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*

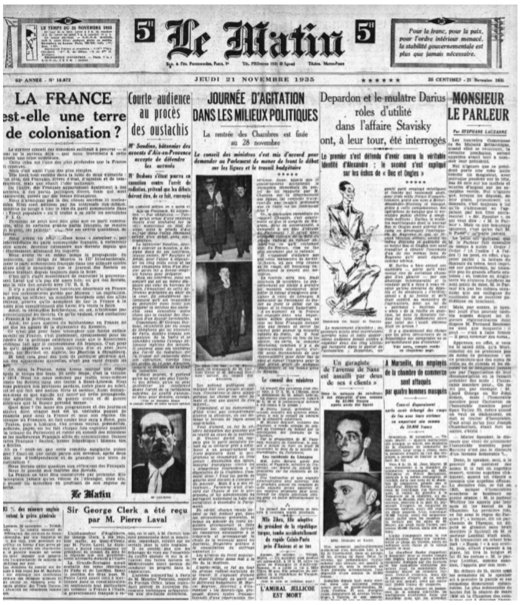
Abstract.

Graph-based semantic measures have been used to solve problems in several domains. They tend to compare semantic entities in order to estimate their similarity or relatedness. While semantic similarity is applicable to hierarchies or taxonomies, semantic relatedness is adapted to ontologies. In this work, we propose a novel semantic relatedness measure, named Rel_{Topic}, within topic ontologies for topic labeling purposes. In contrast to traditional measures, which are dependent on textual resources, Rel_{Topic} considers semantic properties of entities in ontologies. Thus, correlations of nodes and weights of nodes and edges are assessed. The pertinence of Rel_{Topic} is evaluated for topic labeling of old press articles. For this purpose, a topic ontology representing the articles, named Topic-OPA, is derived from open knowledge graphs by applying a SPARQL-based automatic approach. A use-case is presented in the context of the old French newspaper *Le Matin*. The generated topics are evaluated using a dual evaluation approach with the help of human annotators. Our approach shows an agreement quite close to that shown by humans. The entire approach's reuse is demonstrated for labeling a different context of articles, recent (modern) newspapers.

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

1. Introduction

This article presents the works accomplished as part of the ASTURIAS[1] project in the domain of cultural heritage. The main goal of ASTURIAS is to thematically and automatically organize a collection of old press articles with a set of topics (e.g., Politics, Art, Sport, Science, Etc.). One of the specific features of old press is that it does not offer thematic entries (see Figure 1). Articles appear and follow one another without a thematic logic.



5f Le Matin 5f

Pour le franc, pour la paix, pour l'ordre intérieur menacé, la stabilité gouvernementale est plus que jamais nécessaire.

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

La rentrée des Chambres est fixée au 28 novembre

Depardon et le mulâtre Darius rôles d'utilité dans l'affaire Stavisky ont, à leur tour, été interrogés

MONSIEUR LE PARLEUR

Sir George Clerk a été reçu par M. Pierre Laval

L'AMIRAL JELLICOE EST MORT

Fig. 1. Excerpt of *Le Matin*.

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

[1] Structural Analysis and Semantic Indexing of Newspaper Articles (February 2019 - June 2022).

Under these conditions, it remains a tedious task to query sources that report the same events from different points of view in different areas of the newspaper. The scientific challenge is to propose robust approaches for the analysis of texts that are noisy due to the imperfect process of automatic transcription of images into electronic texts. These approaches need also to be multi-thematic, and robust to linguistic evolution over the centuries. The ambition of the ASTURIAS project (whose workflow appears in Figure 2) is to study the digitization process from end to end of the processing chain: WP1- from newspaper images, automatically analyze sections, articles and texts; WP2- extract named entities from these elements; WP3- Topic labeling and hyperlinking the articles based on the analysis made in WP1 and the named entities extracted in WP2.

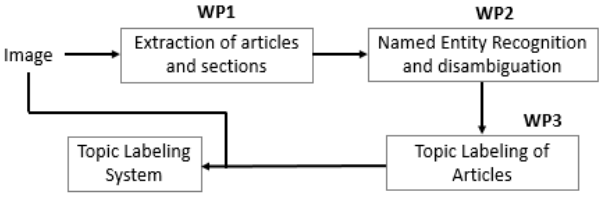

WP1
WP2
WP3
Image
Extraction of articles and sections
Named Entity Recognition and disambiguation
Topic Labeling System
Topic Labeling of Articles

Fig. 2. The pipeline of the project ASTURIAS.

Our work's main goal is to propose a framework that permits automatic labeling of old press articles (WP3). This framework tends to replace humans with software for labeling a vast number of articles that would require too much human effort to do it manually. The task of labeling documents according to their topics has traditionally been addressed either by using classifiers for assigning to the articles a set of predefined topics (e.g., [39], [45]), or by topic detection methods (e.g., probabilistic latent semantic analysis (pLSA) [43], latent Dirichlet allocation (LDA) [44]), which generate topics from textual resources [39]. The main advantage of the first approach is generating clean, and formally-defined research topics [39]. This approach is recommended when a good characterization of the research topics within a domain is available [39]. However, the second approach suffers from a significant limitation of generating topics from scratch leading to noisier and less interpretable results [39].

In this study, we propose applying graph-based semantic relatedness measures that permit assessment of the semantic relatedness of topics in topic ontologies with articles' content. 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 essential tools for designing 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]. Building semantic relatedness measures is a challenging and important research issue since they have to consider several kinds of relations and not only the taxonomic ones [19]. In the literature, apart from Hirst and St-Onge's measure [9], there have been relatively few attempts to develop relatedness measures [19, 37]. Most efforts are directed to design similarity measures such as [7, 8, 10, 11]. For comparing ontological entities, graph-based measures are classified into two basic approaches: *path-based*, which compare the concepts according to properties of paths in graphs, and *node-based*, that use properties of concepts in the ontology graph for comparing concepts. However, these approaches suffer from different limitations.

The major contribution of this work is the design and evaluation of a semantic relatedness measure, named Rel_{Topic}, that considers the semantic properties of entities in topic ontologies. Rel_{Topic} is designed as a combination of 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. Rel_{Topic} will be used for topic labeling of old press articles, which are represented by a set of "not ambiguous" named entities extracted from open data sources (WP2). A second contribution to mention is building a topic ontology named Topic-OPA from the open knowledge graph Wikidata using a SPARQL-based automatic approach. Topic-OPA is required for the application of Rel_{Topic}. Based on Rel_{Topic} and Topic-OPA, we defined the selection process of the most relevant topics for labeling the articles. To demonstrate the performance of our approach, a use-case is presented in the context of *Le Matin*[2], an old French newspaper first published in 1884 and discontinued in 1944. Finally, Topic-OPA and Rel_{Topic} are evaluated using dual evaluation approaches. Our approach's reusability is demonstrated for labeling articles in different contexts, such as recent newspapers.

The remainder of this paper is organized as follows: the research problem is specified in section 2. Section 3 considers the main related works. In section 4, we discuss our semantic relatedness measure Rel_{Topic}. Section 5 introduces Topic-OPA. 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 and section 9 respectively. Finally, section 10 concludes the paper.

2. Problem Definition

To define our research problem, a fundamental hypothesis is considered that articles are represented by a set of "not ambiguous" named entities (e.g. *person*, *organization*, *product* and *location*) extracted from open data sources (coming from WP2 of the project). Thus, the research problem can be defined as follows: Given a corpus of articles A, a set of named entities N (represented by a set of *URIs*) that are collected from A (WP2), and a topical structure T. The problem is to find the most relevant topics from T that label $A_i, \forall A_i \in A$. Based on this perspective, our work (WP3) considers mainly the following issues:

1. *Construction of the topical structure as a predefined set of topics*: 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 process as a relevance assessment*: this process is performed for each $A_i \in A$. It aims to map N_i, the named entities of A_i, to the topics of T to evaluate their relevance. Thus, the mapping process takes as inputs $n, \forall n \in N_i$, and $t, \forall t \in T$, and evaluates if t is relevant to n or not. The relevance is 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 as a topic labeling process*: takes as input the relatedness values of n and $t, \forall A_i \in A$, obtained from the entity-topic mapping process and aims to rank them and select the best topic(s) to label A_i.

