Topic-OPA: A Topic Ontology for Modeling Topics of Old Press Articles

Mirna El Ghosh, Cecilia Zanni-Merk, Nicolas Delestre, Jean-Philippe Kotowicz and Habib Abdulrab
Normandie Université, INSA Rouen, LITIS, 76000 Rouen, France

Keywords: Topic Ontologies, Topic Modeling, Open Knowledge Graphs, SPARQL.

Abstract: Topic ontologies are recently gaining much importance in several domains. Their purpose is to identify the themes necessary to describe the knowledge structure of an application domain. Meanwhile, their development from scratch is hard and time consuming task. This paper discusses the development a topic-specific ontology, named Topic-OPA, for modeling topics of old press articles. Topic-OPA is extracted from the open knowledge graph Wikidata by the application of a SPARQL-based fully automatic approach. The development process of Topic-OPA depends mainly on a set of disambiguated named entities representing the articles. Each named entity is unambiguously identified by a Wikidata URI. In contrast to existent topic ontologies, which are limited to taxonomies, the structure of Topic-OPA is composed of hierarchical and non-hierarchical schemes. The domain application of this work is the old french newspaper *Le Matin*. Finally, an evaluation process is performed to assess the structure quality of Topic-OPA.

1 INTRODUCTION

Topic ontologies are recently gaining significant attention in the ontology engineering community. They are being increasingly used in various domains such as semantic matching (Tang et al., 2009), topic labeling (Allahyari and Kochut, 2017), topic modeling (Sleeman et al., 2018) and evaluating topical search (Maguitman et al., 2010). The purpose of topic ontologies is to represent the main "themes" of a given application domain. The most commonly known approaches for building topic models are the keyword-based construction approaches which are based mainly on text mining and information retrieval techniques (Maguitman et al., 2010). Examples of these approaches are the statistical approaches such as probabilistic latent semantic analysis (pLSA) (Hofmann, 1999) and latent Dirichlet allocation (LDA) (Blei et al., 2003). These approaches depend on the textual content of the articles and consider it as a mixture of topics. Their main drawback is that they risk to retrieve specific topics. Although, it is hard and time consuming to construct an ontology from a large corpus of documents (Maguitman et al., 2010).

The works we are presenting in this article are part of the ASTURIAS[1] project. The main goal of this project is to thematically organize a collection of old press articles with a set of topics (e.g. Politics, Art, Sport, Science, etc.).

In fact, one of the specific features of old press is that it does not offer thematic entries: articles appear and follow one another without a thematic logic. Under these conditions, it remains a tedious task to query sources that report the same events from different points of view in different areas of the newspaper. The scientific challenge is to propose robust approaches for the analysis of texts that are noisy due to the imperfect process of automatic transcription of images into electronic texts. These approaches need also to be multi-thematic, and robust to linguistic evolution over the centuries. The ambition the ASTURIAS project (whose workflow appears in Figure 1) is to study the digitization process from end to end of the processing chain: WP1- from newspaper images, automatically analyze sections, articles and texts; WP2- extract named entities from these elements WP3- Topic labeling and hyperlinking the articles based on the analysis made in 1 and the named entities extracted in WP2.

This article will present our results on building a topic model for WP3 of the ASTURIAS project. In this context, a fundamental hypothesis is that articles are represented by a set of "not ambiguous" named entities (e.g. *person*, *organization*, *product* and *location*) extracted from open data sources (coming from

[1] Analyse STructURelle et Indexation sémantique d'ArticleS de presse - Structural Analysis and Semantic Indexing of Newspaper Articles

El Ghosh, M., Zanni-Merk, C., Delestre, N., Kotowicz, J. and Abdulrab, H.
Topic-OPA: A Topic Ontology for Modeling Topics of Old Press Articles.
DOI: 10.5220/0010147202750282
In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 2: KEOD*, pages 275-282
ISBN: 978-989-758-474-9

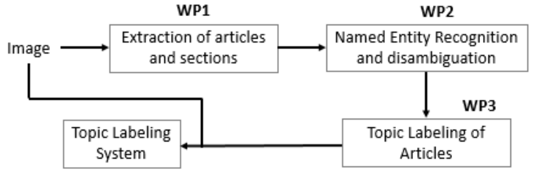

WP1
WP2
WP3
Image
Extraction of articles and sections
Named Entity Recognition and disambiguation
Topic Labeling of Articles
Topic Labeling System

Figure 1: The pipeline of the project ASTURIAS.

