Hindawi
Scientific Programming
Volume 2017, Article ID 7831897, 10 pages
https://doi.org/10.1155/2017/7831897

Research Article

Semantic Annotation of Unstructured Documents Using Concepts Similarity

Fernando Pech,[1] Alicia Martinez,[1] Hugo Estrada,[2] and Yasmin Hernandez[3]

[1]*National Center of Research and Technological Development (CENIDET), Cuernavaca, MOR, Mexico*
[2]*Center for Research and Innovation in Information and Communications Technologies, Ciudad de México, Mexico*
[3]*National Institute of Electricity and Clean Energy (INEEL), Cuernavaca, MOR, Mexico*

Correspondence should be addressed to Fernando Pech; fpech@cenidet.edu.mx

Received 17 June 2017; Revised 2 October 2017; Accepted 8 November 2017; Published 7 December 2017

Academic Editor: José María Álvarez-Rodríguez

There is a large amount of information in the form of unstructured documents which pose challenges in the information storage, search, and retrieval. This situation has given rise to several information search approaches. Some proposals take into account the contextual meaning of the terms specified in the query. Semantic annotation technique can help to retrieve and extract information in unstructured documents. We propose a semantic annotation strategy for unstructured documents as part of a semantic search engine. In this proposal, ontologies are used to determine the context of the entities specified in the query. Our strategy for extracting the context is focused on concepts similarity. Each relevant term of the document is associated with an instance in the ontology. The similarity between each of the explicit relationships is measured through the combination of two types of associations: the association between each pair of concepts and the calculation of the weight of the relationships.

1. Introduction

The rapid growth of the web has generated an enormous amount of information in the form of unstructured documents. Search engines have become common and basic tools for users. However, engines still have difficulties in performing searches because search methods are based on keywords, and they do not capture and do not explore the meaning and context of the need of the user. This challenge has drawn attention of several research groups which are interested in solving the issues associated with information storage and search and retrieval of information in this enormous cumulus of data.

On the other hand, the continuous growth of the Semantic Web has motivated the development of knowledge structures on different domains and applications, like Wikipedia [1], Linked Open Data (LOD) [2], DBpedia [3], Freebase [4], and YAGO [5], among other applications. Additionally, some ontologies for several domains have been developed, such as Snomed CT [6] and UMLS [7] for the medical field and AGROVOC [8] for the agricultural field. An ontology is a formal representation of knowledge, which plays a very important role in the semantic web because of its capability to express meanings and relationships. Ontologies have been valuable in knowledge extraction technologies, especially in the aggregation of knowledge from unstructured documents. Ontologies are a key component of semantic association, which is the process to formalizing knowledge through the linking of words or phrases of plain text (mentions or named entities) with elements of the ontology (concepts or entities).

The semantic annotation of a document consists in finding mappings between text chunks of a document and the instances or individuals in ontology. The annotation plays an important role in a variety of semantic applications, such as generation of linked data, extraction of open information, alignment of ontologies, and semantic search. Specifically, semantic search allows users to express their information needs in terms of the knowledge base concepts. Unlike traditional keyword-based search, semantic search can make use of semantic relationships in the ontology to accomplish new tasks, such as refining user queries with broader or more specific concepts.

The semantic annotation has been applied in different areas of knowledge. For example, it has been applied in biological systems for the identification of biomedical entities such as genes, proteins, and their relationships; also, it has been applied in news analysis for identification of people, organizations, and places.

At the present, semantic annotation strategies are carried out without regard to context [9–11]; these works do not analyze the meaning or semantics of the terms. Generally, authors assume that lexicons are enough to express the meaning of the terms in a document. However, to a large degree, the semantic of a concept depends of the context in which it occurs. Therefore, the identification of meaning could lead to problems of ambiguity. Several research works have demonstrated the complexity of word-sense disambiguation (WSD), where traditionally a term is searched in a data dictionary (e.g., WordNet) [12]. Other approaches have chosen to analyze the context of the terms to improve the annotation process [13]. The problems related to semantic annotation are still an open research topic.