3. Related Works

This section outlines the following related works: graph-based semantic measures, semantic relatedness measures, topic ontologies, and ontology-based labeling of articles.

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. Rada's measure is an example of similarity measures adapted to taxonomies:

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

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]:

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

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, quantifies the *weight* between two concepts as follows [9]:

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

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

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 (i.e., a downward link after an upward link). The particular difficulty of this approach is to determine the direction of each link [19]. The path-based approaches suffer from a significant drawback: they consider all edges equivalent, indicating 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 term's frequency 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 uses the Information Content of the *LCS* as the similarity value:

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

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

$$IC(c) = -logP(c) \tag{5}$$

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

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

Two main limitations are recognized for these approaches: (1) they are based on textual resources, and (2) applicable only on taxonomies.

3.2. Semantic Relatedness Measures

This section outlines significant works in the literature that addressed the design of semantic relatedness measures. However, these measures are strongly dependent on textual resources. Mazuel and Sabouret [19] have proposed a semantic relatedness measure that evaluates the semantic relatedness of two concepts by considering the object properties in ontologies. They differed between two different types of paths. First, the single-relation path in which all the edges have the same type (e.g., hierarchical relations). Second, the mixed-relation path in which different types of relations (hierarchical and non-hierarchical) are involved. The proposed semantic relatedness measure is composed of three main tasks: (1) consider a set of patterns given in [9] to filter the paths which are not semantically correct; (2) use of the information-theoretic definition of semantic similarity given in [10] to weight the hierarchical edges in the graph; (3) compute the weight of non-hierarchical edges. Finally, the relatedness measure is the sum of these tasks. Another work to cite is a context-vector approach proposed in the biomedical domain [37, 41]. This approach aims to compute the semantic relatedness between pair of concepts in the Unified Medical Language System (UMLS)[3]. The context-vector approach is based on a *Gloss Overlaps* (i.e., number of shared words in the definitions of two concepts) approach relied on the WordNet[4] dictionary [42]. The gloss vector approach combines the definitions of concepts with co-occurrence data in a given corpus (e.g., clinical reports). Every word in the definition is replaced by its context vector from the co-occurrence data and relatedness is calculated as the cosine of the angle between the two vectors. Due to the limitation of semantic relations provided in WordNet (*is-a*, *part-of*), the context-vector approach extended the construction of concept definitions by using different relations in the UMLS.

3.3. Topic Ontologies

Topic ontologies are considered a 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. Topic ontologies are being increasingly used in various domains such as semantic matching [12], topic labeling [13], topic modeling [14], evaluating topical search [15] and classification of research articles [40].

[3] https://www.nlm.nih.gov/research/umls/index.html, last visited February 4, 2021

[4] https://wordnet.princeton.edu/, last visited February 5, 2021

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, 39]. However, these approaches are not efficient, hard, and time-consuming to construct an ontology from a large corpus of documents [15]. In the literature, few works have been found about building topic ontologies from knowledge graphs (e.g., [24]) or Web sources (e.g., [38]). In [24], building topic-specific ontologies from open knowledge graphs such as ConceptNet [36] is presented. A query-based interactive approach is applied for extracting entities and relations from the knowledge graph. Three main phases are defined in this approach: construction of the central taxonomy, ontology enrichment, and ontology cleaning. Another approach to cite is Klink-2 [38], which generates ontologies of research topics [40] by integrating multiple web sources. In particular, Klink-2 analyses networks of research entities (including papers, authors, venues, and technologies) to infer three main types of semantic relationships. For instance, the hierarchical relationships between two entities, which can occur in a set of documents, are inferred by considering the similarity between the distributions of co-occurring keywords and their string similarity. Besides, this approach handles the ambiguity of keywords that are associated with a set of noisy relationships.

3.4. Labeling Articles using Topic Ontologies

As a considerable related work, we present the CSO Classifier, an ontology-driven classifier of scholarly articles [39] according to the Computer Science Ontology (CSO) [40]. CSO includes 14K semantic topics and 162K relationships[5]. The CSO Classifier takes as input the text from the metadata associated with a scholarly article (title, abstract, and keywords) and returns a list of CSO research topics. The selection of topics is performed in three steps: (1) identify all topics in the ontology that are explicitly mentioned, or referred, in the paper; (2) identify semantically related topics, that may not be explicitly referred in the article, by utilizing part-of-speech tagging and world embeddings; the word embeddings are used to compute the semantic similarity between the terms in the document and the CSO concepts; (3) enrich the results by including super-areas topics according to CSO.

4. Our Semantic Relatedness Measure

In this section, we propose a hybrid graph-based semantic relatedness measure within topic ontologies. As a contribution to the community of approaches that tend to overcome the limitations of existing measures (e.g., [19]), we designed our measure as a combination of *path-based*, and *node-based* approaches. Thus, we comprehensively consider the semantic properties of nodes and edges:

- *Weighting of edges*: to differentiate between hierarchical and non-hierarchical relations regarding the properties of the paths. This semantic property aims to overcome the limitation of considering all edges equivalent in path-based approaches.
- *Weighting and Correlation of nodes*: to consider semantic properties of concepts independently from textual resources. The weighting and computing the correlation of a concept aim to measure its neighborhood and coverage in the ontology graph respectively [22]. The application of such semantic properties can overcome the limitation of dependency of texts in node-based approaches.

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, $\tau : E \to T$ is a function that maps edges in E to their types in T {subclassOf, part of, used by, ...}, $\omega : V \to \mathbb{R}^+$ is a node-weighting function that maps nodes to their weights and $\delta : E \to \mathbb{R}^+$ 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*

[5] https://w3id.org/cso/downloads, last visited February 4, 2021

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 the general topic ontologies, the allocation of weights for edges depends mainly on the relations types. 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]. Based on the experimentation, the non-hierarchical relations are given a cost $\in [0.1, 0.4]$. In this study, we applied 0.25 being a discriminant value.

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 a 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 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:

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

The *correctness* of the semantic path between two nodes is prescribed based on the constraints proposed in [9]. If a path P($v_i \rightarrow v_j$) changes the direction from *upward* (generalization) to *downward* (specialization) at a point related to a hierarchical link, P($v_i \rightarrow v_j$) 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. Topic-OPA: A Topic Ontology for Modeling Topics of Old Press Articles

In this study, Rel_{Topic} is applied within topic ontologies to compute the relatedness of instances and topics. For this purpose, we need to build a topic ontology that represents the domain of old press articles. In this section, we present Topic-OPA, a topic ontology, harvested from open knowledge graphs, for modeling topics of old press articles.

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 items are described. The data model of Wikidata is based on a directed, labeled graph where entities are connected by edges that are labeled 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*). In the following, we discuss the ontology definition, specification, requirements, and development.

5.1. Ontology Definition

Topic-OPA is defined as a general topic ontology 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.*

5.2. *Ontology Specification and Requirements*

The ontology specification specifies the purpose and the scope of the topic ontology. Concerning the purpose, Topic-OPA is intended to be used as a knowledge base for a topic labeling system in the domain of old press articles. Regarding the scope, Topic-OPA is *application-based domain-dependent* ontology. For example, given a corpus of articles of the year 1920, Topic-OPA is constructed from all the disambiguated named entities representing these articles.

For the requirements [29], Topic-OPA has a *functional* requirement that requires the definition of two different schemes in the ontology: hierarchical and non-hierarchical.

- *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.

Besides, Topic-OPA has a non-functional requirement that considers data *traceability* and *scalability* by mapping the concepts and the relations of Topic-OPA to entities in open knowledge graphs such as Wikidata.

5.3. *Ontology Development: SPARQL-Based Approach*

This section discusses a SPARQL-based approach that aims to harvest topic ontologies from open knowledge graphs. A main requirement for this approach is that the domain application is represented by a set of disambiguated named entities. The proposed approach is composed of three main phases: (1) construction of the hierarchical scheme, (2) construction of the non-hierarchical scheme and (3) ontology enrichment. In this study, the ontology development phases are applied in Wikidata.

5.3.1. *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 ordered 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* (Figure 5).

