Rel\textsubscript{Topic}: A Graph-Based Semantic Relatedness Measure in Topic Ontologies and Its Applicability for Topic Labeling of Old Press Articles

Mirna El Ghosh\textsuperscript{a,}\textsuperscript{*}, Nicolas Delestre\textsuperscript{a}, Jean-Philippe Kotowicz\textsuperscript{a}, Cecilia Zanni-Merk\textsuperscript{a} and Habib Abdulrab\textsuperscript{a}

\textsuperscript{a}LITIS, Normandie Université, INSA Rouen, 76000 Rouen, France

Abstract. Graph-based semantic measures have been used to solve problems in several domains. They tend to compare ontological entities in order to estimate their semantic similarity or relatedness. While semantic similarity is applicable to hierarchies, semantic relatedness is adapted to ontologies. However, designing semantic relatedness measures is a difficult and challenging issue. In this paper, we propose a novel semantic measure within topic ontologies, named Rel\textsubscript{Topic}, for assessing the relatedness of instances and topics. To design Rel\textsubscript{Topic}, we considered topic ontologies as weighted graphs where topics and instances are represented as weighted nodes and semantic relations as weighted edges. The use of Rel\textsubscript{Topic} is evaluated for labeling old press articles. For this purpose, a topic ontology, named Topic-OPA, is derived from open knowledge graphs by the application of a SPARQL-based fully automatic approach. The ontology building process is based mainly on a set of disambiguated named entities representing the articles. To demonstrate the performance of our approach, a use-case is presented in the context of the old French newspaper Le Matin. Our experiments show that Rel\textsubscript{Topic} produces more than 80% relevant labeling topics as compared to the topics assigned by human annotators.

Keywords: Semantic relatedness, Graph-based semantic measures, Weighted graphs, Knowledge Graphs, Topic ontologies, Topic labeling

1. Introduction

Graph-based semantic measures have been used to solve problems in a broad range of domains such as Natural Language Processing (e.g. [1]), Information Retrieval (e.g. [2]), Knowledge Engineering (e.g. [3]), Semantic Web and Linked Data (e.g. [4]) and Bioinformatics (e.g. [5]). They are considered as essential tools for the design of numerous algorithms in which semantics matters [6]. A graph-based semantic measure is a mathematical tool used to estimate the strength of the semantic interaction between entities (concepts or instances) based on the analysis of ontologies [6]. Thus, the application of this measure is strongly dependent on the availability of an ontology that represents the application domain. Two main categories of graph-based semantic measures are distinguished: (1) similarity measures adapted to taxonomies and (2) relatedness measures adapted to semantic graphs composed of different types of relationships [6]. In the literature, few relatedness measures have been designed. Most efforts are directed for designing similarity measures. For comparing ontological entities, graph-based measures are classified into two basic approaches: path-based, that compare the concepts according to properties of paths in graphs, and node-based, that use properties of concepts in the ontology graph for comparing concepts. However, these approaches suffer from different limitations.
The goal of this work is to design and evaluate a semantic relatedness measure within topic ontologies for topic labeling of old press articles. The articles are represented by a set of "not ambiguous" named entities extracted from open data sources (e.g., Wikidata). In order to overcome the limitations of existing approaches, we propose a hybrid measure, named \textit{RelTopic}, by combining node-based and path-based approaches. In contrast to existing measures, our measure tends to assess the relatedness of concepts and instances by considering different types of relations. For the application of \textit{RelTopic}, a topic ontology named Topic-OPA is derived from the open knowledge graph Wikidata using a SPARQL-based fully automatic approach. The ontology building process is launched from the set of the "disambiguated" named entities of the corpus of old press articles to label. Based on \textit{RelTopic} and Topic-OPA, we defined the selection process of the most relevant topics for labeling the articles. In order to demonstrate the performance of Topic-OPA and \textit{RelTopic}, a use-case is presented in the context of \textit{Le Matin}, an old French newspaper first published in 1884 and discontinued in 1944. The topic ontology Topic-OPA and the semantic measure \textit{RelTopic} are evaluated using dual evaluation approaches.

The remainder of this paper is organized as follows; the problem definition is outlined in section 2. Section 3 presents the related works of this study. In section 4, we discuss our semantic relatedness measure \textit{RelTopic}. Section 5 introduces the SPARQL-based approach for building Topic-OPA. The section 6 discusses the topic labeling process. In section 7, we present a use case for labeling the articles of \textit{Le Matin}. We evaluate and discuss the approach in section 8. Finally, section 9 concludes the paper.

2. Problem Definition

While more recent press articles are thematically organized and available for accessing and searching over them, old press articles do not present this feature. Ancient newspapers and journals are particularly (see Figure 1):

(a) presented in few pages: often a large sheet folded into two, with occasionally an inter-leaf, so having 4 to 6 pages;

(b) not organized thematically: thematic pages (i.e., politics, art, sport, etc.) are not available;

(c) articles are presented consecutively.

In the ASTURIAS\(^2\) project (see Figure 2) we need to thematically organize a collection of old press articles with a set of topics (e.g., Politics, Art, Sport, Science, etc.). For this purpose, and since articles cannot be organized thematically according to their positions in the journals, their content will be considered. In the context of ASTURIAS, a fundamental hypothesis is that the articles are represented by "not ambiguous" named entities extracted from open data sources (WP2).

Our research problem can be defined as follows: Given an article \(A\), a set of named entities \(N\) that are collected from \(A\) and represented by a set of URIs, and a topical structure \(T\), the problem is to find the most relevant topics from \(T\) that label \(A\). Based on this perspective, our work (WP3) considers mainly three main issues:

\(1\)https://gallica.bnf.fr/ark:/12148/cb328123058/date, last visited on April 8 2020

\(2\)Analyse STructURelle et Indexation sémantique d’ArticleS de presse.
1. **construction of the topical structure**: takes as input \( N \) the set of disambiguated named entities and constructs \( T \) a convenient topical structure based on \( N \).
2. **named entity-topic mapping**: takes as inputs \( n \in N \) and \( t \in T \) and evaluates if \( t \) is relevant to \( n \) or not.
   The relevance can be examined as a semantic (not syntactic) relatedness. For this purpose, a semantic measure is needed to compute the relatedness.
3. **ranking and selection of most relevant topics**: takes as input the relatedness values of \( n \) and \( t \) and aims to rank them and select the best topic(s) to label \( A \).

3. **Related Works**

In this section, we outline the related works of our study: graph-based semantic measures, topic ontologies and ontology engineering approaches.

3.1. **Graph-Based Semantic Measures**

For comparing ontological entities, graph-based measures are classified into two basic approaches: path-based and node-based. In path-based approaches, concepts are compared according to properties of paths in graphs. The most common property is the shortest path that connects nodes in a given ontology. The shorter the path is, the higher the similarity is. The Rada’s measure is an example of similarity measures adapted to taxonomies:

\[
\text{Sim}_{\text{Rada}}(c_1, c_2) = \frac{1}{1 + \text{dist}_{\text{Rada}}(c_1, c_2)}, \tag{1}
\]

where \( \text{dist}_{\text{Rada}} \) is the shortest path and \( \text{Sim}_{\text{Rada}} \) is the distance to similarity conversion [7].

Although, Leacock and Chodorow’s measure is an example of this category which is designed for WordNet [8]:

\[
\text{Sim}_{\text{LC}}(c_1, c_2) = -\log\left(\frac{\text{len}(c_1, c_2)}{2 \times \text{maxdepth}(c)}\right), \tag{2}
\]

where \( \text{len}(c_1, c_2) \) is the shortest path between \( c_1 \) and \( c_2 \) and \( \text{maxdepth}(c) \) is the maximum depth of \( c, \forall c \in \text{WordNet} \).

In this category of measures, Hirst and St-Onge’s measure, that considers the non-taxonomic links, is identified [9]:

\[
\text{Rel}_{\text{HS}}(c_1, c_2) = C - \text{len}(c_1, c_2) - k \times \text{turns}(c_1, c_2), \tag{3}
\]

where \( C \) and \( k \) are constants \( (C = 8 \text{ and } k = 1) \), and \( \text{turns}(c_1, c_2) \) is the number of times the path between \( c_1 \) and \( c_2 \) changes direction. The main drawback of these approaches is that they consider all edges equivalent, indicating therefore a uniform distance.