The annotation process could be a source of different types of problems, for example, (i) ambiguous annotations, when entities have been assigned to more than one concept in the ontology, (ii) erroneous annotations, when the meaning of a text is not found in the ontology, and, (iii) false annotations, when the annotation does not provide any value for the realization of a semantic search. In this sense, this paper presents a strategy of semantic annotation in unstructured documents. Our approach is based on ontologies and on the extraction of contextual semantic information from entities of the ontology. The semantic context of an entity is determined by their relationships in the ontology. Therefore, we propose to extract the semantic context of the entities by calculating the similarity of association between each pair of concepts and the calculation of the weights of the relationships of the entities. With this strategy, we deal with the problems of ambiguous, erroneous, and false annotations. Our method of semantic annotation is part of a semantic search system in natural language and it has been evaluated with the corpus compiled by Lee and Welsh [14] and DBpedia.

This paper is organized as follows: Section 2 describes the background of our proposal, Section 3 presents the related work, Section 4 presents the architecture of the system, Section 5 presents the evaluation of the proposed approach, and finally, Section 6 provides some conclusions and an outlook for future work.

2. Foundations

This section presents the concepts and foundations of the proposed semantic annotation approach.

2.1. Ontology. An ontology is composed of a schema and instances (see Figure 1). A schema is defined as $\langle C, D, P\rangle$ where C is the set of classes/concepts $C = \{c_1, c_2, \ldots, c_n\}$, D is the set of data types, and P is the set of properties $P = \{p_1, p_2, \ldots, p_n\}$ which are the relationships between classes. Instances represent knowledge and denote an instanced class and their relationships. Instances can be defined as a graph

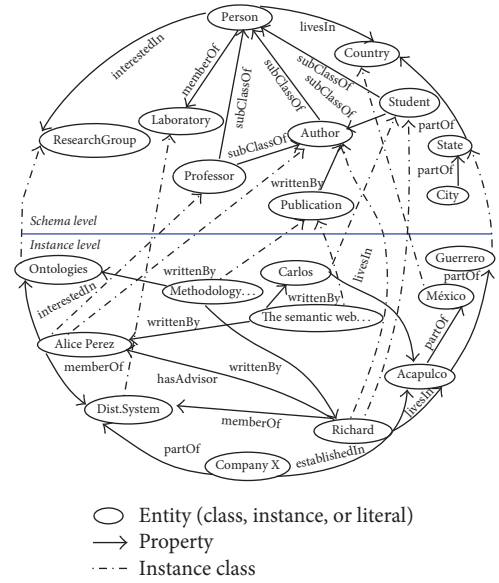

Figure 1: Two-level ontology: schema and instances.

$G = \langle V, E\rangle$, where V is the set of instances, and E is the set of relationships or predicates binding the instances.

In an ontology, classes, properties, data types, and instances are explicitly identified by Uniform Resource Identifiers (URI). In addition, they represent entities within the ontology, which are characterized by their textual description declared in the property *rdfs:label*. This may have lexical variations defined as *rdfs:label* = {"*text1*","*text2*"}.

Figure 1 shows a fragment of an ontology for the research domain. The schema level defines classes such as *Laboratory* and *Professor*, and properties such as *interestedIn*.

The instance level indicates the instantiated schemas. For example, *ontologies* is an instance of the class *ResearchGroup; Methodology,* and *Alice Perez* are related to the property *writtenBy* and belong to the classes *Publication* and *Author*, respectively. The *Acapulco* instance contains its textual description with two lexical variations *rdfs:label* = {"*Acapulco*","*Acapulco de Juárez*"}.

2.2. Semantic Annotation. The semantic annotation is fundamental to obtaining better results in the semantic search because the documents are represented in a conceptual space.

The semantic annotation of a document d consists in linking the terms t in $d = \{t_1, t_2, \ldots, t_n\}$ with the entities in the ontology which describe the content of the term in its textual description best (see Figure 2). Namely, let an entity-term pair be $\langle c, t\rangle$, where c is an entity in the ontology and t is a term/phrase of d, so that there is a mapping between the textual descriptions defined in the label *rdfs:label* of c and t.

In semantic annotation techniques, a document is analyzed in order to identify its relevant terms and to define the importance of each term. There are tools to identify mentions, such as TagMe [15] and Spotlight [16].