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$(Q333091)\}$. 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 an 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.}
```

Thereafter, examples 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, Politician $\sqsubseteq$ Professional}.

5.3.2. Building the Non-Hierarchical Scheme

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

```
CONSTRUCT { ?domain ?property ?range.
  ?range rdfs:label ?rangeLabel.
  ?property rdfs:label ?propertyLabel.}
WHERE {
  VALUES ?property
  { wdt:P1269 wdt:P425 wdt:P101
  wdt:P136 wdt:P527 wdt:P1552 wdt:P1557 wdt:P106
  wdt:P2388 wdt:P2389 wdt:P361 wdt:P710 wdt:P3095
  wdt:P4646 wdt:P641 wdt:P2578 wdt:P366 wdt:P1535}
  {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. Excerpts of these triples are presented in what follows.

NH={(Civil Servant,*field of this occupation*,Civil Service), (Politician,*field of this occupation*,Politics), (Judge $\sqcap$ Magistrate ,*field of this occupation*,Judiciary), (Public Employee,*facet of*,Public Sector $\sqcap$ Government)}

5.3.3. 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. Thereafter, excerpt of the appended hierarchical relations is presented.

H={Political Organization $\sqsubseteq$ Organization, Government $\sqsubseteq$ Political Organization, Judiciary $\sqsubseteq$ Authority, Civil service $\sqsubseteq$ Organization, Politics $\sqsubseteq$ Activity}

6. The Topic Labeling Process

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

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 to contextualize the articles. The disambiguated named entities will be assigned as Topic-OPA instances and thereby 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 5.3.1). 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 $\sqcap$ Jurist $\sqcap$ Lawyer.

6.2. Instance-Topic Mapping: Classification of Topics

Let us consider the article A_i, which is represented by a set of instances I, and T the set of topic concepts of 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_i, $\forall i \in I$, $\exists T_i \subset T$, $\forall t \in T_i$, $Rel_{Topic}(i,t) \geq \overline{Rel_{Topic}}(I,T)$. 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_i. For this purpose, we define the following procedure:

1. Eliminate the non relevant concepts based on three criteria:
   (a) *Level of abstraction*: remove most abstract topic concepts such as, Entity, Occurrent and Knowledge, by considering their depths. In Topic-OPA, these concepts' depths are less than the average of the depths of all the topic concepts.
   (b) *Hypernyms of named entities*: remove 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.
   (c) *Hyponyms of general concepts*: remove 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.
2. Compute the most common topic concepts T_c from $T_n = \sum T_i, \forall i \in I$.
3. Compute the size of T_c.
4. If $|T_c| = 1$, then $T_c = \{t_c\}$ is the unique labeling topic of A_i.
5. Otherwise, if $|T_c| > 1$ calculate the average of the semantic relatedness values $\overline{Rel_{Topic}}(i,t_c)$, for $Rel_{Topic}(i,t_c) \geq Rel_{Topic}(I,T)$, $\forall t_c \in T_c, \forall i \in I$.
6. Define two strategies to rank T_c and to select the top-ranked topic(s) that label A_i: *relatedness-guided* and *centrality-guided*. The relatedness-guided strategy aims to select the most related topic concept(s) according to the relatedness values' average. Meanwhile, the centrality-guided strategy selects the most connected topic concept(s) based on the degree centrality values. Thus, the further considers the content of A_i, and the latter observes the *semantic relevance* of the topic concepts. By applying the dual strategy, we extend the selection of the best topics that label A_i.
   (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.
7. Finally, compute the topic labeling set of A_i, $T_k = T_d \cup T_r$, 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 old French newspaper *Le Matin*. For this purpose, $A = \{A_1, A_2, ..., A_{48}\}$ a corpus of 48 articles, published between 1910 and 1937, is selected. Every article $A_i \in A$ is described by an XML file consisting of N_i a set of disambiguated named entities represented by Wikidata URIs (see Figure 5 for an example). Generally, the named entities representing the articles are the outcome of WP2. However, in this work, they are collected manually following the hypothesis presented in section 2. Besides, T, a set of topics representing all the topic concepts of Topic-OPA, is considered. In Figures 3 and 4, $\{A_1, A_2, ..., A_8\}$ a subset of A is illustrated. Our main goal is to automatically label the articles by applying 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 of Le Matin

For Building Topic-OPA representing *Le Matin*, a set of $N = 392$ named entities representing A is considered, and the SPARQL-based automatic approach (section 5.3) 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[6]. Note that the topic ontology is not curated. We maintained the concepts and relations as obtained by the application of the SPARQL-based approach. Thus, Topic-OPA contains 2073 concepts, 3261 SubClassOf relations and 1135 non-hierarchical relations. In Figures 6 and 7, we depict excerpts of Topic-OPA around the Politics and Medicine topics. The solid lines represent the SubClassOf relations, and the dashed lines represent the non-hierarchical relations.

7.2. Assignment of Disambiguated Named Entities as Instances

For each article $A_i \in A$, the disambiguated named entities are assigned as instances of Topic-OPA. Therefore, $\forall A_i \in A$, A_i is represented by a set of instances I_i. In Table 1, we show the assignment of the named 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 vast number of general concepts. This implies that the average of the relatedness values is low (around 0.28). Such a 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 $threshold(A_i) = -log_{10}(\overline{Rel_{Topic}}(I_i, T))$, in order to shift the average value of the threshold to the interesting range.

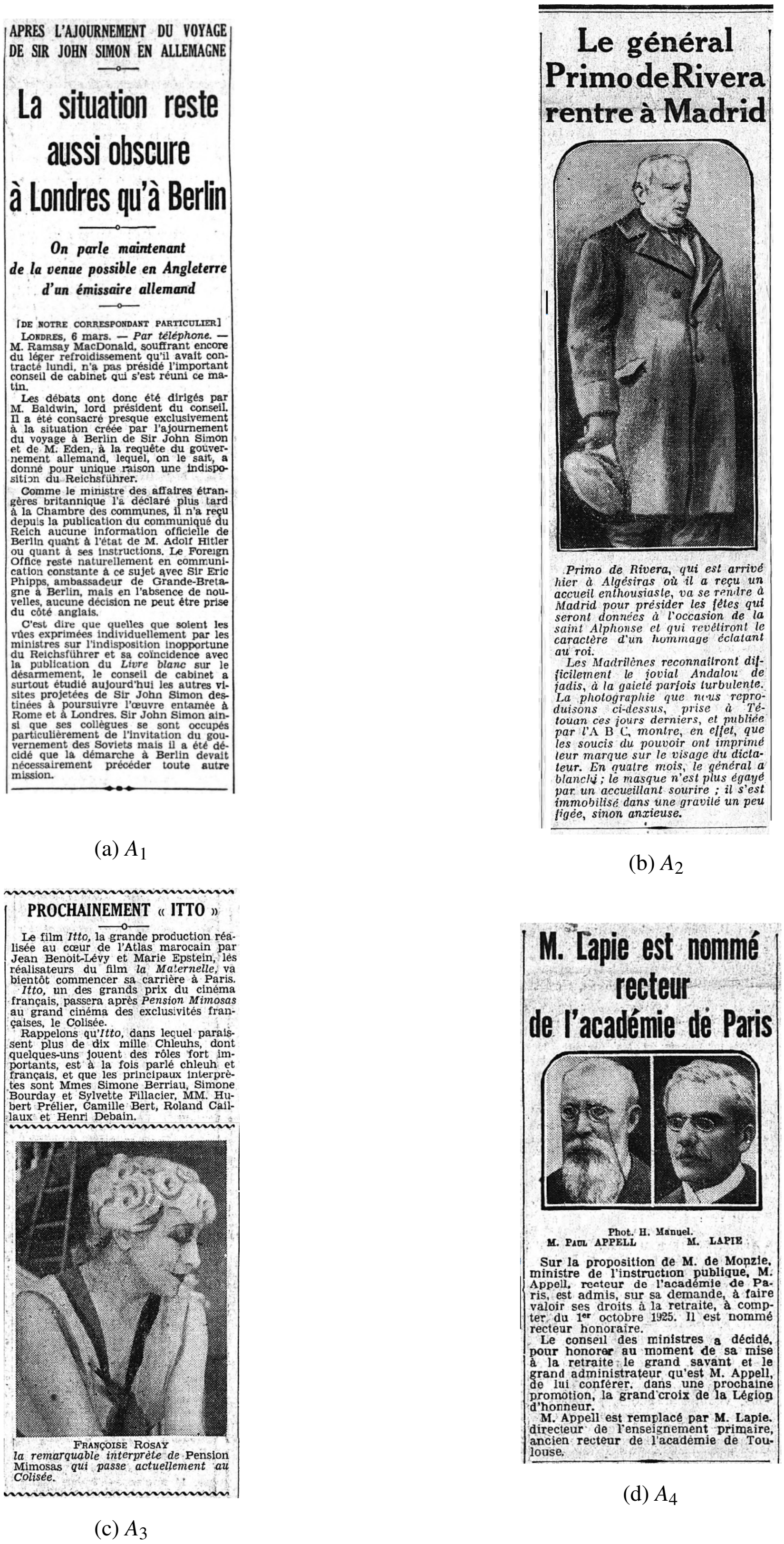
For instance, by referring to the articles A_7 and A_8, the averages of the relatedness values are $\overline{Rel_{Topic}}(I_7, T) = 0.26$ and $\overline{Rel_{Topic}}(I_8, T) = 0.30$. Hence, the threshold values are: $threshold(A_7) = -log_{10}(0.26) = 0.55$ and $threshold(A_8) = -log_{10}(0.30) = 0.52$. 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 $Rel_{Topic}(i,t) \geq -log_{10}(\overline{Rel_{Topic}}(I_i, T))$.