Concerning the node-based approaches, they use properties of concepts in the ontology graph for comparing concepts. The most common property is the Information Content (IC) of nodes which is calculated based on the frequency of the term in a given corpus. IC is a property that denotes how specific and informative a concept is. The most well-known IC measures, which are based on the lowest common subsumer (LCS) property, are Resnik’s [10] and Lin’s [11] measures.
Resnik’s measure simply uses the Information Content of the LCS as the similarity value:

$$Sim_{Resnik}(c_1, c_2) = IC(LCS(c_1, c_2)),$$

where IC of a concept is defined as the negative log of the probability of that concept:

$$IC(c) = -\log P(c)$$

Concerning the Lin’s measure, it is considered as a refinement of Resnik’s measure and is computed as follows:

$$Sim_{Lin}(c_1, c_2) = \frac{2 \times Sim_{Resnik}(c_1, c_2)}{IC(c_1) + IC(c_2)}$$

Three main limitations are recognized for these approaches: (1) they are based on textual resources (2) they do not consider concepts with multiple ancestors and (3) they are applicable only on taxonomies.

### 3.2. 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 [16]. 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 [15]. 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).

Topic ontologies are being increasingly used in various domains such as semantic matching [12], topic labeling [13], topic modeling [14] and evaluating topical search [15]. For topic labeling purposes, the topic model KB-LDA [13] 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.

For topic modeling purposes, IPCC [14] 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 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.

Both topic models are related to textual resources either for comparing their content with the ontology content or for the application of data mining techniques.

### 3.3. 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 [36], Methontology [37] and ON-TO
KNOWLEDGE [38]. 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 [39] and OntoGen [40] 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 [24], the authors discuss the building of topic-specific ontologies from open knowledge graphs such as ConceptNet [41]. 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.

4. Our Semantic Relatedness Measure

Building semantic relatedness measures is a challenging research issue. In this section, we propose a hybrid graph-based semantic relatedness measure within topic ontologies. Aiming to cover the limitations of existing measures, we design our measure as a combination of path-based and node-based approaches. Thus, we comprehensively consider: (1) non-hierarchical relations and differentiate them from hierarchical relations regarding the paths properties, (2) correlation of nodes and (3) comparing instances to concepts.

4.1. Topic Ontologies as Semantic Graphs

For the application of graph-based semantic measures, there is a need to represent ontologies as graphs using a graph-based formalism. In semantic graphs associated to general topic ontologies, we denote topics and instances as nodes and different types of relationships (hierarchical and non-hierarchical) as edges.

Definition 1. We define the semantic graph associated to a general topic ontology as a directed weighted graph $G = (V, E, T, \tau, \omega, \delta)$, where $V$ is a finite set of nodes that represent topics and instances, $E \subseteq V \times V$ is a finite set of edges connecting different pairs of nodes $(v_i, v_j)$ from $V$, $T$ is a finite set of edge types, $\tau : E \rightarrow T$ is a function that maps edges in $E$ to their types in $T$ (subclassOf, partOf, used by, ...), $\omega : V \rightarrow \mathbb{R}^+$ is a node-weighting function that maps nodes to their weights and $\delta : E \rightarrow \mathbb{R}^+$ is an edge-weighting function that assigns weights to edges.

Definition 2. The set of neighbours $N(v_i)$ for a node $v_i \in V$ is represented by the nodes $\{v_j, ..., v_k\}$ that are linked to $v_i$ by the edges $\{e_j, ..., e_k\} \in E$.

Definition 3. The set of hyponyms $H(v_i)$ for a node $v_i \in V$ is represented by the nodes $\{v_h, ..., v_k\}$ that are linked to $v_i$ by the edges $\{e_h, ..., e_k\}$, where $\tau(e_m) = \{\text{subclassOf}\} \lor \{\text{instanceOf}\}, e_m \in \{e_h, ..., e_k\}$.

Definition 4. A path $P(v_i \rightarrow v_j)$ between $v_i, v_j \in V$ is a sequence of nodes and edges $\{v_{i_1}, v_{r_1}, v_{k_1}, ..., v_{i_{m-1}}, v_{k_{m-1}}, v_j\}$ connecting $v_i$ and $v_j$. For every two consecutive nodes $v_k, v_{k+1} \in V$ in $P(v_i \rightarrow v_j)$, there exists an edge $e_k \in E$.

Definition 5. The length of a path $|P(v_i \rightarrow v_j)|$ is obtained by summing up the weights of the edges that constitute the path between $v_i$ and $v_j$. $|P(v_i \rightarrow v_j)| = \sum_{e \in E(P)} \delta(e)$.

Definition 6. The distance $\text{dist}(v_i \rightarrow v_j)$ between $v_i, v_j$ is the minimum length of a path from $v_i$ to $v_j$.

Definition 7. The size of a semantic graph $|G|$ is the total number of nodes in $G$.

4.2. Design of $\text{Rel}_\text{Topic}$

For designing $\text{Rel}_\text{Topic}$, five main phases are defined: (1) weight allocation for nodes, (2) weight allocation for edges, (3) computation of the degree centrality of nodes, (4) computation of the semantic distance and (5) computation of the semantic relatedness.

4.2.1. Weight Allocation for Nodes

Informed by the information-content measures [10, 18], that outlined the adequacy of the log function for node weighting [19], we propose the weight allocation for nodes based on this function. In addition, we took advantage of the neighborhood of nodes and we differentiate between weights for topics and weights for instances. Concerning the topics, weights are formally defined by $\omega(v_i) = -\log\left(\frac{N(v_i)}{|G|}\right)$. For the instances, two main cases are identified:

1. $v_i$ is an instance of a single hyponym node $v_h$.

   In this case, the weight is formally defined by $\omega(v_i) = \omega(v_h)$.
2. \( v_i \) is an instance of multiple hypernym nodes represented by \( H(v_i) = \{v_k, ..., v_m\} \). Here, \( \omega(v_i) = (\omega(v_k))_{v_k \in H(v_i)} \), where \( \omega(v_i) \) is the average of the weights of the hypernms of \( v_i \).

### 4.2.2. Weight Allocation for Edges

Based on the diversity of relations within general topic ontologies, the allocation of weights for edges depends mainly on the types of relations. Therefore, we consider a static weight allocation which reflects the “strength” of a given relation type [19, 20]. Two main types of relations are recognized:

- Hierarchical relations: subclassOf and instance of which are classified as vertical relations with a cost of 1.
- Non-hierarchical: part/whole relations (e.g. part of, has part) and general relations (e.g. facet of, field of work, practiced by, used by). This type of relation is considered being informative and the cost of this edge must be low [19].

Given two nodes \( v_i \) and \( v_{i+1} \) linked by an edge \( e_i \), the weight of \( e_i \) is:

\[
\delta(e_i) = \begin{cases} 
1, & \text{if } \tau(e_i) = \text{subclassOf} \lor \text{instanceOf} \\
0.25, & \text{otherwise}
\end{cases}
\]

(7)

### 4.2.3. Computation of the Degree Centrality for Nodes

The Degree Centrality of a node is considered as a basic indicator for studying networks and is defined as the number of adjacencies [21]. It corresponds to how much surface the node is correlated to in the whole domain of interest [22]. The degree measure is formally defined, for unweighted graphs, by \( D(v_i) = |N(v_i)| \), where \(|N(v_i)|\) is the number of neighbours of the node \( v_i \) [23]. Meanwhile, in weighted graphs, \( D(v_i) = \sum_{j \in N(v_i)} \delta(e_j) \times \omega(v_j) \), where \( e_j = \{v_j, v_i\} \).

