Thesaurus Enhanced Extraction of Hohfeld's Relations from Spanish Labour Law

Patricia Martín-Chozas[1][0000−0001−5416−6370]
Artem Revenko[2][0000−0001−6681−3328]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
pmchozas@fi.upm.es
[2] Semantic Web Company, Vienna, Austria
artem.revenko@semantic-web.com

Abstract. In this paper we describe the design of an experiment to extract Hohfeld's deontic relations from legal texts. Our approach intends to minimise the manual effort in the annotation process by expanding a set of initial annotations with the legal domain knowledge contained in thesauri represented in Semantic Web formats. With such annotations, we perform a set of iterations to train a deep learning relation extraction model. After analysing the results, we will adapt the process to work on the extraction of Hohfeld's potestative relations. We also plan to use that model to recognise relations in unseen legal sub-domains.

Keywords: Relation Extraction · Thesaurus · Terminology · Semantic Web

1 Introduction

New legal documentation is being generated daily, which implies new regulations and laws that need to be processed and, most importantly, *understood*. Several works have already tackled the difficulties in legal information processing, such as [5], which identifies five major aggravating factors: multijurisdictionality, volume, accessibility, updates and consolidation and vagueness of legal document classification.

Natural language processing tools help solving such challenges, and they can reach great performance on many language understanding tasks [25]. Yet, these models require significantly large annotated datasets and language resources to train. We found, however, that legal language resources are scarce, mostly monolingual, and sometimes published in close and proprietary formats. This may be one of the reasons why most Information Extraction systems, and Relation Extraction tools specifically, do not handle legal texts properly and, if they do, they tend to return very general results (see Section 2). Therefore, with the aim of making legal information understandable and easier accessible, in this paper we describe the design of an experiment to extract relations amongst terms in legal texts. We further represent them as part of rich domain-specific multi-lingual resources, that can be ultimately exploited for different use cases.

This work is framed within Lynx[3] project, an Innovation Action funded by the European Union's Horizon 2020, whose goal is to create a Knowledge Graph of legal and

[3] http://lynx-project.eu/

regulatory data to ease the access to information from different jurisdictions, languages and domains. Such a Legal Knowledge Graph (LKG) could be of a great help to comply with current regulations, specially for non-legal-expert users.

Amongst all legal relations the Hohfeld's fundamental legal relations are the most general ones [10]. The Hohfeld's relations, being the highest abstraction of all possible legal relations, may serve the basis for more detailed domain-specific legal relations. In other words, the legal relations appearing in legal sub-domains may be seen as sub-relation of Hohfeld's relations. They are divided in two sets of relations: *deontic relations* (Right, Duty, No-Right and Priviledge) and *potestative relations* (Power, Liability, Disability and Immunity). The term "deontic" refers to a branch of the logic that is responsible for studying the inferential relationships between normative formulas that include the operators of permission (P), obligation (O) and prohibition (F), amongst others [24]. While deontic relations (Figure 1) are those that modify (ordinary) actions, potestative relations modify deontic relations. In this preliminary experiment we will put the focus on the deontic relations, leaving potestative relations for future work.

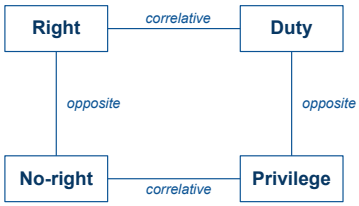

Fig. 1: Hohfeld's Deontic Relations

Taking into account the nature of the deontic relations, we decided that a good starting point would be to analyze the subdomain of labour law, that deals with rights and duties of employers and employees. We have selected one of the most representative texts of Spanish labour law, the Spanish Workers' Statute[4]. In the next steps of this approach we aim at generalizing the models to recognize Hohfeld's relations in different legal areas and in multiple languages. Since in every subarea of law, we will find different instances of Hohfeld's relations, we suggest anyone aiming at Information Extraction from legal texts to start with our general models and fine-tune them for the specific use case. As fine-tuning general requires much less training data, the fine-tuning datasets could be created by the legal experts with little effort and would, therefore, enable the tuning of the model to the specific task at hand.