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, $Rel_{Topic}(i,t) \geq -log_{10}(\overline{Rel_{Topic}}(I_7, T)) \forall i \in I_7 and \forall t \in T$, 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-

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



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 vûes 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

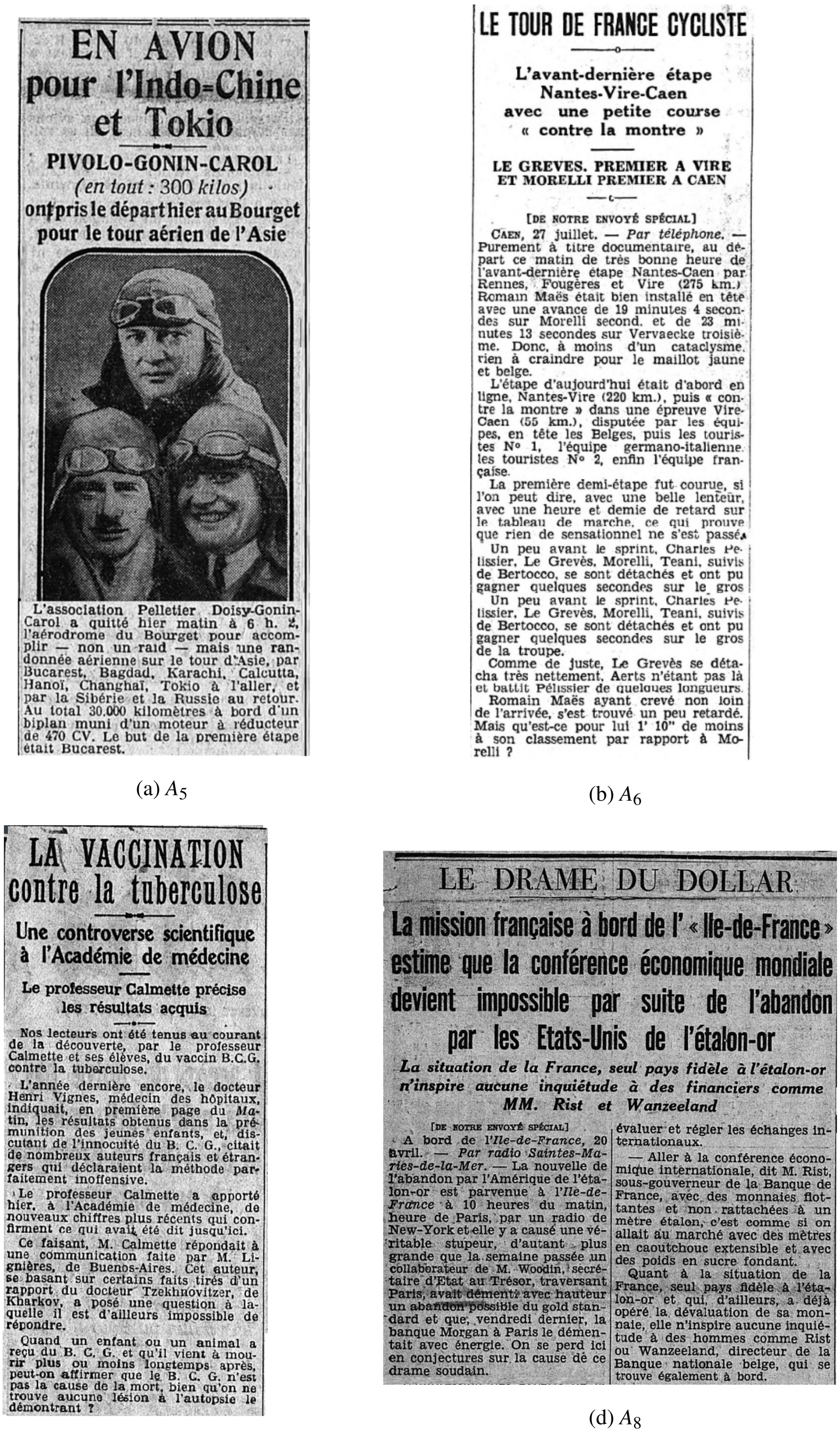
Sur la proposition de M. de Monzie, ministre de l'instruction publique, M. Appell, recteur de l'académie de Paris, est admis, sur sa demande, à faire valoir ses droits à la retraite, à compter du 1^{er} octobre 1925. Il est nommé recteur honoraire.

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. 3. 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. 2, 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 Saintes-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.

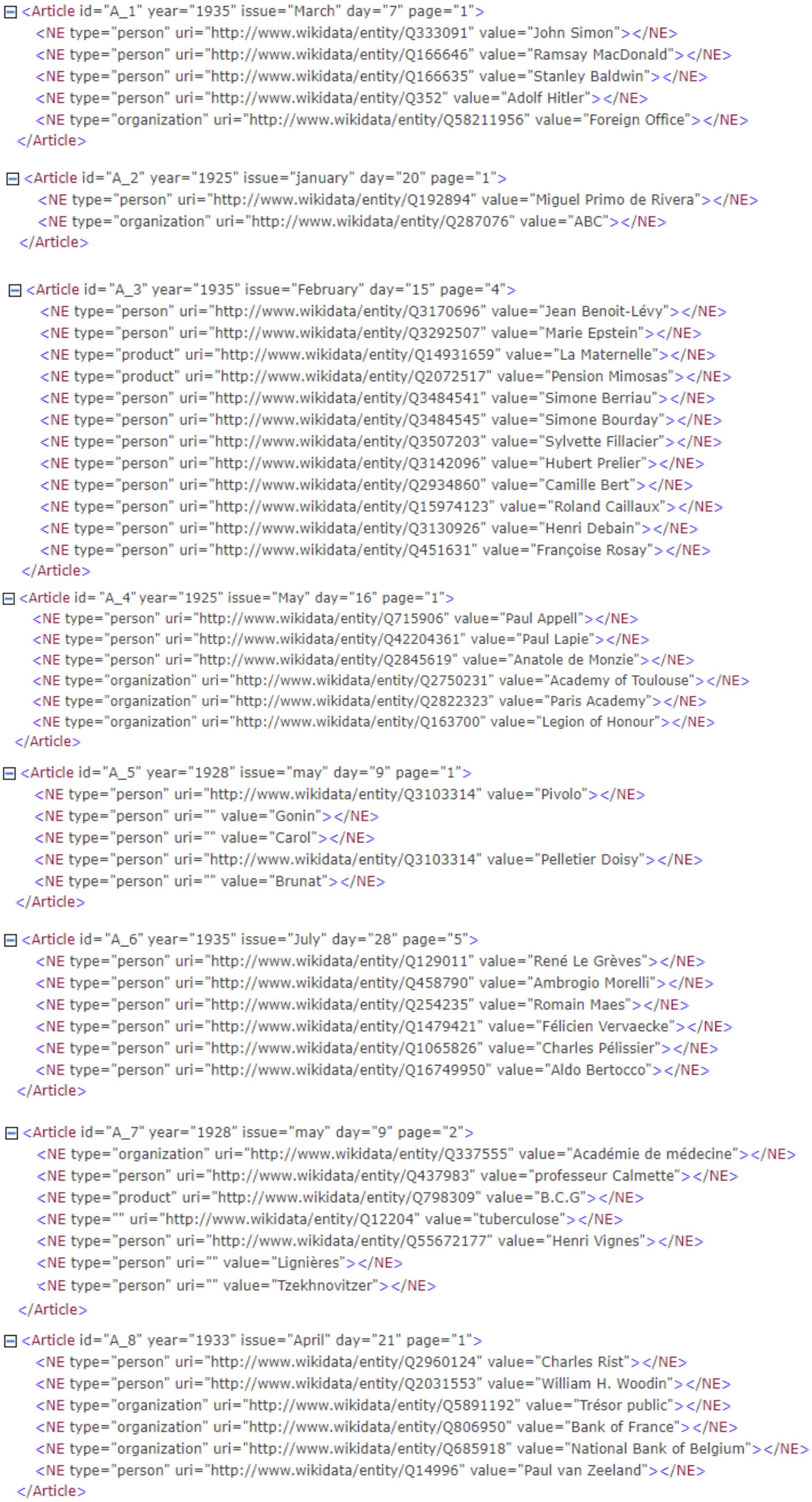
évaluer et régler les échanges internationaux.

