Identifying Causal Relations in Legal Documents with Dependency Syntactic Analysis

Pablo Gamallo
Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS)
University of Santiago de Compostela, Galiza
pablo.gamallo@usc.es

Patricia Martín-Rodilla
Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS)
University of Santiago de Compostela, Galiza
patrica.martin.rodilla@usc.es

Beatriz Calderón
University of Santiago de Compostela, Galiza
beatriz.calderon.cerrato@rai.usc.es

Abstract

This article describes a method for enriching a dependency-based parser with causal connectors. Our specific objective is to identify causal relationships between elementary discourse units in Spanish legal texts. For this purpose, the approach we follow is to search for specific discourse connectives which are taken as causal dependencies relating an effect event (head) with a verbal or nominal cause (dependent). As a result, we turn a specific syntactic parser into a discourse parser aimed at recognizing causal structures.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases Dependency Analysis, Discourse Analysis, Causal Markers, Legal Documents

Digital Object Identifier 10.4230/OASIcs.SLATE.2019.20

Funding This work has received financial support from the DOMINO project (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE), eRisk project (RTI2018-093336-B-C21), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

1 Introduction

Compared to work on syntactic analysis, approaches focused on higher linguistic levels, such as discourse analysis, are scarcer for all languages, including English [10]. In the case of Spanish language, there is still very little work on discourse analysis, which is mainly focused on RST-like models [1]. Among the different approaches on discourse analysis, special attention deserves automatic identification of causal relationships. Causation plays a central role in scientific and social domains, including the legal domain where cause-effect relationships stand for the scaffolding necessary to provide an argumentative diagnosis. By identifying cause-effect relationships in legal documents, it is also possible to identify the agents that play a role in legal acts [7]. Besides, in the legal domain, we are confronted with specific text types and, therefore, with the need to adequate the formalism for representing discourse patterns typical of this domain. [11].

Current Machine Learning (ML) based approaches have shown good results at lexical and syntactic levels, and there is some attempt to work on extraction of discursive components with ML algorithms, with good results [8]. However, existing systems require a large volume of annotated corpus for the training phase, which is time-consuming. Besides, pre-training resources do not focus on the legal domain and on its particular structure and semantics,

so we would have to annotate *ad hoc*. In addition, existing works focus on textual sources in English and, as far as we know, the analysis of this type of discursive structures for Spanish from the ML perspective is not widespread. Because of this, a rule-based work on discursive structure identification conforms, in addition to a breakthrough in the state of the art of automatic discourse treatment in Spanish, a structural basis for further computational treatment from ML perspectives.

The objective of this article is to provide a dependency-based syntactic analyzer with specific rules to identify causal relations between linguistic units in legal domain documents. We convert a specific syntactic analyzer into a discourse parser by adding specific discursive rules adapted to the legal domain. This will be done by identifying causal markers appearing in legal texts, as well as by defining the corresponding grammar rules for each marker and the linguistic/discursive units they put in relation. Our work, therefore, resembles those few approaches that propose to use dependency structure to directly represent the relations between elementary discourse units [9, 13]. Unlike other approaches focused on causality verbs to identify elementary discourse units [12], our main focus are syntactic connectors codified as causal conjunctions and/or locutions.

The enriched syntactic parser was applied to Spanish legal texts and was evaluated using a test document manually annotated. Performance of the system reached 0.65 F-score with high precision and rather low recall. This short article is organized as follows. Section 2 describes the proposed method, which is evaluated in Section 3. Conclusions and future work are addressed in Section 4.

2 The Method

Our method consists of two main tasks: first, we study a set of legal documents so as to identify the main causal connectors and their syntactic behavior, namely what type of units they relate. And second, we improve an existing rule-based syntactic parser by adding specific grammar rules to identify causal relationships between sentences and/or nominal phrases.

