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

Mirna El Ghosh [a,*], Nicolas Delestre [a], Jean-Philippe Kotowicz [a], Cecilia Zanni-Merk [a] and Habib Abdulrab [a]

[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_{Topic}, for assessing the relatedness of instances and topics. To design Rel_{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_{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_{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.

*Corresponding author. E-mail: mirna.elghosh@insa-rouen.fr.

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 Rel_{Topic}, 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 Rel_{Topic}, 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 Rel_{Topic} 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 Rel_{Topic}, a use-case is presented in the context of Le Matin [1], an old french newspaper first published in 1884 and discontinued in 1944. The topic ontology Topic-OPA and the semantic measure Rel_{Topic} 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 Rel_{Topic}. 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 *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*.


Le Matin
JEUDI 21 NOVEMBRE 1935
LA FRANCE est-elle une terre de colonisation ?
Courte audience au procès des oustachis
JOURNÉE D'AGITATION DANS LES MILIEUX POLITIQUES
Depardon et le mulâtre Darius rôles d'utilité dans l'affaire Stavisky
MONSIEUR LE PARLEUR
Sir George Clerk a été reçu par M. Pierre Laval
L'AMIRAL JELLICOE EST MORT

Fig. 1. Excerpt of *Le Matin*.

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.

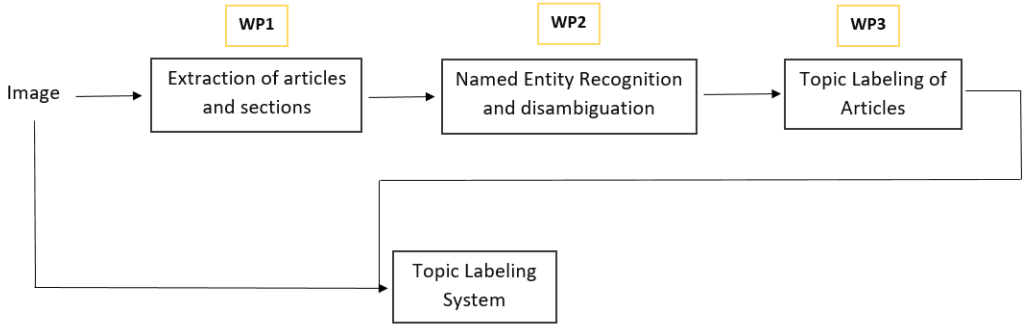


Fig. 2. The pipeline of the project ASTURIAS.

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:

$$Sim_{Rada}(c_1,c_2) = \frac{1}{1+dist_{Rada}(c_1,c_2)}, \qquad (1)$$

where $dist_{Rada}$ is the *shortest path* and Sim_{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]:

$$Sim_{LC}(c_1,c_2) = -log(\frac{len(c_1,c_2)}{2\times maxdepth(c)}), \qquad (2)$$

where $len(c_1,c_2)$ is the *shortest path* between c_1 and c_2 and $maxdepth(c)$ is the maximum depth of $c, \forall c \in WordNet$.
In this category of measures, Hirst and St-Onge's measure, that considers the non-taxonomic links, is identified [9]:

$$Rel_{HS}(c_1,c_2) = C - len(c_1,c_2) - k\times turns(c_1,c_2), \qquad (3)$$

where C and k are constants ($C=8$ and $k=1$), and $turns(c_1,c_2)$ 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:

$$Sim_{Resnik}(c_1, c_2) = IC(LCS(c_1, c_2)), \quad (4)$$

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

$$IC(c) = -logP(c) \quad (5)$$

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

$$Sim_{Lin}(c_1, c_2) = \frac{2 \times Sim_{Resnik}(c_1, c_2)}{IC(c_1) + IC(c_2)} \quad (6)$$

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 discusses 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 pair 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, part of, 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 **hypernyms** $H(v_i)$ 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 $\tau(e_m) = \{subclassOf\} \vee \{instanceof\}, e_m \in \{e_h, ..., e_k\}$.*

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, e_i,..., v_k, e_k, v_{k+1}, e_{k+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. $|P(v_i \rightarrow v_j)| = \sum_{e_i \in E(P)} \delta(e_i)$.*

Definition 6. *The **distance** $dist(v_i \rightarrow v_j)$ 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 Rel_{Topic}

For designing Rel_{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

Inspired 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 $\omega(v_i) = -log(\frac{|N(v_i)|}{|G|})$. For the instances, two main cases are identified:

1. v_i is an instance of a single hypernym 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 $H(v_i) = \{v_h, ...v_m\}$. Here, $\omega(v_i) = \overline{(\omega(v_n))}_{v_n \in H(v_i)}$, where $\overline{(\omega(v_n))}$ is the average of the weights of the hypernyms 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 $cost = 1$.
- 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:

$$\delta(e_i) = \begin{cases} 1, & \text{if } \tau(e_i) = subclassOf \vee instanceof \\ 0.25, & \text{otherwise} \end{cases} \tag{7}$$

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, $D(v_i) = \sum_{v_j \in N(v_i)} \delta(e_j) \times \omega(v_j)$, 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 single hypernym node. Thus, the degree centrality of nodes representing instances is formally defined by: $D(v_i) = \sum_{v_j \in N(v_h)} \delta(e_j) \times \omega(v_j)$, where v_h is the hypernym of v_i, $e_h = \{v_i, v_h\}$, $\tau(e_h) = \{instanceof\}$ and $e_j = \{v_j, v_h\}$.
2. v_i is an instance of multiple hypernym nodes. v_i instance of multiple hypernym nodes that are represented by $H(v_i) = \{v_h, ...v_m\}$, $D(v_i) = \overline{(D(v_n))}_{v_n \in H(v_i)}$, where $\overline{(D(v_n))}$ is the average of the degree centrality of the hypernyms 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 $dist(v_i \rightarrow v_j)$ (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 $\mathrm{P}(v_i \rightarrow v_j)$ is the path between v_i and v_j. The semantic relatedness measure takes these elements as input and returns a numerical description, $Rel_{Topic} \in [0,1]$, that quantifies their relatedness based on the following formula:

$$Rel_{Topic}(v_i, v_j) = \left(\frac{1}{1 + dist(v_i \rightarrow v_j)}\right) + k \times \left(\frac{\log(D(v_i) + D(v_j))}{\omega(v_i) + \omega(v_j)}\right), \tag{8}$$

where $dist(v_i \rightarrow v_j)$ 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:

$$k = \begin{cases} 1, & \text{if } \mathrm{P}(v_i \rightarrow v_j) \text{ is } semantically\ correct \\ 0, & \text{if } \mathrm{P}(v_i \rightarrow v_j) \text{ is } semantically\ incorrect \end{cases} \tag{9}$$

The *correctness* of the semantic path between two nodes is prescribed based on the constraints proposed in [9]. If a path $\mathrm{P}(v_i \rightarrow v_j)$ changes the direction from *upward* (generalization) to *downward* (specialization) at a point related to a hierarchical link, $\mathrm{P}(v_i$

$\rightarrow v_j$) is considered *semantically incorrect*. For instance, given a node v_k in P($v_i \rightarrow v_j$), where $\{v_{k-1}, e_k, v_k\}$ / $\tau(e_k) = \{subClassOf\}$ and $\{v_{k+1}, e_{k+1}, v_k\}$ / $\tau(e_{k+1}) = \{subClassOf\}$. Thereby, all the paths traversing the top of the ontology are penalized.