— 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. 4. 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. 5. Example of named entities extracted from $\{A_1, A_2, ..., A_8\}$.

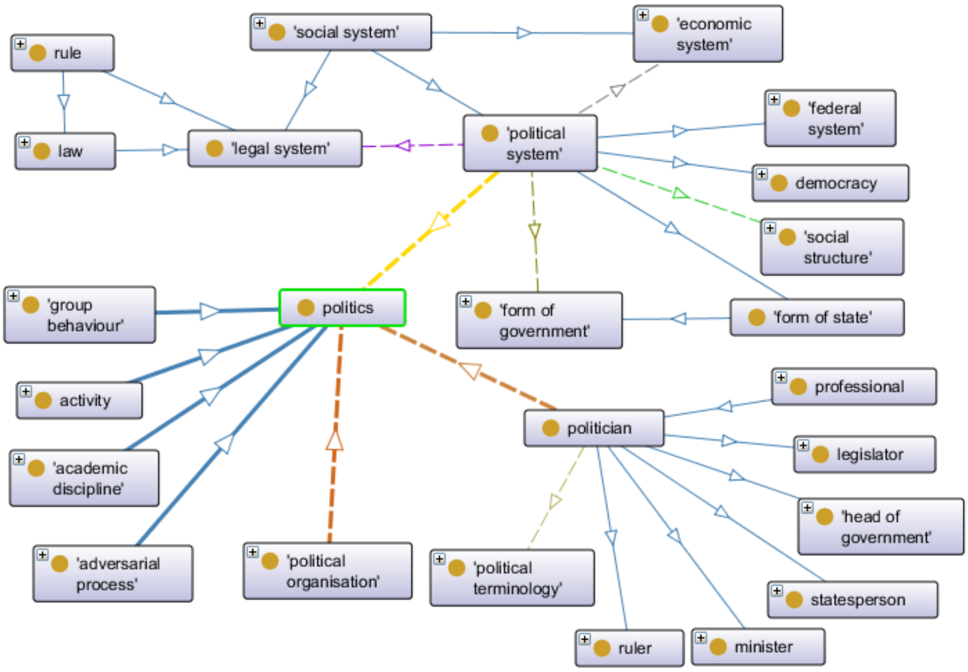


Fig. 6. Excerpt of Topic-OPA around the concept Politics.

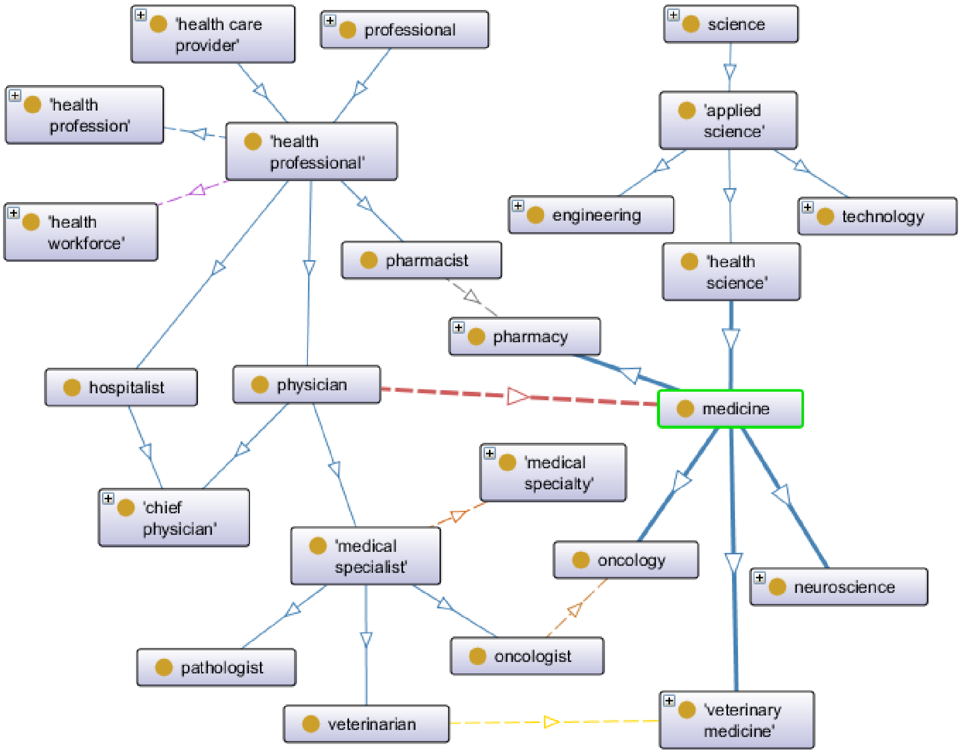


Fig. 7. Excerpt of Topic-OPA around the concept Medicine.

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	$Economist \sqcap Politician \sqcap Lawyer \sqcap Diplomat \sqcap Jurist$

tion process is performed to choose the best topic(s) for labeling A_i. This process has experimented with the 48 articles of *Le Matin*. Table 3 shows an excerpt of the experimental results. It presents thresholds, most common topics, average 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 (section 6.3) for A_7 (Table 2). Note that *step 1* is not shown in the present experimentation.

- By fulfilling *step 1 (b)*, the concepts Academy, National Academy, 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 1 (c)*. Physicist is a hyponym of Person and Research Institute is a hyponym of Organization.
- The aim of *step 2* is to compute the most common topics T_c of A_i. For A_7, $T_c = \{\text{Science} \sqcap \text{Medicine} \sqcap \text{Bacteriology} \sqcap \text{Immunology} \sqcap \text{Virology} \sqcap$

Table 2

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

Instance (i)	Related Topic (t)	$Rel_{Topic}(i,t) \geq -log_{10}(\overline{Rel_{Topic}}(I_7,T))$
Académie Nationale de Médecine	Research Institute	0.80
	Science	0.76
	Academic District	0.69
	Academy	0.80
	National Academy	0.72
Albert Calmette	Physician	0.73
	Medicine	0.66
	Health Professional	0.58
	Immunology	0.64
	Bacteriology	0.64
	Virology	0.64
BCG vaccine	Medication	0.58
	Biopharmaceutical	0.59
	Vaccine	0.75
	Vaccination	0.71
Tuberculose	Disease	0.67
	Health Problem	0.5

Vaccination$\}$. Thus, since $|T_c|= 7$ (*step 3*), *step 4* 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 5* computes the average of relatedness values for each common topic concept $t_c \in T_c$.
- By achieving *step 6* and *step 7*, A_7 is labeled by Vaccination as top-ranked topic having the highest average of relatedness ($\overline{Rel_{Topic}}(I_7, \text{Vaccination}) = 0.71$) as well as the highest *degree centrality* ($D(\text{Vaccination}) = 13.48$).