When the semantic annotations are made without regard to the context, its terms or mentions are linked with the

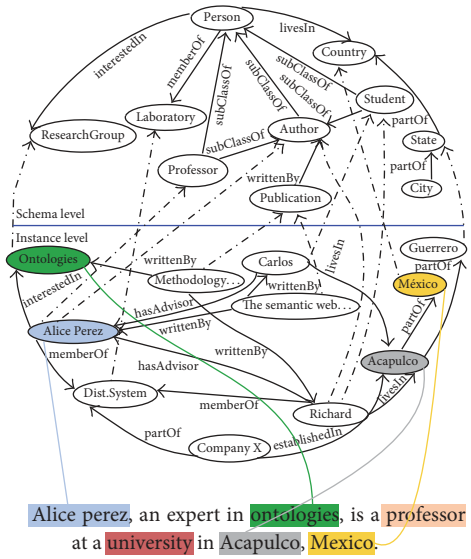

FIGURE 2: Link between terms (mentions) of a document and the ontology entities.

entities in the ontology without taking into account their meaning. This causes ambiguous or erroneous annotations.

Our research work proposes to analyze the context of the annotations in order to identify their meaning through the entities in the ontology, and in this way to avoid ambiguities. In the extraction of the context, the explicit relationships of each entity in the ontology are analyzed. For example, Figure 2 shows the relationship between the *Ontologies* entity and *ResearchGroup* and *Alice Perez*.

3. Related Work

The semantic search involves different components: (i) preprocessing, (ii) semantic query translator, (iii) semantic annotation and indexing, (iv) retrieval of semantic content, and (v) semantic ranking.

Currently, there are several research works with different contributions in the area of the semantic web. Several general-purpose tools have been developed to support the annotation process, and, also, specific domain ontologies and knowledge bases have been proposed by research groups.

General-Purpose Tools. There are several available services for annotation of named entities in documents that could be accessed using RESTful APIs such as the case of OpenCalais [17].

Let us remark that, AlchemyAPI [18] and OpenCalais [17] use context-based statistical techniques to disambiguate the candidate instances to annotate a term. These tools use proprietary vocabularies and ontologies whose instances are linked to DBpedia through the *owl:sameAs* relationship. However, OpenCalais provides some limited linkage to DBpedia. Also, OpenCalais is mainly focused on organizations. This approach has two disadvantages. Firstly, it only explores the surface of the graph for each DBpedia instance considering the labels, abstract, links to Wiki pages, and synonyms. Secondly, this approach annotates a term with only one instance of DBpedia. Therefore, this approach does not exploit the semantic information available in DBpedia to disambiguate the instance annotating a given term.

DBpedia Spotlight [16] is a semantic annotation tool for data entities in a document and it is based on DBpedia for the annotation. Also, this tool provides interfaces for disambiguation, including a Web API which supports XML, JSON, and RFD formats.

Gate [19] is a tool for text engineering to help users in the process of text annotation manually. This tool provides basic processing functionalities, such as recognition of entity named, sentence dividers, markers, and so on.

Ontea [20] is a tool for semantic metadata extraction from documents. This tool uses regular expressions patterns as text analysis tool, and it detects semantically equivalent elements according to the domain ontology defined in the tool. This tool creates a new individual ontology from a defined class and it assigns the detected elements as properties in the ontology class. The patterns of regular expressions are used to annotate the text without format with elements in the ontology.

These approaches have two main drawbacks. On the one hand, they just explore the surface of the graph for each DBpedia instance; they mainly consider label, abstract, links to Wiki pages, and synonyms. Therefore, these approaches do not exploit the semantic information available in DBpedia to disambiguate the instance annotating a given term. Another disadvantage of this work lies in the fact that it discards the relationship, which contains relevant information about a term. That is, they do not enrich the description of relevant terms with the semantic graphs that contain the DBpedia instances related to the context of the document. Some works do face these drawbacks by annotating their documents with graphs extracted from DBpedia.

Specific Domain Tools. There are specific tools for biomedical annotations such as MetaMap [8], Whatizi [21], and Semantator [22]. Most of this approaches and tools are based on a strategy to search terms in thesaurus. These methods consist in finding occurrences of a concept chain in a text fragment using strict coincidence of terms.

Semantic Annotation Approaches Based on Information Retrieval Techniques. Popov and colleagues [23] present KIM, a platform for information and knowledge management, annotation, and indexed and semantic retrieval. This tool provides a scalar infrastructure for personalized information extraction and also for documents management and its corresponding annotations. The main contribution of KIM is the recognition of the named entities according to ontology.

Castells et al. [24] propose an information retrieval model using ontologies for the annotation classification. This model uses an ontology-based schema for the semiautomatic semantic annotation of documents. This research was extended by Fernández et al. [25] to provide natural language queries.