In our work, we take advantage of this measure to quantify the degree centrality of topics and instances. We consider that the degree centrality of an instance is related to the degree centrality of its hypernym node(s). More precisely, for every path \( P(v_i \rightarrow v_k) \), where \( v_i \) is the instance node and \( v_k \) is the topic node, we calculate the degree centrality for \( v_k \) and for the hypernym node(s) of \( v_i \). Two main cases are identified:

1. \( v_i \) is an instance of a single hypernym node. Thus, the degree centrality of nodes representing instances is formally defined by:

\[
D(v_i) = \sum_{e_j \in N(v_k)} \delta(e_j) \times \omega(v_j), \text{where } \omega(v_k) \text{ is the hypernym of } v_i, e_k = \{v_i, v_k\}, \tau(e_k) = \{\text{instanceOf}\} \text{ and } e_j = \{v_j, v_k\}.
\]

2. \( v_i \) is an instance of multiple hypernym nodes, \( v_i \) instance of multiple hypernym nodes that are represented by \( H(v_i) = \{v_k, ..., v_m\} \), \( D(v_i) = (D(v_k))_{v_k \in H(v_i)} \), where \( (D(v_k)) \) is the average of the degree centrality of the hypernms of \( v_i \).

### 4.2.4. Semantic Distance Computation

In order to estimate the relatedness of two nodes \( v_i \) and \( v_j \), there is a need to calculate the semantic distance \( \text{dist}(v_i \rightarrow v_j) \) (i.e. shortest path) between them. In weighted graphs, different approaches can be used to estimate the semantic distance such as Dijkstra [34] and Bellman Ford [35] algorithms. In our study, we have applied Dijkstra’s algorithm.

### 4.2.5. Semantic Relatedness Computation

In this section, we present the computation of the semantic relatedness between instances and topics within topic ontologies. Given two elements in a given topic ontology, an instance \( v_i \) and a topic \( v_j \) and \( P(v_i \rightarrow v_j) \) is the path between \( v_i \) and \( v_j \). The semantic relatedness measure takes these elements as input and returns a numerical description, \( \text{Rel}_{\text{Topic}} \in [0, 1] \), that quantifies their relatedness based on the following formula:

\[
\text{Rel}_{\text{Topic}}(v_i, v_j) = \left( \frac{1}{1 + \text{dist}(v_i \rightarrow v_j)} \right)^k + k \times \left( \frac{\log(D(v_i) + D(v_j))}{\omega(v_i) + \omega(v_j)} \right),
\]

where \( \text{dist}(v_i \rightarrow v_j) \) is the semantic distance between \( v_i \) and \( v_j \), \( \omega(v_i) \) and \( \omega(v_j) \) are the weights of \( v_i \) and \( v_j \) respectively and \( D(v_i) \) and \( D(v_j) \) are the degree centrality of \( v_i \) and \( v_j \) respectively. In this formula, we also assigned a variable \( k \) that takes two possible values:

\[
k = \begin{cases} 
1, & \text{if } P(v_i \rightarrow v_j) \text{ is semantically correct} \\
0, & \text{if } P(v_i \rightarrow v_j) \text{ is semantically incorrect}
\end{cases}
\]

The correctness of the semantic path between two nodes is prescribed based on the constraints proposed in [9]. If a path \( P(v_i \rightarrow v_j) \) changes the direction from upward (generalization) to downward (specialization) at a point related to a hierarchical link, \( P(v_i \rightarrow v_j) \)
The topic ontology is intended to be used as a knowledge graph and acts as a central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wiktionary, and others [25]. Wikidata stores more than 402 million statements about over 45 million entities [26]. 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” [27]. 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).

5. SPARQL-Based Automatic Approach for Building Topic Ontologies

For the application of RelTopic, there is a need for a topic ontology that represents the domain of old press articles. The most commonly known approaches for building topic ontologies are the keyword-based construction approaches which are based mainly on text mining and information retrieval techniques [15]. However, these approaches are not efficient, hard and time consuming to construct an ontology from a large corpus of documents [15]. From this perspective and for simplifying the construction process of Topic-OPA, open knowledge graphs, such as Wikidata, are considered in our study. 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 [24]. 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 [25]. Wikidata stores more than 402 million statements about over 45 million entities [26]. 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” [27]. 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).

5.1. Ontology Specification

The ontology specification clarifies the purpose and the scope of the targeted topic ontology Topic-OPA. The topic ontology is intended to be used as a knowledge base for a topic labeling system aiming to label old press articles. Therefore, given a corpus of articles to label, Topic-OPA is constructed from the disambiguated named entities representing these articles.

Therefore, if the goal is to label the articles of the year 1910 of a given journal/newspaper, Topic-OPA has to be developed from all the named entities representing all the articles of this year. Thereby, Topic-OPA will not be useful nor compatible for labeling articles from recent journals in 2020. Topic-OPA has two significant benefits: (1) to build automated applications such as topic labeling and (2) to develop larger ontologies for more specialized purposes reducing the time and effort needed to develop ontologies from scratch.

5.2. Ontology Requirements

In the ontology engineering domain, the set of requirements that the ontology should satisfy is divided into functional and non-functional requirements [29]. 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.

5.3. 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 instances and mapping to knowledge graphs.

Definition 8. We define a general topic ontology, in which instances and mapping to knowledge graphs are considered, by \( O = (T, I, R, E, \phi) \), with

- \( T \) the set of topic concepts,
5.4. Ontology Building

For building Topic-OPA, a SPARQL-based fully automatic methodology is applied. This methodology, 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.

5.4.1. Building the Hierarchical Scheme: Bottom-Up Approach

The hierarchical scheme of Topic-OPA, which represents the taxonomy of topic concepts, can be formally defined by $H = (T, R, E, \phi)$, where $T$ is the set of topic concepts, $R$ is the unique predicate $\{\text{subClassOf}\}$ used for ordering the topic concepts, $E$ is the set of ordering relations and $\phi$ is the mapping function to Wikidata. In the hierarchy, a root element denoted $T$ is defined as a general subsumer for all the topic concepts, i.e., $\forall t_i \in T, t_i \subseteq T$. For building the hierarchy, a query-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 is launched from a set of named entities $N$ represented by a set of URIs (see Figure 3).

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 $S_T \subseteq T$ the most specific topic concepts of the hierarchy, $\forall t_i \in S_T, \forall_j t_j \subseteq t_i$. 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 $\{\text{P31}, \text{P106}\}$. 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 4). In Wikidata, John Simon is instance of (P31) Human (Q5) and linked to judge, lawyer and politician by the property occupation (P106). Thus, $S_T(n) = \{\text{Judge, Lawyer, Politician}\}$.

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 $\{?t_i ?t_j \subseteq S_T\}$ 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 $\text{rdfs:subClassOf}$) 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 $qH$ 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 rdfs:label ?classLabel. 
  