WP2 of the project). Therefore, our research problem can be defined as follows: Given a corpus of old press articles *A* represented by a set of named entities *N*, that are collected from *A* and identified by a set of *URIs*, a topic model is required for modeling the topics that represent *A*. The topic model will be used for topic labeling the old press articles. From this perspective, we propose a SPARQL-based approach, relying mainly on the set of the disambiguated named entities, for building the topic model. In this regard, open knowledge graphs, such as Wikidata, are considered. The main goal of this paper is to discuss the development process of a topic-specific ontology, named Topic-OPA, by the application of a SPARQL-based fully automatic approach. Topic-OPA is derived from the open knowledge graph Wikidata based on a set of "not ambiguous" named entities representing the articles. A case study is demonstrated for building Topic-OPA from the articles of *Le Matin*[2], an old french newspaper first published in 1884 and discontinued in 1944. Finally, Topic-OPA is evaluated by the application of a structure-based evaluation approach. The rest of the paper is organized as follows: Section 2 presents the main related works of this study. In section 3, we discuss the development process of Topic-OPA. Section 4 presents a case study in the context of *Le Matin*. In section 5, we evaluate Topic-OPA. Finally section 6 concludes the paper.

2 RELATED WORKS

In this section, topic ontologies and ontology engineering approaches are introduced as the main related foundations for our study.

2.1 Topic Ontologies

Topic ontologies are considered as special type of ontologies. Their purpose is to identify the "themes" necessary to describe the knowledge structure of an application domain (Zhao and Meersman, 2005). A topic ontology is represented as a set of topics that are interconnected using semantic relations. Two main types of topic ontologies are defined: *simple*, and *general* (Maguitman et al., 2010). The simple topic ontologies are composed of topics linked by hierarchical relations. Meanwhile, in general topic ontologies, *transverse* relations are included to link different topics in a non-hierarchical scheme. For representing general topic ontologies, the following components are commonly defined:

- *Topics:* concepts of the topic ontology (e.g. Sport, Art, Politics).
- *Predicates:* types of relationships defining the semantic relations which can be established between ontology concepts. Multiple predicates are defined in general topic ontologies: hierarchical (e.g. *subClassOf*) and non-hierarchical (e.g. *studied by*, *part of*, etc.)
- *Relationships:* concrete links among ontology concepts which will be used to characterize paths in graphs. They are distinguished according to their predicate and the couple of elements they link. They can be represented as a triplet (s,p,o) where s the subject, o the object and p the predicate that links s and o (e.g. Literature *subClassOf* Art, Art *part of* Culture).

2.1.1 KB-LDA Topic Model

For topic labeling purposes, the topic model KB-LDA (Allahyari and Kochut, 2017) is developed based on combining topic models with ontological concepts in a single framework. KB-LDA used the semantic knowledge graph of concepts in an ontology (e.g. DBpedia) and their diverse relationships with unsupervised probabilistic topic models for generating automatic topic labels. The topic labeling process is performed based on the semantic similarity between the entities included in text documents and a suitable portion of the ontology. For this purpose a semantic graph is constructed from the concepts of the ontology and their classification hierarchy as labels for topics.

2.1.2 IPCC Topic Model

For topic modeling purposes, IPCC (Sleeman et al., 2018) is a domain-specific topic ontology used for grounding a topic model in the domain of climate research. The topic ontology is "seeded" with predefined key word phrase concepts which are obtained from domain-specific sources such as domain experts, and by data mining semi-structured sources. Natural Language Processing techniques have been used to extract the meaningful key word phrase concepts

[2] https://gallica.bnf.fr/ark:/12148/cb328123058/date, last visited on April 8 2020

from these sources. While, the topic modeling process is applied on textual resources such as, reports and research papers, the ontology concepts are used for weighting concepts founded in these resources. Furthermore, the topic ontology is enriched with the concepts associated with the textual resources and the generated topics.

2.2 Ontology Engineering Approaches

In the ontology engineering domain, several approaches have been proposed for building ontologies from scratch or by reusing other existing ontologies. The most known approaches are *Uschold and Gruninger* (Uschold and Gruninger, 1996), *Methontology* (Fernández-López et al., 1997) and *ON-TO-KNOWLEDGE* (Sure et al., 2004). These approaches focus on an iterative process of ontology building and are composed of common phases such as *specification*, *conceptualization*, *formalization*, *application* and *evaluation*. In addition, approaches such as Text2Onto (Cimiano and Völker, 2005) and OnToGen (Fortuna et al., 2007) aim to generate ontologies semi-automatically with the help of user interference. These approaches exploit textual resources and rely on natural language processing techniques. However, few works have been found in the literature about building ontologies from knowledge graphs. In (Böhm and Ortiz, 2018), the authors discusses the building of topic-specific ontologies from open knowledge graphs such as ConceptNet (Speer et al., 2017). A query-based interactive approach is applied for extracting entities and relations from the knowledge graph. Based on the extraction process as well as the interaction of the user, the central taxonomy of the topic ontology is constructed. Furthermore, adding complex concepts is processed to enrich the ontology. Finally, a clean-up phase is performed in order to modify or to add new concepts to the taxonomy.