Berlanga et al. [26] propose a semantic annotation/query strategy for a corpus using several knowledge bases. This

method is based on a statistical framework where the concepts of the knowledge bases and the corpus documents are homogeneously represented through statistical models of language. This enables the effective semantic annotation of the corpus.

Nebot and Berlanga [27] explore the use of semantic annotation in the biomedical domain. They present a scalable method to extract domain-independent relationships. They propose a probabilistic approach to measure the synonymy relationship and also a method to discover abstract semantic relationships automatically.

Fuentes-Lorenzo et al. [28] propose a tool to improve the quality of results of the Web search engines, performing a better classification of the query results.

In the literature we can find several approaches to optimize query results. Swoogle [29] is a raster based system to discover, index, and query RDF documents. SemSearch [30] is another search engine relying on semantic indexes and is based on Sesame [31] and Lucene. The ranking algorithm was specifically designed for the extraction of ontologies through annotation. In [32] a search engine is proposed to infer the context of Web pages and also to create links to relevant Web pages. Lopez et al. [33] developed an information retrieval system based on ontologies. This system takes as input a natural language query and converts it to semantic entities using a question-answering system. PowerAqua [33] is a system to recover and to classify documents through TF-IDF measures [34].

4. Semantic Annotation Architecture

This paper presents a novel semantic annotation approach based on ontologies for the improvement of information search in unstructured documents. We present an approach to annotation that enriches and semantically describes the content of a document using the similarity of entities of an ontology. Specifically calculating (1) the association between each concept pair and (2) the relationships weight.

The goals of our approach are (a) to link the entities with their meaning in order to be annotated and (b) to provide a framework for semantic searches using natural language processing. The semantic annotation approach extracts the semantic context through the similarity analysis calculating the association of the explicit relationships and the weight of the relationships of the entities involved. Figure 3 shows an overview of our proposed solution for the semantic annotation.

4.1. Documents Indexing. Commonly, Natural Language Processing (NLP) is used for the analysis of unstructured documents, and also for the recognition and extraction of mentions or named entities [35].

In this approach, the indexing of unstructured web documents generates inverted indexes, which contain the set of terms to be compared with the entities in the ontology. We propose an algorithm for the indexing of documents using Lucene. The output of this algorithm is an inverted index containing the list of terms or keywords and a set of documents where the terms appear.

Therefore, the algorithm provides a mapping from terms to documents and a mechanism for annotating search results. Also, it obtains the position of the information: the list of terms IDs, the association with the ID of the document, and its position.

4.2. Entity Identifications. Given a document d and a knowledge base, the objective of this phase is to extract the textual descriptions and the semantic context of all the information about d from the knowledge base.

Identification of Mentions. Documents are analyzed to detect terms. Generally, this process is known as acknowledgment of mentions or named entities [35]. A mention is a term/phrase in the text which may correspond to an entity in the knowledge base.

From the ontological point of view, an entity can denote classes, relationships, or instances. Entities can represent people, organizations, locations, and so on. There are different tools to define entities, like Spotlight [16] and TagMe [15], among others. TagMe uses Wikipedia as a dictionary of terms for mentions detection. We have used this tool with the same purpose.

TagMe analyzes the input text and detects mentions using a dictionary of entities/words (surface form). For each word, it registers the set of entities recognized by that name. This dictionary is constructed by extracting the words from four sources: Wikipedia papers, redirected pages, Wikipedia page titles, and other variants.

Words with few occurrences and single-character words are discarded. Finally, an additional filtering to discard words with low link probability is done (e.g., less than 0.001). The link probability is defined as stated in

$$p\text{link}(m) = P(\text{link} \mid m)\,\frac{\text{link}(m)}{\text{freq}(m)}, \tag{1}$$

where link(m) is the number of times the mention m appears as a link and freq(m) denotes the number of times the mention m occurs in Wikipedia.

The detection of mentions is carried out by comparing the n-grams (until $n = 6$) of the document.

4.2.1. Extraction of Instances. Each mention detected in document d is searched in the ontology, and if an instance matches its textual description, it is extracted from the label *rdfs:label*. All the values contained in *rdfs:label* (lexical variations) are considered as labels that are later compared in the document index.

Figure 4 shows a fragment of the *México* entity code containing URI, class, and textual description with two lexical variations *México* and *Estados Unidos Mexicanos*.

