Cognitive Linguistic Representation of Legal Events

Towards a semantic-based legal information retrieval

Anderson Bertoldi, Rove Chishman, Sandro José Rigo, Thaís Domênica Minghelli
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Brazil
andersonbertoldi@yahoo.com, rove@unisinos.br, rigo@unisinos.br, thaisdomenica@hotmail.com

***Abstract*—An important role of an attorney in Brazil is to search Brazilian courts databases in order to find precedent decisions to base their requests on. This paper discusses the initial efforts that have been made towards the development of a legal knowledge base, composed by semantic frames, to improve Brazilian courts information retrieval systems. Linguistic methods are applied to recognize possible legal event structures to be described in legal documents. Afterwards, based on the linguistic theory of Frame Semantics, the participants and props of legal events are described. This is a work in progress that will involve both legal and linguistic description as well as system development. With this legal knowledge base, the results expected are the improvement of the legal information retrieval system of Brazilian courts databases, using semantic representation of Brazilian legal events.**

Keywords-knowledge representation; semantic modeling; semantic frames; legal information retrieval.

I. Introduction

This paper presents a project in progress whose aim is to represent legal knowledge to replicate the possible cognitive connections that a law specialist makes in the moment he/she is analyzing a legal document. This project applies the Frame Semantics theory [1][2] for the semantic modeling of legal events. Legal events are represented as semantic frames and relations among these semantic frames are established in order to reproduce the connections of knowledge that specialists have to make to understand legal documents.

In Brazil, usually, before filing a lawsuit, attorneys search the online databases of Brazilian courts for similar cases. In this search they look for precedent decisions to decide whether a lawsuit has chance to be accepted for the judge and to base the request that originates the suit. Despite the efforts to improve their Information Retrieval (IR) systems, Brazilian court databases still do not work with semantic annotation of their documents, which could improve the search results. As a consequence, lawyers, when looking for precedent decisions, have to deal with a huge amount of documents. The proposal of this project is to develop a knowledge base composed by semantic frames describing the legal knowledge related to legal terms/words. These resources will be used to annotate a legal corpus aimed to be applied together with an automatic corpus annotation tool.

To discuss this topic, this paper is structured as follows. Section II presents the problem that motivates this research. As IR systems of Brazilian courts return a huge amount of documents, lawyers spend too much time reading documents to separate meaningful to non-meaningful documents. Section III presents the solution proposed in this project. Using linguistic methods to describe the meaning of the legal terms/words, this project expects to develop a set of semantic tags for legal text annotation. Section IV discusses the expected results as well as the steps that have to be accomplished in order to build a semantic-based legal information retrieval system. Section V presents the related work in which this research is based on. Section VI discusses the possible contributions of this work with respect to legal information retrieval system of Brazilian courts and with respect to the previous work in the areas of Frame Semantics and legal information retrieval.

II. The Problem

When specialists search the online databases of Brazilian courts, such as the State Court of Appeal [3] or the Federal Court of Appeal [4], looking for precedent decisions to base the lawsuit request on, they need to provide a combination of words to get better results. Let us consider, for example, an attorney who is working on a divorce case. He needs to make a request of child support and, therefore, he looks for the jurisprudence of the State Court of Appeal. He will need to provide to the IR system a combination of words, such as *alimentos* (*food*) and *divórcio* (*divorce*) or *alimentos* (*food*) and *pensão* (*alimony*).

This example demonstrates that the IR systems of the Brazilian courts search their databases by string patterns, not by the meaning of the legal terms/words. The search result is a huge amount of legal documents stored in the court database. The specialist has to spend a considerable part of his/her time just reading documents that were returned to find which documents are really meaningful for his/her intents.

The assumption explored in this project is that describing the meaning of the legal terms/words and establishing relations among these terms, with the support of FrameNet [5] formalism, the result will be a knowledge base that can provide valuable resource to the IR system improvements in order to return better results. The framework to implement and use this knowledge base comprises the development of additional components that expand both the documents indexing and the term searching operations of the existing IR systems.

III. The Solution Proposed

The solution proposed by this project is to build a knowledge base describing the meaning of the legal terms/words through semantic frames. The theory concerned about semantic frames is called 'Frame Semantics' and was proposed by the linguist Charles Fillmore, inspired by the previous work of Marvin Minsky [6].

