Semantic Annotation and Retrieval of Parliamentary Content: A Case Study on the Spanish Congress of Deputies

Iván Cantador
ivan.cantador@uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

Lara Quijano-Sánchez
lara.quijano@uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

ABSTRACT

In this paper, we present an ontology-based annotation and retrieval approach for parliamentary content, such as debate transcripts and law proposals. Exploiting a number of domain ontologies, semantic web technologies and information retrieval techniques, our approach extracts topics, concepts and named entities (e.g., names of politicians and political parties) appearing in input documents. The domain ontologies were designed to support multilinguality, and were built from the United Nations taxonomy of sustainable development goals. The approach was instantiated with a text corpus extracted from the Spanish Congress of Deputies and is being integrated into an e-government platform.

CCS CONCEPTS

• **Applied computing** → **Computing in government**; • **Information systems** → **Ontologies**; **Information extraction**; **Information retrieval**.

KEYWORDS

parliamentary content, semantic annotation, argument extraction, ontology-based information retrieval

1 INTRODUCTION

Managing and publicly providing digital libraries on parliamentary activity are essential tasks for open government, promoting democracy, enhancing transparency, and facilitating accountability. However, the amount of multimedia content recording the debates and proposals generated by parliaments is huge and ever-increasing. This together with the unstructured nature of such content makes its organization, access and retrieval challenging.

As stated in [12], metadata facilitates the classification, storage and retrieval of e-government resources. It summarizes the available contents, allows users to manage, find and access the resources, helps understanding and determining if the corresponding information meet particular requirements, and, thanks to a consistent description of the data, promotes its sharing and exchange.

Public administrations are aware of the advantages of sharing open government data with regard to transparency, stakeholder collaboration, improved services, and new economic activities [20]. Hence, in the last decade, there has been a large increment of initiatives to publish and interlink government data and services. This has been facilitated by the use of semantic web technologies and standards, and the generation of Linked Open Data (LOD) [24].

In this context, researchers have developed approaches to automatically generate semantic annotations for parliamentary content of diverse types, such as laws, political programs, and transcriptions compiling interventions of parliament members in plenary meetings. The majority of these approaches have focused on identifying certain parliament entities such as political groups and representative members [23], and a limited number of topics addressed within the input text documents [8]. Moreover, in general, they only support content in a single language [20].

Aiming to address these limitations, we present a first version of an ontology-based approach that makes use of information retrieval techniques and semantic web technologies to annotate and retrieve parliamentary contents in multiple languages. More specifically, our approach is built upon a knowledge base composed of ontologies covering the United Nations taxonomy of sustainable development goals, which are related to a variety of domains, such as education, economy, natural resources, climate change, and social rights. The approach identifies concepts (i.e., classes) and instances (i.e., class individuals) of the above ontologies in input text documents, by means of information retrieval techniques applied to indices created from multilingual labels of the ontology concepts. The extracted concepts do not only represent several levels of thematic annotations, but also allow computing ontology-based similarities that enhance the retrieval of semantically related search and recommendation results beyond keyword-based matching.

As a proof of concept, the approach has been instantiated and preliminary evaluated on a text corpus managed by Parlamento2030, an online platform that monitors parliamentary activity in the Spanish Congress of Deputies. A user study on search tasks shows the benefits of the semantically enhanced annotation and retrieval results provided by our approach.

The reminder of the paper is structured as follows. In Section 2, we survey related work on both retrieval of parliamentary content and semantic retrieval for e-government applications. In Section 3, we introduce the case study addressed with the Parlamento2030 platform. Next, in Sections 4 and 5 we present the proposed approach, distinguishing between its knowledge base building, semantic annotation, and ontology-based retrieval methods. Then, in Section 6, we present preliminary results from a user study on search tasks.

Parlamento2030, https://www.parlamento2030.es/

2 RELATED WORK

Our approach is built upon an ontology-based framework for annotating and retrieving parliamentary content. Hence, in this section, we survey related work by distinguishing between specific approaches to information retrieval in parliamentary contexts and more general approaches to semantic annotation and retrieval in e-government applications.

2.1 Retrieval of Parliamentary Content

In [22], Sánchez-Nielsen et al. conceptualized a smart system where citizens interact with elected representatives in Parliaments, access to Parliament proceedings, and subscribe to new parliamentary contents. Among other issues, the authors proposed the use of semantic technologies and recommender systems to incorporate models for concept and knowledge representation, manual and automatic annotation of textual content from plenary sessions, fragmentation of audiovisual content, provision of customized feeds and information retrieval, and production of automatic reports. Recently, in [23], the authors presented an approach to automatically annotate video transcriptions of the debates occurred in plenary meetings of the Canary Islands Parliament, in Spain. Specifically, the annotations –expressed as RDF semantic data– were associated to concepts belonging to an ad hoc ontology that modeled legislation, representation and parliamentary activity concepts, such as legislatures, legislative proposals, political groups, representative members, sessions, interventions and votes. Exploiting the generated annotations, the authors developed a prototype aimed to retrieve video clips that fulfill a user's specified need on the parliamentary activity, as well as contextual information for content understanding.

In a series of works ([4, 5, 8, 26]), De Campos, Fernández-Luna, Huete and colleagues investigated information retrieval approaches for the Parliament of Andalusia, in Spain. In [8], the authors presented an XML digital library automatically created from the official documents –published in PDF– of the parliament session diaries, i.e., the transcriptions of all the deputies' interventions in plenary and commission sessions. The generated XML files contained simple metadata, such as session date, starting and ending times, agenda points, addressed topics, and vote results. Making use of such XML structures, in [4], the authors developed a system to support the manual annotation of certain parts of the transcriptions with their corresponding segments in videos recorded during the parliament sessions. Exploiting generated annotations, a search engine for structured documents based on Bayesian Networks and Influence Diagrams was tested to retrieve parts of the transcriptions and videos relevant to a given query. More recently, in [26], the authors enhanced their search engine with personalization capabilities. In particular, motivated by the need of maintaining the user's privacy in political contexts, the retrieval model was adapted to a number of content-based stereotype profiles, which were built through several term and category weighting techniques. Lastly, in [5], instead of addressing personalized search, the authors focused on an information filtering task, where the members of parliament receive personalized recommendations of those documents that may be relevant for them. Through empirical comparison, the evaluated information retrieval- and machine learning-based methods did not show significant performance differences.