4.2.2. Extraction of the Instances Semantic Context. In this process, the semantic context of the instances is extracted to be analyzed in detail. The explicit relationships in the *URI* are also analyzed. Several strategies have been proposed to evaluate the proximity of entities according to their semantic characteristics [21]. The use of the semantic measure based on graphs allows us to compare concepts, terms, and instances.

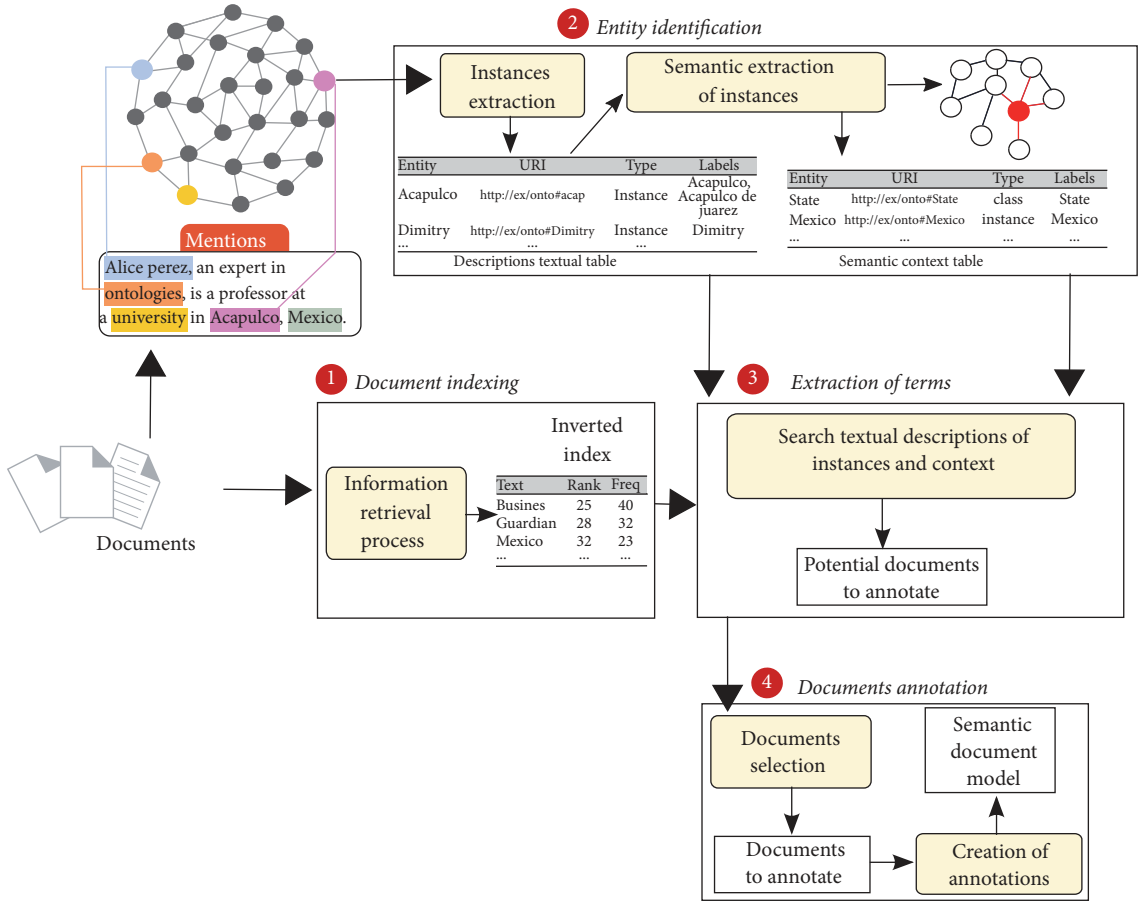

Figure 3: Methodological proposal for semantic annotation process.

```
<owl:NamedIndividual
   rdf:about="http://example/ontea#Mexico">
      <rdf:type rdf:resource="http://example/ontea#Country"/>
      <rdfs:label rdf:datatype="http://www.w3.org/2001/
                                XMLSchema#string">
          Mexico, Estados Unidos Mexicanos
      </rdfs:label>
</owl:NamedIndividual>
```

Figure 4: Code fragment of an instance in the ontology.

This measure is represented as an edge in a semantic graph in order to determine the relationship strength among the ontology concepts.

Therefore, this research work uses the semantic measure as a strategy to measure the strength of the explicit relationship between entities. Two types of measures are considered: (1) the association between each concept pair and (2) the relationships weight. Each measure reflects the similarity degree or relationship between the ontology entities according to its meaning.