A semantic frame is a schematic structure which describes the role of the participants and props of an event or state [1]. This schematic structure is evoked by lexical units. According to Frame Semantics, lexical units work like a trigger that makes the speaker retrieve in his/her mind related concepts that help to understand the meaning of a certain concept. For instance, in order to understand the meaning of 'buying' the speaker needs to understand the meaning of 'selling'. The lexical units 'buy' and 'sell' evoke a frame of commercial transaction. To understand the concept of commercial transaction, speakers must understand concepts that are related to a commercial transaction, such as 'seller', 'buyer', 'goods', and 'money'.

The proposal presented here is to develop legal semantic frames based on the study of legal documents and to describe the meaning of legal lexical units relating lexical units to semantic frames. After storing semantic frames and lexical units in a knowledge base, this knowledge base could be used as a component of a legal IR system, providing knowledge reasoning capability to IR systems.

The methodology adopted to develop legal semantic frames is majorly based in non-automatic linguistics methods. First, a collection of legal texts is compiled and, based on linguistic methods, the most relevant lexical units of the texts are described, relating them to a specific semantic frame. For instance, the lexical unit *acusação* (*charges*) evokes a semantic frame of 'Charging.' Fig. 1 shows an example of relations among semantic frames for Charging frame.

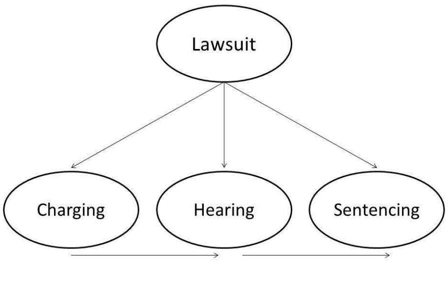

Figure 1. Relations among semantic frames.

Concepts related to Charging frame include the 'Prosecutor,' the authority in Brazilian legal system responsible for taking a suspect to the court, the 'Judge,' the authority responsible in the Brazilian legal system for saying if the suspect can be suited or not, the 'Suspect,' the person that is suspect to have committed an infraction or a crime, and the 'Charges,' the infraction or the crime committed. Once the legal events are described as semantic frames, it is possible to establish relations among semantic frames, pointing which legal action comes first. In this moment, the project has had a moderate progress having described about ten legal frames related to the lawsuit process.

The Lawsuit frame has as participants and props 'Type of Action,' which indicates the type of lawsuit that was filled against a defendant (administrative, criminal, familiar), 'Author,' who is the person that goes to the court with a request, 'Defendant,' who is the person that is been suited, and 'Concrete case,' which is the legal base that gives the author the right to make a legal request. Fig. 2 shows a schematic representation of the Lawsuit frame.

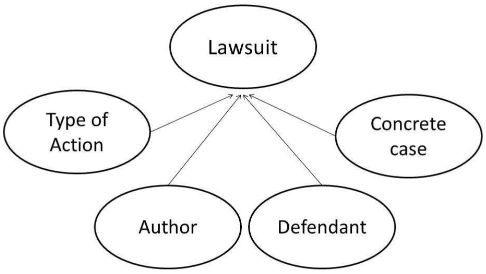

Figure 2. Participants and props of Lawsuit frame

After having developed a set of legal semantic tags organized by semantic frames, the project intends to annotate legal texts to develop a learning corpus. The assumption is that if the legal databases of the Brazilian legal courts present a more fine-grained semantic annotation, the IR system could be able to return documents related closely to the topic that the specialist needs when he searches court databases for precedent decisions. Fig. 3 shows an example of semantically annotated sentences of legal documents.

Antonio Jair da Costa$_{\text{[AUTHOR]}}$ **ajuizou** ação$_{\text{[TYPE OF ACTION]}}$ contra o Instituto Nacional Do Seguro Social – INSS$_{\text{[DEFENDANT]}}$.

[*Antonio Jair da Costa filed a lawsuit against the National Institute of Social Security – INSS*.]

Figure 3. Example of annotated sentences

Fig. 3 shows an example of annotation based on semantic frame. The participants of the lawsuit are pointed with the semantic tags AUTHOR and DEFENDANT. The tag TYPE_OF_ACTION indicates if the lawsuit is, for instance, criminal or civil.

What this project intends is to describe the knowledge that a specialist has about legal terms/words when he/she is reading a legal document. The representation of these legal events in a knowledge base, with the FrameNet [5] foundation could reproduce the cognitive connections that specialists make when they are reading a legal document. Once this database was developed, the semantic tags could be used for documents annotation, following others frame-based annotation projects [7][8][9].