Differently to the previous works, in [11], Kaptein and Marx attempted to extract high level semantic annotations from parliamentary debates, aiming to summarize and visualize the narrative structure of meetings as tables of topics and intervention graphs. The authors developed a search engine that exploited the generated annotations enabling the provision of entry points to documents (in XML), grouping of search results, and faceted data exploration. In a user study on a large dataset with official transcripts of meetings from the Dutch Parliament, users reported that, in comparison to a standard document retrieval systems, the proposed search engine provided a better overview of the data.

In addition to identifying certain parliament entities (e.g., political groups and representative members), as done in [23], we also aim to extract thematic annotations in several semantic levels, namely topics and domains. Similarly to the approaches presented in [4, 8, 11], ours processes transcriptions stored in text documents. However, instead of generating annotations in XML, our system produces RDF tuples, which could be linked to external semantic web repositories. We note that the RDF annotations of [23] refer to a limited vocabulary of legislation, representation, and parliamentary activities, whereas our annotations correspond to concepts from a variety of domains. The personalization of search results, as proposed by [5, 26], is left as future work.

2.2 Semantic Retrieval in e-Government

The Olimpo system [10] presents one of the first reported approaches to semantic search for an e-government application. In particular, the system applied case-based reasoning and information retrieval techniques to find documents similar to an input list of documents provided by the user through an iterative process. Implemented for searching United Nation (UN) security resolutions, the system made use of the structured representation of documents to identify and extract from the documents a variety of attributes, such as subjects, dates, institution acronyms, country names, number of decisions, and text parts with higher occurrence of indicative expressions of the resolutions. More recently, Liu and Hu [12] also presented an approach to extract metadata from e-government information resources, and exploit it to provide semantic search functionalities. In this case, the identification of attributes was conducted by means of lexical analysis (i.e., part-of-speech labeling), stop words removal, and term frequency-based filtering for the Chinese language. The generated metadata was stored in XML, but the authors proposed its transformation to RDF.

In the previous works, the exploited metadata consisted of a limited set of attributes without semantic structure or relationships between them. Differently, others have proposed the use of ontologies as knowledge representation frameworks whose interlinked concepts conform the fundamental elements of the document annotations used by the search engines. In this context, semantic web technologies, e.g., RDF and OWL, represent a popular trend in the literature.

Amato, Mazzeo and Picariello [1] presented a system that exploits ontologies and NLP techniques to annotate e-government multimedia documents. The ontologies contained both domain

knowledge and lexical vocabularies for Italian and English languages. In the paper, the authors did not provide descriptions of the domains ontologies. They, nonetheless, implemented an information retrieval prototype where a user study was conducted on a small collection of 60 criminal law and juridical documents, showing preliminary accuracy results of the annotation process. In [17], Moreira et al. presented POWER, an ontology of political processes designed to track politicians, political organizations, and elections in social media. The authors also presented EMPOWERED, a framework to populate the POWER ontology with information extracted from various resources. The authors assessed the framework on the Portuguese Government and National Elections Committee websites, retrieving 3.6K politicians, 3K terms related to political institutions, and 74 political associations from mandates taken place since 1976. They did not explain the developed semantic annotation method, and proposed to exploit the annotations for information retrieval tasks –i.e., expert finding and question answering– and to align the ontologies with Linked Open Data repositories, such as DBpedia, YAGO and FOAF. Motivated by the benefits of LOD to support interoperability between European administrations and to improve the information access for citizens across Europe, in [20], Narducci, Palmonary and Semeraro presented CroSeR, a cross-language e-government service retriever for different European languages. The underlying semantic annotation algorithm of the system enriched the short descriptions of the services with labels extracted from Wikipedia concepts related to the services. In particular, it used Explicit Semantic Analysis [9], which disambiguates a word meaning through a semantic similarity with Wikipedia concepts. The authors carried out an empirical evaluation consisting of an information retrieval task on a catalogue of 2.4K services in five different languages –namely Dutch, Belgian, German, Norwegian and Swedish–, and comparing CroSeR with well known semantic annotators, such as Wikipedia Miner [16] and DBpedia Spotlight [15]. More recently, in [14], the authors presented OntoloGov, a system that supports interoperability between e-government repositories, and provides semantic search functionalities. More specifically, the system performed a metadata extraction process, made use of ontology-based contextual user profiles, and applied case based-reasoning to support knowledge retrieval.

As done in [10], we consider content generated by UN, but, instead of focusing on vocabularies associated to its security resolutions, we use its taxonomy of sustainable development goals, covering various domains. Similarly to the approach presented in [12], we make use of natural language processing techniques and resources for the semantic annotation process. Moreover, as done in [1] and [17], we propose ontology-based representations. However, the former did not provide descriptions of its domain ontologies, and the latter used a limited ontology modeling political processes. Lastly, as in [20], supporting multilinguality and generating LOD represent two of the principal requirements for our system. Differently to that work, we left the rigorous evaluation of generic semantic annotation and retrieval methods for the future.

3 CASE STUDY

Our approach is aimed to be integrated into Parlamento2030, an online platform that monitors parliamentary activity in the Spanish Parliament. For such integration, the approach has been implemented and preliminary evaluated on a dataset managed by Parlamento2030. In Section 4, we describe such dataset and the knowledge base we have generated from it. Before that, we introduce the Spanish Parliament and Parlamento2030 tool.

3.1 The Spanish Congress of Deputies

The *Cortes Generales* (lit. General Courts) are the Spanish Parliament. Established and regulated in the Constitution of Spain, they form a bicameral legislative chamber, consisting of the Senate (*Senado*) –i.e., the upper house– and the Congress of Deputies (*Congreso de los Diputados*) –i.e., the lower house.