5. SPARQL-Based Automatic Approach for Building Topic Ontologies

For the application of Rel_{Topic}, 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* $O = \langle T, I, R, E, \phi \rangle$*, with*

- T *the set of topic concepts,*

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

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 $H = \langle T, R, E_{\sqsubseteq}, \phi \rangle$, 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 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 $S_T \subset T$ 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 = {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, $S_T(n)$={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 $q_H(t_i)/t_i \in S_T$ 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 5, an example of triples extracted based on S_T(John Simon).

H={Judge $\sqsubseteq$ Magistrate, Magistrate $\sqsubseteq$ Official $\sqcap$ Jurist, Official $\sqsubseteq$ Civil Servant, Civil Servant $\sqsubseteq$ Public Employee, Public Employee $\sqsubseteq$ Employee, Lawyer $\sqsubseteq$ Jurist, Politician $\sqsubseteq$ Professional}.

5.4.2. 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

```
<?xml version="1.0" encoding="UTF-8"?>
<Article id="A_1" year="1935" issue="March" day="7" page="1">
    <NE type="person" uri="http://www.wikidata/entity/Q333091" value="John Simon"></NE>
    <NE type="location" uri="http://www.wikidata/entity/Q183" value="Allemagne"></NE>
    <NE type="location" uri="http://www.wikidata/entity/Q84" value="Londres"></NE>
    <NE type="location" uri="http://www.wikidata/entity/Q64" value="Berlin"></NE>
    <NE type="person" uri="http://www.wikidata/entity/Q166646" value="Ramsay MacDonald"></NE>
    <NE type="person" uri="http://www.wikidata/entity/Q166635" value="Baldwin"></NE>
    <NE type="person" uri="http://www.wikidata/entity/Q352" value="Adolf Hitler"></NE>
    <NE type="organization" uri="http://www.wikidata/entity/Q58211956" value="Foreign Office"></NE>
</Article>
```

Fig. 3. Example of named entities extracted from article A_1 (see figure 8).

```
Selection of most specific concepts based on the named entities:
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
```

Fig. 4. Definition of the most specific concepts based on the named entities of A_1.

```
CONSTRUCT the hierarchy of topics starting from / judge / in a bottom-up strategy.
Add /http://www.wikidata.org/entity/Q212238(civil servant)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q212238(civil servant)/ ->  subClassOf  -> public employee
Add /http://www.wikidata.org/entity/Q16533(judge)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q16533(judge)/ ->  subClassOf  -> magistrate
Add /http://www.wikidata.org/entity/Q215627(person)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q185351(jurist)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q185351(jurist)/ ->  subClassOf  -> person linked to the law
Add /http://www.wikidata.org/entity/Q599151(official)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q599151(official)/ ->  subClassOf  -> civil servant
Add /http://www.wikidata.org/entity/Q4594605(magistrate)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q4594605(magistrate)/ ->  subClassOf  -> jurist
Add /http://www.wikidata.org/entity/Q4594605(magistrate)/ ->  subClassOf  -> official
Add /http://www.wikidata.org/entity/Q80363469(public employee)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q80363469(public employee)/ ->  subClassOf  -> employee
CONSTRUCT the hierarchy of topics starting from / lawyer / in a bottom-up strategy.
Add /http://www.wikidata.org/entity/Q40348(lawyer)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q40348(lawyer)/ ->  subClassOf  -> jurist
CONSTRUCT the hierarchy of topics starting from / politician / in a bottom-up strategy.
Add /http://www.wikidata.org/entity/Q82955(politician)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q82955(politician)/ ->  subClassOf  -> professional
```

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 { ?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 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).

NH={(Civil Servant,*field of this occupation*,Civil Service), (Politician,*field of this occupation*,Politics), (Lawyer,*field of this occupation*,Law), (Jurist,*field of this occupation*,Jurisprudence), (Judge ⊓ Magistrate ,*field of this occupation*,Judiciary), (Public Employee,*facet of*,Public Sector ⊓ Government)}

5.4.3. Ontology Enrichment

In this phase, an ontology enrichment process is performed 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 7).

H={Political Organization ⊑ Organization, Government ⊑ Political Organization, Judiciary ⊑ Authority, Civil service ⊑ Organization, Politics ⊑ Activity}

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 $\overline{Rel_{Topic}}(I,T)$. Therefore, we consider t is relevant to i if $Rel_{Topic}(i,t) \geq \overline{Rel_{Topic}}(I,T)$.

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, $\forall i \in I$, $\exists T_i \subset T$, $\forall t \in T_i$, $Rel_{Topic}(i,t) \geq \overline{Rel_{Topic}}(I,T)$. The matter now is to rank the topics according to these values and select the most relevant topic(s) $t_k \in T_k \subset T_i$ for labeling A. for this purpose, we define the following procedure:

```
CONSTRUCT the non-hierarchical scheme of the ontology
Add /http://www.wikidata.org/entity/Q212238(civil servant)/-> field of this occupation-> civil service
Add /http://www.wikidata.org/entity/Q82955(politician)/-> field of this occupation-> politics
Add /http://www.wikidata.org/entity/Q40348(lawyer)/-> field of this occupation-> court proceeding
Add /http://www.wikidata.org/entity/Q40348(lawyer)/-> field of this occupation-> law
Add /http://www.wikidata.org/entity/Q185351(jurist)/-> field of this occupation-> jurisprudence
Add /http://www.wikidata.org/entity/Q16533(judge)/-> field of this occupation-> judiciary
Add /http://www.wikidata.org/entity/Q4594605(magistrate)/-> field of this occupation-> judiciary
Add /http://www.wikidata.org/entity/Q80363469(public employee)/-> facet of-> public sector
Add /http://www.wikidata.org/entity/Q80363469(public employee)/-> facet of-> government
```

Fig. 6. Example of triples for building the non-hierarchical scheme of Topic-OPA.

```
ENRICH the ontology with new topics
Add /http://www.wikidata.org/entity/Q7188(government)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q7188(government)/ ->  subClassOf  -> political organisation
Add /http://www.wikidata.org/entity/Q43229(organization)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q43229(organization)/ ->  subClassOf  -> agent
Add /http://www.wikidata.org/entity/Q488383(object)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q7210356(political organisation)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q7210356(political organisation)/ ->  subClassOf  -> organization
Add /http://www.wikidata.org/entity/Q105985(judiciary)/ ->  AS  ->  Class
Add /http://www.wikidata.org/entity/Q105985(judiciary)/ ->  subClassOf  -> authority
```

Fig. 7. Example of the enrichment of the hierarchical scheme of Topic-OPA.

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 A_1, 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 A_1, Political Activist is related to the instance John Simon. However, Political Activist is not an hypernym 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 $\overline{Rel_{Topic}}(i,t_c)$, 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 $\overline{Rel_{Topic}}(i,t_c)$,
      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 $A = \{A_1, A_2, ..., A_{48}\}$ 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, $\{A_1, A_2, ..., A_8\}$ 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[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 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.

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

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 entities representing the articles $\{A_1, A_2, ..., A_8\}$.

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 $\overline{Rel_{Topic}}(I_i, T)$.

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 $threshold(A_i) = -log(\overline{Rel_{Topic}}(I_i, T))$, 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 $\overline{Rel_{Topic}}(I_7, T) = 0.26$ and $\overline{Rel_{Topic}}(I_8, T) = 0.30$. 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 $Rel_{Topic}(i,t) \geq -log(\overline{Rel_{Topic}}(I_i, T))$.