[4] https://www.boe.es/eli/es/rdlg/2015/10/23/2/con

2 Related Work

Since the scope of this approach is still very open, the related work revised is equally wide. We refer the readers to [17] that exposes in detail the latest advances several information extraction techniques, including works in Relation Extraction.

Throughout the literature, we can find many relation extraction experiments based on very different technologies. Some of them are based on Knowledge Bases, such as [28], that is based on Freebase (currently deprecated) [6], and is aimed at inferring answers to questions in natural language. A similar work, [23], employs two different KBs, PATTY [18] to identify DBpedia [3] predicates that allow translating natural language questions into SPARQL queries to reason over entities. Other works employ linguistic approaches, such as [1] that applies deep linguistic patterns to infer relations over the English Wikipedia; and [22], that presents Falcon, a tool that identifies entities in short texts and create relations based on KBs and linguistic patterns.

Recent advances in deep learning methodologies [12, 11] have significantly improved the state of the art results on well-established relation extraction benchmarks such as TACRED [29] or SemEval 2010 Task 8 [9]. These models use the contextualised pre-trained representation of word-pieces to obtain high quality semantic information about different words in context. Best performing models, for example, SpanBERT [13] and REDN [15], use not only the individual embeddings of tokens, but also spans of entities, their lengths, and aggregated embeddings of contexts to get better performance.

Based on the analysis of works cited in the previous paragraphs, we claim that relations between terms can be of different nature, going beyond hypernymy or synonymy. We therefore intend to discover domain-specific relations amongst them, adding extra information to each element, such as the superclass of the terms involved (subject and object) and the kind of Hohfeld's relation expressed by the predicate.

3 Envisioned approach

3.1 Corpus

As mentioned in Section 1, our study is based on the Spanish Workers' Statute, that is published in the Official State Gazette website[5]. This corpus is divided into three main sections named as "titles". The first title covers individual labour relations; the second title covers the rights of collective representation and workers' assemblies inside companies, and the third title covers collective bargaining and collective agreements. In total, the three sections gather 92 articles, containing approximately 50.000 tokens.

With the current state of analysis we estimate the density of relations in the Spanish labour law to be 3.65 relations per article. This number is considered a lower boundary, since the estimation is calculated over explicit relations, i. e. those relations that can be attributed to a particular verb in the sentence, but we also expect to retrieve suggestions of implicit relations predicted by the model.

To get an idea of the number of entities contained in the corpus, we performed statistical terminology extraction with TBXTools[6], which applies its own algorithm based

[5] `https://www.boe.es/`

[6] https://sourceforge.net/projects/tbxtools/files/

on the calculation of *n-grams* (the combination of n words appearing in the corpus) and on the *normalisation* of terms [20] [19]. The list of ranked extracted terms, including multi-word expressions, is revised manually to remove noisy results. After this analysis, we can count with a total of 614 terms, that are considered the arguments of our relations. These terms do not include Named Entities, so we also consider it as a lower boundary. Both the corpus and the entity list are publicly available[7] – the results of the experiments will also be progressively uploaded.

3.2 Methodology

In the first step a small excerpt for the legal corpus is manually annotated. We use the well-established legal thesauri[8][9], generated within the frame of the Lynx project, and the manually verified terminology to produce candidates relations. Since every type of relation of our interest has domain and range restrictions defined manually, we can filter candidate relations by applying the restrictions. Hence, we can efficiently generate candidate entity pairs for each relation, for instance, amongst *employee* and *contract* in Example 1. The total size of acquired manually verified relations at this stage is in the order of 100 samples. These annotations include both entity and relation annotations (see Example 1), that enable the specification of the relations of interest, including domain and range restrictions of all relation types.

Example 1. **Context** El *trabajador* podrá rescindir el *acuerdo* y recuperar su libertad de trabajo en otro empleo (The *worker* may rescind the *agreement* and regain his freedom to work in another job).
Entities *trabajador (worker)*: `LegalEntity`, *acuerdo (agreement)*: `LegalDocument`.
Relation Type Right.

Context El *empresario* deberá informar por escrito al *trabajador* sobre las condiciones de trabajo (The *employer* must inform the *worker* by written notification about the working conditions).
Entities *empresario (employer)*: `LegalEntity`, *trabajador (worker)*: `LegalEntity`.
Relation Type Duty.