In the Congress of Deputies, parliamentarians form working groups –called commissions– in different areas of interest (e.g., economy, education, healthcare, agriculture, etc.), and discuss initiatives related to the commissions they belong to. The proposals generated by the commissions are presented and debated in plenary sessions. Certain proposals are then formalized as law projects, which have to be voted and approved by a majority of the deputies for their implementation as laws. The Senate, which represents the territorial regions in Spain and supervises the work done by the Congress of Deputies, does not propose laws, but revises and suggests changes to the law projects provided by the deputies.

As a form of government transparency, both law proposals and Parliament session diaries are available online as HTML and PDF documents. The Parlamento2030 platform crawls, scraps and categorizes such diaries to provide search functionalities, which we describe next.

3.2 The Parlamento2030 platform

Forming part of *Salvador Soler Foundation*, CIECODE, *Centro de Investigación y Estudios sobre Coherencia y Desarrollo* (lit. Center for Research and Studies on Coherence and Development), aims to analyze public policies and private practices of developed countries, inform about their effects on developing countries, and make proposals to move towards a more egalitarian society and fair world.

In particular, CIECODE implements innovative projects on accessing to political information. Among them, Parlamento2030 is an online platform that monitors parliamentary activity in the Spanish Congress of Deputies, aiming to promote an active, informed and demanding citizenship and a responsible political class subject to public scrutiny. It is an adaptation of the TIPI Ciudadano toool and is built upon the CIECODE Political Watch open-source framework.

The Parlamento2030 tool scans all the political activities of the Congress of Deputies by crawling the activity transcriptions publicly available at the congress website, and automatically categorizing the crawled content according to their relationships with 17 priority thematic areas for poverty, social justice and sustainable

Fundación Salvador Soler, https://unmundosalvadorsoler.org
CIECODE, https://www.ciecode.es
TIPI Ciudadano, https://tipiciudadano.es
CIECODE Political Watch, https://github.com/politicalwatch
Congreso de los Diputados, http://www.congreso.es

development. More specifically, it annotates the parliamentary content by keyword matching using a vocabulary with more than 3K Spanish terms provided by expert individuals and organizations in each of the areas.

The platform has a search engine with which the user can refine information filtering queries based on multiple criteria, such as author, date, theme and keyword (see Figure 1). It also offers a personalized system of alerts that allows a user to be up to date on the political news of her topics of interest. Parlamento2030's code and data can be freely accessed and downloaded.

Our approach is built and tested on a Parlamento2030 dataset composed of the above mentioned 17 thematic vocabularies and a collection of structured transcriptions of parliamentary activity. In the next section, we describe the dataset, as well as the semantic knowledge base and annotations our approach generates from such dataset.

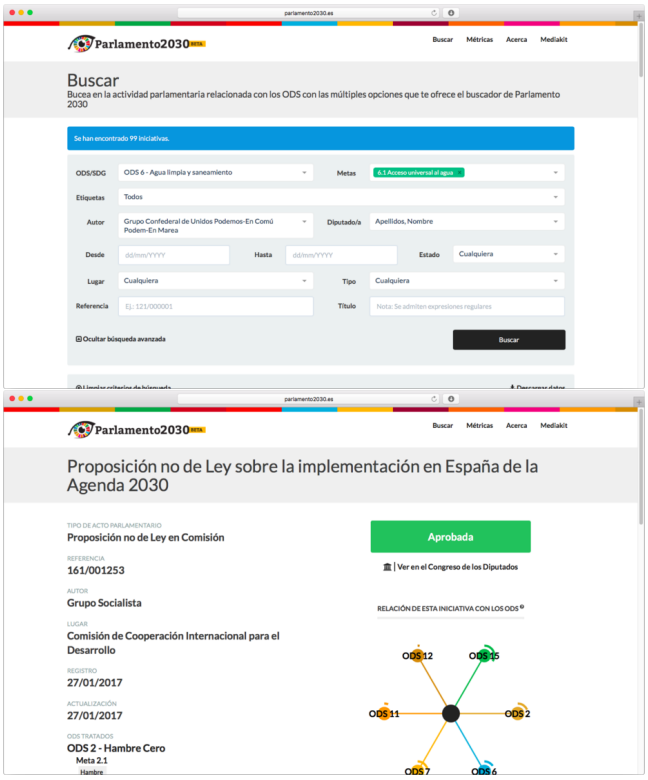

Figure 1: Screenshots of Parlamento2030 search form and result pages.

4 KNOWLEDGE BASE AND SEMANTIC ANNOTATIONS

In this section, we present the knowledge base building and semantic annotation methods we developed. Before, we introduce the original Parlamento2030 dataset on which we run the above methods.

Table 1: Examples of concepts and keywords of the dataset.

Domain	Concept	Keywords
Poverty	Unemployment and vulnerability	unemployment and vulnerability, unemployed and vulnerability, vulnerable unemployed
	Global poverty rate	global poverty rate, worldwide poverty rate, international poverty rate
Education	Reading and math skills	reading and math skills, reading and math proficiency
	STEM careers and women	STEM careers and women, STEM degrees and women, STEM and women
	ANECA	ANECA, National Agency for Quality Assessment and Accreditation

4.1 Original dataset

The Parlamento2030 system performs keyword matching heuristics based on regular expressions to identify the topics of the textual content periodically published in the website of the Spanish Congress of Deputies. The regular expressions were manually generated and curated by the experts that built the domain vocabularies.

The topics correspond to the 17 Sustainable Development Goals (SDGs) established by United Nations: no poverty (G1), zero hunger (G2), good health and well-being (G3), quality education (G4), gender equality (G5), clean water and sanitation (G6), affordable and clean energy (G7), decent work and economic growth (G8), industry, innovation and infrastructure (G9), reduced inequality (G10), sustainable cities and communities (G11), responsible consumption and production (G12), climate action (G13), life below water (G14), life on land (G15), peace and justice strong institutions (G16), and cooperation and alliances (G17). These goals are aligned with Agenda 2030, the global action plan and commitment to eradicate poverty and achieve sustainable development by 2030 worldwide.