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, $Rel_{Topic}(i,t) \geq -log(\overline{Rel_{Topic}}(I_7, T)) \forall i \in I_7 and \forall t \in T$, 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-

APRES L'AJOURNEMENT DU VOYAGE DE SIR JOHN SIMON EN ALLEMAGNE

La situation reste aussi obscure à Londres qu'à Berlin

On parle maintenant de la venue possible en Angleterre d'un émissaire allemand

[DE NOTRE CORRESPONDANT PARTICULIER]

LONDRES, 6 mars. — *Par téléphone.* — M. Ramsay MacDonald, souffrant encore du léger refroidissement qu'il avait contracté lundi, n'a pas présidé l'important conseil de cabinet qui s'est réuni ce matin.

Les débats ont donc été dirigés par M. Baldwin, lord président du conseil. Il a été consacré presque exclusivement à la situation créée par l'ajournement du voyage à Berlin de Sir John Simon et de M. Eden, à la requête du gouvernement allemand, lequel, on le sait, a donné pour unique raison une indisposition du Reichsführer.

Comme le ministre des affaires étrangères britannique l'a déclaré plus tard à la Chambre des communes, il n'a reçu depuis la publication du communiqué du Reich aucune information officielle de Berlin quant à l'état de M. Adolf Hitler ou quant à ses instructions. Le Foreign Office reste naturellement en communication constante à ce sujet avec Sir Eric Phipps, ambassadeur de Grande-Bretagne à Berlin, mais en l'absence de nouvelles, aucune décision ne peut être prise du côté anglais.

C'est dire que quelles que soient les vues exprimées individuellement par les ministres sur l'indisposition inopportune du Reichsführer et sa coïncidence avec la publication du *Livre blanc* sur le désarmement, le conseil de cabinet a surtout étudié aujourd'hui les autres visites projetées de Sir John Simon destinées à poursuivre l'œuvre entamée à Rome et à Londres. Sir John Simon ainsi que ses collègues se sont occupés particulièrement de l'invitation du gouvernement des Soviets mais il a été décidé que la démarche à Berlin devait nécessairement précéder toute autre mission.

(a) A_1

Le général Primo de Rivera rentre à Madrid

Primo de Rivera, qui est arrivé hier à Algésiras où il a reçu un accueil enthousiaste, va se rendre à Madrid pour présider les fêtes qui seront données à l'occasion de la saint Alphonse et qui revêtiront le caractère d'un hommage éclatant au roi.

Les Madrilènes reconnaîtront difficilement le jovial Andalou de jadis, à la gaieté parfois turbulente. La photographie que nous reproduisons ci-dessus, prise à Tétouan ces jours derniers, et publiée par l'A B C, montre, en effet, que les soucis du pouvoir ont imprimé leur marque sur le visage du dictateur. En quatre mois, le général a blanchi ; le masque n'est plus égayé par un accueillant sourire ; il s'est immobilisé dans une gravité un peu figée, sinon anxieuse.

(b) A_2

PROCHAINEMENT « ITTO »

Le film *Itto*, la grande production réalisée au cœur de l'Atlas marocain par Jean Benoit-Lévy et Marie Epstein, les réalisateurs du film *la Maternelle*, va bientôt commencer sa carrière à Paris.

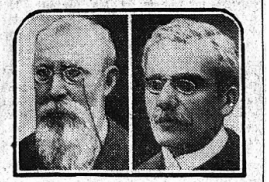
Itto, un des grands prix du cinéma français, passera après *Pension Mimosas* au grand cinéma des exclusivités françaises, le Colisée.

Rappelons qu'*Itto*, dans lequel paraissent plus de dix mille Chleuhs, dont quelques-uns jouent des rôles fort importants, est à la fois parlé chleuh et français, et que les principaux interprètes sont Mmes Simone Berriau, Simone Bourday et Sylvette Fillacier, MM. Hubert Prélier, Camille Bert, Roland Caillaux et Henri Debain.

FRANÇOISE ROSAY
la remarquable interprète de Pension Mimosas *qui passe actuellement au Colisée.*

(c) A_3

M. Lapie est nommé recteur de l'académie de Paris



Phot. H. Manuel.
M. PAUL APPELL M. LAPIE

Sur la proposition de M. de Monzie, ministre de l'instruction publique, M. Appell, recteur de l'académie de Paris, est admis, sur sa demande, à faire valoir ses droits à la retraite, à compter du 1er octobre 1925. Il est nommé recteur honoraire.

Le conseil des ministres a décidé, pour honorer au moment de sa mise à la retraite le grand savant et le grand administrateur qu'est M. Appell, de lui conférer, dans une prochaine promotion, la grand'croix de la Légion d'honneur.

M. Appell est remplacé par M. Lapie, directeur de l'enseignement primaire, ancien recteur de l'académie de Toulouse.

(d) A_4

Fig. 8. Example of articles from *Le Matin*

EN AVION
pour l'Indo-Chine
et Tokio

PIVOLO-GONIN-CAROL
(en tout : 300 kilos)
ont pris le départ hier au Bourget
pour le tour aérien de l'Asie

L'association Pelletier Doisy-Gonin-Carol a quitté hier matin à 6 h. ½, l'aérodrome du Bourget pour accomplir — non un raid — mais une randonnée aérienne sur le tour d'Asie, par Bucarest, Bagdad, Karachi, Calcutta, Hanoï, Changhaï, Tokio à l'aller, et par la Sibérie et la Russie au retour. Au total 30.000 kilomètres à bord d'un biplan muni d'un moteur à réducteur de 470 CV. Le but de la première étape était Bucarest.

(a) A_5

LE TOUR DE FRANCE CYCLISTE

L'avant-dernière étape
Nantes-Vire-Caen
avec une petite course
« contre la montre »

LE GREVES. PREMIER A VIRE
ET MORELLI PREMIER A CAEN

[DE NOTRE ENVOYÉ SPÉCIAL]

CAEN, 27 juillet. — *Par téléphone.* — Purement à titre documentaire, au départ ce matin de très bonne heure de l'avant-dernière étape Nantes-Caen par Rennes, Fougères et Vire (275 km.) Romain Maës était bien installé en tête avec une avance de 19 minutes 4 secondes sur Morelli second, et de 23 minutes 13 secondes sur Vervaecke troisième. Donc, à moins d'un cataclysme, rien à craindre pour le maillot jaune et belge.

L'étape d'aujourd'hui était d'abord en ligne, Nantes-Vire (220 km.), puis « contre la montre » dans une épreuve Vire-Caen (55 km.), disputée par les équipes, en tête les Belges, puis les touristes N° 1, l'équipe germano-italienne, les touristes N° 2, enfin l'équipe française.

La première demi-étape fut courue, si l'on peut dire, avec une belle lenteur, avec une heure et demie de retard sur le tableau de marche, ce qui prouve que rien de sensationnel ne s'est passé.

Un peu avant le sprint, Charles Pelissier, Le Grevès, Morelli, Teani, suivis de Bertocco, se sont détachés et ont pu gagner quelques secondes sur le gros

Un peu avant le sprint, Charles Pelissier, Le Grevès, Morelli, Teani, suivis de Bertocco, se sont détachés et ont pu gagner quelques secondes sur le gros de la troupe.

Comme de juste, Le Grevès se détacha très nettement, Aerts n'étant pas là et battit Pélissier de quelques longueurs.

Romain Maës ayant crevé non loin de l'arrivée, s'est trouvé un peu retardé. Mais qu'est-ce pour lui 1' 10" de moins à son classement par rapport à Morelli ?