Context La *duración del contrato* no podrá ser inferior a *seis meses* (The *duration of the contract* must not be less than *six months*).
Entities *duración del contrato (duration of the contract)* : `LegalEntity`, *seis meses (six months)*: `Duration`.
Relation Type No-right.

Context Asimismo, el *Gobierno* podrá otorgar *subvenciones*, desgravaciones y otras medidas (Likewise, the *Government* may grant *subsidies*, tax breaks and other measures).
Entities *Gobierno (Government)*: `LegalEntity`, *subvenciones (subsidies)*: `LegalConcept`.
Relation Type Privilege.

[7] `https://github.com/pmchozas/term_relex`
[8] `https://zenodo.org/record/3843561`
[9] `http://lkg.lynx-project.eu/kos`

At this point, we go with second step of our methodology, that is the initial training dataset to train the Relation Extraction model – modelV0.1. For the training, we use R-BERT [27] model. This models takes into account the aggregated entity spans as well as the embeddings of the whole context to classify the relations. Though the model is not reaching the best scores, it is quite competitive, robust and easy to use. Several implementations are openly available[10].

Once the model is trained, we reach the final step, where we can use the model to predict new relations. As the training set is still small, we expect the model to produce many incorrect predictions. These predictions are verified manually to expend the training set and re-train the model (see Figure 2).

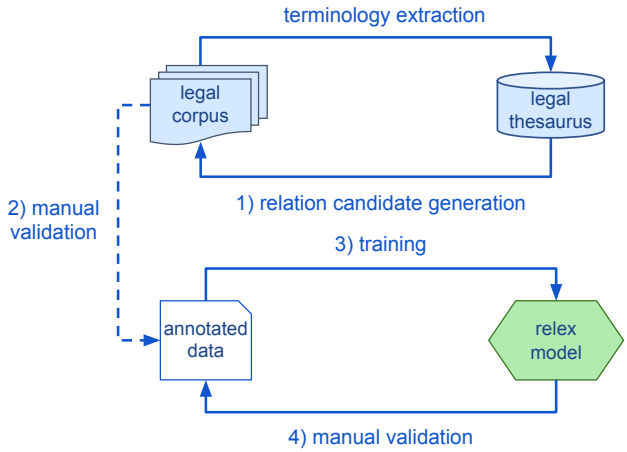

Fig. 2: Our envisioned methodology is composed of four steps: 1) relation candidate generation, 2) manual validation, 3) training 4) manual validation, and then again 3) (re-)training. The whole process can be iterated as many times as needed.

As mentioned in the introduction, the idea is to include these Hohfeld's relations into the knowledge graph represented in Semantic Web formats. We find several works that tackle the representation of Hohfeld's relations in Semantic Web formats. One of the most well-known legal ontologies including such concepts is LegalRuleML [2], a markup language able to represent the particularities of the legal normative rules. On the other hand, we can find the Provision Model [4] that was extended in [7], to cover Hohfeld's relations. Both of them include properties to represent deontic relations (see Table 1) and can be of a great help to represent those found in this experiment.

[10] for example, `https://github.com/monologg/R-BERT`

Table 1: Properties representing Deontic operators as per LegalRuleML and Provision Model ontologies.

Hohfeld's Deontic Relations	**LegalRuleML**	**Provision Model**
Right	`lrml:Right`	`prv:Right`
Duty	`lrml:Obligation`	`prv:Duty`
No-right	`lrml:Prohibition`	`prv:Prohibition`
Privilege	`lrml:Permission`	`prv:Permission`

3.3 Evaluation

For the evaluation of the performance of our model we will use well established metrics such as precision (P), recall (R) and F_1 score. Let the *gold standard* be the correct manually annotated data. Let the *true positives (TP)* be all the correctly predicted relations; *false positives (FP)* – incorrectly predicted relations; *false negatives (FN)* – those cases when a relation is not predicted, though it does exist in the gold standard; *true negatives (TN)* – the relation is not predicted and it does not exist in the gold standard. Then $P = \frac{TP}{TP+FP}$, $R = \frac{TP}{TP+FN}$ and $F_1 = 2 * \frac{P*R}{P+R}$. These measures are well established and widely used for evaluation of different classification models, for example, on the aforementioned benchmarks TACRED [29] and SemEval 2010 Task 8 [9]. The best models on these datasets currently reach the scores of 74.8% F_1 on TACRED[11] and above 91% F_1 on SemEval[12].