2.1 Causal Markers

In the current work, we focus on cause-effect relationships established through causal markers, where the cause can be either a verb clause or a nominal phrase, and the effect is a verb clause. Implicit causal relations are not considered since they are not introduced by any connector. We only take into account causal relations identified by means of explicit markers such as in the work by [14], where the authors introduced a set of *concessive* discursive markers (e.g., *though*, *but*, etc) aimed at selecting opposition relationships between discursive units for both English and Spanish languages.

In our framework, the cause is syntactically codified as the subordinate or dependent unit, while the effect is the head. The connector represents an element specifying the dependency relation. Causal connectors (including those having a consecutive meaning) are conjunctions or locutions classified into two main subcategories:

Verbal Cause: connectors putting in relation a verbal effect with a verbal cause, for instance, “porque”, “pues”, “puesto que”, “por lo que”, etc (all of them might be translated into English as *because*),

Nominal Cause: prepositional locutions connecting a verbal effect with a nominal cause, for instance, “en virtud de” (*by virtue of*) or “en razón de” (*because of*).

Table 1 List of Spanish causal conjunctions and locutions manually identified in legal texts.

Verbal Cause	porque, pues, puesto que, dado que, visto que, considerando que, teniendo en cuenta que, ya que, en consecuencia, por tanto, por consiguiente, por eso, por lo cual, por lo tanto, por lo que
Nominal Cause	a causa de, a fuerza de, a propuesta de, a petición de, a efecto de, a los efectos de, con motivo de, en razón de, en virtud de, en vista de, por causa de, por culpa de, por razón de, de acuerdo con, gracias a

It is worth noting that *verbal cause* connectors allow us to identify inter-sentence relations while *nominal cause* locutions identify intra-sentence relations. In both cases, the related elements represent elementary discourse units. By using Spanish legal documents, we have identified and defined 30 causal and consecutive locutions/conjunctions, 15 of them being verbal cause connectors, and 15 nominal cause ones. We make use of a broad sense of causal and consecutive relations, including *finality* values in some cases.

2.2 Dependency Rules

To implement causal rules, we use the grammar formalism, DepPattern [5], which is suited to define syntactic depedencies between any syntactic unit. These formal grammars are compiled into finite-state transducers working as dependency parsers [4, 3]. To carry out our objective, we used as starting point the freely available Spanish grammar of the DepPattern project.[1] Lemmatization and PoS tagging has been performed with the multilingual toolkit, Linguakit [2].

A DepPattern grammar is constituted by a set of dependent rules. Every rule aims at identifying a specific dependent-head relation by means of a pattern of part-of-speech (PoS) tags. Any dependency rule is constituted by two elements:

- a pattern of PoS tags,
- the name of a dependent-head relation found within the pattern.

Let us see an example:

DobjR: VERB NOUN

The two elements of a rule are separated by a semicolon. The first element is "DobjR", which stands for the name of a specific dependency relation, namely "a direct object (Dobj) appearing to the right (R) of the head". The second element is a sequence of PoS tags: VERB and NOUN, respectively represent the "head" and the "dependent" units. This rule is applied after having identifying the dependent units of the noun within the nominal phrase and the dependent units of the verb within the verbal phrase.

Each rule is implemented as a transducer that recognizes head-dependent token relations and removes the *dependent* tokens from the input sequence. It is applied from left to right until it reaches the end of the input sentence. The successive application of these rules simplifies and reduces the search space of the next rule to be applied [4]. For instance, to process the sequence "to have a nice day", the parser applies first the rules that identify the dependency between the adjective "nice" (dependent) and "day" (head), and that between the determiner "the" (dependent) and "day" (head), before applying the direct object rule linking the verbal head "have" to the dependent noun, "day". At each rule application, the dependent unit is removed from the search space as each dependent word only must have one head.