3 SPARQL-BASED AUTOMATIC APPROACH FOR BUILDING TOPIC-OPA

For building topic ontologies, the most commonly known approaches are the keyword-based construction approaches which are based mainly on text mining and information retrieval techniques (Maguitman et al., 2010). However, these approaches are not efficient, hard and time consuming to construct an ontology from a large corpus of documents (Maguitman et al., 2010). From this perspective and for simplifying the construction process of Topic-OPA, open knowledge graphs are considered. Generally, knowledge graphs are very large and contain many entities that are too general or specific to be successfully used as topics for topic labeling (Böhm and Ortiz, 2018). Meanwhile, they can be leveraged to build with moderate efforts small to medium-sized meaningful topic ontologies. As a knowledge graph, we selected Wikidata. It is a free and open knowledge graph and acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wiktionary, and others (Erxleben et al., 2014). Wikidata stores more than 402 million statements about over 45 million entities (Malyshev et al., 2018). Today, more than 60 million of items are described. The data model of Wikidata is based on a directed, labelled graph where entities are connected by edges that are labelled by "properties" (Bielefeldt et al., 2018). Thus, the system distinguishes two main types of entities: *items* and *properties*. Items are uniquely identified by a "Q" followed by a number, such as Paris (Q90). Properties describe detailed characteristics of an item and represented by a "P" followed by a number, such as *instance of* (P31). Entities are represented by *URIs* (e.g. http://www.wikidata.org/entity/Q90 for Paris and http://www.wikidata.org/entity/P31 for *instance of*).

3.1 Ontology Specification

The ontology specification clarifies the scope and the purpose of the targeted topic ontology Topic-OPA. Topic-OPA is a topic-specific ontology intended for modeling the topics of old press articles. Thus, the scope is limited to old newspapers and journals which are not organized thematically as the recent ones. Therefore, given a corpus of articles in 1920, Topic-OPA is constructed from the disambiguated named entities representing these articles (see Figure 2 for an example of named entities representing the articles depicted in Figure 7). Thereby, Topic-OPA will not be useful for labeling articles in 2020. Concerning the purpose, Topic-OPA is intended to build automated applications such as topic labeling systems. Although, it can be used to develop larger ontologies for more specialized purposes reducing the time and effort needed to develop ontologies from scratch.

3.2 Ontology Requirements

In the ontology engineering domain, the set of requirements that the ontology should satisfy is divided into *functional* and *non-functional* requirements

(Fernández et al., 2009). The functional requirements define what needs to be expressed by the ontology model. Meanwhile, the non-functional requirements specify how an ontology needs to be designed in order to be applicable. For Topic-OPA, the main functional requirement is that it needs to be composed of two different schemes:

- *Hierarchical Scheme:* consists of hierarchical relations such as *subClassOf* that permit the inference of knowledge in the ontology graph.
- *Non-hierarchical Scheme:* involves non-hierarchical relations such as *related, part of, used by, etc.* that have an important implication in the semantic relationships between the concepts.

Concerning the non-functional requirements, we consider data *traceability* and *scalability* by mapping the concepts and the relations of the topic ontology to entities in open knowledge graphs such as Wikidata.

3.3 Ontology Definition

In our work, we are interested in general topic ontologies which are composed of hierarchical and non hierarchical schemes. In the following, we define these ontologies by considering mapping to knowledge graphs.

Definition 8. *We define a* ***general topic ontology****, in which mapping to knowledge graphs is considered, by* $O = \langle T, R, E, \phi \rangle$*, with*

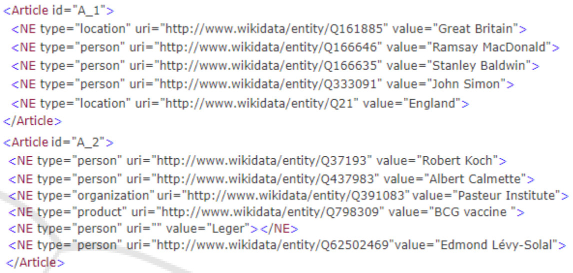
- T *the set of topic concepts,*
- R *the set of predicates:* $\{$*subClassOf, instance of, part of, use, related by, etc.*$\}$*,*
- E *the set of relationships:* $E \subseteq T \times R \times T$
- ϕ *the mapping of* T *and* R *to entities in a knowledge graph* K*.*

3.4 Ontology Building

For building Topic-OPA, a SPARQL-based fully automatic approach is applied. This approach, which aims to harvest Topic-OPA from the open knowledge graph Wikidata, is composed of three main phases: (1) construction of the hierarchical scheme, (2) construction of the non-hierarchical scheme and (3) ontology enrichment.