Concept Pairwise Association. An entity is explicitly related to other concepts in the ontology. To measure the association strength between each pair of concepts c_1 and c_2, we compare each pairwise by calculating similarity. Figure 2 shows the *Acapulco* entity with four explicitly related concepts (*Carlos*, *Guerrero*, *México*, and *Richard*).

The association strength between each pairwise can be measured taking into account different characteristics, such as the shortest path between concepts pairwise, the depth of their common ancestor, and information content [36].

We have adopted the Resnik approach [37] to measure the similarity between two concepts c_1 and c_2 according to the information content, using the formula

$$\mathrm{Sim}\left(p\left(c_1, c_2\right)\right) = \frac{\mathrm{IC}\left(\mathrm{MSCA}\left(c_1, c_2\right)\right)}{\mathrm{IC}\left(c_1\right) + \mathrm{IC}\left(c_2\right)}, \tag{2}$$

where $\mathrm{MSCA}(c_1, c_2)$ denotes the common ancestor of c_1 and c_2 with the higher information content. IC is the information content calculated for each node c in the ontology, whereas the more specific the node in the ontology is, the greater its information content is. There are different metrics to calculate IC [36].

Generally, these metrics are intrinsic. Namely, they are based on the topological information of the ontology and consider the instances occurrence. This approach considers the occurrence of an instance x quantified as $l(x) = \log_2 \mathrm{pr}(x)$, which has been reformulated as stated in

$$\mathrm{IC} = -\log_2 \frac{I(D(c))}{I(C)}, \tag{3}$$

where $I(D(c))$ denotes the number instances of the concept c and $I(C)$ represents the number of instances on the ontology.

From the ontology in Figure 2 which contains 1000 resources including the entities *Person*, *Publication*, and *ResearchGroup,* we can see a group of 600 people interested in some research group *(ResearchGroup)* and 100 people *(Author)* who wrote some publications *(Publication)*. The information content in *interestedIn* and *writtenBy* is obtained as stated in

$$\begin{aligned}
&\mathrm{IC}(interestedIn(Person, ResearchGroup)) \\
&\quad = -\log_2 \mathrm{pr}(interestedIn(Person, ResearchGroup)) \\
&\quad = -\log_2 \frac{600}{1000} = -\log_2 0.6 \approx 0.73, \\
&\mathrm{IC}(writtenBy(Publication, Author)) \\
&\quad = -\log_2 \mathrm{pr}(writtenBy(Publication, Author)) \\
&\quad = -\log_2 \frac{100}{1000} = \log_2 0.1 \approx 3.32.
\end{aligned} \tag{4}$$

The information content in a property represents the strength of the discrimination among the relationships. However, this is not enough to determine the meaning of the entity. We propose to measure the weight of each property linked to a concept c.

Relationships Weight. Based on information theory, the amount of information contained in a random variable over another variable is measured by mutual information. This strategy has been proposed by Cover [38] and we have adapted it to measure the relationship strength of pairwise c_1 and c_2.

$$\mathrm{MI}(p(c_1, c_2)) = \sum \sum \mathrm{pr}(c_1, c_2) \cdot \log_2 \frac{\mathrm{pr}(c_1, c_2)}{\mathrm{pr}(c_1) \cdot \mathrm{pr}(c_2)}, \tag{5}$$

where $\mathrm{pr}(c_1, c_2)$ is the probability of relationship e belonging to a set of properties of c_1 and c_2. $\mathrm{pr}(c_1)$ is the probability of relationship belonging to set of properties of c_1, whereas $\mathrm{pr}(c_2)$ is the probability of relationship e belonging to set of properties of c_2.

Figure 5 shows the relationships *writtenBy*, *memberOf*, *hasAdvisor*, and *livesIn* belonging to *Richard* entity in the ontology. The instances of these relationships are shown in Figure 6.

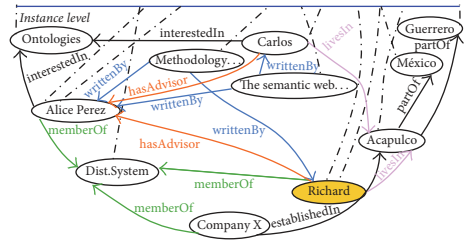

FIGURE 5: Relationships *memberOf*, *writtenBy*, *hasAdvisor*, and *livesIn* in the ontology.

