fin/w3csw

traction, choice of ontology terms, definition of hierarchy and relations as well as formalization of the ontology in the Protégé[12] editor. Experts in linguistics, law and knowledge representation did this task.

The next step was to find verb`s definition for a better semantic description. The participation of research group´s law experts was essential in the determination and description of the events to better represent the domain knowledge. We also made use of a Legal Vocabulary dictionary (Silva, 2009) to clarify the meaning of terms related to legal events. Considering the verbs extracted from the corpus and their meaning, a list of legal events evoked by each one was defined. Having clearly defined verbs and events, we moved on to the semantic analysis based on Lexical Semantics (Cruse, 2000) to establish a taxonomic relation between events and verbs.

The relations of hyponymy and synonymy stood out, guiding the organization in terms of ontology. In addition, we performed a parsing of sentences to verify participants involved in the event. The last part of this process was to include detected events, participants and verbs in Protégé[11] ontology editor. Closing the study phase of the linguistic corpus, the domain ontology was structured including legal events found in the analyzed corpus. That ontology resulted in 95 axioms, being 51 logical axioms, 41 classes (3 main and 38 subclasses), with 7 axioms of class equivalence.

4.2 Linguistic rules description

In this study case our objective was to automatically identify the legal events *Denúncia* (formal charges), *Absolvição* (acquittal), *Condenação* (conviction) and *Interrogatório* (questioning). These are the main events described in the domain ontology created for the experiment.

The linguistic analysis of phrases intends to identify linguistic patterns, which will lead to the creation of the linguistic rules used to identify these legal events. This process will be illustrated in details through the analysis of the phrase in Figure 2, which was extracted from one of the case study documents. This phrase describes one example of the *Denúncia* event.

The excerpt from Figure 2 presents a simple linguistic pattern typical of phrases containing the event Formal Charges in its verbal form. So, we identified that the presence of the verb *Denunciar* (present formal charges, in English) is an indication of the presence of the event.

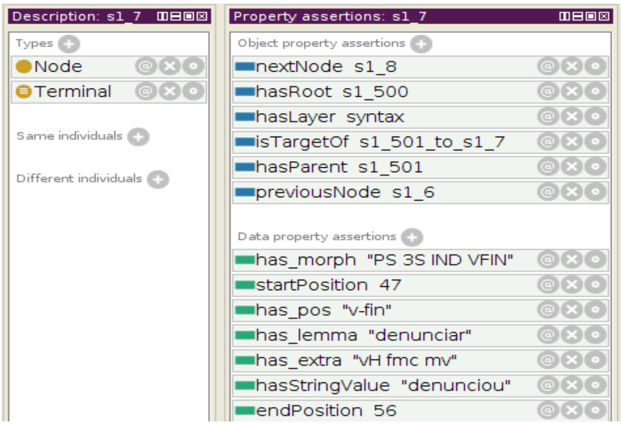
However, we must seek other linguistic marks, because the verb alone is not sufficient to conclude the presence or absence of the event. In the sentence being analyzed in Figure 2, we see that the agent of the verb is the Prosecutor, indicating that the verb expresses the meaning we want to identify.

> *"O Ministério Público denunciou NNNN como incurso nas sanções do artigo 121, § 2o, inciso IV, do Código Penal."*
>
> *(in English: "*Prosecutors charged NNNN according to the article 121, paragraph 2, item (IV), of the Penal Code.")

Figure 2. Excerpt referring to *DENÚNCIA* (formal charges).

The above findings lead us to define that phrases containing the verb *DENUNCIAR* whose agent is the Prosecutor refer the event *DENÚNCIA* (formal charges). These conclusions will be represented in the form of linguistic rules. The information required for the elaboration of the linguistic rules are generated by the Portuguese language parser Palavras [Bick 2000], which provides various information ranging from the sentence analyzed through labeling and classifying words and phrases.

The Figure 3 shows part of the the linguistic information generated by the Palavras parser, but now represented in OWL language through the POWLA data model. In the Figure 3 we can see the integration of the syntactic and structural information. This structural information aims to represent, for example, the relations of the term described as "s1_7" with other phrase components, such as the components described as nextNode, previousNode, hasRoot, isTargetOf and hasParent. These components are part of the annotation layer of the POWLA data model.