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

  In addition, A_4 and A_6 are labeled by dual topics by fulfilling *step 6* and *step 7*. 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 Comparison

The first part of this section evaluates Topic-OPA being an application-based ontology. The second part assesses the performance of Rel_{Topic} by evaluating the results of the entire framework (Topic-OPA + Rel_{Topic} + the topic labeling process). Furthermore, we consider applying the whole approach for labeling recent press articles. Finally, Rel_{Topic} is compared to alternative graph-based semantic measures.

8.1. *Evaluation of Topic-OPA*

In the literature, various approaches for evaluating ontologies are recognized. These approaches are categorized depending on what kind of ontologies are being evaluated and for what purpose [30]. Examples of these approaches are [31]: *gold standard-based*, *corpus-based*, *application-based*, and *criteria-based*. 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. Thereby, *gold standard-based* and *corpus-based* approaches are eliminated for the following reasons. The former aims to compare the developed ontology with a previously created *reference ontology*. However, having a suitable gold ontology can be challenging since it should be created under similar conditions with similar goals to the developed ontology. The latter is eliminated since it is strongly dependent on textual resources. Therefore, the *application-based* and *criteria-based* approaches are applied to evaluate the performance and the semantic accuracy of Topic-OPA.

8.1.1. *Application-Based Evaluation*

The application-based approach evaluates the performance of ontologies in a specific task. Topic-OPA is employed for labeling old press articles by using it as a knowledge base. Technically, the semantic relatedness measure Rel_{Topic} is applied to the graph structure of Topic-OPA. Rel_{Topic} performs a "browsing" of the hierarchical and non-hierarchical structure of

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	0.71	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

Topic-OPA. It inspects nodes and edges, their properties, such as weights and depths, and the correlation of nodes, which is defined by the *degree centrality*. Thus, 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 a pertinent ontology if the results obtained by Rel_{Topic} are accurate.

8.1.2. Structure-Based Evaluation

The structure-based approach quantifies how far an ontology adheres to specific desirable criteria (e.g., size and complexity). This approach is recommended as an efficient approach for evaluating the learned ontologies [32]. Several measures have been recognized for the structure-based evaluation such as *Knowledge coverage and popularity* measures (e.g., number of classes and number of properties) and *structural* measures (e.g., maximum depth, average depth, depth variance, Etc.) [31]. The application of these measures relies on the assumption that is *a richly populated ontology, with higher depth and breadth variance, is more likely to provide reliable semantic content*. In contrast to *Knowledge coverage and popularity* measures, 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 quantified the following structural measures by considering the taxonomic structure of Topic-OPA: (1) *Maximum depth*, that represents the length of the longest *taxonomic* branch in the ontology, is measured as the number of concepts from the root node to the leaves of the taxonomy ($maximumdepth = 28$); (2) *Average depth* is computed as the average length of all *taxonomic* branches ($averagedepth = 6$); (3) *Depth variance*, which is the dispersion with respect to the average depth, is computed as the standard mathematical variance ($depthvariance = 6.38$). We conclude that the majority of the topic concepts within Topic-OPA are dispersed homogeneously within the core level. This result implies two essential points. First, it will be challenging for Rel_{Topic} to distinguish between the different concepts located at the same depth to select the best labeling topics. Second, 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 of measuring how well this measure can label a given corpus of articles. Thereby, we evaluate the performance of the whole framework (Topic-OPA + Rel_{Topic} + the topic labeling process) using a dual evaluation approach: (1) a *quantitative* evaluation that compares the automatic labeling to human labeling [13] and (2) a *qualitative* evaluation that appraises the generated topics regarding their semantic interpretability [28].

8.2.1. Quantitative Evaluation

For evaluating the relevance of the generated topics, a *quantitative* evaluation is used by considering human-based labeling [13] and rating [46] methods. For this purpose, A the corpus of 48 articles from *Le Matin*, which is previously used in the use case (section 7), is considered. Since humans can be in contradiction for evaluating specific articles, three different annotators are involved for labeling and rating each article, $A_i \in A$. Concerning the labeling process, the textual content of the articles $A_i \in A$ is assigned to the human annotators. The humans who were blind to Topic-OPA and the results generated by Rel_{Topic}, have read the articles and assigned the labeling topics based on the content (Table 4). Based on human labeling, an inter-annotator evaluation is established to compute the agreement among the annotators for each $A_i \in A$. A comparison of the assigned topics (performed in the context of Wikidata) has shown an agreement of 87.5%: 46% for *exact* topics (e.g., A_3, A_5, A_6), 26% for *specific/general* (e.g., A_7, A_8), and 15.5% for *semantically related* (e.g., A_1, A_2, A_4). Furthermore, we compared Rel_{Topic} topics (see Table 3) with those assigned by humans. Our approach manifested an agreement of 82% with human labeling: 42% for *exact* topics (e.g., A_5, A_6), 34% for *specific/general* (e.g., A_3, A_7, A_8), and 6% for *semantically related* (e.g., A_1, A_4).

Table 4

Excerpt of human labeling of the articles represented in Table 3

A_i	Annotator1	Annotator2	Annotator3
A_1	International Politics	International Relations	International Politics
A_2	Politics	Foreign Policy	Politics
A_3	Cinema	Art, Cinema	Cinema
A_4	Higher Education	Politics, Education	Politics, Science
A_5	Aviation	Event, Exploration, Aviation	Aviation
A_6	Cycling	Sport, Cycling	Cycling
A_7	Medicine	Science, Medicine	Science, Vaccination
A_8	Economics	Economics	Finance

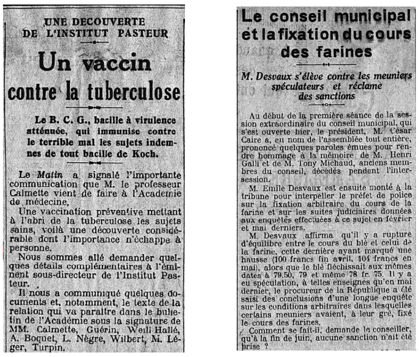
For the rating method, the humans are asked to rate Rel_{Topic} labels for each $A_i \in A$ using the following scores [46]: 3 for *very good* labels; 2 for *reasonable* labels; 1 for *semantically related* labels, but not considered as good topics; 0 for *inappropriate* labels. As a result (see Table 5 for an example), 36% of the Rel_{Topic} topics are assessed as *very good*, 40% as *reasonable*, 14% as *semantically related*, and 10% as inappropriate.

Table 5

Excerpt of human rating of the articles represented in Table 3

Article	Annotator1	Annotator2	Annotator3
A_1	2	2	2
A_2	0	0	2
A_3	2	3	2
A_4	3	2	2
A_5	3	2	3
A_6	3	3	3
A_7	2	2	3
A_8	3	3	2

In the following, we analyze the cases where Rel_{Topic} produced general or irrelevant labels considering the validity of the named entities. In this context, two main issues are observed: (1) the existence of not disambiguated named entities; (2) the lack of some types of named entities. For this purpose, two additional articles are considered A_9 and A_{10} (Figure 8).



UNE DECOUVERTE
DE L'INSTITUT PASTEUR

Un vaccin contre la tuberculose

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

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

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

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

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

(a) A_9, June 27, 1924

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 ?

(b) A_{10}, June 20, 1922

Fig. 8. *Le Matin*

The Existence of Not Disambiguated Named Entities In the presented use-case (section 7), 20 articles have been represented by some named entities that are not disambiguated (e.g., A_5, A_7). In this section, we discuss the influence of these named entities on the relevance of the automatically generated labeling topics. First, we analyzed two articles A_7 (Figure 4) and A_9

(Figure 8a). A_7 consists of 5 disambiguated named entities and 2 that are not disambiguated. Despite this default, Rel_{Topic} assigned Vaccination (Table 3) as a specific topic compared to Medicine assigned by humans. Second, we considered article A_9 which consists of 10 disambiguated named entities and 2 that are not disambiguated. By applying Rel_{Topic}, A_9 is labeled by Science (*step 4* of the topic labeling process). However, Medicine is assigned by the human annotators. 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 Marcel Léger). Robert Wilbert is a Veterinarian[7] and Marcel Léger is a Epidemiologist, Microbiologist and Bacteriologist[8]. 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 labeling relevance degree of A_9.