3.4.1 Building the Hierarchical Scheme: Bottom-up Strategy

The hierarchical scheme of Topic-OPA, which represents the taxonomy of topic concepts, can be formally defined by $H = \langle T, R, E_{\sqsubseteq}, \phi \rangle$, where T is the set of topic concepts, R is the unique predicate $\{$*subClassOf*$\}$ used for ordering the topic concepts, $E_{\sqsubseteq}$ is the set of ordering relations and ϕ is the mapping function to Wikidata. In the hierarchy, a root element denoted $\top$ is defined as a general subsumer for all the topic concepts, i.e., $\forall t_i \in T, t_i \sqsubseteq \top$. For building the hierarchy, a SPARQL-based bottom-up approach is applied. The development process starts with a definition of the most specific topic concepts of the hierarchy and continues by extracting the more general concepts. The approach started from a set of named entities N represented by a set of *URIs* (see Figure 2).



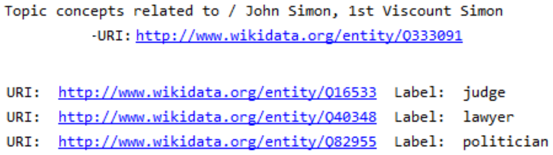
```
<Article id="A_1">
  <NE type="location" uri="http://www.wikidata/entity/Q161885" value="Great Britain">
  <NE type="person" uri="http://www.wikidata/entity/Q166646" value="Ramsay MacDonald">
  <NE type="person" uri="http://www.wikidata/entity/Q166635" value="Stanley Baldwin">
  <NE type="person" uri="http://www.wikidata/entity/Q333091" value="John Simon">
  <NE type="location" uri="http://www.wikidata/entity/Q21" value="England">
</Article>
<Article id="A_2">
 <NE type="person" uri="http://www.wikidata/entity/Q37193" value="Robert Koch">
 <NE type="person" uri="http://www.wikidata/entity/Q437983" value="Albert Calmette">
 <NE type="organization" uri="http://www.wikidata/entity/Q391083" value="Pasteur Institute">
 <NE type="product" uri="http://www.wikidata/entity/Q798309" value="BCG vaccine ">
 <NE type="person" uri="" value="Leger"></NE>
 <NE type="person" uri="http://www.wikidata/entity/Q62502469"value="Edmond Lévy-Solal">
 </Article>
```

Figure 2: Example of named entities representing articles A_1 and A_2 depicted in Figure 7.

Definition of the Most Specific Topic Concepts. At this phase, a SELECT SPARQL query, relying mainly on N and the Knowledge graph K, is applied to define $S_T \subset T$ the most specific topic concepts of the hierarchy, $\forall t_i \in S_T, \nexists t_j / t_j \sqsubseteq t_i$. The SELECT query $q(n, r)$ takes as inputs a named entity $n \in N$ and a property $r \in K$ and returns set of topic concepts. For the application of q, we defined two main relation types $\{$P31, P106$\}$. The property *instance of (P31)* is used for all the named entities to retrieve their superclasses. Meanwhile, for the named entities that are instances of Human (Q5), which is a very general topic, applying the property *occupation* (P106) is required to fetch more specific topic concepts. In the following, the syntax of q is presented. We denote by *entityId*, the Wikidata ID of the named entity which is extracted from the URI.

```
SELECT ?specificTopic WHERE {
wd:entityId ?property ?specificTopic.
VALUES ?property {wdt:P31 wdt:P106}}
```

As an example, let us consider a named entity $n = \{\text{John Simon}(Q333091)\}$ (see Figure 3). In Wikidata, John Simon is *instance of* (P31) Human (Q5) and linked to judge, lawyer and politician by the property *occupation* (P106). Thus, $S_T(n)$={Judge, Lawyer, Politician}.



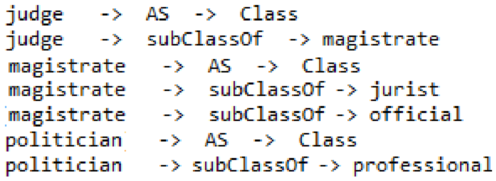
```
Topic concepts related to / John Simon, 1st Viscount Simon
          -URI: http://www.wikidata.org/entity/Q333091

URI:  http://www.wikidata.org/entity/Q16533  Label:  judge
URI:  http://www.wikidata.org/entity/Q40348  Label:  lawyer
URI:  http://www.wikidata.org/entity/Q82955  Label:  politician
```

Figure 3: Definition of the most specific concepts based on the named entities of A_1.