Description: s1_7

Types
- Node
- Terminal

Same individuals

Different individuals

Property assertions: s1_7

Object property assertions
- nextNode s1_8
- hasRoot s1_500
- hasLayer syntax
- isTargetOf s1_501_to_s1_7
- hasParent s1_501
- previousNode s1_6

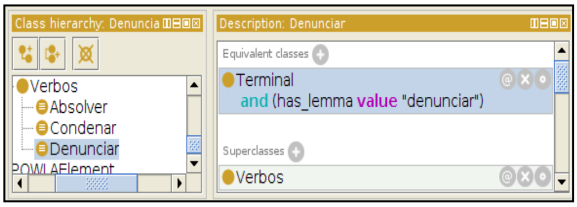
Data property assertions
- has_morph "PS 3S IND VFIN"
- startPosition 47
- has_pos "v-fin"
- has_lemma "denunciar"
- has_extra "vH fmc mv"
- hasStringValue "denunciou"
- endPosition 56

Figure 3. Linguistic information in OWL using POWLA data model.

[12] http://protege.stanford.edu/

The use of OWL language to represent the linguistic information of the text documents makes possible to use Description Logic or SWRL to formalize the linguistic rules. A linguistic rule to identify the verb *denunciar*, for example, can be described in a simple way: all the individuals of the Terminal class containing the term which canonical form (lemma) is *denunciar* can be considered examples of this form.

The rule for the identification of examples of the *denunciar* verb can be defined in a Description Logic axiom as illustrated in Figure 4. The POWLA's data property has_lemma, which corresponds to the tag lemma of TIGER-XML, contains the canonical form of the word. This makes it possible to define that any individual Terminal class, whose has_lemma property is *denunciar*, is also an instance of the class *Denunciar*.



Terminal
and (has_lemma value "denunciar")

Figure 4. Linguistic Rule to identify the verb denunciar

The other essential element for assessing the presence of the event is the agent of the verb. By definition, we know that the agent of the *denunciar* verb should be *Ministério Público* (Prosecutor, in English). The linguistic rule for identification of this entity on text is also simple and can be represented by another Description Logic axiom. The Figure 5 shows this linguistic rule using Manchester syntax (Horridge, 2006).

```
Class: Ministério_Público
Equivalent to:
        Terminal
        and (has_lemmavalue "Ministério_Público")
```

Figure 5. DL linguistic rule to identify *Ministério Público* (*Prosecutor*).

Now that we have the linguistic rules for the identification of the two main components of *Denúncia* event, we can define the linguistic relations between them to verify if the event is referenced at the analyzed phrases. As this rule is more complex and requires a more expressive set of elements, it is formalized in SWRL. Figure 6 shows an example of the SWRL rule, that use the information generated by linguistic parser. This rule uses both structural (hasParent, isSourceOf, hasTarget and hasChild) and syntactical (has_label) information.

```
1. Denunciar(?verbo),
2. hasParent(?verbo, ?fcl),
3. isSourceOf(?fcl, ?relation),
4. has_label(?relVAux, "S"^^string),
5. hasTarget(?relVAux, ?np),
6. hasChild(?np,?mp)
7. Ministério_Público(?mp)
8. -> Denúncia(?fcl)
```

Figure 6. SWRL linguistic rule to identify *Denúncia* (present formal charge).

At lines 1 and 7 in the Figure 6 are illustrated the use of the previously defined Description Logic rules (*Denunciar* and *Ministério_Público*), in an approach that foster the reuse of some basic rules in order to build more complex ones. In the line 8 of the rule is expressed the obtained conclusion: the *Denúncia* event is present in the phrase analyzed. The defined linguistic rules are inserted in an OWL file apart from the domain ontology, therefore maintaining the separation between the ontology containing concepts and the other one containing linguistic rules.

To perform this experiment were generated 12 linguistic rules, aiming to identify the main events of interest. These rules allow also the treatment of linguistic aspects, such as, for instance, the use of passive voice. The phase described in this section is a very simple one, but the methodology allow the treatment of complex linguistic structures as well. For instance, the implemented rules can deal with relations beyond verb and subject ones, exploring the linguistic information generated by the Palavras parser. Also the rules make use of the domain ontology components, both in order to generate the resulting knowledge base and to relate specific concepts.

