Ontology Population and Alignment for the Legal Domain: YAGO, Wikipedia and LKIF

Cristian Cardellino[1], Milagro Teruel[1], Laura Alonso Alemany[1], and Serena Villata[2]

[1] University of Córdoba, Argentina
[2] Université Côte d'Azur, CNRS, Inria, I3S, France

Abstract. We present a methodology and framework to align ontologies through annotation of texts, and we show how this methodology applies successfully to the legal domain. This method reduces the difficulty of aligning ontologies, because annotators are asked to associate two labels from different inventories to a concrete example, which requires a simple judgment. In a second phase, those correspondences are consolidated into a proper alignment. The resulting alignment is a partial connection between diverse ontologies. By annotating judgments of the European Court of Human Rights, we have aligned an ontology of the legal domain, LKIF, to YAGO, we have thus populated LKIF in order to train a legal Named Entity Recognizer and Classifier with examples from the Wikipedia that are trivially mapped to LKIF classes. The resulting resources and the best practices we defined supported a step towards automation in legal informatics.

1 Introduction

A number of ontologies have been developed so far for the legal domain [1, 2], but they mainly deal with higher-level abstract concepts, and do not include the concrete entities organized by those concepts. For all these reasons, legal Natural Language Processing (NLP) applications are still underdeveloped with respect to the growing needs of legal scholars and common users dealing with legal documents.

In this paper, we propose a methodology to bridge the gap between higher-level concepts in ontologies and entities present in legal texts by using more abstract ontologies to annotate concrete entities occurring in texts. In the annotation process, abstract ontologies provide generalizations for concrete concepts, and concrete concepts populate abstract ontologies, which makes them useful for tasks like Information Retrieval (IR), Question Answering (QA), or Information Extraction (IE).

We annotate entities with two or more ontologies, using as a backoff a general-domain ontology, i.e., YAGO [3], and the LKIF legal ontology [2]. As a result, the ontologies that are used for the annotation end up to be aligned. Ontology alignment [4] is a very challenging task. The process of finding semantically equivalent concepts in two different conceptualizations of the same domain is very difficult for humans, even if they are adequately trained. The annotation task alleviates this difficulty by making decisions more concrete. In this task, human experts detect mentions of the relevant concepts in naturally occurring text and assign them to a concept of each of the ontologies to be aligned, which is much more natural for the annotators.

2 Legal documents annotation and ontology alignment

The annotation-based alignment process is summarized as follows. Given a target domain, we *(i)* gather a corpus of documents representative of the domain, and one or more ontologies specific for that domain; *(ii)* manually identify entities in the text; *(iii)* tag each entity with either *1)* the most specific concept in the domain ontology, if it exists, or *2)* the most specific concept from another domain ontology, or *3)* the most specific concept in YAGO or Wikipedia; *(iv)* find the most specific concept in YAGO or, if the concept is not in YAGO, in Wikipedia. We take into account that the most specific concept may be *the actual entity*.

After the annotation process, we revise the resulting alignments to check whether they are sound, and we resolve possible conflicts among the annotators. In case the assigned YAGO node has a granularity that is too fine-grained for the concept assigned from the domain-specific ontology, we establish the mapping between that concept and the most adequate ancestor of the selected YAGO node. When some equivalent concept is found, we establish the alignment using the OWL primitives `equivalentClass` and `subClassOf`. Relations are not aligned, only classes. An example is shown below.

domain-specific

The [Court]$_{PublicBody}$ is not convinced by the reasoning of the [combined divisions of the Court of Cassation]$_{PublicBody}$, because it was not indicated in the [judgment]$_{Decision}$ that [Eitim-Sen]$_{LegalPerson}$ had carried out [illegal activities]$_{Crime}$ capable of undermining the unity of the [Republic of Turkey]$_{LegalPerson}$.

YAGO

The [Court]$_{\text{wordnet trial court 108336490}}$ is not convinced by the reasoning of the [combined divisions of the Court of Cassation]$_{\text{wordnet trial court 108336490}}$, because it was not indicated in the [judgment]$_{\text{wordnet judgment 101187810}}$ that [Eitim-Sen]$_{\text{wordnet union 108233056}}$ had carried out [illegal activities]$_{\text{wordnet illegality 104810327}}$ capable of undermining the unity of the [Republic of Turkey]$_{wordnetcountry108544813}$.

By doing this, Named Entities (NE) are associated to concepts from both the domain ontology (e.g., LKIF) and Wikipedia, and thus an alignment is effectively established between both. This alignment allows to transfer properties from one ontology to the other, leading to a relevant result for inference and reasoning tasks. Being of importance for NLP applications, like NERC or IE, this alignment also provides the domain ontology (i.e., LKIF) with manually annotated examples from the Wikipedia. Wikipedia provides a fair amount of naturally occurring text where some entity mentions are manually tagged and linked to the DBpedia ontology. We consider as tagged entities the spans of text that are an anchor for a hyperlink whose URI is one of the entities that have been mapped through the annotation process.