  H=\{Judge, Magistrate, Magistrate, Official, Jurist, Official, Civil Servant, Civil Servant, Public Employee, Public Employee, Employee, Lawyer, Jurist, Politician, Professional\}.
```

In Figure 5, an example of triples extracted based on $S_T(\text{John Simon})$.

5.4.2. Building the Non-Hierarchical Scheme

The non-hierarchical scheme of Topic-OPA can be formally defined by $NH = (T, R, E, \phi)$, 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 $\phi$ the mapping function. In this phase, the non-hierarchical relations are extracted
Fig. 3. Example of named entities extracted from article A1 (see figure 8).

Fig. 4. Definition of the most specific concepts based on the named entities of A1.

Fig. 5. Example of triples for building the hierarchical scheme of Topic-OPA.

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 the hierarchy of topics starting from / judge / in a bottom-up strategy.

CONSTRUCT the hierarchy of topics starting from / lawyer / in a bottom-up strategy.

CONSTRUCT the hierarchy of topics starting from / politician / in a bottom-up strategy.
The results obtained by executing $q_{NH}$ are represented by triples denoted $(d, p, r)$, where $d$ the domain, $p$ the predicate and $r$ the range. In Figure 6, an example of non-hierarchical relations extracted based on the previously added concepts (see Figure 5).

The named entities are categorized in: persons, locations, organizations and products. For the labeling process, we are interested mainly in: persons, organizations and products. The named entities of the type locations will be used in further works for contextualizing the articles. The disambiguated named entities will be assigned as instances of Topic-OPA and thereby they will be added as nodes to the ontology graph. Although, the instance of relations are added as hierarchical edges to the graph. Concerning the named entities associated to locations, they will be used later for contextualizing the articles (e.g. regional, local and international news).

For adding the instances, we took advantage of the properties instance of (P31) and occupation (P106) in Wikidata to select the appropriate classes in Topic-OPA (for the same reason explained in section 5.4.1). For example, in Wikidata, John Simon (Q352) is an instance of Human (Q5) and related, by field of occupation (P245), to politician, jurist and lawyer. Therefore, in Topic-OPA, John Simon is instance of Politician ⊓ Jurist ⊓ Lawyer.

6. The Topic Labeling Process

In this section, we define the topic labeling process which is based mainly on $Rel_{Topic}$ and Topic-OPA. Given an article $A$ represented by a set of non ambiguous named entities $N$, the topic labeling process of $A$ is composed of three main phases: (1) assign $N$ as instances of Topic-OPA, (2) apply an instance-topic mapping process and (3) rank and select the best topics to label $A$. 

6.1. Named Entities As Instances of Topic-OPA

The named entities are categorized in: persons, locations, organizations and products. For the labeling process, we are interested mainly in: persons, organizations and products. The named entities of the type locations will be used in further works for contextualizing the articles. The disambiguated named entities will be assigned as instances of Topic-OPA and thereby they will be added as nodes to the ontology graph. Although, the instance of relations are added as hierarchical edges to the graph. Concerning the named entities associated to locations, they will be used later for contextualizing the articles (e.g. regional, local and international news).

For adding the instances, we took advantage of the properties instance of (P31) and occupation (P106) in Wikidata to select the appropriate classes in Topic-OPA (for the same reason explained in section 5.4.1). For example, in Wikidata, John Simon (Q352) is an instance of Human (Q5) and related, by field of occupation (P245), to politician, jurist and lawyer. Therefore, in Topic-OPA, John Simon is instance of Politician ⊓ Jurist ⊓ Lawyer.

6.2. Instance-Topic Mapping: Classification of Topics

Let us consider again the article $A$, which is represented by a set of instances $I$, and $T$ a set of topic concepts from Topic-OPA, the instance-topic mapping process is performed as a binary classification process between $I$ and $T$. For each $(i,t)$, $\forall i \in I \ and \ \forall t \in T$, we evaluate if $t$ is a relevant topic for $i$ or not. For this purpose, we apply $Rel_{Topic}$ that, as evoked earlier, returns a numerical relatedness value $\in [0,1]$ for each couple $(i,t)$. For classifying the results, there is a need to fix a threshold. In this context, an ideal threshold is the average of all the relatedness values $Rel_{Topic}(i,T)$. Therefore, we consider $t$ is relevant to $i$ if $Rel_{Topic}(i,t) \geq Rel_{Topic}(i,T)$.

6.3. Ranking and Selection of Labeling Topics

The ranking and selection of labeling topics is accomplished based on the results of the instance-topic mapping process. For $A$, $\forall i \in I, \exists T_i \subset T, \forall t \in T_i$, $Rel_{Topic}(i,t) \geq Rel_{Topic}(i,T)$. The matter now is to rank the topics according to these values and select the most relevant topic(s) $T_k \subset T_i$ for labeling $A$. For this purpose, we define the following procedure:
1. eliminate the most abstract topic concepts such as, Entity, Occurrent and Knowledge, by considering their depths. In Topic-OPA, the depths of these concepts are less than the average of the depths of all the topic concepts.

2. eliminate the topic concepts that are hypernyms of the named entities. For instance, by referring to A1, John Simon is a Politician, thereby concepts such as Professional, Worker, Person, Agent and Individual are eliminated.

3. eliminate the topic concepts that are hyponyms of Person, Organization, Product and Location. For instance, by referring to A1, Political Activist is related to the instance John Simon. However, Political Activist is not an hyponym of John Simon but a subClassOf Person. Thus, it will be eliminated being an hyponym of Person.

4. compute the most common topic concepts $T_c$ from $T_n = \sum T_i, \forall i \in I$.

5. compute the size of $T_c$.

6. if $|T_c| = 1$, then $T_c = \{t_c\}$ is the unique labeling topic of A.

7. otherwise, calculate the average of the semantic relatedness values $\bar{Rel}_{Topic}(i,t_c)$, for $Rel_{Topic}(i,t_c) \geq Rel_{Topic}(I,T), \forall t_c \in T_c, \forall i \in I$.

8. define two strategies to rank $T_c$ and to select the top-ranked topic(s) that label A: relatedness-guided and centrality-guided. The relatedness-guided strategy aims to select the most related topic concept(s) according to the average of the relatedness values. Meanwhile, the centrality-guided strategy tends to select the most connected topic concept(s) based on the degree centrality values. Thus, the further considers the content of A and the latter observes the semantic relevance of the topic concepts. By applying the dual strategy, we extend the scope of the selection of the best topics that label A.

(a) the relatedness-guided strategy is composed of:

i. ranking the topic concepts $t_c, \forall t_c \in T_c$ according to the average of the relatedness values $\bar{Rel}_{Topic}(i,t_c)$,

ii. selecting the topic concept(s) $t_r \in T_r \subseteq T_c$ having the highest value.

(b) the centrality-guided strategy is composed of:

i. computing the degree centrality of $t_c, \forall t_c \in T_c$,

ii. ranking the topic concepts $t_c, \forall t_c \in T_c$ according to their degree centrality,

iii. selecting the topic concept(s) $t_d \in T_d \subseteq T_c$ having the highest value.
9. finally, compute the topic labeling set of A, $T_{k} = T_{d} \cup T_{r}$, as a combination of the results of the centrality-guided and the relatedness-guided strategies.

7. Use-Case: Le Matin

In this section, we present a case study for labeling the articles of the old French newspaper Le Matin. For this purpose, we have chosen $A = \{A_1, A_2, \ldots, A_39\}$ a corpus of 48 articles published between 1910 and 1937, and $T$ a set of topics which represent all the topic concepts of Topic-OPA. In Figures 8 and 9, $\{A_1, A_2, \ldots, A_8\}$ a subset of $A$ is illustrated. Each article $A_i \in A$ is represented by a set of named entities $N_i$. The disambiguated named entities are represented by URIs extracted from the open knowledge graph Wikidata (see Figure 10). Our main goal is to label automatically the articles by the application of our proposed semantic relatedness measure $Rel_{Topic}$.

In order to achieve the goal, we need to construct the topic ontology Topic-OPA from these articles. Furthermore, the following processes are performed: (1) the assignment of the named entities as instances of Topic-OPA, (2) the instance-topic mapping process and (3) the ranking and selection process.

7.1. Topic-OPA

For Building Topic-OPA, a set of $N = 392$ named entities representing $A$ is considered and the SPARQL-based automatic approach (see section 5.4) is applied. As a result, we obtained a topic ontology, as a subset of Wikidata, which is accessible and manageable in ontology editors such as Protégé.²