For each of these goals, in cooperation with CIECODE members, experts on different domains elaborated a list of targets, i.e., issues of interest and relevant problems to be addressed by governments and public administrations. Each target was split into concepts, which are represented by a set of keywords in Spanish. More specifically, the Parlamento2030 system uses a vocabulary of more than 3K concepts, where each concept has associated a number of regular expressions that generate keywords according to singular and plural forms, morphological deviations (e.g., masculine and feminine forms, articles and prepositions linked to nouns), adjectives and adverbs, abbreviations and acronyms. For example, for the Quality education ("Educación de calidad") target, the concept "ayudas oficiales al desarrollo en educación" (lit. official development grants in education) has keywords such as "ayudas oficiales al desarrollo en educación", "ayuda oficial al desarrollo en educación", "ayuda oficial al desarrollo educativo," and "AOD en educación." For a better comprehension, Table 1 shows some examples of concepts and keywords of the dataset translated into English.

https://www.un.org/sustainabledevelopment

4.2 Knowledge base

Our approach uses a semantic knowledge base built from the Parlamento2030 dataset. In particular, the knowledge base integrates domain ontologies that follow the schema shown in Figure 2. The schema has 3 main classes, namely `Goal`, `Target` and `Concept`.

The `Goal` subclasses and individuals correspond to the 17 United Nations SDGs. The `Target` subclasses and individuals are associated to domain targets established in Parlamento2030 for each SDG. Lastly, the `Concept` subclasses and individuals correspond to topics (concepts) and keywords assigned to each target in Parlamento2030. A `Concept` subclass may have subclasses, forming partial taxonomies that represent the knowledge of the covered domains; see in Figure 3 the partial view of the ontology with some concepts on the Education domain. Moreover, through the `rdfs:label` property, the `Concept` individuals have one or more terms in different languages. These terms correspond to the keywords exploited in Parlamento2030; see examples in Table 1.

Each class, individual and property of the ontologies is identified by a language- and semantics-independent URI. For instance, `tipi:target#Target_4_6` is the URI of target *T4.6 Literacy*, related to goal *G4 Quality education*, whose URI is `tipi:goal#Goal_4`. To support multiliguality, every class and individual can have multiple String labels through the `rdfs:label` property. Hence, for example, the individual of T4.6 class may have label values such as "literacy" and "alfabetización" for English and Spanish, respectively.

Lastly, the individuals of `Goal`, `Target` and `Concept` subclasses are linked by means of the `tipi:hasGoal` and `tipi:hasTarget` properties. Following the above mentioned examples, the individual `tipi:concept#Concept_942` (i.e., digital literacy) has target `tipi:target#Target_4_6` (i.e., literacy), which is related to goal `tipi:goal#Goal_4` (i.e., quality education). The individuals of `Concept` subclasses are created from the singular and acronym forms of every list of keywords available in Parlamento2030.

The building of classes and individuals was conducted automatically. The organization of the `Concept` subclasses (including the creation of new inner subclasses), in contrast, was done manually by experts, using the graphical user interface of the Protégé tool [19]. The whole knowledge base is composed of 169 targets, 3.6K concepts and 10.3K terms.

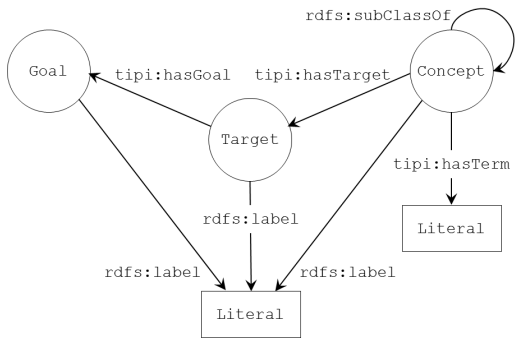

Figure 2: Ontology schema diagram.

The prefix `tipi:` has to be replaced by `http://ciecode.es/tipi/ontology/`

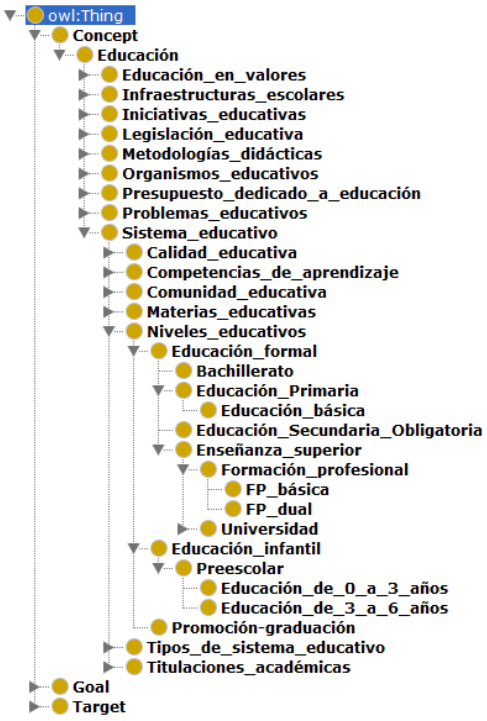

Figure 3: Partial view of the Education domain ontology, visualized by means of its `rdfs:label` values in Spanish.

4.3 Semantic Annotations

Figure 4 shows the architecture of the implemented semantic annotation method, which makes use of two indices: an index for the input parliamentary documents and an index for our ontological knowledge base.