(b) A_6

LA VACCINATION
contre la tuberculose

Une controverse scientifique
à l'Académie de médecine

Le professeur Calmette précise
les résultats acquis

Nos lecteurs ont été tenus au courant de la découverte, par le professeur Calmette et ses élèves, du vaccin B.C.G. contre la tuberculose.

L'année dernière encore, le docteur Henri Vignes, médecin des hôpitaux, indiquait, en première page du *Matin*, les résultats obtenus dans la prémunition des jeunes enfants, et, discutant de l'innocuité du B. C. G., citait de nombreux auteurs français et étrangers qui déclaraient la méthode parfaitement inoffensive.

Le professeur Calmette a apporté hier, à l'Académie de médecine, de nouveaux chiffres plus récents qui confirment ce qui avait été dit jusqu'ici.

Ce faisant, M. Calmette répondait à une communication faite par M. Lignières, de Buenos-Aires. Cet auteur, se basant sur certains faits tirés d'un rapport du docteur Tzekhnovitzer, de Kharkov, a posé une question à laquelle il est d'ailleurs impossible de répondre.

Quand un enfant ou un animal a reçu du B. C. G. et qu'il vient à mourir plus ou moins longtemps après, peut-on affirmer que le B. C. G. n'est pas la cause de la mort, bien qu'on ne trouve aucune lésion à l'autopsie le démontrant ?

(c) A_7

LE DRAME DU DOLLAR

La mission française à bord de l'« Ile-de-France »
estime que la conférence économique mondiale
devient impossible par suite de l'abandon
par les Etats-Unis de l'étalon-or

La situation de la France, seul pays fidèle à l'étalon-or n'inspire aucune inquiétude à des financiers comme MM. Rist et Wanzeeland

[DE NOTRE ENVOYÉ SPÉCIAL]

A bord de l'*Ile-de-France*, 20 avril. — *Par radio Sainte-Maries-de-la-Mer.* — La nouvelle de l'abandon par l'Amérique de l'étalon-or est parvenue à l'*Ile-de-France* à 10 heures du matin, heure de Paris, par un radio de New-York et elle y a causé une véritable stupeur, d'autant plus grande que la semaine passée un collaborateur de M. Woodin, secrétaire d'Etat au Trésor, traversant Paris, avait démenti avec hauteur un abandon possible du gold standard et que, vendredi dernier, la banque Morgan à Paris le démentait avec énergie. On se perd ici en conjectures sur la cause de ce drame soudain.

évaluer et régler les échanges internationaux.

— Aller à la conférence économique internationale, dit M. Rist, sous-gouverneur de la Banque de France, avec des monnaies flottantes et non rattachées à un mètre étalon, c'est comme si on allait au marché avec des mètres en caoutchouc extensible et avec des poids en sucre fondant.

Quant à la situation de la France, seul pays fidèle à l'étalon-or et qui, d'ailleurs, a déjà opéré la dévaluation de sa monnaie, elle n'inspire aucune inquiétude à des hommes comme Rist ou Wanzeeland, directeur de la Banque nationale belge, qui se trouve également à bord.

(d) A_8

Fig. 9. Example of articles from the selected corpus of *Le Matin*

```
<Article id="A_1" year="1935" issue="March" day="7" page="1">
  <NE type="person" uri="http://www.wikidata/entity/Q333091" value="John Simon"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q166646" value="Ramsay MacDonald"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q166635" value="Stanley Baldwin"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q352" value="Adolf Hitler"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q58211956" value="Foreign Office"></NE>
</Article>

<Article id="A_2" year="1925" issue="january" day="20" page="1">
  <NE type="person" uri="http://www.wikidata/entity/Q192894" value="Miguel Primo de Rivera"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q287076" value="ABC"></NE>
</Article>

<Article id="A_3" year="1935" issue="February" day="15" page="4">
  <NE type="person" uri="http://www.wikidata/entity/Q3170696" value="Jean Benoit-Lévy"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q3292507" value="Marie Epstein"></NE>
  <NE type="product" uri="http://www.wikidata/entity/Q14931659" value="La Maternelle"></NE>
  <NE type="product" uri="http://www.wikidata/entity/Q2072517" value="Pension Mimosas"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q3484541" value="Simone Berriau"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q3484545" value="Simone Bourday"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q3507203" value="Sylvette Fillacier"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q3142096" value="Hubert Prelier"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q2934860" value="Camille Bert"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q15974123" value="Roland Caillaux"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q3130926" value="Henri Debain"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q451631" value="Françoise Rosay"></NE>
</Article>
<Article id="A_4" year="1925" issue="May" day="16" page="1">
  <NE type="person" uri="http://www.wikidata/entity/Q715906" value="Paul Appell"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q42204361" value="Paul Lapie"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q2845619" value="Anatole de Monzie"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q2750231" value="Academy of Toulouse"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q2822323" value="Paris Academy"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q163700" value="Legion of Honour"></NE>
</Article>
<Article id="A_5" year="1928" issue="may" day="9" page="1">
  <NE type="person" uri="http://www.wikidata/entity/Q3103314" value="Pivolo"></NE>
  <NE type="person" uri="" value="Gonin"></NE>
  <NE type="person" uri="" value="Carol"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q3103314" value="Pelletier Doisy"></NE>
  <NE type="person" uri="" value="Brunat"></NE>
</Article>

<Article id="A_6" year="1935" issue="July" day="28" page="5">
  <NE type="person" uri="http://www.wikidata/entity/Q129011" value="René Le Grèves"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q458790" value="Ambrogio Morelli"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q254235" value="Romain Maes"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q1479421" value="Félicien Vervaecke"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q1065826" value="Charles Pélissier"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q16749950" value="Aldo Bertocco"></NE>
</Article>

<Article id="A_7" year="1928" issue="may" day="9" page="2">
  <NE type="organization" uri="http://www.wikidata/entity/Q337555" value="Académie de médecine"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q437983" value="professeur Calmette"></NE>
  <NE type="product" uri="http://www.wikidata/entity/Q798309" value="B.C.G"></NE>
  <NE type="" uri="http://www.wikidata/entity/Q12204" value="tuberculose"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q55672177" value="Henri Vignes"></NE>
  <NE type="person" uri="" value="Lignières"></NE>
  <NE type="person" uri="" value="Tzekhnovitzer"></NE>
</Article>
<Article id="A_8" year="1933" issue="April" day="21" page="1">
  <NE type="person" uri="http://www.wikidata/entity/Q2960124" value="Charles Rist"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q2031553" value="William H. Woodin"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q5891192" value="Trésor public"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q806950" value="Bank of France"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q685918" value="National Bank of Belgium"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q14996" value="Paul van Zeeland"></NE>
</Article>
```

Fig. 10. Example of named entities extracted from $\{A_1, A_2, ..., A_8\}$.