Extraction of Hierarchies. The aim of this phase is to build the taxonomy of topic concepts H. The building process starts from the most specific to the most general concepts. For this purpose, a CONSTRUCT SPARQL query $q_H(t_i)/t_i \in S_T$ and associated to $\phi(t_i)$, is applied to fetch the parent classes of t_i aiming to build a RDF graph of the hierarchy. In this context, each query returns three different types of triples: (1) to define the ontology classes, (2) to create the taxonomic relations (inspired by usage in RDF *rdfs:subClassOf*) and (3) to label the ontology classes. All triples are denoted by (s, p, o), where s the subject, p the predicate and o the object. In the following, the syntax of q_H is presented. We denote by *topicId* the Wikidata ID of $t_i \in S_T$.

```
CONSTRUCT {  ?class a owl:Class.
?class rdfs:subclassOf ?superclass.
?class rdfs:label ?classLabel.
?property rdfs:domain ?class.
?property rdfs:label ?classLabel.}
WHERE { wd:topicId wdt:P279* ?class.
?class wdt:P279 ?superclass.
?class rdfs:label ?classLabel.}
```

In Figure 4, an example of triples extracted based on S_T(John Simon).



```
judge    ->  AS  ->  Class
judge    ->  subClassOf  -> magistrate
magistrate   ->  AS  ->  Class
magistrate   ->  subClassOf -> jurist
magistrate   ->  subClassOf -> official
politician   ->  AS  ->  Class
politician   -> subClassOf -> professional
```

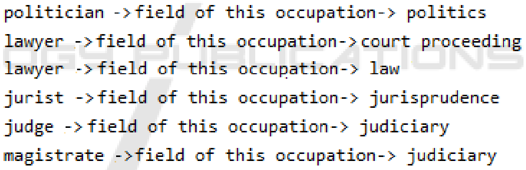
Figure 4: Example of triples for building the hierarchical scheme of Topic-OPA.

3.4.2 Building the Non-hierarchical Scheme

The non-hierarchical scheme of Topic-OPA can be formally defined by $NH = \langle T, R, E, \phi \rangle$, where T is the set of topic concepts, R is the finite set of predicates, $E \subseteq T \times R \times T$ is the set of *transverse* relationships among the topics and ϕ the mapping function. In this phase, the non-hierarchical relations are extracted from Wikidata for building NH. These relations are represented by the definition of the domain/range of the properties that will be added to the graph as edges between domains and ranges. For this purpose, a CONSTRUCT query $q_{NH}(t_i)/t_i \in T$ and associated to $\phi(t_i)$, is applied to fetch all the triples where t_i are domains or ranges. In this context, the selection of properties is restricted to a predefined list based on their relevance in different domains (e.g. *field of work (P101)*, *has part* (P527), *has quality* (P1552), *part of* (P361), *practiced by* (P3095), etc.). In the following, the syntax of q_{NH} is presented. We denote by *topicId* the Wikidata ID of $t_i \in T$.

```
CONSTRUCT { ?domain ?property ?range.
?range rdfs:label ?rangeLabel.
?property rdfs:label ?propertyLabel.}
WHERE { VALUES ?property {
wdt:P1269 wdt:P425 wdt:P101
wdt:P136 wdt:P527 wdt:P1552 wdt:P1557 wdt:P106
wdt:P2388 wdt:P2389 wdt:P361 wdt:P710 wdt:P3095
wdt:P4646 wdt:P641 wdt:P2578 wdt:P366 wdt:P1535
wdt:P2283 wdt:P1889}
{wd:topicId ?property ?range.
?range rdfs:label ?rangeLabel.}}
```

The execution of q_{NH} produced a list of triples denoted by (d, p, r), where d the domain, p the predicate and r the range. Furthermore, these triples are parsed and added to the structure of Topic-OPA for building the non-hierarchical scheme. In Figure 5, an example of non-hierarchical relations extracted based on the previously added concepts (see Figure 4).

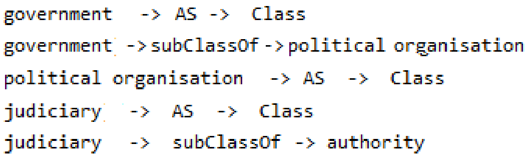


```
politician ->field of this occupation-> politics
lawyer ->field of this occupation->court proceeding
lawyer ->field of this occupation-> law
jurist ->field of this occupation-> jurisprudence
judge ->field of this occupation-> judiciary
magistrate ->field of this occupation-> judiciary
```

Figure 5: Example of triples for building the non-hierarchical scheme of Topic-OPA.

3.4.3 Ontology Enrichment

After building H and NH, we apply in this phase an enrichment process based on NH. The application of q_{NH} has imported new concepts to the ontology such as Government, Judiciary and Politics, among many others. Therefore, these concepts will be added to the hierarchy as well as their parent classes by applying the query q_H (see Figure 6).