As an example, let us calculate the relationship weight between *Richard* and *Methodology*, which is *writtenBy*, and it is computed as stated in

$$\begin{aligned}
&\mathrm{MI}(writtenBy(Publication, Author)) \\
&\quad = I(Methodology, Richard) \\
&\qquad \cdot \log_2 \left(\frac{I(Methodology, Richard)}{I(Methodology) \cdot I(Richard)} \right) \\
&\qquad + I(Metodology, AlicePerez) \\
&\qquad \cdot \log_2 \left(\frac{I(Methodology, AlicePerez)}{I(Methodology) \cdot I(AlicePerez)} \right) \\
&\qquad + I(TheSemanticWeb, AlicePerez) \\
&\qquad \cdot \log_2 \left(\frac{I(TheSemanticWeb, AlicePerez)}{I(TheSemanticWeb) \cdot I(AlicePerez)} \right) \\
&\qquad + I(TheSemanticWeb, Carlos) \\
&\qquad \cdot \log_2 \left(\frac{I(TheSemanticWeb, Carlos)}{I(TheSemanticWeb) \cdot I(Carlos)} \right) \\
&\quad = \frac{1}{4} \cdot \log_2 \left(\frac{1/4}{(1/2) \cdot (1/4)} \right) + \frac{1}{4} \\
&\qquad \cdot \log_2 \left(\frac{1/4}{(1/2) \cdot (1/2)} \right) + \frac{1}{4} \cdot \log_2 \left(\frac{1/4}{(1/2) \cdot (1/2)} \right) \\
&\qquad + \frac{1}{4} \cdot \log_2 \left(\frac{1/4}{(1/2) \cdot (1/4)} \right) = 0.5.
\end{aligned} \tag{6}$$

It should be noted that a relationship can have many instances. Consequently, calculating the relationships weight would have a high computational cost. Thus, we calculate the mutual information as stated in

$$\mathrm{MI}(e) = \approx \log_2 \left(\frac{1/[I(e)]}{(1/I(c_1)) \cdot (1/I(c_2))} \right), \tag{7}$$

where $[I(e)]$ represents all relationships e in the relationships set, $I(c_1)$ represents all relationships in c_1 (subject), and $I(c_2)$ represents all relationships in c_2 (object).

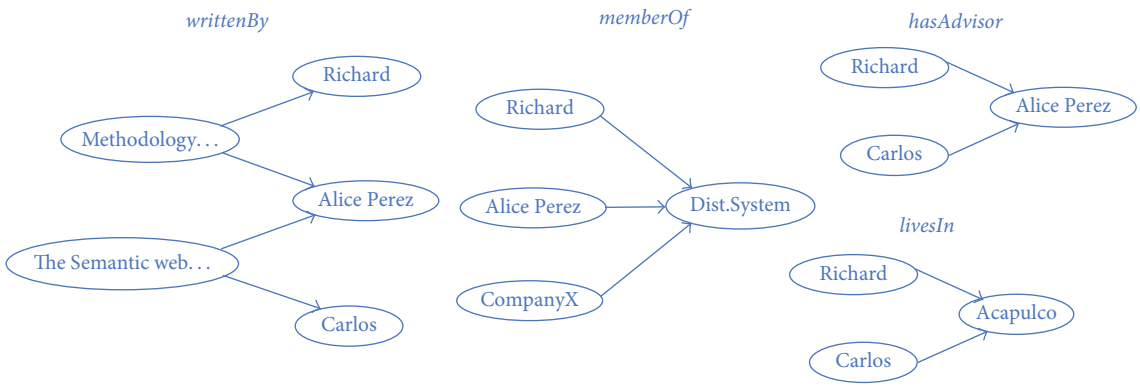

Figure 6: Examples of *writtenBy*, *memberOf*, and *hasAdvisor* entities and *livesIn* property.

Table 1: Document annotation.