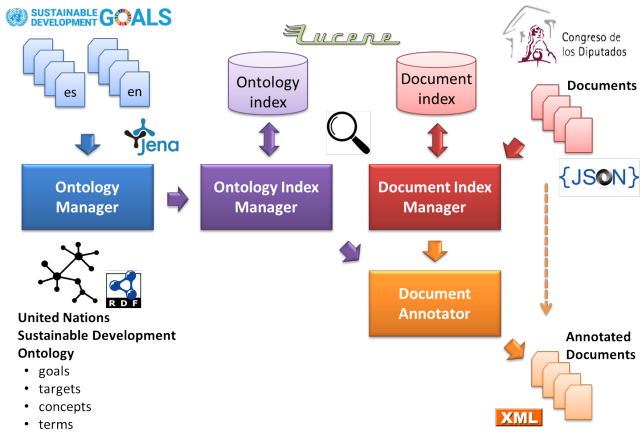

Figure 4: Semantic annotation framework.

Parlamento2030 crawls and scraps the documents published in the Spanish Congress of Deputies website, providing JSON files with the content generated by parliamentarians –i.e., debates and law proposals– in plain text. From these files, our method builds

an index (on the top right of the figure) using the Apache Lucene library. This index is used by a document index manager to retrieve a ranking of documents for a given keyword-based query. The documents are indexed by title, content and language separately. Indexed terms are weighted with TF-IDF values computed on the whole corpus of documents. For a query, the result list are limited to the 100 documents and the ranking scores of the documents are normalized to sum 1.

Our method builds a second Lucene index (in the middle of the figure) to efficiently access to the information available in the domain ontologies. In this case, each ontology entity –i.e., class or individual– is indexed by domain (goal), target, concept name, concept keyword, RDF label, and language. The index also stores the URI of each entity. Similarly to the document index, there is an ontology index manager that retrieves a ranked list of entities for a given keyword-based query.

The ontology index manager uses an ontology manager (on the left of the figure) to efficiently obtain all the ontology entities and their data. The latter was indeed the component that created the UN sustainable development ontologies from the Parlamento2030 dataset, and stored them in a RDF repository using the Apache Jena framework.

Once the document and ontology indices are built, a document annotation component (on the bottom right of the figure) generates XML files with the semantic annotations of the input documents. The annotation process is as follows (for a given language). The annotator launches on the document index several queries $q_{e,k}$ for each ontology entity e. The queries are composed of the k terms associated to the entity name, keywords and labels, and thus generate several ranked lists of documents that contain the terms associated to e. Next, the obtained ranking scores $s_{e,k}(d)$ of the documents are aggregated into weights $w_e(d)$, which measure the relevance of each ontology entity e for each document d:

$$w_e(d) = \sum_{k \in \text{terms}(d)} s_{e,k}(d)$$

These weights are normalized by document, and are considered as semantic annotations of the documents at entity (concept) level.

Through the relationships of the ontologies, we can compute weights $w_e(t)$ and $w_t(g)$, respectively establishing the relevance of entity e for target t and the relevance of target t for goal g. As a first implementation, we set $w_e(t) = 1$ if entity e (or the class of e) is related to target t by means of the `tipi:hasTarget` property, and $w_t(g) = \sum_{t \to g} w_e(t)$ for those targets t that are related to goal g by means of the `tipi:hasGoal` property.

Using the weights $w_e(t)$ and $w_t(g)$, our method computes weights $w_t(d)$ and $w_g(d)$ that will be considered as semantic annotations of the documents at target and goal levels:

$$w_t(d) = \sum_{e \to t} w_e(t) \cdot w_e(d)$$

$$w_g(d) = \sum_{t \to g} w_t(g) \cdot w_t(d)$$

Apache Lucene, https://lucene.apache.org
Apache Jena, https://jena.apache.org/

As an illustrative example, the next XML fragment shows some semantic annotations generated for an input document about actions planned by the Hydrographic Confederation, formed by Spanish-Portuguese committees for the management and care of the Duero river. From keywords such as “embalses” and “presas” (lit. reservoirs and dams), our method extracts “Recurso hídrico” (lit. water resource), “Gestión integrada del agua” (lit. integrated water management), and “Agua limpia y saneamiento” (lit. clean water and sanitation), as semantic annotations at concept, target and goal levels.

```
<doc_annotation_list>
  <doc_id>81c8bb406e52a1e8009619ee0632705c7f366ca6</doc_id>
  <doc_title>Actuaciones previstas por la Confederacion
             Hidrografica ...</doc_title>
  <term_annotation_list>
    <term_annotation>
      <term>embalses</term>
      <weight>0.5</weight>
    </term_annotation>
    <term_annotation>
      <term>presas</term>
      <weight>0.5</weight>
    </term_annotation>
  </term_annotation_list>
  <concept_annotation_list>
    <concept_annotation>
      <concept_uri>tipi_concept:Concept_147</concept_uri>
      <concept_name>Recurso hidrico</concept_name>
      <weight>1.0</weight>
     </concept_annotation>
  </concept_annotation_list>
  <target_annotation_list>
    <target_annotation>
      <target_uri>tipi_target:Target_6_5</target_uri>
      <target_name>Gestion integrada del agua</target_name>
      <weight>1.0</weight>
    </target_annotation>
  </target_annotation_list>
  <goal_annotation_list>
    <goal_annotation>
      <goal_uri>tipi_goal:Goal_6</goal_uri>
      <goal_name>Agua limpia y saneamiento</goal_name>
      <weight>1.0</weight>
    </goal_annotation>
  </goal_annotation_list>
</doc_annotation_list>
```

We note that, in addition to the explained thematic annotations, our method also extracts annotations related to named entities, such as proper nouns of people (e.g., parliamentarians), organizations (e.g., government agencies, political parties) and places.

5 ONTOLOGY-BASED RETRIEVAL

In this section, we present the developed semantic search method. Before, we describe the ontology-based document representation model used by the method.

5.1 Document Representation

As we will explain in the next subsection, our search method is built upon the well known Vector Space Model, VSM [21]. Hence, to describe content documents, i.e., parliamentary debate transcripts and law proposals, we make use of a vector representation. However, instead of the classical information retrieval representation based on terms, we propose to use semantically related concepts as units of information. The explicit relationships between concepts allow

computing semantic relatedness, and address the VSM limitation of term vector pairwise orthogonality (i.e., linear independence), an unrealistic assumption where any pair of terms do not relate to each other [25, 27]. Using concepts also allows avoiding ambiguities of polysemic terms and applying semantic inference through the concepts relationships at retrieval stage [7].