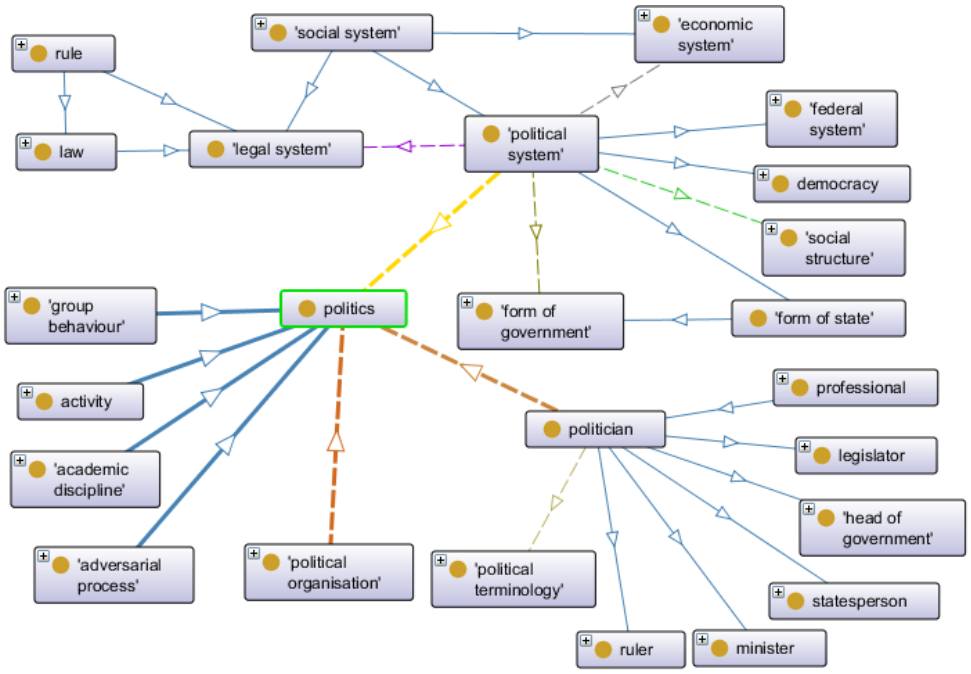


Fig. 11. Excerpt of Topic-OPA around the concept Politics.

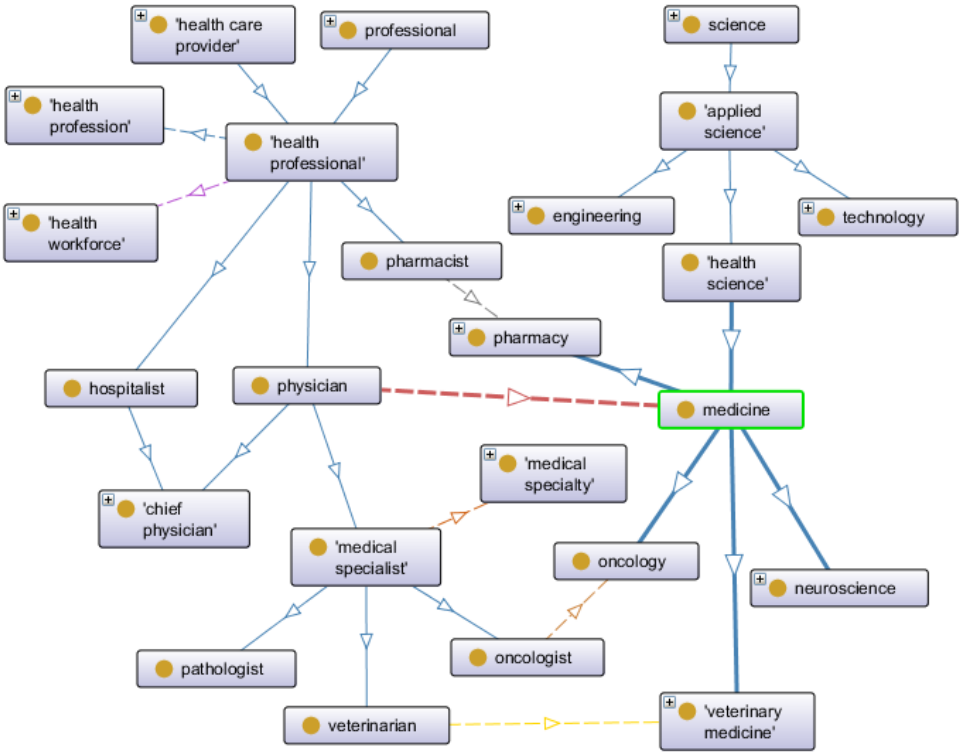


Fig. 12. Excerpt of Topic-OPA around the concept Medicine.

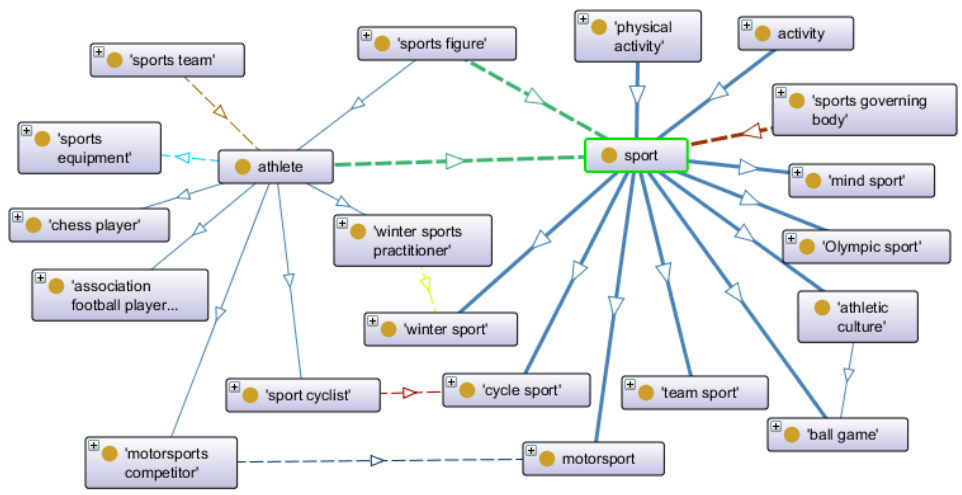


Fig. 13. Excerpt of Topic-OPA around the concept Sport

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 Rel_{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, $T_c = \{\mathsf{Science} \sqcap \mathsf{Medicine} \sqcap \mathsf{Physics} \sqcap \mathsf{Bacteriology} \sqcap \mathsf{Immunology} \sqcap \mathsf{Virology} \sqcap \mathsf{Vaccination}\}$. 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 ($\overline{Rel_{Topic}}(I_7, \mathsf{Vaccination}) = 0.69$) as well as the highest degree centrality ($D(\mathsf{Vaccination}) = 13.48$).
  Although, A_2 is labeled by the topic Military Affairs having the highest average of relatedness ($\overline{Rel_{Topic}}(I_2, \mathsf{Military\ Affairs}) = 0.67$) as well as by the topic War having the highest degree centrality ($D(\mathsf{War}) = 22.22$).
  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 Rel_{Topic}. In addition, we compare Rel_{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