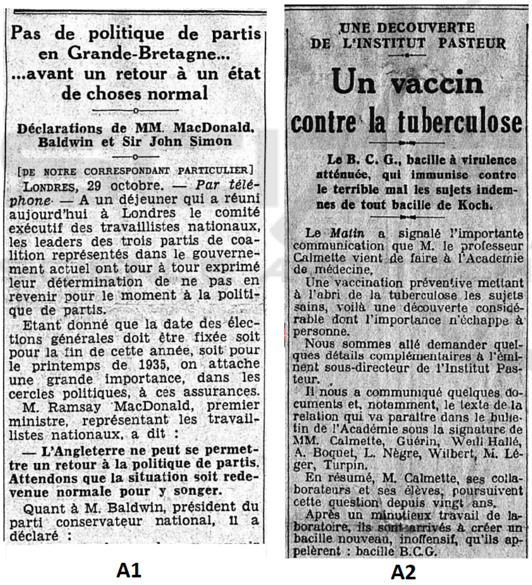


```
government   -> AS ->  Class
government  ->subClassOf->political organisation
political organisation   -> AS  ->  Class
judiciary   ->  AS  ->  Class
judiciary   ->  subClassOf  -> authority
```

Figure 6: Example of the enrichment of the hierarchical scheme of Topic-OPA.

4 CASE STUDY: *LE MATIN*

In this section, we introduce the application of the SPARQL-based approach for developing Topic-OPA in the context of the old newspaper *Le Matin*. For this purpose, we have chosen A a corpus of 48 articles published between 1910 and 1937 (see Figure 7 for an example). For Building Topic-OPA, a set of $N = 392$ named entities representing A is considered (see Figure 2). As a result, we obtained a topic ontology, as a subset of Wikidata, which is accessible and manageable in ontology editors such as Protégé[3]. Note that the topic ontology is not curated. We maintained the concepts and relations which are obtained by the application of the fully automatic approach. Thus, Topic-OPA contains 2073 concepts, 3261 SubClassOf relations and 1135 non-hierarchical relations. In Figure 8, we depict an excerpt of Topic-OPA around the Politics topic. The solid lines represent the SubClassOf relations and the dashed lines represent the non-hierarchical relations.



Pas de politique de partis en Grande-Bretagne...
...avant un retour à un état de choses normal

Déclarations de MM. MacDonald, Baldwin et Sir John Simon

[DE NOTRE CORRESPONDANT PARTICULIER]

LONDRES, 29 octobre. — *Par téléphone* — A un déjeuner qui a réuni aujourd'hui à Londres le comité exécutif des travaillistes nationaux, les leaders des trois partis de coalition représentés dans le gouvernement actuel ont tour à tour exprimé leur détermination de ne pas en revenir pour le moment à la politique de partis.

Etant donné que la date des élections générales doit être fixée soit pour la fin de cette année, soit pour le printemps de 1935, on attache une grande importance, dans les cercles politiques, à ces assurances.

M. Ramsay MacDonald, premier ministre, représentant les travaillistes nationaux, a dit :

— **L'Angleterre ne peut se permettre un retour à la politique de partis. Attendons que la situation soit redevenue normale pour y songer.**

Quant à M. Baldwin, président du parti conservateur national, il a déclaré :

UNE DECOUVERTE DE L'INSTITUT PASTEUR

Un vaccin contre la tuberculose

Le B. C. G., bacille à virulence atténuée, qui immunise contre le terrible mal les sujets indemnes de tout bacille de Koch.

Le *Matin* a signalé l'importante communication que M. le professeur Calmette vient de faire à l'Academie de médecine.

Une vaccination préventive mettant à l'abri de la tuberculose les sujets sains, voilà une découverte considérable dont l'importance n'échappe à personne.

Nous sommes allé demander quelques détails complémentaires à l'éminent sous-directeur de l'Institut Pasteur.

Il nous a communiqué quelques documents et, notamment, le texte de la relation qui va paraître dans le bulletin de l'Académie sous la signature de MM. Calmette, Guérin, Weill Hallé, A. Boquet, L. Nègre, Wilbert, M. Léger, Turpin.

En résumé, M. Calmette, ses collaborateurs et ses élèves, poursuivent cette question depuis vingt ans.

Après un minutieux travail de laboratoire, ils sont arrivés à créer un bacille nouveau, inoffensif, qu'ils appelèrent : bacille B.C.G.

Figure 7: Example of articles from *Le Matin*.

5 ONTOLOGY EVALUATION

Generally, the ontology evaluation approaches are divided into four main categories (Fernández et al., 2009): (1) *gold standard-based* that aims to compare the developed ontology with a previously created reference ontology known as the gold standard; (2) *corpus-based* that tends to compare the developed ontology with the content of a text corpus that covers a given domain significantly ; (3) *application-based* that considers the evaluation of ontologies according to their performance in applications; (4) *structure-based* that quantifies structure-based properties such as the size and the complexity of ontologies.