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 Figures 11, 12 and 13, we depict excerpts of Topic-OPA around the Politics, Medicine and Sport topics. The solid lines represent the SubClassOf relations and the dashed lines represent the non-hierarchical relations.

7.2. Assignment of Disambiguated Named Entities as Instances

For each article $A_i \in A$, the disambiguated named entities are assigned as instances of Topic-OPA. Therefore, $\forall A_i \in A$, $A_i$ is represented by a set of instances $I_i$. In Table 1, we show the assignment of the named concepts representing the articles $\{A_1, A_2, \ldots, A_8\}$.

7.3. Instance-Topic Mapping

The instance-topic mapping process is performed between each article $A_i \in A$, which is represented by a set of instances $I_i$, and $T$ the set of topic concepts of Topic-OPA. The process is executed as a binary classification process between $I_i$ and $T$. For each $(i, t)$, $\forall i \in I_i$ and $\forall t \in T$, we evaluate if $t$ is a relevant topic for $i$ or not. For this purpose, we apply $Rel_{Topic}$ that takes as inputs all the instances $i \in I_i$ and the topic concepts of $t \in T$. In order to classify the results, we need to apply the specified threshold which is the average of all the relatedness values $Rel_{Topic}(i, T)$.

However, since Topic-OPA is not curated, it contains a huge number of general concepts. This means that the average of the relatedness values is low (around 0.28). Such low value of the threshold makes the overall performance of the classification process be degraded. Experimentation has shown that a threshold of about 0.5 provides good and relevant results. Therefore, we propose to use $threshold(A_i) = -log(Rel_{Topic}(i, T))$, in order to shift the average value of the threshold to the interesting range.

For instance, by referring to the articles $A_7$ and $A_8$, the averages of the relatedness values are $Rel_{Topic}(I_7, T) = 0.26$ and $Rel_{Topic}(I_8, T) = 0.30$. Hence, the threshold values are: $threshold(A_7) = -log(0.26) = 0.55$ and $threshold(A_8) = -log(0.30) = 0.52$. By applying these threshold values, we seek to select the most related topic concepts for each article. Therefore, we consider $t$ is relevant to $i$ if $Rel_{Topic}(i, t) \geq -log(Rel_{Topic}(i, T))$.

Table 2 shows the experimental results of the mapping process of $A_7$ to Topic-OPA. In this table, an excerpt of the instances, the relevant topics and the relatedness values, $Rel_{Topic}(i, t) \geq -log(Rel_{Topic}(I_i, T)) \forall i \in I_7$ and $\forall t \in T$, are presented.

7.4. Ranking and Selection of Labeling Topics

Given a set of relevant topics for each instance $i \in I_i$ representing an article $A_i \in A$, a ranking and selec-

²https://protege.stanford.edu/, last visited 23 July 2020
Fig. 8. Example of articles from *Le Matin*:

(a) *A*₁

(b) *A*₂

(c) *A*₃

(d) *A*₄
EN AVION pour l’Indo-Chine et Tokio
PIVOLO-GONIN-CAROL
(entré à 1, 300 kilomètres)
empris le départ avec le départ du Bourget
pour le tour aérien de l’Asie.

LA VACCINATION contre la tuberculose
Une controverse scientifique à l’Académie de médecine
Le professeur Galletto précise les résultats acquis.

LE TOUR DE FRANCE CYCLISTE
L’avant-dernière étape
Nantes-Vire-Caen
avec une petite course "contre la montre"

LE Drame du DOLLAR
La mission française à bord de l’« Île-de-France»
estime que la conférence économique mondiale
devient impossible par suite de l’abandon
de la France par les États-Unis de l’étalon-or

Fig. 9. Example of articles from the selected corpus of Le Matin.
Fig. 10. Example of named entities extracted from \{A_1, A_2, \ldots, A_8\}.
Fig. 11. Excerpt of Topic-OPA around the concept Politics.

Fig. 12. Excerpt of Topic-OPA around the concept Medicine.
tion process is performed in order to choose the best topic(s) for labeling $A_i$. This process is experimented on the 48 articles of Le Matin. Table 3 shows an excerpt of the experimental results. It presents thresholds, most common topics, average of the relatedness values, degree centrality, relatedness-guided topics and centrality-guided topics. The column Selected Topics indicates the best topics produced by $\text{Rel}_{\text{Topic}}$.

In the following, we describe the execution of the ranking and labeling procedure (see section 6.3) for $A_7$ (see Table 2). Note that step 1 is not shown in the present experimentation.

- By fulfilling step 2, the concepts Academy, National Academy, Learned Society, Physician, Health Professional, Immunologist, Medication, Vaccine, Biopharmaceutical and Disease are eliminated. For instance, Physician and Immunologist are eliminated being hypernyms of the instance Albert Calmette.

- Furthermore, concepts such as Physicist and Research Institute are eliminated by fulfilling step 3. Physicist is a hyponym of Person and Research Institute is a hyponym of Organization.

- The aim of step 4 is to compute the most common topics $T_c$ of $A_i$. For $A_7$, $T_c = \{\text{Science} \cap \text{Medicine} \cap \text{Physics} \cap \text{Bacteriology} \cap \text{Immunology} \cap \text{Virology} \cap \text{Vaccination}\}$. Thus, since $|T_c| = 7$ (step 5), step 6 is not executed for $A_7$. Meanwhile, it is implemented for $A_3$, $A_5$ and $A_8$ which are labeled by the topics Art, Aviation and Economics respectively.

- step 7 computes the average of relatedness values for each common topic concept $t_c \in T_c$.

- By achieving step 8 and step 9, $A_7$ is labeled by Vaccination as top-ranked topic having the highest average of relatedness ($\text{Rel}_{\text{Topic}}(I_7, \text{Vaccination}) = 0.69$) as well as the highest degree centrality ($D(\text{Vaccination}) = 13.48$).

Although, $A_2$ is labeled by the topic Military Affairs having the highest average of relatedness ($\text{Rel}_{\text{Topic}}(I_2, \text{Military Affairs}) = 0.67$) as well as by the topic War having the highest degree centrality ($D(\text{War}) = 22.22$).

In addition, $A_4$ and $A_6$ are labeled by dual topics by fulfilling step 8 and step 9. The topics Higher Education and Science are selected as best topics for labeling $A_4$. The topics Cycle Sport and Cycling are the top-ranked topics for labeling $A_6$.

8. Evaluation and Discussion

In this section, we evaluate the topic ontology Topic-OPA as well as the proposed semantic relatedness measure $\text{Rel}_{\text{Topic}}$. In addition, we compare $\text{Rel}_{\text{Topic}}$ to alternative graph-based measures.

8.1. Evaluation of Topic-OPA

In the literature, various approaches for the evaluation of ontologies have been recognized depending on what kind of ontologies are being evaluated and for
what purpose [30]. Generally, the ontology evaluation approaches are divided into four main categories [31]:

1. **gold standard-based**: which is known also as ontology alignment, aims to compare the developed ontology with a previously created reference ontology known as the gold standard. However, having a suitable gold ontology can be challenging, since it should be created under similar conditions with similar goals to the developed ontology.

2. **corpus-based**: tends to compare the developed ontology with the content of a text corpus that covers significantly a given domain. The basic approach is defined as follows: (1) perform an automated extraction of concepts and relations from the corpus and (2) apply a mapping between these concepts and relations and those of the developed ontology.

3. **application-based**: considers the evaluation of the ontologies that are intended for a particular application. Thus, a given ontology is only evaluated according to its performance in this application, regardless of all structural characteristics. Therefore, a “good” ontology is an ontology that helps to produce better results of a specific task.

### Table 1: Assignment of the named entities of the subset articles of A as instances of Topic-OPA.

<table>
<thead>
<tr>
<th>Article</th>
<th>Named Entity</th>
<th>Instance of</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>John Simon</td>
<td>Politician ⊓ Lawyer ⊓ Judge</td>
</tr>
<tr>
<td></td>
<td>Ramsay MacDonald</td>
<td>Politician ⊓ Journalist ⊓ Diplomat</td>
</tr>
<tr>
<td></td>
<td>Adolf Hitler</td>
<td>Politician ⊓ Soldier ⊓ Stateperson ⊓ Writer ⊓ Painter</td>
</tr>
<tr>
<td></td>
<td>Eric Phipps</td>
<td>Politician ⊓ Diplomat</td>
</tr>
<tr>
<td></td>
<td>Anthony Eden</td>
<td>Politician ⊓ Diplomat</td>
</tr>
<tr>
<td></td>
<td>Stanley Baldwin</td>
<td>Politician ⊓</td>
</tr>
<tr>
<td></td>
<td>Foreign Office</td>
<td>Foreign Affairs Ministry</td>
</tr>
<tr>
<td>A2</td>
<td>Miguel Primo de Rivera</td>
<td>Politician ⊓ Military Personnel</td>
</tr>
<tr>
<td></td>
<td>ABC</td>
<td>New York Times</td>
</tr>
<tr>
<td></td>
<td>Jean Bessin-Loisy</td>
<td>FilmDirector ⊓ FilmProducer ⊓ Screenwriter ⊓ Actor</td>
</tr>
<tr>
<td></td>
<td>Marie Epstein</td>
<td>FilmDirector ⊓ FilmProducer ⊓ Screenwriter ⊓ Actor</td>
</tr>
<tr>
<td></td>
<td>La Maternelle</td>
<td>Film</td>
</tr>
<tr>
<td></td>
<td>Pension Mimosas</td>
<td>Film</td>
</tr>
<tr>
<td></td>
<td>Simone Bertrau</td>
<td>FilmActor ⊓ Actor</td>
</tr>
<tr>
<td></td>
<td>Simone Bourdais</td>