Formally, let $O = \{\mathcal{E}, \mathcal{R}\}$ be an ontology composed of entities $\mathcal{E} = \{C, \mathcal{I}\}$ that can be classes C or individuals $\mathcal{I}$, and relationships $\mathcal{R} : \mathcal{E} \rightarrow \mathcal{E} \cup \mathcal{L}$ that link pairs of entities or entities with literals (e.g., numeric or string). Let $e_1, e_2, \ldots, e_N \in \mathcal{E}$ belong to a N-dimensional Euclidean space. A document d is represented as a vector $\mathbf{d} = (w_{d,1}, \ldots, w_{d,N}) \in \mathbb{R}^N$, where the weight $w_{d,n}$ corresponds to the semantic annotation score computed as explained in Section 4.3.

In this paper, the considered relationships are the properties of our knowledge base (Section 4.2), and the entities of a given document correspond to its semantic annotations extracted by our approach (Section 4.3). Next, we explain how the relationships are used by the proposed search method.

5.2 Search Method

Our search method is based on the Generalized Vector Space Model, GVSM [27] proposed by Tsatsaronis and Panagiotopoulou [25], which incorporates a semantic relatedness (*SR*) measure into the term-based vector similarity of the VSM.

Let $\mathbf{d} = (w_{d,1}, \ldots, w_{d,N}) \in \mathbb{R}^N$ and $\mathbf{q} = (w_{q,1}, \ldots, w_{q,N}) \in \mathbb{R}^N$ be the weight vectors of document d and query q, respectively. The value q_n is set to 1 if entity e_n appears in the input query q, and to 0 otherwise. Considering a typical keyword-based search scenario, we map the keywords of the user's query to entities. This is done by the OntologyManager explained in Section 4.3 (Figure 4), by exact matching of the keywords with the entity terms (see Figure 2).

We define the similarity between d and q as follows:

$$sim(d, q) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{d,i} \cdot w_{q,j} \cdot SR(e_i, e_j)}{\sqrt{\sum_{i=1}^{N} w_{d,i}^2} \cdot \sqrt{\sum_{i=1}^{N} w_{q,i}^2}}$$

To implement this similarity, we propose to use the semantic relatedness metric proposed by Corella and Castells in [3], which we explain next. Being $\mathcal{E}$ the set of existing ontology entities, the semantic relatedness *SR* between two entities $e_1, e_2 \in \mathcal{E}$ is measured in terms of their *distance* in the ontology hierarchy as follows. Let e_0 be the closest ancestor (super class) to e_1 and e_2 in the ontology hierarchy, and let $h_1 = 1 + dist(e_1, e_0)$ (and $h_2 = 1 + dist(e_2, e_0)$) be 1 plus the number of levels between e_1 (and e_2) and e_0 in the ontology hierarchy. We define the semantic relatedness between entities e_1 and e_2 as:

$$SR(e_1, e_2) = \left(1 - \frac{\alpha}{h(O)} \cdot \frac{|h_1 - h_2|}{h_1 + h_2}\right) \cdot \frac{1}{\min(h_1, h_2)} \cdot \left(1 - \frac{\max(h_1, h_2) - 1}{h(O)}\right)$$

The formula has three factors:

- The first factor measures the distance between e_1 and e_2 as a proportion of the depth $h(O)$ of the ontology hierarchy. The $\alpha \in [0, 1]$ parameter allows ensuring a minimum non-zero *SR* value, even for the most dissimilar entities, so that *SR* ranges in certain interval $[a, 1]$ with $a > 0$. The authors set $\alpha = 0.8$ in their experiments.
- The second factor increases *SR* proportionally to the closeness of the two entities to their common ancestor e_0. Let us consider cases where $h_1 = h_2$, for which the first factor of *SR* equals 1. The second factor allows establishing a higher *SR* value for the case $h_1 = h_2 = x$ than the case $h_1 = h_2 = y$ if $x < y$.
- The third factor decreases *SR* when e_1 and e_2 are in the same branch of the ontology class hierarchy; that is, they are not sibling classes or instances.

An ontology hierarchical level can be either a `rdfs:subclassOf` relationship between two classes, or a `rdf:type` relationship between an individual and its class.

6 EXPERIMENTS

To show the semantic capabilities of the proposed search method, we tested 3 different variants of the query-document similarity given in Section 5.2. The first similarity, sim_{key} is implemented with $SR(e_1, e_2) = 1, \forall e_1, e_2 \in \mathcal{E}$, i.e., it does not exploit ontological relationships between entities, and thus it is equivalent to the classical VSM. This version is limited to matching of query and document key terms. The second similarity, sim_{sem_all}, considers the $SR(e_1, e_2)$ value of every entity in the document and query vectors. It thus exploits any ontological relationship underlying the query and document. This version favors the retrieval of a large number of documents belonging to the domain of the query. Lastly, the third similarity, sim_{sem_max}, only applies the maximum $SR(e_1, e_2)$ value for the query and document entities. This version also extends the keyword matching, but focuses on the strongest relationships according to the closeness of the entities in the domain ontology.

For a better comprehension of the experimental results, we present next a real searching example for the query "educación especial" (lit. especial education).

Running this query on our Parlamento2030 corpus, the sim_{key} method only retrieved 3 documents $-d_1, d_2, d_3-$ having the term "educación especial." These documents had been annotated with `tipi:concept#Concept_929` (i.e., *especial education* concept), which, in the ontology, has associated terms such as *specific needs education*, *aided education*, and *exceptional education*. The method omitted these terms since it is based on the matching of the query keywords.