5 Results Analysis

For the development of the case study presented here, two corpuses were elaborated, being originated from documents returned by a query performed at jurisprudence search tool available at State Superior Court TJRS website. The first one was called learning corpus, because it was used to elaborate the linguistic rules used in the experiment. This corpus consists of 10 judgments, covering the decisions published by 4 different judges. The number of sentences in the corpus is 1.861 and the number of words is 6.142.

The testing corpus had the same origin that the learning corpus, but this time 200 documents were selected and the judgments of the learning corpus were not used. The testing corpus had 39.895 sentences, 618.892 words, covering decisions taken by 19 judges.

In order to identify the events in the text documents, the domain ontology and the linguistic rules ontology are merged with the OWL file containing the linguistic information from the original text, described in POWLA format. Then the Pellet reasoner is triggered, resulting in the evaluation of the rules and in the identification of the existing events.

All the steps are performed in the context of the Sauron system. This system is fully implemented, with all the features necessary to the proposed methodology.

Comparing the results of our approach against the manually parsed set of the text documents, we have the precision and recall results shown in Table 1.

The precision metric stands for the number of correctly identified events, given the number of identified ones. The recall metric stands for the number of events identified correctly, given the total number of existent events.

The good results in the precision of events identification can be associated with the use of rules based on linguistic information. The previous documents study by experts and the broad use of parser generated linguistic information allows the creation of linguistic rules with good accuracy. The recall results present also a good outcome.

Further analysis of the text documents and the linguistic rules applied shows that these results can be improved. In our analysis, they are dependent on the available rules and, therefore, the inclusion of some additional specific rules can improve these results.

Table 1. Results from test set legal documents

Brazilian Legal Event	Equivalent English Term	Precision (%)	Recall (%)
Denúncia	Formal charges	100	92
Absolvição	Acquittal	96	90
Condenação	Conviction	98	84
Interrogatório	Questioning	100	100

The performance achieved in terms of recall indicated that the solution proposed here has a good level of generalization, which is enough to use them in the real world applications. The proportion between the learning and the testing corpora used in this experiment, and the results presented in Table 1, indicate that the suggested methodology can be successfully used on a wider scale.

The tests were conducted in a computer with 32 Gbytes of memory, equipped with Xeon processor and running Windows Server operational system. The mean size of the processed documents after the conversion to POWLA data model increase in 318% and their mean is 231 Kbytes size. The computational effort to run the reasoning system is feasible, since the mean time to process the documents is 79 seconds.

6 Conclusions and Future Work

The approach presented here indicates important perspectives, evidenced in the aspects of accuracy and recall observed in experiment. These results are associated with the integration between the linguistic and computational phases, allowing effective results and flexibility.

This work is in continuous development, with experiments planned to provide the model verification in some different domain, such as the educational domain and the medical domain.

References

Amardeilh, F., Laublet, P. and Minel, J. L. 2005. Document Annotation and Ontology Population from Linguistic Extractions, Proceedings of Knowledge Capture (KCAP), Banff.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D. and Patel-Schneider, P. F. 2003. The Description Logic Handbook: Theory, Implementation and Applications, Cambridge University Press.

Bick, E. 2000. The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework, Aarhus University Press.

Bick, E. 2005. Gramática Constritiva na Análise Automática da Sintaxe Portuguesa in T. B. Sardinha, A Língua Portuguesa no Computador, Mercado das Letras, Campinas.

Brighi, R. and Palmirani, M. 2009. Legal Text Analysis of the Modification Provisions: A Pattern Oriented Approach, Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL ’09, pages 238–239, ACM, New York, NY, USA.

Bruninghaus, S. and Ashley, K. 2001. Improving the Representation of Legal Case Texts with Information Extraction Methods, Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL ’01), pages 42-51, ACM Press.

Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S. and Soria, C. 2005. Automatic Semantics Extraction in Law Documents, Proceedings of the 10th International Conference on Artificial Intelligence and Law, ICAIL ’05, pages 133–140, ACM, New York, NY, USA.