The Influence of the lack of named entities types 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 lack of some types on the relevance of generated topics. For instance, article A_{10} (Figure 8b) is composed of 6 persons and 2 products and the majority of persons are politicians (Table 6). Thereby, A_{10} is labeled by Politics (*step 4* of the topic labeling process). However, based on the content and the subject of A_{10}, the human annotators have assigned the topic Economics. In this context, we recognized that most politicians, with the absence of organizations or persons related to economics, have affected the labeling results' pertinence.

Table 6

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*

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*[9]. 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 (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 (e.g., $depth(Entity) = 1$, $depth(Object) = 2$ 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 (e.g., Politics, Art, Science, etc.) are located at depths greater than or equal to the average of depths in Topic-OPA (e.g., *depth*(Politics)=5, *depth*(Art)=4 and *depth*(Science)=5). Although, these topics are more general than the specific concepts (e.g., Contract Law, Pharmacy, etc.) which are located at higher levels of depth (e.g., *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.

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

[8] http://www.pathexo.fr/documents/notices/leger.html, last visited April 27, 2020

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

8.3. Evaluation of Topic Labeling using Rel_{Topic} in Recent Press Articles

To evaluate the performance of topic labeling using Rel_{Topic}, we have applied the entire approach on different context of articles such as recent newspapers (e.g., *Le Monde*[10], *Le Figaro*[11], *Liberation*[12]). For this purpose, a corpus of 36 recent articles is considered. The named entities representing these articles are defined and disambiguated manually using Wikidata. The total number of named entities (disambiguated and not disambiguated) is 738. As for old press articles, three types of named entities are considered (*person*, *organization*, and *product*) having a cardinality of 443. In contrast to old press articles, the recent articles are thematically classified using commonly known topics such as Sport, Politics, Art, and Science. The articles of the given corpus are composed of four categories: 9 articles are labeled with Sport, 10 with Science, 9 with Politics and 8 with Art.

To automatically label these articles with Rel_{Topic}, the following phases, which are defined in our approach, are fully applied:

1. Construct a topic ontology representing the application domain named Topic-RPA (Topic ontology for Recent Press Articles). Thus, the SPARQL-based approach (section 5.3) is applied based on the articles' disambiguated named entities (the number of disambiguated named entities is 371). As a result, we obtained Topic-RPA, a not curated topic ontology composed of 2616 concepts, 1584 object properties, and 4132 SubClassOf relations. In contrast to Topic-OPA, Topic-RPA contains contemporary concepts such as Computer Science, Telecommunication, and Electronic Journal. Meanwhile, semantic properties such as the *average depth* is identical in both ontologies ($averageofdepth = 4$). Concerning the ontology size, Topic-RPA is larger than Topic-OPA (+25%).
2. Application of the topic labeling process using Rel_{Topic} (section 6). This phase is composed of three basic steps: (1) assign named entities as instances of Topic-RPA, (2) apply an instance-topic mapping process, and (3) rank and select the best topics that label the recent articles. As a result, Rel_{Topic} has labeled correctly 70% of the articles with exact topics. The inefficiency of Rel_{Topic} to label 30% of the articles (20% of them are *politics* articles) is due to the following major reasons. First, the considerable threshold values close to 0.65 (since the ontology is not curated and thus contains a large number of abstract or *noisy* concepts) make it challenging to select the most commonly related topics for some articles. Second, the named entities that are not disambiguated and the lack of some types of entities have provoked cases similar to old press articles (section 8.2.1).

To conclude, our proposed approach has generated promising results in recent press endorsing its reusability for labeling different textual resource contexts. In this regard, applying the approach in different contexts or domains is independent of languages. It is based mainly on disambiguated named entities detached from any language.

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

In this section, we compare Rel_{Topic} (Equation (8)) with alternative graph-based measures. Specifically, we choose path-based measures since node-based measures are dependent on textual resources, which are out of the scope of our study. To analyze the importance of semantic relatedness regarding semantic similarity, we compared Rel_{Topic} to Sim_{Rada} (Equation (1)). Thus, Sim_{Rada} is applied to the whole graph of Topic-OPA, including the hierarchical and non-hierarchical schemes. Besides, a comparison with the most commonly known semantic relatedness measure Rel_{HS} (Equation (3)) is addressed. For applying Rel_{HS}, there is a need to compute each link's direction change (hierarchical and non-hierarchical) through all the paths. However, this computation is considered a difficult task [19]. To simplify, we computed the direction changes of the hierarchical edges only. Furthermore, we compared the results of applying these measures regarding the instance-topic mapping process on A. Table 7 shows an excerpt of the results of mapping A_7 to Topic-OPA. The results imply that t_i the topic concepts related to $i, \forall i \in I_i$ (I_i are the instances associated to $A_i \in A$) are identified by Rel_{Topic} as well as by Sim_{Rada} and Rel_{HS} (e.g., Education and Research are related to $i =$ Académie Nationale de Médecine in A_7, Figure 9). However, the use of Rel_{Topic} and Rel_{HS} makes also evident the identification of the topics that are not related

[10] https://www.lemonde.fr/

[11] https://www.lefigaro.fr/

[12] https://www.liberation.fr/

to $i, \forall i \in I_i$ due to the considerable gap among the relatedness values (e.g., Economics and Business are not related to $i =$ Académie Nationale de Médecine in A_7). Besides, the results obtained by Rel_{Topic} and Rel_{HS} are close. Nevertheless, the computation of semantic relatedness using Rel_{Topic} is undemanding regarding the edges' direction changes. For an accurate comparison, the relatedness values of Rel_{HS} ($\in [0,8]$, Equation (3)) are converted to $[0,1]$ (division by 8).

9. Discussion

This study's main contribution is the design of a novel graph-based semantic relatedness measure, named Rel_{Topic}, for topic labeling purposes. By proposing Rel_{Topic} as a hybrid measure, we contributed to overcoming node-based and edge-based approaches' limitations. Rel_{Topic} considers hierarchical and non-hierarchical relations and inspects the semantic properties of entities within topic ontologies. Thus, we considered the correlation of nodes to overcome the dependency of measures to textual resources. Besides, we separated hierarchical and non-hierarchical edges using different weights to overcome the limitation of equality of edges in path-based approaches. Rel_{Topic} takes as inputs two entities (e.g., instances and concepts) and returns a numerical value representing their relatedness according to a topic ontology. In this work, Rel_{Topic} is applied mainly for labeling old press articles by assessing the relatedness of instances (named entities) and topic concepts in the topic ontology. Besides Rel_{Topic} is reused for labeling different articles, recent newspapers. The reusability of Rel_{Topic} for purposes requiring the computation of semantic relatedness between entities in a given ontology is demonstrated. However, for this purpose, two main factors are mandatory: (1) a reasonable characterization of the application domain using a domain ontology and (2) a definition of the input entities that should be included in the ontology (e.g., concept-concept, instance-concept).

This study's second contribution is developing the general topic ontology Topic-OPA using a SPARQL-based automatic approach. Topic-OPA is harvested from open knowledge graphs (e.g., Wikidata) based on a set of disambiguated named entities representing the application domain. Topic-OPA is a domain-dependent topic ontology since it is developed from the named entities of the given domain. Nevertheless, if Topic-OPA is developed from all the named entities of the application domain (e.g., *Le Matin*), it could be reused as a topic ontology for labeling old press articles of any journal or newspaper belonging to the same period of time. We assume that approximately the same types of persons (e.g., politician, diplomat, actor, physician, botanist, etc.), organizations (e.g., bank, public treasury, academy, etc,), or products (e.g., vaccine, film, etc.) are available during a comparable period of time (e.g., 1910-1945). Besides, the SPARQL-based approach is reusable (as shown in section 8.3 in the case of recent newspapers) for harvesting ontologies from open knowledge graphs, requiring the starting named entities representing the domain of discourse.