3.4 Envisioned use case

The use case that we propose for this experiment is based on one of the pilots of the aforementioned Lynx project. Lynx Pilot 2[13], supported by Cuatrecasas[14], a globally well-known Spanish law firm, describes a platform that helps lawyers effectively identify relevant documents related to the cases they are handling. This platform is built on top of the Legal Knowledge Graph, which connects legal sources from different legal orders, countries or languages in the field of labour law, enabling the retrieval of complex information with a single query.

Based on this pilot, we propose a use case that delves a little deeper into the extraction of information: instead of identifying documents, we propose to directly identify what are the rights and the duties of a certain employee or employer under certain working conditions. We envision an interface, similar to OpenIE[15], where the user only needs to add a few parameters, such as the type of relation (duty, right...) and the type of agent

[11] `https://paperswithcode.com/sota/relation-extraction-on-tacred` accessed on April 19, 2021

[12] `https://paperswithcode.com/sota/relation-extraction-on-semeval-2010-task-8` accessed on April 19, 2021

[13] https://lynx-project.eu/project/pilot2

[14] https://www.cuatrecasas.com/

[15] https://openie.allenai.org/

(employer, employee...). First, we propose this solution at the national level, but as part of future work it is to explore whether this technique allows us to extract this type of fine grained information between jurisdictions and languages. The ultimate aim is to provide non legal experts with easily understandable pieces of information, avoiding the time-consuming task of browsing through heterogeneous legal documentation. A preliminary diagram of the user interface and architecture shown in Figure 3.

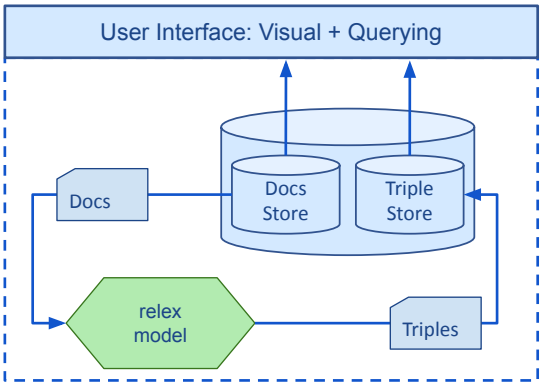

Fig. 3: Envisioned architecture and user interface.

4 Conclusions and future work

In this experiment, we train a model to extract instances of Hohfeld's deontic relations from Spanish labour law. Our methodology involves the usage of legal thesauri to perform entity annotation in an automatic way, therefore saving manual effort. The initial training set of relations has to be annotated manually, however we use the (inaccurate) predictions from the preliminary versions of the trained model to prepare samples for manual checking and, therefore, bootstrapping the training dataset. This way we efficiently use manual effort to quickly improve the model in a few iterations.

In the next steps of our experiment we aim at using transfer learning techniques [21, 16] and in particular cross-lingual transfer learning [8] to generalize the model and learn representations of Hohfeld's relations in different legal domains and in different languages. We aim at comparing the performance of multi-lingual [26] vs monolingual models for the specified task. Another interesting direction is to explore the usage of modern Language Models tuned on specific legal corpora, for example, the PatentBert [14]. These models might show better performance due to its learnt understanding of legal expressions.

Finally, we will do an experiment of deducing the general deontic relations to domain specific entities. We will use the most general trained deontic multilingual models to recognize relations in unseen domains, for example, for contract analysis and com-

pliance checking. Afterwards, we will proceed to explore the automatic extraction of potestative relations, covering the two sets of Hohfeldian legal concepts.

Acknowledgements

This work has received funding from the EU's Horizon 2020 Research and Innovation programme through the contracts Lynx (grant agreement No. 780602) and Prêt-à-LLOD (grant agreement No. 825182), and from the Spanish Ministry of Economy, Industry and Competitiveness through the Datos4.0 contract (TIN2016-78011-C4-4-R).