The process of text annotation requires extensive support to ensure consistency among annotators and reproducibility of the results. To achieve that, we developed precise guidelines (i.e., best practices) for the annotators, and an annotation interface. The guidelines were roughly based on the LDC guidelines for annotation of Named Entities[3], but adapted to the annotation of legal concepts. To carry out the annotation, we

[3] `http://nlp.cs.rpi.edu/kbp/2014/ereentity.pdf`

adapted an annotation interface for NERC[4], resulting in a new annotation interface for the legal domain[5]. For the annotation, we:

1. Upload a number of documents to be annotated with the ontology;
2. Load the concepts in the domain-specific ontology (e.g., LKIF), and annotate, s.t.
   (a) When the annotator finds an entity in the text, she selects the first word and identifies the span of the entity,
   (b) The entity is assigned a label from the domain-specific ontology, which is chosen from a drop-down menu that contains all the concepts in the ontology (Figure 1). This label is the most concrete concept for that entity in the ontology.
   (c) Then, it is assigned the adequate concept in YAGO, i.e., the exact canonical name of the entity that is mentioned. Concepts that are used for the first time to annotate are manually written in the box for the labels, and from then on they are available for further uses in the drop-down menu. For instance, the entity "Government" in the text is annotated with the LKIF class `Public_Body` and the Wikipedia URI `https://en.wikipedia.org/wiki/Government_of_Spain`, since the exact entity could not be found in YAGO (Figure 1).
   (d) If an entity of interest cannot be property labeled with the concepts in the domain ontology or with a YAGO URI, the annotator looks for that concept in Wikipedia. The new label is manually written in the text box for the corresponding label, and it is available from then on in the drop-down menu.

The proposed methodology has been applied to the LKIF legal ontology [2] over the judgments of the European Court of Human Rights (ECHR). We annotated excerpts from 5 judgments of the ECHR, totalling 19,000 words.[6]We identified 1,500 entities, totalling 3,650 words.[7] Out of a total of 69 classes in the selected portion of the LKIF ontology, 30 could be mapped to a YAGO node, either as children or as equivalent classes. 55% of the classes of LKIF could not be mapped to a YAGO node, because they were too abstract (e.g., *Normatively Qualified*), there was no corresponding YAGO node circumscribed to the legal domain (e.g., *Mandate*), there was no specific YAGO node (e.g., *Mandatory Precedent*), or the YAGO concept was overlapping but not roughly equivalent (e.g., "*agreement*" or "*liability*").

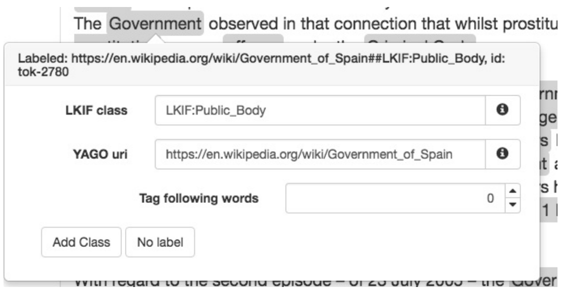
From the YAGO side, 47 classes were mapped to a LKIF class, with a total of 358 classes considering their children, and a total of 174,913 entities. We retrieved 4'5 million occurrences of these entities within the Wikipedia text. However, not all of these classes were equally populated with mentions. The number of mentions per class is highly skewed, with only half of YAGO classes having any mention whatsoever within the Wikipedia text. Of these 122 populated YAGO classes, only 50 were heavily populated, with more than 10,000 mentions, and 11 had less than 100 mentions. When

[4] `https://github.com/mayhewsw/ner-annotation`

[5] `https://github.com/MIREL-UNC/ner-annotation`

[6] The annotated texts and the resulting alignment are available at `https://github.com/PLN-FaMAF/legal-ontology-population`.

[7] Four different annotators trained for legal document annotations, and three judgments were annotated by two annotators independently (inter-annotator agreement $\kappa = .4$ to $\kappa = .61$ where most of the disagreement regards the recognition of concepts, not their classification).



The Government observed in that connection that whilst prostitu

Labeled: https://en.wikipedia.org/wiki/Government_of_Spain##LKIF:Public_Body, id: tok-2780

LKIF class: LKIF:Public_Body

YAGO uri: https://en.wikipedia.org/wiki/Government_of_Spain

Tag following words: 0

Fig. 1. The annotation of the entity *Government* in LKIF and Wikipedia.

it comes to particular entities, more than half of the entities had less than 10 mentions in the text, only 15% had more than 100 and only 2% had more than 1000.

We evaluated our NERC for the legal domain both on Wikipedia and on the judgments of the ECHR. Results are good, but the approach is sensitive to domain change (Table 1). For more details about the NERC, we refer the reader to [5].

approach	accuracy	precision	recall	F1
test on Wikipedia, trained on Wikipedia	.95	.76	.64	.69
test on ECHR, trained on Wikipedia	.89	.16	.08	.08
test on ECHR, trained on ECHR	**.95**	.76	.76	.75

Table 1. Results for NERC on the test portion of the Wikipedia corpus or the ECHR, trained with Wikipedia examples or with the annotations for the ECHR. Accuracy figures take into consideration the majority class of non-NEs, but precision and recall are an average of all classes (macro-average) except the majority class of non-NEs.

To conclude, we have presented a methodology to enhance domain-specific ontologies of the legal domain by aligning them to the YAGO general-domain ontology. The alignment is driven by examples of concepts in naturally occurring texts, facilitating the selection of the most adequate concept for the annotators.

References

1. Gangemi, A., Sagri, M.T., Tiscornia, D.: A constructive framework for legal ontologies. Law and the Semantic Web (2005) 97–124
2. Hoekstra, R., Breuker, J., Bello, M.D., Boer, A.: The lkif core ontology of basic legal concepts. In: Proceedings of LOAIT-2007. (2007)
3. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A core of semantic knowledge. In: Proceedings of WWW-2007, ACM (2007) 697–706
4. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer-Verlag (2007)
5. Cardellino, C., Teurel, M., Alemany, L.A., Villata, S.: A low-cost, high-coverage legal named entity recognizer, classifier and linker. In: Proceedings of ICAIL-2017. (2017)