Finally, a significant contribution is applying an ontology-based automatic topic labeling approach for labeling press articles. This process, which is composed of a topic ontology and the semantic relatedness measure Rel_{Topic}, is generalizable for implementing labeling activities for any text, including newspaper or magazine articles. As demonstrated in this work, the entire approach is applied for labeling articles in two different contexts, old and recent press. A primary requirement for the approach reuse is the availability of the named entities representing the text to be labeled. These entities, which are independent of any language, will permit a topic ontology building representing the domain. Thus, Rel_{Topic} will assess the relatedness of each text's named entities to the topics of the topic ontology. Finally, a selection process of the best topics is performed to label the textual resources.

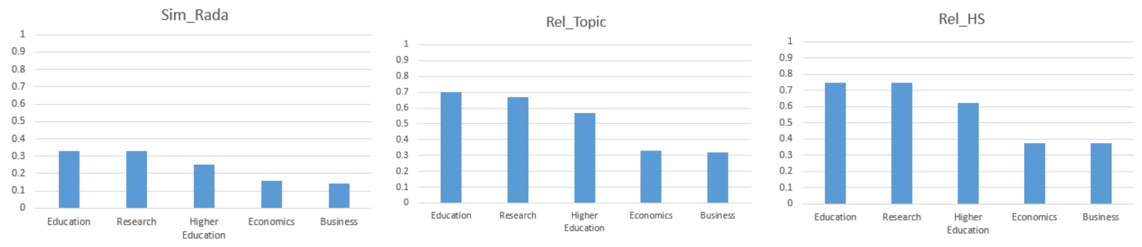
In this context, an important question arises. What if there is a lack of named entities, or if they are ambiguous or inexact? This situation contemplates the validity of this work's general hypothesis (see section 2). In section 8.2.1, we analyzed the influence of two issues on the generated results: (1) existence of not disambiguated named entities and (2) lack of some types of named entities. Both of them had an impact on the generated topics. However, this impact is relative depending on the named entities representing the text to label. For example, the first issue has affected the generality of the assigned topic (e.g., Science is given instead of Medicine). Meanwhile, the second issue has affected the relevance of the assigned topic (e.g., Politics is given instead of Economics).

To resume, the relevance of the whole framework's outcome is a crucial measure of the validity of this work's hypothesis. Thus, the given named entities representing the articles are valid if Rel_{Topic} and the whole framework achieves relevant labeling topics. This as-

Table 7

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

Instance (i)	**Topic Concepts (t_i)**	Sim_{Rada}	$Rel_{Topic}(i,t_i)$	Rel_{HS}	$Rel_{HS}/8$
Académie Nationale de Médecine	Research Institute	0.5	0.81	7	0.87
	Education	0.33	0.7	6	0.75
	Research	0.33	0.76	6	0.75
	Higher Education	0.25	0.57	5	0.62
	Business	0.14	0.32	3	0.37
	Economics	0.16	0.33	3	0.37
Albert Calmette	Physician	0.5	0.73	7	0.87
	Medicine	0.33	0.66	6	0.75
	Immunology	0.33	0.64	6	0.75
	Science	0.25	0.59	4	0.5
	Business	0.16	0.31	3	0.37
Alfred Boquet	Physician	0.5	0.74	7	0.87
	Veterinary Medicine	0.33	0.69	6	0.75
	Science	0.25	0.61	4	0.5
	Business	0.16	0.31	3	0.37
BCG vaccine	Vaccination	0.33	0.71	6	0.75
	Medication	0.33	0.68	6	0.75
	Health Care	0.25	0.48	4	0.5
	Business	0.14	0.3	2	0.25



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

Fig. 9. Comparison of the results of the instance-topic mapping of A_7 (Académie Nationale de Médecine).

sumption is demonstrated in two different contexts, old and recent press.

10. Conclusion

The task of automatically labeling newspaper articles according to a predefined set of topics is a challenging research issue, specifically in cultural heritage. A pertinent characterization of the application domain is required for this purpose. In the context of the ASTURIAS project, which aims to label a vast number of old press articles automatically, we envisaged graph-based semantic measures. These measures have shown effective results in different areas such as knowledge engineering, Semantic Web, and Natural Language Processing. Graph-based semantic measures are composed of similarity and relatedness measures. The former class is adapted to taxonomies and widely investigated in the community. The latter class is adapted to ontologies, and few attempts have been found in the literature to design such measures. Designing semantic relatedness measures is a challenging research task. Nevertheless, they are valuable since they inspect the semantic properties of entities in ontologies.

In this study, we proposed a novel semantic relatedness measure, named Rel_{Topic}, within topic ontologies for topic labeling of old press articles. In contrast to existing measures, Rel_{Topic} considers *hierarchical* and *non-hierarchical* relations and assesses the relatedness between instances and concepts. To apply Rel_{Topic}, we considered topic ontologies as weighted graphs where nodes and edges are given positive numerical weights. Besides, Rel_{Topic} considers the *degree centrality* of nodes, which reflects the node's surface of connection with regards to the rest of the ontology. For the application of Rel_{Topic}, a topic ontology, named Topic-OPA, representing the domain of old press articles, is harvested from Wikidata by applying a SPARQL-based automatic approach.

The proposed approach is evaluated using a dual evaluation approach. First, a quantitative evaluation is performed with the help of three different annotators. The human annotators have assigned labels to a corpus of 48 articles from *Le Matin*. Our approach has shown an agreement quite close to that shown by humans for exact, specific, or general topics. Furthermore, the annotators have rated the results of Rel_{Topic} regarding their relevance. We obtained 76% of the generated topics are rated as *very good* and *reasonable*. The second phase of the evaluation consists in applying a qualitative approach that appraised the semantic interpretability of the automatically generated topics. We noticed that the topic labels within Topic-OPA are highly correlated and located at the ontology's core level. Additionally, the reuse of the entire approach is demonstrated for labeling recent newspaper articles. Promising results are achieved endorsing the reusability of the labeling approach using Rel_{Topic} in different domains. Finally, we compared Rel_{Topic} to alternative graph-based semantic measures. The strength of Rel_{Topic} is its capability to clearly identify the related topics from the non-related topics with an undemanding computation of direction changes of paths.

In future works, we will be interested in the following tasks. First, the contextualization of the articles is envisaged taking into account the named entities of type *location* (e.g., A_1 could be labeled with International Politics, A_3 with Local or French Art and A_6 with French Sport). In this study, we do not consider the topic ontology's curation; we maintained the ontology structure and content, including the abstract and specific concepts, as derived from Wikidata. In further work, we will apply a curation process to clean and leverage Topic-OPA. Furthermore, we will study the application of Rel_{Topic} on the leveraged version of Topic-OPA and assess the generated labeling topics' quality.

Acknowledgments

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

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] R. Speer, J. Chin and C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: AAAI, 2017, pp. 4444–4451.

[37] T. Pedersen, S. V.S. Pakhomov, S. Patwardhan and C.G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, Journal of Biomedical Informatics, Volume 40, Issue 3, June 2007, pp. 288–299.

[38] F. Osborne and E. Motta, Klink-2: Integrating Multiple Web Sources to Generate Semantic Topic Networks. In: The Semantic Web Conference- ISWC 2015. Lecture Notes in Computer Science, Springer International Publishing, Cham, 2015, pp. 408-–424.

[39] A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne and E. motta: The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles. In: International Conference on Theory and Practice of Digital Libraries. springer, Cham, 2015.

[40] A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne and E. motta: The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas. In: The Semantic Web – ISWC, 2018.

[41] Y. Liu, B.T. Mclennes, T. Pedersen, G. Melton-Meaux and S. Pakhomov. Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 2012, pp. 363–372.

[42] S. Banerjee and T. Pedersen. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, 2003, pp. 805–810.

[43] T. Hofmann. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM New York, pp 50-–57, 1999.

[44] Y. Andrew, M.D. Blei and M.I. Jordan, Latent dirichlet allocation. The Journal of Machine Learning Research 3 (2003), 993—1022.

[45] E. Chernyak, An Approach to the Problem of Annotation of Research Publications. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15. pp. 429-–434. ACM Press, USA, 2015.

[46] J.H. Lau, K. Grieser, D. Newman and T. Baldwin, Automatic Labelling of Topic Models. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1536–1545. Association for Computational Linguistics, USA, 2011.