The sim_{sem_all} method, in contrast, was able to retrieve a total of 45 documents, all of them belonging to the education domain. As for the first method, the documents d_1, d_2 and d_3 were retrieved in the top 3 ranking positions. After them, in the results list, there were 4 documents about *access to education* and *school bullying*. In the ontology, these two concepts are sibling of the *especial education* concept. In particular, they are subclasses of the *educational problems* concept. Next, in the ranking, there were 7 documents $-d_5, \ldots, d_8-$ highly related to the query. These documents were annotated with the `tipi:concept#Concept_937` (i.e., *Specific Needs of Educational Support, SNES*). This concept is a direct child (subclass) of the *especial education* concept in the ontology. Its relationship with *especial education* is stronger than *access to education* and *school bullying*. However, its associated *SR* values have less influence on the ranking scores than the TF-IDF values of the document entities.

Table 2: Number of indexed documents with parsed content.

Document type	Num.
Non-law proposals in plenary session	40
Law proposals by parliamentary groups	13
Law proposals by deputies	2
Law proposals by autonomous cities and communities	10
Popular legislative initiatives	4
Proposals to reform the congress rules	3

Table 3: Average precision $P@N$ values for each method.

Method	$P@5$	$P@10$	$P@15$	$P@20$
sim_{key}	0.633	0.483	0.422	0.358
sim_{sem_all}	**0.733**	0.550	0.500	0.492
sim_{sem_max}	**0.733**	**0.683**	**0.656**	**0.600**

Boosting the semantic relatedness in the ranking process, the sim_{sem_max} method generated a result list where documents $d_5, \ldots, d_8$ appear after the first 3 positions.

Testing the methods on the Parlamento2030 corpus, we found other similar examples. We do not present them due to space limitations. Moreover, aiming to show a more rigorous evaluation of the methods, we performed a preliminary experiment, described next.

The Parlamento2030 corpus had 3534 documents. From them, only 72 had textual content parsed. Table 2 shows the types of these documents. The remainder documents had links to associated PDF files from the Congress of Deputies. We indexed and annotated all the documents analyzing their titles and their content if available. A total of 435 documents were annotated with 642 concept annotations, meaning a coverage of 12.3% of the corpus. The Parlamento2030 system has topic tags for 427 documents obtained through their language-dependent regular expressions.

The obtained Parlamento2030 dataset did not have sets of queries and relevance judgements, so we opted to conduct a user study. Three experts participated in the study. After revising the document collection, each of them stated 5 queries. Then, they were requested to assess the 20-top results provided by the 3 search methods for the 15 queries, stating whether each retrieved document was *non relevant*, *relevant* or *highly relevant* for the corresponding query. The Fleiss' kappa inter-rater correlation coefficient was $\kappa = 0.984$, meaning an almost perfect agreement between the assessors' judgements. The experts also evaluated the correctness of the annotations of the documents, measuring an accuracy of 98.7% for the semantic annotation process.

Table 3 reports the average precision values $P@N$ of the three methods for the top $N = 5, 10, 15, 20$ results. *Relevant* and *highly relevant* assessments were considered as positive judgements. The reported values show that the proposed semantic methods outperform the VSM-based sim_{key} method. Moreover, they show that limiting the ontological expansion of matched concepts by the sim_{sem_max} improves the accuracy achieved by sim_{sem_all}.

7 CONCLUSIONS

In this paper, we have presented a novel approach to ontology-based annotation and retrieval of parliamentary content. The approach was built upon domain ontologies that cover a large number of topics related to the United Nations taxonomy of sustainable development goals. As a proof of concept, the approach was instantiated and preliminary evaluated on a corpus extracted from the Spanish Congress of Deputies, and used by the Parlamento2030 platform. The approach is in a preliminary stage. Next, we describe several future research lines.

The ontological schema of the approach is composed of three main classes, namely `Goal`, `Target` and `Concept`, and three properties to relate the class individuals, namely `hasGoal`, `hasTarget` and `subClassOf` (i.e., `subConceptOf`). As shown in the paper, exploiting these properties leads to semantically enhanced search results. However, more specific properties between particular entities could be considered as well. For instance, we may have inter-domain properties that relate concepts –such as `Job security` (from G1: No poverty) `equivalentTo` `Precarious work` (G8: Decent work economic growth)–, and targets –such as `Promoting renewable energy` (G7: Affordable and clean energy) `impactsOn` `Decreasing pollution` (G13: Climate change), which `impactsOn` `Disease prevention` (G3: Good health well being). Exploiting these and other types of relationships could enrich the semantic inference capabilities of the retrieval method, which would be able to better find related documents. Properties and relationships may be defined manually or extracted automatically by mining external knowledge bases.

Moreover, our semantic annotation method is able to extract named entities, such as proper nouns of people, organizations and places. We may extend the ontology to model these issues. For instance, we may have classes and properties describing and relating cities and administrative divisions. Then, we could exploit this information to find government initiatives according to particular demographic, sociocultural, political and economic attributes of target locations, or in terms of geographic relationships between such locations, e.g., educational initiatives proposed for certain surrounding area.

Modeling and extracting time aspects of parliamentary content are left as future work. Temporal traceability of addressed goals and targets, and time-based government initiative similarities are examples of interesting functionalities for a parliamentary information retrieval system. We also envision the extension of the annotation method to work at argumentation level [2, 6, 13, 18], aiming to automatically extract argumentative structures from input texts. In particular, we propose to develop a method able to identify arguments, as well as components from them, such as facts, statements and predictions. The extracted arguments would serve as summaries of existing debates and proposals, and may be used as instruments to measure and monitor ideological and activity dynamics in the Parliament.

Our approach supports multilinguiality, but we only implemented it with a vocabulary in Spanish. A further step would be to translate the generated ontology entities into other languages, starting with English. For both the ontologies and semantic annotations, we plan to link their entities to external knowledge bases, following the

Linked Open Data initiative. In this context, we would make all the generated resources publicly available as RDF repositories.

The proposed information retrieval method, which makes use of a novel semantic relatedness metric between ontology entities, could be adapted to provide personalized search and recommendation. For such purpose, as done in previous work [5, 22, 26], we will have to model individual or stereotype-based user profiles, according to privacy and application issues.