References

1. Akbik, A., Broß, J.: Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In: www workshop. vol. 48 (2009)
2. Athan, T., Governatori, G., Palmirani, M., Paschke, A., Wyner, A.: Legalruleml: Design principles and foundations. In: Reasoning Web International Summer School. pp. 151–188. Springer (2015)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007)
4. Biagioli, C.: Law making environment: model based system for the formulation, research and diagnosis of legislation. Artificial Intelligence and Law (1996)
5. Boella, G., Humphreys, L., Martin, M., Rossi, P., van der Torre, L., Violato, A.: Eunomos, a legal document and knowledge management system for regulatory compliance. In: Information systems: crossroads for organization, management, accounting and engineering, pp. 571–578. Springer (2012)
6. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250 (2008)
7. Francesconi, E.: Semantic model for legal resources: Annotation and reasoning over normative provisions. Semantic Web **7**(3), 255–265 (2016)
8. Gracia, J., Fäth, C., Hartung, M., Ionov, M., Bosque-Gil, J., Veríssimo, S., Chiarcos, C., Orlikowski, M.: Leveraging linguistic linked data for cross-lingual model transfer in the pharmaceutical domain. In: Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web – ISWC 2020. pp. 499–514. Springer International Publishing, Cham (2020)
9. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Séaghdha, D.Ó., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 33–38 (2010)
10. Hohfeld, W.N.: Some fundamental legal conceptions as applied in judicial reasoning. Yale Lj **23**, 16 (1913)
11. Hu, R., Singh, A.: Transformer is all you need: Multimodal multitask learning with a unified transformer (2021)
12. Huang, Y.Y., Wang, W.Y.: Deep residual learning for weakly-supervised relation extraction (2017)
13. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics **8**, 64–77 (2020)

14. Lee, J.S., Hsiang, J.: Patentbert: Patent classification with fine-tuning a pre-trained bert model. arXiv preprint arXiv:1906.02124 (2019)
15. Li, C., Tian, Y.: Downstream model design of pre-trained language model for relation extraction task. arXiv preprint arXiv:2004.03786 (2020)
16. Ma, J., Cheng, J.C., Lin, C., Tan, Y., Zhang, J.: Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmospheric Environment **214**, 116885 (2019)
17. Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I.: Information extraction meets the semantic web: a survey. Semantic Web (Preprint), 1–81 (2020)
18. Nakashole, N., Weikum, G., Suchanek, F.: Patty: A taxonomy of relational patterns with semantic types. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1135–1145 (2012)
19. Oliver, A., Vàzquez, M.: Tbxtools: a free, fast and flexible tool for automatic terminology extraction. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (2015)
20. Oliver, T., Vàzquez, M.: A free terminology extraction suite. In: Proceedings of the Twenty-ninth International Conference on Translating and the Computer (2007)
21. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)
22. Sakor, A., Mulang, I.O., Singh, K., Shekarpour, S., Vidal, M.E., Lehmann, J., Auer, S.: Old is gold: linguistic driven approach for entity and relation linking of short text. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 2336–2346 (2019)
23. Singh, K., Mulang', I.O., Lytra, I., Jaradeh, M.Y., Sakor, A., Vidal, M.E., Lange, C., Auer, S.: Capturing knowledge in semantically-typed relational patterns to enhance relation linking. In: Proceedings of the Knowledge Capture Conference. pp. 1–8 (2017)
24. Von Wright, G.H.: Deontic logic. Mind (1951)
25. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2019)
26. Wang, Z., Mayhew, S., Roth, D., et al.: Cross-lingual ability of multilingual bert: An empirical study. arXiv preprint arXiv:1912.07840 (2019)
27. Wu, S., He, Y.: Enriching pre-trained language model with entity information for relation classification (2019)
28. Xu, K., Reddy, S., Feng, Y., Huang, S., Zhao, D.: Question answering on freebase via relation extraction and textual evidence. arXiv preprint arXiv:1603.00957 (2016)
29. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 35–45 (2017)