[1] `https://github.com/citiususc/DepPattern`

For the current work, the Spanish grammar was provided with a set of rules aimed at identifying causal relations. All causal connectors we have identified in the previous task were declared and classified in the enhanced version of the Spanish grammar so as to be used in specific rules. Table 2 shows two specific cause/effect rules. The first one is the dependency relation "CausalR" which detects the relationship between two verbs linked by a causal connector (CONJ<loc:verbal_cause>), and where the dependent/subordinated verb (the verbal cause) appears to the right of the head verb (the effect). This rule is able to identify the causal relation between the two verbs appearing in the language sequence of the second column. Symbol "[Fc]?" represents an optional comma. CONJ<loc:verbal_cause> means that the rule is using connectors belonging to the *verbal cause* subcategory. It is worth noting that this rule is only applied after having identified all the dependent elements (complements, auxiliars, and modifiers) of the two verbs heading the related sentences. So, once this rule is applied, we are able to build the entire syntactic tree starting from the root verb, which is the head of the cause/effect dependency.

The second rule in Table 2 is used to identify causal relations starting with the connector and the dependent noun (nominal cause), followed by a verbal head (effect). The notation CONJ<loc:nominal_cause> means that the rule is using connectors belonging to the *nominal cause* subcategory.

Table 2 The first column shows DepPattern rules to identify causal dependencies using conjunctions or locutions as markers. In the second column, we show a corpus-based sequence the rule was applied to. The identified units are in bold: cause, effect and connector.

Cause/effect rules	**Language sequences**
CausalR: VERB [Fc]? CONJ<loc:verbal_cause> VERB	"Galicia, compendio de universalidad, **quiere participar** con plena dignidad y protagonismo en el concierto de las culturas, **por lo que** en este texto se **asumen** mandatos, criterios y principios recogidos en las diversas cartas" (*Galicia, a compendium of universality,* ***wants to participate*** *with full dignity and protagonism in the concert of cultures,* ***so*** *this text* ***assumes*** *mandates, criteria and principles contained in the various letters*)
CausalL: CONJ<loc:nominal_cause> NOUN [Fc]? VERB	"**A los efectos de** esta **ley**, **integran** el patrimonio artístico de Galicia las manifestaciones pictóricas, escultóricas, cinematográficas, fotográficas, musicales y de las restantes artes plásticas, de especial relevancia, de interés para Galicia" (***For the purposes of*** *this* ***law****, the artistic heritage of Galicia* ***includes*** *pictorial, sculptural, cinematographic, photographic, musical and other plastic arts of special relevance, of interest to Galicia.*

Table 3 Performance of the system in the process of recognizing causal relations.

total	*tp*	*fp*	*fn*	*precision*	*recall*	*F-score*
91	45	2	46	0,96	0,50	0,65

The output of the system can be either in CoNLL-X format [6] or in dependency triples (similar to Standford dependencies).[2] The grammar updated with causal rules is freely available from the DepPattern project.[3]

3 Evaluation

In order to evaluate our system, a legal document, namely the law 5/2016 on Galician cultural inheritance, containing 13k words, was manually annotated. More precisely, a linguist identified all causal connectors and their related discursive units. The document was syntactically analyzed and the performance of the results were measured. Precision is the number of correct decisions (true positives) divided by all decisions taken by the system (true positives + false positives). Recall is the number of correct decisions (true positives) divided by the total number of causal relationships found in the document (true positives + false negatives). Finally, F-score is the harmonic average of the precision and recall.

Table 3 shows the results of the evaluation. 91 causal relations were manually found in the test document. In order to compute precision, recall, and F-score, we count the number of true positives (tp), false positives (fp) and false negatives (fn). The system achieves 0.95 precision, while only 0,50 recall. F-score thus reaches 0,65.

According to the results depicted in Table 3, the system has a very high degree of accuracy, but coverage fails. Among the main problems of coverage, we must highlight the cases of very ambiguous connectors that were not introduced into the grammar. For instance, the ambiguous conjunction "como" (*as*) is missing as a causal marker in the grammar, even if it is used with this meaning in the test document. Preposition "por" (*by*) is also used as a causal connector but it has not adding to the grammar as such due to its very ambiguous behavior.