<td>Actor</td>
</tr>
<tr>
<td></td>
<td>Sylvette Filaxier</td>
<td>Actor</td>
</tr>
<tr>
<td></td>
<td>Hubert Peler</td>
<td>Actor</td>
</tr>
<tr>
<td></td>
<td>Camille Bert</td>
<td>Actor</td>
</tr>
<tr>
<td></td>
<td>Roland Caillaux</td>
<td>Actor ⊓ Painter</td>
</tr>
<tr>
<td></td>
<td>Henri Debain</td>
<td>FilmActor ⊓ FilmDirector</td>
</tr>
<tr>
<td></td>
<td>Françoise Rosay</td>
<td>Actor ⊓ StageActor ⊓ FilmActor</td>
</tr>
<tr>
<td></td>
<td>Paul Appell</td>
<td>UniversityTeacher ⊓ Mathematician</td>
</tr>
<tr>
<td>A3</td>
<td>Academy of Toulouse</td>
<td>Academic District</td>
</tr>
<tr>
<td></td>
<td>Paris Academy</td>
<td>Academic District</td>
</tr>
<tr>
<td></td>
<td>Legion of Honour</td>
<td>Order</td>
</tr>
<tr>
<td>A4</td>
<td>Georges Polletier d’Oisy</td>
<td>AircraftPilot</td>
</tr>
<tr>
<td>A5</td>
<td>Rene Le Grèves</td>
<td>SportCyclist</td>
</tr>
<tr>
<td></td>
<td>Ambroise Morelli</td>
<td>SportCyclist</td>
</tr>
<tr>
<td></td>
<td>Romain Maesi</td>
<td>SportCyclist</td>
</tr>
<tr>
<td></td>
<td>Félicien Vertazecck</td>
<td>SportCyclist</td>
</tr>
<tr>
<td></td>
<td>Charles Pélissier</td>
<td>SportCyclist</td>
</tr>
<tr>
<td></td>
<td>Aldo Bertocco</td>
<td>SportCyclist</td>
</tr>
<tr>
<td>A6</td>
<td>Académie Nationale de Médecine</td>
<td>Academy ⊓ NationalAcademy</td>
</tr>
<tr>
<td></td>
<td>Albert Calmette</td>
<td>Physician ⊓ Bacteriologist ⊓ Immunologist ⊓ Virologist</td>
</tr>
<tr>
<td></td>
<td>BCG vaccine</td>
<td>Vaccine</td>
</tr>
<tr>
<td></td>
<td>Tuberculosis</td>
<td>Disease ⊓ NotifiableDisease ⊓ EndemicDisease</td>
</tr>
<tr>
<td></td>
<td>Charles Rist</td>
<td>Economist ⊓ Banker</td>
</tr>
<tr>
<td>A7</td>
<td>William H. Woodin</td>
<td>Politician ⊓ Businessperson</td>
</tr>
<tr>
<td></td>
<td>Trésor public</td>
<td>PublicTreasury</td>
</tr>
<tr>
<td></td>
<td>Bank of France</td>
<td>Bank ⊓ CentralBank ⊓ Business</td>
</tr>
<tr>
<td>A8</td>
<td>National Bank of Belgium</td>
<td>CentralBank</td>
</tr>
<tr>
<td></td>
<td>Paul van Zeeland</td>
<td>Economist ⊓ Politician ⊓ Lawyer ⊓ Diplomat ⊓ Jurist</td>
</tr>
</tbody>
</table>
Table 2
Excerpt of the instance-topic mapping process between $A_7$ and $T$.

<table>
<thead>
<tr>
<th>Instance (i)</th>
<th>Related Topic (t)</th>
<th>$\text{Rel}_	ext{topic}(i,t) \geq -\log(\text{Rel}_	ext{topic}(I_7,T))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Académie Nationale de Médecine</td>
<td>Research Institute</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Academic District</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Academy</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Learned Society</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>National Academy</td>
<td>0.72</td>
</tr>
<tr>
<td>Albert Calmette</td>
<td>Physicist</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Medicine</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Physician</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Health Professional</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Physic</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Immunologist</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Immunology</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Bacteriology</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Virology</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Medication</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Biopharmaceutical</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Vaccine</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Vaccination</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Disease</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Table 3
Ranking and selection of labeling topics.

<table>
<thead>
<tr>
<th>$A_i$</th>
<th>Threshold</th>
<th>Most Common Topics ($t_c$)</th>
<th>$\text{Rel}_	ext{topic}(I_7,t_c)$</th>
<th>Degree Centrality</th>
<th>Relatedness-Guided</th>
<th>Centrality-Guided</th>
<th>Selected Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>0.55</td>
<td>Politics</td>
<td>0.68</td>
<td>29.17</td>
<td>Politics</td>
<td>Politics</td>
<td>Politics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Political Activism</td>
<td>0.56</td>
<td>6.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_2$</td>
<td>0.55</td>
<td>Military Affairs</td>
<td>0.67</td>
<td>6.94</td>
<td>Military Affairs</td>
<td>War</td>
<td>Military Affairs-War</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Political Activism</td>
<td>0.62</td>
<td>6.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>War</td>
<td>0.59</td>
<td>22.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_3$</td>
<td>0.59</td>
<td>Art</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Art</td>
</tr>
<tr>
<td>$A_4$</td>
<td>0.52</td>
<td>Higher Education</td>
<td>0.58</td>
<td>15.28</td>
<td>Higher Education</td>
<td>Science</td>
<td>Higher Education-Science</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Science</td>
<td>0.55</td>
<td>23.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_5$</td>
<td>0.61</td>
<td>Aviation</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Aviation</td>
</tr>
<tr>
<td>$A_6$</td>
<td>0.55</td>
<td>Cycle Sport</td>
<td>0.68</td>
<td>13.20</td>
<td>Cycle Sport</td>
<td>Cycling</td>
<td>Cycle Sport-Cycling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cycling</td>
<td>0.59</td>
<td>27.38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sport</td>
<td>0.55</td>
<td>13.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_7$</td>
<td>0.58</td>
<td>Vaccination</td>
<td>0.69</td>
<td>13.48</td>
<td>Vaccination</td>
<td>Vaccination</td>
<td>Vaccination</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacteriology</td>
<td>0.64</td>
<td>7.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Immunology</td>
<td>0.64</td>
<td>7.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medicine</td>
<td>0.58</td>
<td>7.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Virology</td>
<td>0.64</td>
<td>7.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_8$</td>
<td>0.51</td>
<td>Economics</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Economics</td>
</tr>
</tbody>
</table>

4. criteria-based: quantifies how far an ontology adheres to certain desirable criteria. For instance, computing structure-based properties such as the size and the complexity of a given ontology. Although, approaches such as [31] study measures such as the average taxonomic depth and the relational density of nodes.

In order to choose the “best” evaluation approach, there is a need to define the motivation behind evaluating a developed ontology [31]. In our study, as evoked
earlier, Topic-OPA is an application-based ontology that is intended to be used in a topic labeling system for classifying and labeling a given set of old press articles. Therefore, the labeling process is affected by two main factors: (1) the ranking and labeling algorithms and (2) the topic ontology being used as knowledge base. In this context, we propose to evaluate Topic-OPA using a dual evaluation approach: application-based and criteria-based. Our decision is founded on the following assumptions:

- 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 for our study: it implies to evaluate the usability of Topic-OPA being an application-based ontology.
- the criteria-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 [32].

### 8.1.1. Application-Based Evaluation

Topic-OPA is employed in the topic labeling system of the old press articles by using it as a knowledge base. Technically, the semantic relatedness measure $Rel_{Topic}$ is applied on the graph structure of Topic-OPA. $Rel_{Topic}$ performs a “browsing” of the hierarchical and the non-hierarchical structure of Topic-OPA. It inspects nodes and edges, their properties, such as weights and depths, as well as the correlation of nodes which is defined by the degree centrality. Therefore, the results obtained by using $Rel_{Topic}$ for the classification and the labeling tasks determine the feasibility of Topic-OPA. For this purpose, the application-based evaluation of Topic-OPA is a function of the evaluation of $Rel_{Topic}$ (see section 8.2). Therefore, Topic-OPA is considered as a “good” ontology if the results obtained by using $Rel_{Topic}$ are accurate.

### 8.1.2. Structure-Based Evaluation

The structure-based evaluation aims to assess the quality of Topic-OPA. 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.) [31]. 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. Actually, the Knowledge coverage and popularity measures, which are commonly used in the ontology evaluation literature, do not show a significant relationship with the ontological accuracy [33]. However, the structural measures are positively correlated with the semantic accuracy of the knowledge modeled in the ontology [33].

In the context of Topic-OPA, we quantify some structural measures, by considering the taxonomic structure of Topic-OPA, as follows:

- **maximum depth**: represents the length of the longest taxonomic branch in the ontology. It is measured as the number of concepts from the root node to the leaves of the taxonomy. In Topic-OPA, $\text{maximumdepth} = 28$.