Cederberg, S. Widdows, D. 2003. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4 (CONLL '03), Vol. 4. Association for Computational Linguistics, Stroudsburg, PA, USA, 111-118.

Chiarcos, C. 2012 a. Interoperability of Corpora and Annotations, in C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, Linked Data in Linguistics, Representing and Connecting Language Data and Language Metadata, pages 161–179, Springer, Heidelberg.

Chiarcos, C. 2012. POWLA: Modeling Linguistic Corpora in OWL/DL, Proc. 9th Extended Semantic Web Conference (ESWC), Heraklion, Crete.

Cruse, D. A. 2000. Meaning in Language: an Introduction to Semantics and Pragmatics, Oxford University Press, New York.

Daya C. Wimalasuriya and Dejing Dou, “Ontology-based information extraction: an introduction and a survey of current approaches,” Journal of Information Science, vol. 36, no. 3, pp. 306–323, 2010.

Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R. and Wang, H. H. 2006. The Manchester OWL Syntax, Proc. of the OWL Experiences and Directions Workshop (OWLED) at the ISWC.

Jijkoun, V., Rijke, M. and Mur, J. 2004. Information extraction for question answering: improving recall through syntactic patterns. In Proceedings of the 20th international conference on Computational Linguistics (COLING '04).

König, E. and Lezius, W. 2003. The TIGER Language - A Description Language For Syntax Graphs, Formal definition, Technical Report.

Lenci, A., Montemagni, S., Pirrelli, V. and Venturi, G. 2007. NLP-Based Ontology Learning From Legal Texts - A Case Study, Proceedings of LOAIT , 2007.

LIDDY, R. NaturalLanguage Processing, Library and Information Science, Marcel Drecker Inc. New York, USA, 2a Ed. 2003.

Minghelli, T. D. 2011. A Relação de Meronímia em uma Ontologia Jurídica, Dissertação de Mestrado, UNISINOS, São Leopoldo.

Moens, M., Uyttendaele, C. and Dumortier, J. 1999. Information Extraction from Legal Texts: The Potential of Discourse Analysis, International Journal of Human-Computer Studies 51 (1999), 1155–1171.

Mazzei, A., Radicioni, D. P. and Brighi, R. 2009. NLP-Based Extraction of Modificatory Provisions Semantics, Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL), pages 50–57, Barcelona, Spain.

McGuinness, D. and van Harmelen, F. 2004. OWL Web Ontology Language Overview, W3C recommendation, http://www.w3.org/TR/owl-features (accessed 31 August 2012).

Oard, D. W., Baron, J. R., Hedin, B., Lewis, D. D. and Tomlinson, S. 2010.Evaluation of Information Retrieval for E-Discovery, Artificial Intelligence and Law.

Palmirani, M., Ceci, M., Radicioni, D. and Mazzei, A. 2011.“FrameNet Model of the Suspension of Norms”, In Proceedings of ICAIL 2011, pp. 189–93, New York: ACM Press.

Rillof, E., Information Extraction as a stepping stone toward story understanding. Understanding Language Understanding: computational models of reading.p.435-460, MIT Press. 1999Silva, P. 2009. Vocabulário jurídico, Forense, Rio de Janeiro.

Saravanan, M., Ravindran, B. and Raman, S. 2009. Improving Legal Information Retrieval Using an Ontological Framework, Springer Science Business Media B.V.

Soysal, E. Cicekli, I. Baykal, N. 2010. Design and evaluation of an ontology based information extraction system for radiological reports. Comput. Biol. Med. 40, 11-12 (November 2010), 900-911.

Sang, E. T. K., Hofmann, K. 2009. Lexical patterns or dependency patterns: which is better for hypernym extraction?. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL '09). Association for Computational Linguistics, Stroudsburg, PA, USA, 174-182.

Wimalasurya, D. C., Dou, D., Using Multiple Ontologies in Information Extraction.CIKM, 09, Hongkong, p.235-245, 2009. ACM Press.

Wyner, A. and Peterson, W. 2011. Rule Extraction from Regulations, The 24th International Conference on Legal Knowledge and Information Systems (Jurix).