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

Article	Named Entity	Instance of
A_1	John Simon	$Politician \sqcap Lawyer \sqcap Judge$
	Ramsay MacDonald	$Politician \sqcap Journalist \sqcap Diplomat$
	Adolf Hitler	$Politician \sqcap Soldier \sqcap Stateperson \sqcap Writer \sqcap Painter$
	Eric Phipps	$Politician \sqcap Diplomat$
	Anthony Eden	$Politician \sqcap Diplomat$
	Stanley Baldwin	$Politician$
	Foreign Office	$ForeignAffairsMinistry$
A_2	Miguel Primo de Rivera	$Politician \sqcap MilitaryPersonnel$
	ABC	$DailyNewspaper$
A_3	Jean Benoit-Lévy	$FilmDirector \sqcap FilmProducer \sqcap Screenwriter$
	Marie Epstein	$FilmDirector \sqcap FilmProducer \sqcap Screenwriter \sqcap Actor$
	La Maternelle	$Film$
	Pension Mimosas	$Film$
	Simone Berriau	$FilmActor \sqcap Actor$
	Simone Bourday	$Actor$
	Sylvette Fillacier	$Actor$
	Hubert Prelier	$Actor$
	Camille Bert	$Actor$
	Roland Caillaux	$Actor \sqcap Painter$
	Henri Debain	$FilmActor \sqcap FilmDirector$
	Françoise Rosay	$Actor \sqcap Singer \sqcap StageActor \sqcap FilmActor$
A_4	Paul Appell	$UniversityTeacher \sqcap Mathematician$
	Academy of Toulouse	$AcademicDistrict$
	Paris Academy	$AcademicDistrict$
	Legion of Honour	$Order$
A_5	Georges Pelletier d'Oisy	$AircraftPilot$
A_6	René Le Grèves	$SportCyclist$
	Ambrogio Morelli	$SportCyclist$
	Romain Maes	$SportCyclist$
	Félicien Vervaecke	$SportCyclist$
	Charles Pélissier	$SportCyclist$
	Aldo Bertocco	$SportCyclist$
A_7	Académie Nationale de Médecine	$Academy \sqcap NationalAcademy$
	Albert Calmette	$Physician \sqcap Bacteriologist \sqcap Immunologist \sqcap Virologist$
	BCG vaccine	$Vaccine$
	Tuberculose	$Disease \sqcap NotifiableDisease \sqcap EndemicDisease$
A_8	Charles Rist	$Economist \sqcap Banker$
	William H. Woodin	$Politician \sqcap Businessperson$
	Trésor public	$PublicTreasury$
	Bank of France	$Bank \sqcap CentralBank \sqcap Business$
	National Bank of Belgium	$CentralBank$
	Paul van Zeeland	$Economist \sqcap Politician \sqcap Lawyer \sqcap Diplomat \sqcap Jurist$

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 2

Excerpt of the instance-topic mapping process between A_7 and T.

Instance (i)	**Related Topic (t)**	$Rel_{Topic}(i,t) \geq -log(\overline{Rel_{Topic}}(I_7,T))$
Académie Nationale de Médecine	Research Institute	0.60
	Science	0.50
	Academic District	0.69
	Academy	0.80
	Learned Society	0.63
	National Academy	0.72
Albert Calmette	Physicist	0.69
	Medicine	0.58
	Physician	0.73
	Health Professional	0.58
	Physics	0.63
	Immunologist	0.70
	Immunology	0.64
	Bacteriology	0.64
	Virology	0.64
BCG vaccine	Medication	0.58
	Biopharmaceutical	0.59
	Vaccine	0.75
	Vaccination	0.69
Tuberculose	Disease	0.67

Table 3

Ranking and selection of labeling topics.

A_i	**Threshold**	**Most Common Topics (t_c)**	$\overline{Rel_{Topic}}(I_i,t_c)$	**Degree Centrality**	***Relatedness-Guided***	***Centrality-Guided***	**Selected Topics**
A_1	0.55	Politics	0.68	29.17	Politics	Politics	Politics
		Political Activism	0.56	6.94			
A_2	0.55	Military Affairs	0.67	6.94	Military Affairs	War	Military Affairs-War
		Political Activism	0.62	6.94			
		War	0.59	22.22			
A_3	0.59	Art	-	-	-	-	Art
A_4	0.52	Higher Education	0.58	15.28	Higher Education	Science	Higher Education-Science
		Science	0.55	23.62			
A_5	0.61	Aviation	-	-	-	-	Aviation
A_6	0.55	Cycle Sport	0.68	13.20	Cycle Sport	Cycling	Cycle Sport-Cycling
		Cycling	0.59	27.38			
		Sport	0.55	13.89			
A_7	0.58	Vaccination	0.69	13.48	Vaccination	Vaccination	Vaccination
		Bacteriology	0.64	7.64			
		Immunology	0.64	7.64			
		Medicine	0.58	7.64			
		Virology	0.64	7.64			
A_8	0.51	Economics	-	-	-	-	Economics

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, $maximumdepth = 28$.
- *average depth*: is the average length of all *taxonomic* branches. In Topic-OPA, $averagedepth = 6$.
- *depth variance*: is the dispersion with respect to the average depth, computed as the standard mathematical variance. In Topic-OPA, $depthvariance = 6.38$ 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[4] and Marcel Léger is a Epidemiologist, Microbiologist and Bacteriologist[5]. 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.

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'Académie 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.

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

[4] https://journals.openedition.org/primatologie/2816?lang=enftn1, last visited April 27, 2020.

[5] http://www.pathexo.fr/documents/notices/leger.html?width=800height=500, last visited April 27, 2020

```
<Article id="A_9" year="1924" issue="June" day="27" page="1">
  <NE type="person" uri="http://www.wikidata/entity/Q37193" value="Robert Koch"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q437983" value="Albert Calmette"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q337555" value="Academie Nationale de Medecine"></NE>
  <NE type="organization" uri="http://www.wikidata/entity/Q391083" value="Pasteur Institute"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q1029190" value="Camille Guerin"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q4722350" value="Alfred Boquet"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q215987" value="Leopold Negre"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q53347259" value="Benjamin Weill-Halle"></NE>
  <NE type="product" uri="http://www.wikidata/entity/Q798309" value="BCG vaccine "></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q3421172" value="Raymond Turpin"></NE>
  <NE type="divers" uri="http://www.wikidata/entity/Q12204" value="Tuberculose"></NE>
  <NE type="person" uri="" value="Wilbert"></NE>
  <NE type="person" uri="" value="Leger"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q50822285" value="Louis Devraigne"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q62502469" value="Edmond Lévy-Solal"></NE>
</Article>
```

Fig. 15. The named entities of A_9.

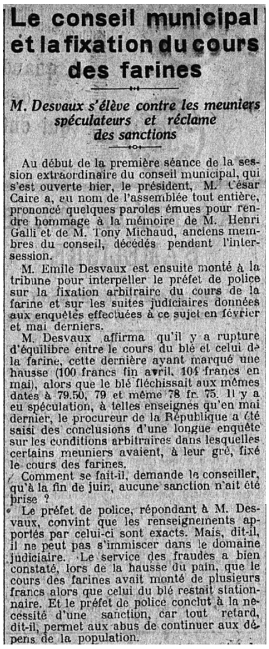
Table 4

Ranking and selection of labeling topics for A_9.

Article	Threshold	Most Common Topics	Selected Topics
A_9	0.53	Science	Science

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 A_{10} (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 A_{10}. Table 7 shows that A_{10} is labeled by the unique most common topic Politics. However, based on the content and the subject of A_{10}, 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.



Le conseil municipal et la fixation du cours des farines

M. Desvaux s'élève contre les meuniers spéculateurs et réclame des sanctions

Au début de la première séance de la session extraordinaire du conseil municipal qui s'est ouverte hier, le président, M. César Caire a, au nom de l'assemblée tout entière, prononcé quelques paroles émues pour rendre hommage à la mémoire de M. Henri Galli et de M. Tony Michaud, anciens membres du conseil, décédés pendant l'intersession.

M. Emile Desvaux est ensuite monté à la tribune pour interpeller le préfet de police sur la fixation arbitraire du cours de la farine et sur les suites judiciaires données aux enquêtes effectuées à ce sujet en février et mai derniers.