- **average depth**: is the average length of all taxonomic branches. In Topic-OPA, $\text{averagedepth} = 6$.
- **depth variance**: is the dispersion with respect to the average depth, computed as the standard mathematical variance. In Topic-OPA, $\text{depthvariance} = 3.88$ which is almost equivalent to the average depth.

We conclude that the majority of the topic concepts within Topic-OPA are dispersed homogeneously within the core level. This implies two essential points: (1) it will be a challenging task to $Rel_{Topic}$ to distinguish between the different concepts that are located at the same depth in order to select the best ones as labeling topics and (2) in a semantic context, the hierarchical structure of Topic-OPA is a balanced taxonomy, in which the majority of taxonomic edges have almost the same depth.

### 8.2. Evaluation of $Rel_{Topic}$

The evaluation of $Rel_{Topic}$ consists in measuring how well this measure can label a given article. For this purpose, we apply a dual evaluation approach com-
posed of: a quantitative evaluation, which consists in comparing the automatic labeling to human labeling [13] and (2) a qualitative evaluation which aims to appraise the generated topics regarding their semantic interpretability [28].

8.2.1. Quantitative Evaluation

For evaluating the results obtained by $Rel_{Topic}$, a quantitative evaluation is used by considering human labeling [13]. The human labeling task consists in involving human annotators to label each article $A_i \in A$ with a unique or multiple topics. The human annotators, which were blind to the topics of Topic-OPA as well as to the results generated by $Rel_{Topic}$, have read the articles and assigned the labeling topics based on their subject and content. Furthermore, regarding the human labeling topics, we evaluated the top-ranked topics provided automatically by $Rel_{Topic}$ by classifying them into two main categories: Good and Not Related. The Good category includes the scores Exact, General and Specific for classifying the labels generated by $Rel_{Topic}$.

Meanwhile, the Not Related category comprises the topics that are not relevant to the manually assigned topics. Out of 48 articles from Le Matin, 9 articles are labeled with Not Related topics and 39 are labeled with Good labels. In the Good category, 25 articles are labeled with Exact labels and 14 articles are labeled with either General or Specific labels. These results imply that our method is globally performant with a precision = 0.81. In the following, we discuss two main issues that affected the relevance of the generated labels: (1) the existence of not disambiguated named entities and (2) the typology of the named entities.

The Influence of the Existence of Not Disambiguated Named Entities on the Labeling Results

In the presented use-case (see section 7), 20 articles have been represented by some named entities that are not disambiguated (i.e. $A_5$, $A_7$). In this section, we discuss the influence of these named entities on the relevance of the automatically generated labeling topics. For this purpose, we consider two cases: (1) the labels having a General score according to the human labeling and (2) the labels that are Not Related.

Concerning the first case, we analyzed two articles $A_7$ (see Figure 9) and $A_9$ (see Figure 14). Article $A_7$ consists of 5 disambiguated named entities and 2 that are not disambiguated (see figure 10). Despite this, $Rel_{Topic}$ assigned a Specific labeling comparing to human labeling (Medicine) by selecting Vaccination as best topic (see Table 3). Let’s now consider article $A_9$ which consists of 10 disambiguated named entities and 2 that are not disambiguated (see Figure 15). By the application of $Rel_{Topic}$, $A_9$ is labeled by Science (see Table 4). The generated topic has been given a General score regarding the topic Medicine that is assigned by the human annotators.

In table 4, we show that Science is selected as unique most common topic for labeling $A_9$ (see step 6 of the ranking and selection procedure). By surveying the results of the instance-topic mapping phase and the computation of the common related topics, we found that Medicine is commonly related 8 times. Meanwhile, Science is commonly related 10 times.

In addition, we have inspected the named entities that are not disambiguated in $A_9$ (Robert Wilbert and/or Marcel Léger). Robert Wilbert is a Veterinarian and Marcel Léger is an Epidemiologist, Microbiologist and Bacteriologist. We conclude that the existence of these not disambiguated named entities has eliminate Medicine from the most common topics set. Thereby, they have affected the relevance of the topic labeling of $A_9$.

---

https://journals.openedition.org/primatologie/2816?lang=en#ftn1

https://www.pathexo.fr/documents/notices/leger.html

Fig. 14. Excerpt from $A_9$, Le Matin 1924, June 27.

4 https://journals.openedition.org/primatologie/2816?lang=en#ftn1

5 http://www.pathexo.fr/documents/notices/leger.html?width=800height=500
The Influence of the Typology of the named entities on the Relevance of the Topic Labeling  

As evoked earlier, in this study, we are interested in three main types of named entities: person, organization and product.

In this section, we discuss the influence of the typology of the named entities on the relevance of the topic labeling. Specifically, we address the articles that are evaluated with Not Related scores. For instance, article A10 (see Figure 16) is composed of 6 persons and 2 products (see Figure 17) and the majority of persons are politicians (see Table 5).

In Table 6, we present the experimental results of the instance-topic mapping process of A10. Table 7 shows that A10 is labeled by the unique most common topic Politics. However, based on the content and the subject of A10, the human annotators have assigned the topic Economics. In this context, we recognize that the majority of politicians with the absence of organizations or persons related to economics have affected the pertinence of the labeling results.

Fig. 16. Example of article (A10) from Le Matin 1922, June 20
Table 6
Excerpt of the instance-topic mapping process between $A_{10}$ and $T$.

<table>
<thead>
<tr>
<th>Instance (i)</th>
<th>Related Topic (t)</th>
<th>$Rel_{Rel}(i,t) \geq Rel_{Rel}(i,T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>César Caire</td>
<td>Law</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Jurisprudence</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Jurist</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Lawyer</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Politics</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Political Activist</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Journalism</td>
<td>0.69</td>
</tr>
<tr>
<td>Henri Galli</td>
<td>Politics</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Political Activist</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Journalism</td>
<td>0.69</td>
</tr>
<tr>
<td>Ambroise Renda</td>
<td>Politics</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Political Activist</td>
<td>0.68</td>
</tr>
<tr>
<td>Alexandre Luquet</td>
<td>Politics</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Political Activism</td>
<td>0.61</td>
</tr>
<tr>
<td>Flour</td>
<td>Food</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Ingredient</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Food Ingredient</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Cooking</td>
<td>0.66</td>
</tr>
<tr>
<td>Wheat</td>
<td>Food</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Ingredient</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Food Ingredient</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Cooking</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 7
Ranking and selection of labeling topics for $A_{10}$.

<table>
<thead>
<tr>
<th>Article</th>
<th>Threshold</th>
<th>Most Common Topics</th>
<th>Selected Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{10}$</td>
<td>0.58</td>
<td>Politics</td>
<td>Politics</td>
</tr>
</tbody>
</table>

8.2.2. Qualitative Evaluation

The qualitative evaluation assesses the labeling topics generated by $Rel_{Topic}$ according to their semantic quality [28]. In linguistics, the topic, or theme, of a sentence is what is being talked about. In a semantic context, defining a labeling topic within topic ontologies is not an easy task. In fact, a topic ontology consists of various concepts including the labeling topics. Meanwhile, it is difficult to find or define these topics. In our experiment, by the application of $Rel_{Topic}$ for labeling the old press articles (see Table 3), we perceived three essential characteristics that define the semantic quality of a labeling topic:

- **highly correlated**: a concept with high degree centrality designates a large surface of connection with the concepts within the ontology. For instance, Politics, War, Science, Art and Sport have respectively 29.17, 22.22, 23.62, 31.34 and 13.89 values of degree centrality. Meanwhile, concepts such as Activity, Occupation and Group Behaviour have respectively 8.68, 9.81 and 7.63 values of degree centrality.