Lastly, an exhaustive evaluation of the proposed approach has to be conducted. In this sense, we plan to perform both offline experiments and online user studies. For the latter case, the approach will be integrated into the Parlamento2030 platform.

ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Science and Innovation (PID2019-108965GB-I00). The authors acknowledge CIECODE for providing the dataset used in this work, and thank its director –Javier Pérez– and members –Pablo Martín, Belén Agüero and Irene Matín– for their help and support in the project. They also thank Alejandro Bellogín for his help on the regular expression processing.

REFERENCES

[1] Flora Amato, Antonino Mazzeo, Vincenzo Moscato, and Antonio Picariello. 2009. A system for semantic retrieval and long-term preservation of multimedia documents in the e-government domain. *International Journal of Web and Grid Services* 5, 4 (2009), 323–338.

[2] Claire Cardie, Cynthia R Farina, Matt Rawding, and Adil Aijaz. 2008. An eRulemaking corpus: Identifying substantive issues in public comments. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

[3] Miguel Ángel Corella and Pablo Castells. 2006. A heuristic approach to semantic web services classification. In *Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 598–605.

[4] Luis M De Campos, Juan M Fernández-Luna, Juan F Huete, and Carlos J Martín-Dancausa. 2008. An integrated system for accessing the digital library of the Parliament of Andalusia: Segmentation, annotation and retrieval of transcriptions and videos. In *Proceedings of the 8th International Workshop on Pattern Recognition in Information Systems - Volume 1*. SciTePress, 38–47.

[5] Luis M de Campos, Juan M Fernández-Luna, Juan F Huete, and Luis Redondo-Expósito. 2017. Comparing machine learning and information retrieval-based approaches for filtering documents in a parliamentary setting. In *Proceedings of the 11th International Conference on Scalable Uncertainty Management*. Springer, 64–77.

[6] Vlad Eidelman and Brian Grom. 2019. Argument Identification in Public Comments from eRulemaking. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law*. 199–203.

[7] Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, and Enrico Motta. 2011. Semantically enhanced Information Retrieval: An ontology-based approach. *Journal of Web Semantics* 9, 4 (2011), 434–452.

[8] Juan M Fernández-Luna, Juan F Huete, Manuel Gómez, and Carlos J Martín-Dancausa. 2008. Development of the XML digital library from the parliament of Andalucía for intelligent structured retrieval. In *Proceedings of the 17th International Symposium on Methodologies for Intelligent Systems*. Springer, 417–423.

[9] Evgeniy Gabrilovich and Shaul Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* 34 (2009), 443–498.

[10] Hugo C Hoeschl, Tânia Cristina D Bueno, Andre Bortolon, Eduardo S Mattos, Marcelo S Ribeiro, Irineu Theiss, and Ricardo Miranda Barcia. 2004. An intelligent search engine for electronic government applications for the resolutions of the United Nations Security Council. In *Building the E-Service Society*. Springer, 23–41.

[11] Rianne Kaptein and Maarten Marx. 2010. Focused retrieval and result aggregation with political data. *Information retrieval* 13, 5 (2010), 412–433.

[12] Xiaoxing Liu and Changxia Hu. 2012. Research and design on e-government information retrieval model. *Procedia Engineering* 29 (2012), 3170–3174.

[13] Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management* 56, 6 (2019), 102055.

[14] Antonio Martín and Carlos León. 2015. Semantic framework for an efficient information retrieval in the e-government repositories. In *Handbook of Research on Democratic Strategies and Citizen-Centered E-Government Services*. IGI Global, 192–213.

[15] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*. 1–8.

[16] David Milne and Ian H Witten. 2008. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. 509–518.

[17] Silvio Moreira, David Batista, Paula Carvalho, Francisco M Couto, and Mário J Silva. 2011. POWER - Politics Ontology for Web Entity Retrieval. In *Proceedings of the 23rd International Conference on Advanced Information Systems Engineering*. Springer, 489–500.

[18] Gaku Morio. 2018. Annotating Online Civic Discussion Threads for Argument Mining. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 546–553.

[19] Mark A Musen. 2015. The protégé project: A look back and a look forward. *AI Matters* 1, 4 (2015), 4–12.

[20] Fedelucio Narducci, Matteo Palmonari, and Giovanni Semeraro. 2013. Cross-language semantic retrieval and linking of e-gov services. In *Proceedings of the 12th International Semantic Web Conference*. Springer, 130–145.

[21] Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18, 11 (1975), 613–620.

[22] Elena Sánchez-Nielsen and Francisco Chávez-Gutiérrez. 2008. Personalized and on-demand retrieval of parliamentary proceedings with social feedback on elected representatives. In *Proceedings of the 21st Annual Conference on Legal Knowledge and Information Systems*. IOS Press, 53–62.

[23] Elena Sánchez-Nielsen, Francisco Chávez-Gutiérrez, and Javier Lorenzo-Navarro. 2019. A semantic parliamentary multimedia approach for retrieval of video clips with content understanding. *Multimedia Systems* 25, 4 (2019), 337–354.

[24] Nigel Shadbolt, Kieron O'Hara, Tim Berners-Lee, Nicholas Gibbins, Hugh Glaser, Wendy Hall, et al. 2012. Linked open government data: Lessons from data.gov.uk. *IEEE Intelligent Systems* 27, 3 (2012), 16–24.

[25] George Tsatsaronis and Vicky Panagiotopoulou. 2009. A generalized vector space model for text retrieval based on semantic relatedness. In *Proceedings of the Student Research Workshop at EACL 2009*. 70–78.

[26] Eduardo Vicente-López, Luis M de Campos, Juan M Fernández-Luna, and Juan F Huete. 2016. Use of textual and conceptual profiles for personalized retrieval of political documents. *Knowledge-Based Systems* 112 (2016), 127–141.

[27] SK Michael Wong, Wojciech Ziarko, and Patrick CN Wong. 1985. Generalized vector spaces model in information retrieval. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 18–25.