M. Desvaux affirma qu'il y a rupture d'équilibre entre le cours du blé et celui de la farine, cette dernière ayant marqué une hausse (100 francs fin avril, 104 francs en mai), alors que le blé fléchissait aux mêmes dates à 79.50, 79 et même 78 fr. 75. Il y a eu spéculation, à telles enseignes qu'en mai dernier, le procureur de la République a été saisi des conclusions d'une longue enquête sur les conditions arbitraires dans lesquelles certains meuniers avaient, à leur gré, fixé le cours des farines.

Comment se fait-il, demande le conseiller, qu'à la fin de juin, aucune sanction n'ait été prise ?

Le préfet de police, répondant à M. Desvaux, convint que les renseignements apportés par celui-ci sont exacts. Mais, dit-il, il ne peut pas s'immiscer dans le domaine judiciaire. Le service des fraudes a bien constaté, lors de la hausse du pain, que le cours des farines avait monté de plusieurs francs alors que celui du blé restait stationnaire. Et le préfet de police conclut à la nécessité d'une sanction, car tout retard, dit-il, permet aux abus de continuer aux dépens de la population.

Fig. 16. Example of article (A_{10}) from *Le Matin* 1922, June 20

Table 5

Assignment of the disambiguated named entities of A_{10} as instances of Topic-OPA.

Article	Named Entity	Instance of
A_{10}	César Caire	*Jurist* $\sqcap$ *Lawyer*
	Henri Galli	*Politician* $\sqcap$ *Journalist*
	Emile Desvaux	–
	Ambroise Rendu	*Politician*
	Alexandre Luquet	*Politician*
	Flour	*FoodIngredient*
	Wheat	*FoodIngredient*

```
<Article id="A_9" year="1922" issue="june" day="20" page="2">
  <NE type="person" uri="http://www.wikidata/entity/Q15974098" value="Cesar Caire"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q3131155" value="Henri Galli"></NE>
  <NE type="person" uri="" value="Tony Michaud"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q18918113" value="Emile Desvaux"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q16026405" value="Ambroise Rendu"></NE>
  <NE type="person" uri="http://www.wikidata/entity/Q13080262" value="Alexandre Luquet"></NE>
  <NE type="product" uri="http://www.wikidata/entity/Q36465" value="Farine"></NE>
  <NE type="product" uri="http://www.wikidata/entity/Q25618328" value="Ble"></NE>
</Article>
```

Fig. 17. Named entities extracted from A_{10}.

Table 6

Excerpt of the instance-topic mapping process between A_{10} and T.

Instance (i)	Related Topic (t)	$Rel_{Topic}(i,t) \geq \overline{Rel_{Topic}}(I_{10},T)$
César Caire	Law	0.63
	Jurisprudence	0.64
	Jurist	0.71
	Lawyer	0.72
Henri Galli	Politics	0.74
	Political Activist	0.68
	Journalism	0.69
Ambroise Rendu	Politics	0.74
	Political Activist	0.68
Alexandre Luquet	Politics	0.74
	Political Activism	0.61
Flour	Food	0.75
	Ingredient	0.77
	Food Ingredient	0.73
	Cooking	0.66
Wheat	Food	0.75
	Ingredient	0.77
	Food Ingredient	0.73
	Cooking	0.66

Table 7

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

Article	Threshold	Most Common Topics	Selected Topics
A_{10}	0.58	Politics	Politics

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*[6]. 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.

[6] https://en.wikipedia.org/wiki/Topic_and_comment, last visited April 28, 2020

8.3. Comparison of Rel_{Topic} with Alternative Graph-Based Measures

In this section, we compare Rel_{Topic} 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 Sim_{Rada} (see Equation (1)) and Sim_{LC} (see Equation (2)) are only applicable to taxonomies. Meanwhile, Rel_{HS} (see Equation (3)) is the most appropriate since it is a relatedness measure applicable in ontologies. However, applying Rel_{HS} is not an easy task due to the difficulty of the computation of the direction changes of edges (hierchical and non-hierarchical) through all the paths. For this purpose, and since Rel_{Topic} is based on the computation of shortest paths (see Equation (8)), we selected Sim_{Rada} for the comparison. In this regard, we applied Sim_{Rada} 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_7 to Topic-OPA. The results imply that the related topic concepts to a given article $A_i \in A$ are clearly identified by Rel_{Topic} as well as by Sim_{Rada}. However, the use of Rel_{Topic} 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 Rel_{Topic}, within topic ontologies. In contrast to existing measures, Rel_{Topic} considers *non-hierarchical* relations and assesses the relatedness between instances and concepts. In order to apply Rel_{Topic}, 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 Rel_{Topic}, 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 we have applied Rel_{Topic} 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 Rel_{Topic} 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 Rel_{Topic}. 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 Rel_{Topic}, 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 Rel_{Topic}. Meanwhile, the *not related* category defines the topics that are not relevant to those given by the annotators.

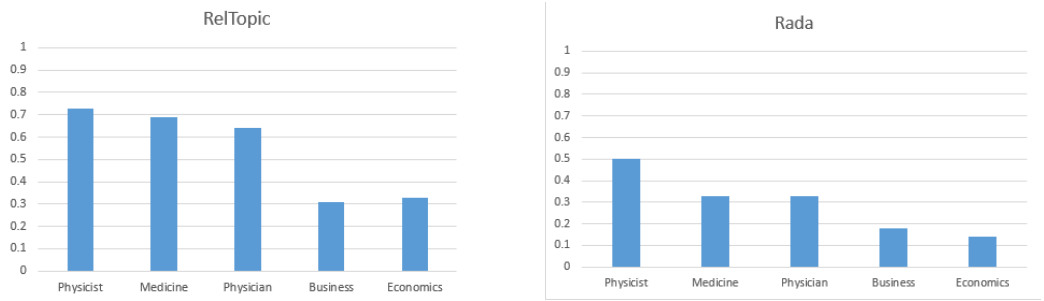
Furthermore, we discussed the issues that affected the relevance of the automatic labeling process. Specifically, we addressed the existence of some named entities that are not disambiguated. We analyzed also the problem of the typology of the named entities and its influence on the quality of the generated 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.

Instance (i)	Topic Concepts (t)	$Rel_{Topic}(i,t)$	Sim_{Rada}
Académie Nationale de Médecine	Research Institute	0.60	0.25
	Science	0.50	0.2
	Academic District	0.69	0.33
	Academy	0.80	0.5
	Economics	0.37	0.1
Albert Calmette	Physicist	0.69	0.33
	Medicine	0.58	0.33
	Physician	0.73	0.5
	Business	0.31	0.16
	Economics	0.33	0.125
BCG vaccine	Medication	0.58	0.33
	Vaccine	0.75	0.5
	Vaccination	0.69	0.33
	Economics	0.36	0.11
	Business	0.33	0.14
Tuberculose	Disease	0.67	0.5
	Health Problem	0.5	0.33
	Work of Art	0	0.1



(a) Rel_{Topic} (b) Sim_{Rada}

Fig. 18. Comparison of the results of mapping Albert Calmette to topics in T

the general and the specific levels. Finally, we compared Rel_{Topic} to alternative path-based semantic measures such as Sim_{Rada}. The comparison showed that both measures, Rel_{Topic} and Sim_{Rada}, can identify the relevant topics but the use of Rel_{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 Rel_{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

[1] S. Fernando and M. Stevenson, A semantic similarity approach to para-phrase detection, in: Proceedings of Computational Linguistics Colloquium, U.K.,2008, pp. 45–52.