- **core concept**: the depth of concepts in ontologies indicates their degree of generality. In Topic-OPA, abstract concepts, such as Entity, Agent, Object, Product and Occurrence are located at depths less than the average of depths in Topic-OPA which is equal to 4 (i.e. $depth(Entity)=1$, $depth(Object)=2$ and $depth(Occurrence)=3$). These concepts are not recommended as labeling topics due to their abstraction interpretability. Meanwhile, the majority of the labeling topics that are produced by our relatedness measure (i.e. Politics, Art, Science, etc.) are located at depths greater than or equal to the average of depths in Topic-OPA (i.e. $depth(Politics)=5$, $depth(Art)=4$ and $depth(Science)=5$). Although, these topics are more general than the specific concepts (i.e. Contract Law, Pharmacy, etc.) which are located at higher depths (i.e. $depth(Contract Law)=7$ and $depth(Pharmacy)=9$).

- **not a hypernym of named entities**: a labeling topic is not linked hierarchically to the named entities. Therefore, it is not a subclass of Person, Organization, Location or Product.
8.3 Comparison of RelTopic with Alternative Graph-Based Measures

In this section, we compare RelTopic with alternatives graph-based measures. Specifically, we choose path-based measures since node-based measures are dependent on textual resources which are out of scope of our study. In this context, path-based measures such as SimRada (see Equation (1)) and SimLC (see Equation (2)) are only applicable to taxonomies. Meanwhile, RelHS (see Equation (3)) is the most appropriate since it is a relatedness measure applicable in ontologies. However, applying RelHS is not an easy task due to the difficulty of the computation of the direction changes of edges (hierarchical and non-hierarchical) through all the paths. For this purpose, and since RelTopic is based on the computation of shortest paths (see Equation (8)), we selected SimRada for the comparison. In this regard, we applied SimRada to the whole graph of Topic-OPA including the hierarchical and non-hierarchical schemes. Thereby, we compared the results of the application of the instance-topic mapping process on $A$. In Table 8, we show an excerpt of the results of mapping $A$ to Topic-OPA. The results imply that the related topic concepts to a given article $A_i \in A$ are clearly identified by RelTopic as well as by SimRada. However, the use of RelTopic makes also evident the identification of the topics that are not related to $A_i$ due to the considerable gap among the relatedness values (see Figure 18 for an example).

9. Conclusion

In this study, we addressed the problem of labeling old press articles by proposing a novel semantic relatedness measure, named RelTopic, within topic ontologies. In contrast to existing measures, RelTopic considers non-hierarchical relations and assesses the relatedness between instances and concepts. In order to apply RelTopic, we considered topic ontologies as weighted graphs where nodes and edges are given positive numerical weights. In addition, the measure takes into consideration the degree centrality of nodes which reflects the level of connection of the node with regards to the rest of the ontology. For the application of RelTopic, there is a need for a topic ontology, named Topic-OPA, that expresses the domain of old press articles.

For building Topic-OPA, a SPARQL-based fully automatic approach is applied to derive the ontology from Wikidata. This approach is grounded on a set of “disambiguated” named entities extracted from the set of articles to be labelled. A use-case of 48 articles, in the context of the old French newspaper Le Matin, is also presented. We developed Topic-OPA from the named entities representing these articles and have applied RelTopic for the topic labeling. By analyzing the automatically generated topics, 81% are considered as “good” compared to the topics given manually by human annotators. These results are encouraging, because they mean that RelTopic has been able to correctly choose the “good” topics in Topic-OPA, despite its size and the fact that its structure makes that almost all the candidate topics are at the same level of abstraction.

For the evaluation process, we evaluated the topic ontology Topic-OPA as well as the relatedness measure RelTopic. Topic-OPA is evaluated using a dual evaluation approach composed of application-based and structure-based approaches. The application-based evaluation is a function of the evaluation of the results of the topic labeling task. Meanwhile, the structure-based evaluation revealed the homogeneous dispersion of the concepts in Topic-OPA.

For RelTopic, we have used also a dual evaluation approach composed of a quantitative and qualitative assessment. On the one hand, with the help of human annotators, which read and labeled the articles based on their content, we have compared the automatically generated results to human labeling. Two main categories of scores are defined: Good and Not Related. The Good category implies the Exact, General and Specific labels that are selected as best topics by RelTopic. Meanwhile, the not related category defines the topics that are not relevant to those given by the annotators.

Furthermore, we discussed the issues that affected the relevance of the automatic labeling process. Specifically, we addressed the existence of some named entities that are not disambiguated. We analyzed also the problem of the typology of the named entities and its influence on the quality of the generated labeling topics. On the other hand, a qualitative evaluation approach is prescribed to assess the quality of topics in a semantic context. In this approach, we relied mainly on the degree centrality of topic concepts as a basic indicator.

In addition, we considered the depths of the labeling topics which reflect their level of generality. Thus, a labeling topic within a topic ontology is considered as a core concept located at a generality level between
Table 8

Excerpt of the results of the instance-topic mapping process of \( A_7 \) to \( T \).

<table>
<thead>
<tr>
<th>Instance (i)</th>
<th>Topic Concepts (t)</th>
<th>( \text{Rel}_{\text{Topic}}(i,t) )</th>
<th>( \text{Sim}_{\text{Rada}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Académie Nationale de Médecine</td>
<td>Research Institute</td>
<td>0.60</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>0.30</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Academic District</td>
<td>0.69</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Academy</td>
<td>0.80</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Economics</td>
<td>0.37</td>
<td>0.1</td>
</tr>
<tr>
<td>Albert Calmette</td>
<td>Physicist</td>
<td>0.69</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Medicine</td>
<td>0.58</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Physician</td>
<td>0.73</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Business</td>
<td>0.31</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Economics</td>
<td>0.33</td>
<td>0.125</td>
</tr>
<tr>
<td>BCG vaccine</td>
<td>Medication</td>
<td>0.58</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Vaccine</td>
<td>0.75</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Vaccination</td>
<td>0.69</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Economics</td>
<td>0.36</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>Business</td>
<td>0.33</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Disease</td>
<td>0.67</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Health Problem</td>
<td>0.5</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Work of Art</td>
<td>0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Fig. 18. Comparison of the results of mapping Albert Calmette to topics in \( T \) (a) \( \text{Rel}_{\text{Topic}} \) (b) \( \text{Sim}_{\text{Rada}} \)

the general and the specific levels. Finally, we compared \( \text{Rel}_{\text{Topic}} \) to alternative path-based semantic measures such as \( \text{Sim}_{\text{Rada}} \). The comparison showed that both measures, \( \text{Rel}_{\text{Topic}} \) and \( \text{Sim}_{\text{Rada}} \), can identify the relevant topics but the use of \( \text{Rel}_{\text{Topic}} \) makes also evident the identification of the not related topics.

In future works, we will be interested in the contextualisation of the articles taking into account the named entities of type location (i.e. \( A_1 \) could be labeled with International Politics, \( A_3 \) with Local or French Art and \( A_6 \) with French Sport). In addition, we will try to resolve the identified problems related to the existence of "non disambiguated" named entities with the goal of improving the accuracy of the whole labeling process.

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 work, we will try to apply a curation process aiming to clean and leverage Topic-OPA. Furthermore, we will study the application of \( \text{Rel}_{\text{Topic}} \) on the leveraged version of Topic-OPA and analyze the quality of the generated labeling topics.

Acknowledgments

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

References