This project has been developed by a multidisciplinary group, counting with linguists, lawyers, and computer science specialists. Despite all the steps described until this point being manually-based, further steps will include the knowledge base implementation, the manual annotation of legal corpora, the development of an automatic document annotation tool, and the integration of the knowledge base to the IR system of the Brazilian court.

The automatic document annotation tool is aimed to allow improvements in the indexing process of the IR system. Since the semantic frames and the semantic tags being described allow an approach not syntactically based, but rather semantically based, it consists in the first step to include the results of the developed knowledge base in the IR system operation. Therefore, the expected outcome of this tool is the possibility of document indexing operations based on the semantic frames related to legal knowledge, which is a more precise approach than the word-based indexing traditionally observed in IR systems.

The integration of these resources with the IR system will be done through the implementation of semantic treatment modules to be applied together with the existing document indexing module and information retrieval module. In the first one, the improvements obtained are related to the use of semantic frames in the indexing process. In the second one, the outcomes are due to the change in the traditional search operations, shifting from word-based search operations to semantic-based operations.

IV. EXPECTED RESULTS

The target of this project is, first, to apply a linguistic theory to legal knowledge modeling and, second, to improve a practical implementation. This project expects to develop a semantic-based legal IR system to help specialist in searching legal documents in the online Brazilian court databases. An important step in this direction is to find a partner court that is willing to provide legal documents for this research project and to support the implementation of a frame-based IR system. Some negotiation process was started with The National Council of Justice (CNJ) [10] to provide the legal documents for this research project.

Considering the amount of manual work that will be needed to develop a legal frame-based knowledge database, the expectation is that in five years the linguistic and conceptual part of the project could be ready. Once the search in the legal database can be done by the content of the documents, the specialist will receive documents more related to the intents of their search, saving a precious time spent only to look for precedent decisions.

V. RELATED WORK

Since Frame Semantics [1][2] was proposed, a number of studies and applications were developed. The first project to apply the principles of Frame Semantics was FrameNet [5]. FrameNet is a computational lexicographic project that has been developing a lexical database describing the meaning of English lexical units relating them to semantic frames. After the development of FrameNet, many projects started to develop FrameNets for different languages. Here are just some examples: Japanese FrameNet [11], Spanish FrameNet [12], German FrameNet [13], Swedish FrameNet [14], and Brazilian FrameNet [15][16].

Another application of FrameNet is in semantic annotation. Gildea and Jurafsky propose an automatic method of using FrameNet semantic tags for automatic annotation [7]. Padó and Lapata suggest an automatic method to annotate multilingual corpora using FrameNet semantic tags [17][18]. The Salsa project opts for manual corpus annotation with semantic frames [8]. Other FrameNet applications include the use of frame-based lexicons for foreign language education [19] and sentiment analysis [20].

In the legal domain, Venturi [9] applies the FrameNet semantic tags to the annotation of legislative texts. According to Venturi's findings [9], despite being created from English lexicon, FrameNet semantic labels can be applied for semantic annotation of Italian legislative texts with no significant mismatches. In previous work [21][22], the authors point the necessity to develop a set of frames for Brazilian legal system. Differently from legislative texts, semantic frames and tags to annotate Brazilian court decision change significantly. This is the reason why this project suggests the development a set of semantic frames for Brazilian legal system.

VI. CONCLUSION AND FUTURE WORK

This paper was concerned to present a semantic-based information retrieval project. Only the linguistic part of the knowledge base development was addressed here. This project has been developed for a multidisciplinary research group integrated by linguists, lawyers and computer scientists. The legal knowledge database development requires a huge amount of lexicographic work to select the legal terms/words that will be described, as well as conceptual work to design semantic frames to represent the meaning of these terms/words.

The expected contribution of this work is to improve the legal information retrieval of court databases, optimizing the time specialists spend looking for meaningful documents on legal databases. Moreover, this work tries to find a solution for a practical problem using linguistic studies for knowledge representation. This project is an initiative towards a semantic-based information retrieval system, trying to meet the needs of the Brazilian society.

ACKNOWLEDGMENT

The authors acknowledge the financial support given by the Brazilian research support agencies CNPQ, CAPES and FAPERGS. The authors thank the valuable comments of the reviewers.

REFERENCES

[1] C. J. Fillmore, “Frame Semantics,” in Linguistics in the Morning Calm, The Linguistic Society of Korea, Ed. Seoul: Hanshin, 1982, pp. 111-137.