In order to choose the "best" evaluation approach, there is a need to define the motivation behind evaluating a developed ontology (Fernández et al., 2009). In our study, as evoked earlier, Topic-OPA is intended to be used as a knowledge base in a topic labeling system. Thus, it is considered as an application-based ontology. In this context, Topic-OPA can be evaluated using application-based and structure-based approaches for the following reasons:

- the gold standard-based approach is not applicable: Topic-OPA is developed as a subset of Wikidata. Thus, the best reference ontology for Topic-OPA is Wikidata itself. However, it is impossible to use Wikidata as a gold standard ontology because of its size. In addition, since Topic-OPA is built for and from a given corpus of press articles, it cannot be compared with other ontologies that should be created under similar conditions with similar goals.
- the corpus-based approach is eliminated: the textual resources are out of scope of our study. As evoked earlier, our hypothesis is based on a set of disambiguated named entities extracted from open knowledge bases such as Wikidata.
- the application-based approach is the best evaluation approach: it implies to evaluate the usability of Topic-OPA being an application-based ontology. This evaluation will be performed in further works after embedding Topic-OPA in the topic labeling system.
- the structure-based approach is a useful evaluation approach for assessing the structure-based properties of Topic-OPA. This approach is recommended as an efficient approach for evaluating the learned ontologies (Dellschaft and Staab, 2008).

Several measures have been recognized for the structure-based evaluation such as *Knowledge coverage and popularity* measures (i.e. number of classes and number of properties) and *structural* measures (i.e. maximum depth, average depth, depth variance, etc.) (Fernández et al., 2009). The application of these measures relies on an assumption that is *a richly populated ontology, with higher depth and breadth variance is more likely to provide reliable semantic content*. The structural measures are positively correlated with the semantic accuracy of the knowledge modeled in the ontology (Sanchez et al., 2015). In the

[3] https://protege.stanford.edu/, last visited 23 July 2020

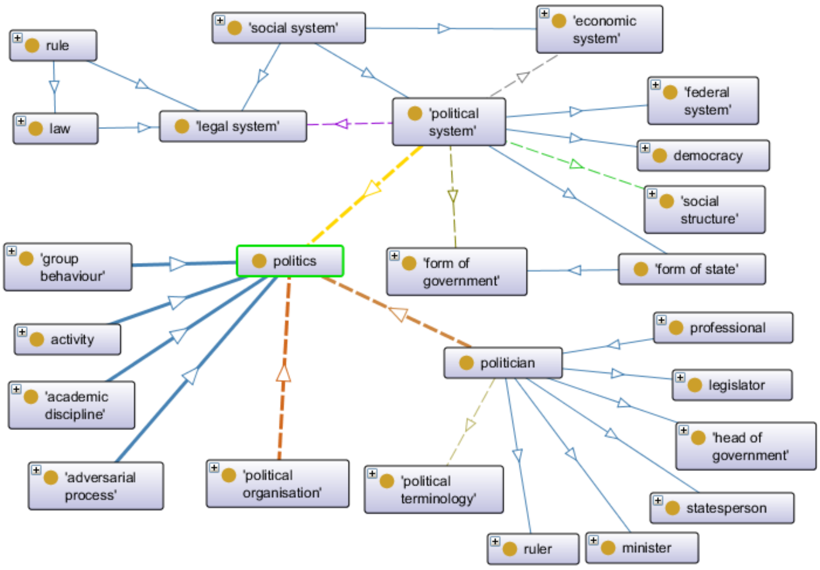


Figure 8: Excerpt of Topic-OPA around the concept Politics.

context of Topic-OPA, we quantified some structural measures, by considering its taxonomic structure, as follows:

- *maximum depth=28:* represents the length of the longest *taxonomic* branch in the ontology..
- *average depth=6:* is the average length of all *taxonomic* branches.
- *depth variance=6.38:* is the dispersion with respect to the average depth, computed as the standard mathematical variance.

We conclude that Topic-OPA is a richly learnt ontology. However, the majority of the topic concepts are dispersed homogeneously within the core level of Topic-OPA. This implies two main issues: (1) the hierarchical structure of Topic-OPA is a balanced taxonomy, in which the majority of taxonomic edges have almost the same depth and (2) it will be a challenging task to expose the topic concepts which are relevant for topic labeling the articles.