[2] N. Fiorini, S. Ranwez, J. Montmain and V. Ranwez, USI: a fast and accurate approach for conceptual document annotation, BMC Bioinformatics, 2015, DOI: 10.1186/s12859-015-0513-4.
[3] J. Euzenat and P. Shvaiko, Ontology Matching: Second Edition, Springer-Verlag, Berlin Heidelberg (DE), 2013.
[4] C. D'Amato, Similarity-based Learning Methods for the Semantic Web, Phd thesis, Universita degli Studi di Bari, 2007.
[5] P.H. Guzzi, M. Mina, C. Guerra and M. Cannataro, Semantic similarity analysis of protein data: assessment with biological features and issues, Briefings, in: Bioinformatics, **13:5** (2012), 569—585. https://doi.org/10.1093/bib/bbr066.
[6] S. Harispe, S. Ranwez, S. Janaqi and J. Montmain, Semantic Similarity from Natural Language and Ontology Analysis, Synth. Lect. Hum. Lang. Technol,**8** (2015), 1—254.
[7] R. Rada, H. Mili, E. Bicknell and M. Blettner, Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man and Cybernetics, **19** (1989), 17–30.
[8] C. Leacock and M. Chodorow, Filling in a sparse training space for word sense identification, ms, 1994.
[9] G. Hirst, D. St-Onge, Lexical chains as representations of context for the detection and correction of malapropisms, WordNet: An Electronic Lexical Database, 1998.
[10] P. Resnik, Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, J. Artif. Intell. Res., **11** (1998), 95–130.
[11] D. Lin, An Information-Theoretic Definition of Similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, ICML, 1998, pp. 296–304.
[12] Y. Tang, P.D. Baer, G. Zhao and R. Meersman, On Constructing, Grouping and Using Topical Ontology for Semantic Matching, in: Meersman R, Herrero P, Dillon T, Proceedings of OTM 2009 Workshops (On the Move to Meaningful Internet Systems), **5872**, Springer Berlin, Heidelberg, 2009, pp. 816–825.
[13] M. Allahyari and K. Kochut, A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling, International Journal of Advanced Computer Science and Applications, **8:9** (2017), pp. 335–349.
[14] J. Sleeman, T. Finin and M. Halem, Ontology-Grounded Topic Modeling for Climate Science Research, in: Proceedings of Semantic Web for Social Good Workshop, ISWC, 2018.
[15] A.G. Maguitman, R.L. Cecchini, C.M. Lorenzetti and F. Menczer, Using Topic Ontologies and Semantic Similarity Data to Evaluate Topical Search, in: Proceedings of Conferencia Latino-americana de Informática, 2010.
[16] G. Zhao and R. Meersman, Architecting Ontology for Scalability and Versatility, in: R. Meersman and Z. Tari, ed., On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE, OTM 2005, Lecture Notes in Computer Science, **3761**, Springer, Berlin, Heidelberg, 2005.
[17] I. Hulpus, N. Prangnawarat and C. Hayes, Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation, in: M. Arenas et al., ed., Proceedings of the Semantic Web - ISWC 2015, LNCS vol. 9366, Springer, Cham, 2015, pp. 442–457.
[18] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: 14th International Joint Conference on Artificial Intelligence, 1995, pp. 448—453.
[19] L. Mazuel and N. Sabouret, Semantic Relatedness Measure Using Object Properties in an Ontology, in: A. Sheth et al., ed., The Semantic Web - ISWC 2008, Lecture Notes in Computer Science, **5318**, Springer, Berlin, Heidelberg.
[20] J. Jiang and D. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proc. on International Conference on Research in Computational Linguistics, Taiwan, 1997, pp. 19-–33.
[21] T. Opsahl, F. Agneessens and J. Skvoretz, Node centrality in weighted networks: Generalizing degree and shortest paths, Social Networks, **32:3** (2010), 245–251.
[22] J. Heitzig, N. Marwan, Y. Zou, J. Donges and J. Kurths, Consistently weighted measures for complex network topologies, Eu-rop. Phys. J. B. **85** (2010), 1–16.
[23] J. Sosnowska J and O. Skibski, Attachment centrality for weighted graphs, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 416–422.
[24] K. Böhm and M. Ortiz, 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), 2018.
[25] F. Erxleben, M. Günther, Krötzsch, J. Mendez and D. Vrandečić, Introducing Wikidata tothe linked data web, in: Proceedings 13th Int. Semantic Web Conf. (ISWC'14), LNCS, 2014, pp. 50-–65.
[26] S. Malyshev, M. Krotzsch, L. Gonzalez, J. Gonsior and A. Bielefeldt, 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), LNCS, Springer, 2018, pp. 376–394.
[27] A. Bielefeldt, J. Gonsior, and M. Krotzsch, 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, 2018.
[28] Y. Zuo, J. Zhao and K. Xu, Word network topic model: a simple but general solution for short and imbalanced texts, Knowledge and Information Systems, **48** (2016), 379–398.
[29] M.C. Suárez-Figueroa, A. Gómez-Pérez and B. Villazón-Terrazas, How to Write and Use the Ontology Requirements Specification Document, in: R. Meersman, T. Dillon, Herrero, ed., On the Move to Meaningful Internet Systems: OTM 2009, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2009.
[30] J. Brank, M. Grobelnik and D. Mladenić, A survey of ontology evaluation techniques, in: Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005), 2005.
[31] M. Fernández, C. Overbeeke, M. Sabou and E. Motta, What makes a good ontology? A case-study in fine-grained knowledge reuse, In The semantic web, pp. 61–75, Springer Berlin Heidelberg, 2009.
[32] K. Dellschaft and S. Staab, 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, 2008, pp. 253-272.
[33] D. Sanchez, M. Batet, S. Martinez and J.D. Ferrer, Semantic variance: An intuitive measure for ontology accuracy evaluation, Engineering Applications of Artificial Intelligence, **39** (2015), 89–99.

[34] E.W. Dijkstra, A note on two problems in connexion with graphs, Numerische Mathematik. **1** (1959), pp. 269–271. doi:10.1007/BF01386390.
[35] R. Bellman, On a routing problem, Quarterly of Applied Mathematics, 16 (1958), pp. 87-–90. doi:10.1090/qam/102435.
[36] M. Uschold and M. Gruninger, Ontologies: principles, methods and applications, The Knowledge Engineering Review, **11** (1996), 93–136.
[37] M. Fernández-López, A. Gómez-Pérez and N. Juristo, METHONTOLOGY: From Ontological Art Towards Ontological Engineering, in AAAI, 1997.
[38] Y. Sure, S. Staab and R. Studer, On-To-Knowledge Methodology (OTKM), in: S. Staab, R. Studer, ed., Handbook on Ontologies, International Handbooks on Information Systems, Springer, Berlin, Heidelberg, 2004.
[39] P. Cimiano and J. Völker, Text2Onto, in: A. Montoyo, R. Muńoz and E. Métais, ed., Natural Language Processing and Information Systems, NLDB 2005, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2005, pp. 227-238.
[40] B. Fortuna, M. Grobelnik and D. Mladenic, OntoGen: Semi-automatic Ontology Editor, in: M.J. Smith and G. Salvendy, ed., Human Interface and the Management of Information, Interacting in Information Environments, Human Interface 2007, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2007, pp. 309–318.
[41] R. Speer, J. Chin and C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: AAAI, 2017, pp. 4444–4451.