[2] C. J. Fillmore, “Frames and the semantics of understandings,” in Quaderni di Semantica, vol. 6, no. 2, 1985, pp. 222-254.

[3] Rio Grande do Sul State Court of Appeal/Tribunal de Justiça do Estado do Rio Grande do Sul. [Access: 2014, 03]. Available at: www.tjrs.jus.br.

[4] Regional Federal Court of Appeal/Tribunal Regional Federal da 4ª Região. [Access: 2014, 03]. Available at: www.trf4.gov.br.

[5] C. J. Fillmore, C. R. Jonhson, and M. R. L. Petruck, “Background to FrameNet,” International Journal of Lexicography, vol. 16, no. 3, Sep. 2003, pp. 235-250.

[6] M. Minsky, “A framework for representing knowledge,” Artificial Intelligence Meno no. 306, Cambridge: Massachusetts Institute of Technology, 1974.

[7] D. Gildea and D. Jurafsky, “Automatic labelling of semantic roles,” Computational Linguistics, vol. 28, no. 3, Sep. 2002, pp. 245-288.

[8] A. Burchardt et al., “Using FrameNet for the semantic analysis of German: annotation, representation, and automation,” in Multilingual FrameNets in computational lexicography: methods and applications, H. C. Boas, Ed., Berlin/New York: Mouton de Gruyter, pp. 209-244, 2009.

[9] G. Venturi, “Semantic annotation of Italian legal texts: a frame-based approach,” Constructions and Frames, vol. 3, no. 1, 2011, pp. 46-79.

[10] Coselho Nacional de Justiça/National Council of Justice [Access: 2014, 03]. Available at: www.cnj.jus.br.

[11] K. H. Ohara, “Frame-based contrastive lexical semantics in Japanese FrameNet: the case of risk and kakeru,” in Multilingual FrameNets in computational lexicography: methods and applications, H. C. Boas, Ed., Berlin/New York: Mouton de Gruyter, pp. 163-182, 2009.

[12] C. Subirats, “Spanish FrameNet: a frame-semantic analysis of the Spanish lexicon,” in Multilingual FrameNets in computational lexicography: methods and applications, H. C. Boas, Ed., Berlin/New York: Mouton de Gruyter, pp. 136-162, 2009.

[13] H. C. Boas, “Semantic frames as interlingual representations for multilingual lexical databases,” in International Journal of Lexicography, vol. 18, no. 4, Dec. 2005, pp. 445-478.

[14] L. Borin, M. Forsberg, and B. Lyngfelt, “Close Encounters of the fifth kind: some linguistic and computational aspects of the Swedish FrameNet ++ project,” in Veredas, vol. 17, no. 1, 2013, pp. 28-43.

[15] M. M. M. Salomão, “FrameNet Brasil: a work in progress,” in Calidoscópio, vol. 7, no. 3. Sep./Dec. 2009, pp. 171-182.

[16] T. T. Torrent and M. Ellsworth, “Behind the labels: criteria for defining analytical categories in FrameNet,” in Veredas, vol. 17, no. 1, 2013, pp. 44-65.

[17] S. Padó and M. Lapata, “Cross-lingual projection of role-semantic information,” in Proc. Of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, Vancouver: Association for Computational Linguistics, 2005, pp. 859-866.

[18] S. Padó, Cross-lingual annotation projection models for role-semantic information. PhD Thesis. Saarbrücken: Universität des Saarlandes, 2007.

[19] H. C. Boas and R. Dux, “Semantic frames for foreign language education: towards a German frame-based online dictionary,” in Veredas, vol. 17, no. 1, 2013, pp. 82-100.

[20] J. Ruppenhofer, “Extending FrameNet for sentiment analysis,” in Veredas, vol. 17, no. 1, 2013, pp. 66-81.

[21] A. Bertoldi and R. Chishman, “Developing a frame-based lexicon for the Brazilian legal language,” in Proc. Of International Workshop on Artifitial Intelligence Approaches to the Complexity of Legal Systems (AICOL-III), AI Approaches to the Complexity of Legal Systems – Models and Ethical Challenges for Legal Systems, Legal Language and Legal Ontologies, Argumentation and Software Agents, vol. 7639, Berlin/ Heidelberg: Springer-Verlag, 2012, pp. 256-270.

[22] A. Bertoldi and R. Chishman, “Applying Frame Semantics for the Description of the Brazilian Law.” in Veredas, vol. 17, no. 1, 2013, pp. 117-133.