6 CONCLUSIONS AND FUTURE WORK

This paper discussed the building process of a topic-specific ontology, named Topic-OPA, for representing the contents of a set of old press articles that needs to be labeled with a set of topics. Topic-OPA will be used for associating topics to each old press article. The development of Topic-OPA relies mainly on a set of disambiguated named entities representing the articles. In this regard, a SPARQL-based fully automatic approach is applied, based on the disambiguated named entities, for harvesting Topic-OPA from the open knowledge graph Wikidata. The proposed approach is composed of four main phases: (1) collect the most specific topic concepts which are located at the lowest level of Topic-OPA, (2) build the hierarchical scheme based on these concepts, (3) construct the non-hierarchical scheme based on the hierarchical scheme and (4) enrich the ontology with the concepts imported by the non-hierarchical relations. A case study is presented in the context of the old french newspaper *Le Matin* for building Topic-OPA from a corpus of 48 articles. By the application of the SPARQL-based approach, a richly learnt topic-specific ontology is obtained. Furthermore, a structure-based evaluation approach is applied to assess the quality of the structure of Topic-OPA. We found that the majority of the topic concepts are located at the core level of Topic-OPA. This implies that a challenging task will take place for defining the topic concepts which will be used for labeling the articles. In this study, we do not consider the curation of the topic ontology after the automatic building process. We maintained the ontology structure and content, including the abstract and specific concepts, as derived from Wikidata. In future works, we will apply a curation process aiming to clean and leverage Topic-OPA. Furthermore, Topic-OPA will be embedded in a topic labeling system for automatic topic labeling of old press articles. A specific semantic relatedness measure, named Rel_{Topic}, has been proposed in order to associate the good topic of Topic-OPA to a specific press article taking into consideration the set of named entities of it. Unfortunately, we have not been able to present it in this paper, because of lack of space. However, it is worth highlighting that the preliminary results on the use of Rel_{Topic} associated with Topic-OPA are encouraging, as its use presents a precision higher than 80% on a corpus of old press articles that were labeled by human experts.

ACKNOWLEDGMENTS

This work is funded by the Normandy Region (France) and the European Union with the European Regional Development Fund (FEDER).

REFERENCES

Allahyari, M. and Kochut, K. (2017). A knowledge-based topic modeling approach for automatic topic labeling. *International Journal of Advanced Computer Science and Applications*, 8(9):335–349.

Böhm, K. and Ortiz, M. (2018). A tool for building topic-specific ontologies using a knowledge graph. In *Proceedings of the 31st International Workshop on Description Logics co-located with 16th International Conference on Principles of Knowledge Representation and Reasoning (KR 2018)*.

Bielefeldt, A., Gonsior, J., and Krotzsch, M. (2018). Practical linked data access via sparql: The case of wikidata. In *Proceedings of the WWW2018 Workshop on Linked Data on the Web (LDOW-18), CEUR Workshop Proceedings*.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Cimiano, P. and Völker, J. (2005). Text2onto. In *Natural Language Processing and Information Systems, NLDB 2005*, pages 227–238. Springer, Berlin, Heidelberg.

Dellschaft, K. and Staab, S. (2008). Strategies for the evaluation of ontology learning. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, Frontiers in Artificial Intelligence and Applications*, pages 253–272.

Erxleben, F., Günther, M., Krötzsch, Mendez, J., and Vrandečić, D. (2014). Introducing wikidata to the linked data web. In *Proceedings 13th Int. Semantic Web Conf. (ISWC'14), LNCS*, pages 50–65.

Fernández, M., Overbeeke, C., Sabou, M., and Motta, E. (2009). What makes a good ontology? a case-study in fine-grained knowledge reuse. In *The semantic Web*, pages 61–75. Springer, Berlin, Heidelberg.

Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: From ontological art towards ontological engineering. In *AAAI*.

Fortuna, B., Grobelnik, M., and Mladenic, D. (2007). Ontogen: Semi-automatic ontology editor. In *Human Interface and the Management of Information, Interacting in Information Environments, Human Interface 2007*, pages 309–318. Springer, Berlin, Heidelberg.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM New York.

Maguitman, A., Cecchini, R., Lorenzetti, C., and Menczer, F. (2010). Using topic ontologies and semantic similarity data to evaluate topical search. In *Proceedings of Conferencia Latino-americana de Informática*.

Malyshev, S., Krotzsch, M., Gonzalez, L., Gonsior, J., and Bielefeldt, A. (2018). Getting the most out of wikidata: Semantic technology usage in wikipedia's knowledge graph. In *Proceedings of the 17th International Semantic Web Conference (ISWC'18)*, pages 376–394. Springer.

Sanchez, D., Batet, M., Martinez, S., and Ferrer, J. (2015). Semantic variance: An intuitive measure for ontology accuracy evaluation. *Engineering Applications of Artificial Intelligence*, 39:89–99.

Sleeman, J., Finin, T., and Halem, M. (2018). Ontology-grounded topic modeling for climate science research. In *Proceedings of Semantic Web for Social Good Workshop, ISWC*.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451.

Sure, Y., Staab, S., and Studer, R. (2004). On-to-knowledge methodology (otkm). In *Handbook on Ontologies, International Handbooks on Information Systems*. Springer, Berlin, Heidelberg.

Tang, Y., Baer, P., Zhao, G., and Meersman, R. (2009). On constructing, grouping and using topical ontology for semantic matching. In *Proceedings of OTM 2009 Workshops (On the Move to Meaningful Internet Systems)*. Springer Berlin, Heidelberg.

Uschold, M. and Gruninger, M. (1996). Ontologies: principles, methods and applications. *The Knowledge Engineering Review*, 11:93–136.

Zhao, G. and Meersman, R. (2005). Architecting ontology for scalability and versatility. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE. OTM 2005*.