4 Conclusions

This paper describes a rule-based method to introduce causal dependencies into a dependency parsing by updating DepPattern grammars. The performance of the system reaches 0.65 F-score with high precision but still low recall due to the fact that ambiguous connectors are missing in the grammar.

In future work, the grammar will be provided with more specific rules and enriched with new connectors. Moreover, the Spanish grammar will be adapted to other languages, namely Portuguese, Galician and English, which are the languages that DepPattern supports in the current version of the formalism. In addition, we will create RST-like structures from the dependency output, by converting the dependencies between causal units into full constituents. For this purpose, we will establish functional equivalence between *head* and *nucleus* as well as *dependent* and *satellite*. Finally, we will also include a visualization module so as to allow users to look up linguistic patterning and discourse structure in a more friendly way.

[2] `https://nlp.stanford.edu/software/stanford-dependencies.shtml`
[3] `https://github.com/citiususc/DepPattern/tree/master/grammars/grammar-devel-es`

References

1 Iria da Cunha. A Symbolic Corpus-based Approach to Detect and Solve the Ambiguity of Discourse Markers. *Research in Computing Science*, 70:95–106, 2013.

2 P. Gamallo, M. Garcia, C. Piñeiro, R. Martinez-Castaño, and J. C. Pichel. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244, 2018. doi:10.1109/SNAMS.2018.8554689.

3 Pablo Gamallo. Dependency Parsing with Compression Rules. In *Proceedings of the 14th International Workshop on Parsing Technology (IWPT 2015)*, pages 107–117, Bilbao, Spain, 2015. Association for Computational Linguistics.

4 Pablo Gamallo and Marcos Garcia. Dependency parsing with finite state transducers and compression rules. *Information Processing & Management*, Available online 5 June 2018, 2018.

5 Pablo Gamallo and Isaac González. A Grammatical Formalism Based on Patterns of Part-of-Speech Tags. *International Journal of Corpus Linguistics*, 16(1):45–71, 2011.

6 Johan Hall and Jens Nilsson. CoNLL-X shared task on multilingual dependency parsing. In *The tenth CoNLL*, 2006.

7 Rinke Hoekstra and Joost Breuker. Commonsense Causal Explanation in a Legal Domain. *Artificial Intelligence Law*, 15(3):281–299, 2007. doi:10.1007/s10506-007-9033-5.

8 Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41(3):385–435, 2015. doi:10.1162/COLI_a_00226.

9 Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. Text-level Discourse Dependency Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi:10.3115/v1/P14-1003.

10 Amália Mendes and Iria del Río. Using a Discourse Bank and a Lexicon for the Automatic Identification of Discourse Connectives. In Aline Villavicencio, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonçalo Oliveira, and Gustavo Henrique Paetzold, editors, *Computational Processing of the Portuguese Language*, pages 211–221, Cham, 2018. Springer International Publishing.

11 Marie-Francine Moens, Caroline Uyttendaele, and Jos Dumortier. Intelligent Information Extraction from Legal Texts. *Information & Communications Technology Law*, 9(1):17–26, 2000. doi:10.1080/136008300111583.

12 Chaveevan Pechsiri and Asanee Kawtrakul. Mining Causality for Explanation Knowledge from Text. *J. Comput. Sci. Technol.*, 22(6):877–889, 2007. doi:10.1007/s11390-007-9093-8.

13 Kenji Sagae. Analysis of Discourse Structure with Syntactic Dependencies and Data-Driven Shift-Reduce Parsing. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 81–84, Paris, France, October 2009. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W09-3813.

14 Maite Taboada and María de los Ángeles Gómez-González. Discourse markers and coherence relations: Comparison across markers, languages and modalities, 2012.