Entity	Document	Weight
http://ex/onto#State	D1	0.5
http://ex/onto#State	D2	0.2
http://ex/onto#State	D87	0.67
http://ex/onto#Mexico	D1	0.45
http://ex/onto#Mexico	D23	0.6

Combining Association and Relationship Weights. The combination of weights requires considering several methods of aggregation, such as average, addition, and multiplication. A weighted sum as combination method to adjust the influence of each factor on the total weight was selected. Finally, to combine the association between each pair of concepts (see (2)) and the weights of the relationships (see (7)), we calculate the final weight to obtain the entities context, as stated in

$$W\left(P\left(c_i, c_j\right)\right) = \alpha \cdot \mathrm{Sim}\left(c_1, c_2\right) + \beta \cdot \mathrm{MI}\left(p\left(c_1 c_2\right)\right), \quad (8)$$

where $0 \leq \alpha, \beta \leq 1$. Sim and MI were normalized to be in the 0, 1 range by unit-based normalization [13], stated in

$$\frac{\mathrm{Sim} - \min_{p \in P}\mathrm{Sim}}{\max_{p \in P}\mathrm{Sim} - \min_{p \in P}\mathrm{Sim}}, \quad \frac{\mathrm{MI} - \min_{p \in P}\mathrm{MI}}{\max_{p \in P}\mathrm{MI} - \min_{p \in P}\mathrm{MI}}. \quad (9)$$

4.3. Terms Extraction and Documents Annotation. The textual descriptions of instances and entities semantic context obtained in the previous stage are searched in the inverted index to extract and generate a documents' annotation table containing the ontology entity, the belonging document, and its weight (see Table 1).

The annotations weight is done by means of TF-IDF algorithm. Term frequency (TF) is the local weighting factor reflecting the importance of a term within a document. Document frequency (DF) is the global weighting factor considering the importance of a term within the document collection. Inverse document frequency (IDF) calculates the frequency of a document within the collection. TF and IDF are calculated using the formulas stated in (10) and (11).

$$\mathrm{TF} = \frac{\mathrm{freq}_{x,d}}{\max_y \mathrm{freq}_{y,d}}, \quad (10)$$

where $\mathrm{freq}_{x,d}$ is the number of occurrences of term x within document d and $\max_y \mathrm{freq}_{y,d}$ is the number of occurrences of all terms within document d.

$$\mathrm{IDF} = \log \frac{|N|}{df}, \quad (11)$$

where $|N|$ is the total number of documents in the collection and df represents the documents where term x appears. The weight dx for x in d is the combination of TF $*$ IDF.

Finally, the annotations are represented in the form of serialized triplets in JSON-LD.

5. Evaluation

Pearson and Spearman correlation were used in order to measure the agreement with the human judgments. Pearson correlation measures the linear correlation between two variables, uses the ranges, orders numbers of each group of subjects, and compares those ranges. Spearman is a correlation measure between two continuous random variables.

Experimental Setup

Ontology and KIM Platform Knowledge Base [23]. This ontology has 271 classes, and 120 relationships and attributes. Some declared classes are of general importance such as People, Organizations, Government, and Location. The knowledge base consists of 200,000 instances, 50,000 locations, 130,000 organizations, 6,000 people, and more.

DBpedia [3]. DBpedia is general-purpose and multilingual in nature and has comprehensiveness. For this reason, it was selected for our experimentation. The English version contains 685 classes and 2795 properties; and the knowledge base is more than 4 million instances. DBpedia contains multiple classification systems, such as YAGO, Wikipedia Categories, and the hierarchical subgraph of the DBpedia Ontology. The

Table 2: Summary of Corpus LP50 annotations with KIM and DBpedia.

#doc	Words	Mention detection	Linked KIM	Linked DBpedia
(1)	80	13	8	30
(2)	98	21	10	37
(3)	98	17	7	34
(4)	106	24	4	42
(5)	80	13	9	47
(6)	97	15	14	43
(7)	97	27	8	39
(8)	82	24	10	35
(9)	126	12	7	28
(10)	76	23	11	41
(11)	83	17	7	31
(12)	67	15	8	38
(13)	103	4	10	21
(14)	105	16	9	24
(15)	90	17	12	45
(16)	75	18	11	41
(17)	73	15	8	29
(18)	62	16	7	25
(19)	103	27	13	33
(20)	122	19	11	25
(21)	94	18	6	31
(22)	61	12	6	22
(23)	72	13	7	23
(24)	54	13	5	16
(25)	57	13	5	29

Wikipedia Category system has the highest coverage of entities among all three options. To overcome these issues, we use the Wikipedia Category Hierarchy by Kapanipathi et al. [39].

Data Sets. LP50 are data sets of documents compiled by Lee and Welsh [14], which was used for our experimentation. LP50 is composed of 50 general-purpose news documents with lengths between 50 and 126 words.

Lucene. The Lucene's documents were indexed to generate a documents index that includes the list of mentions and documents where they appear. Also, the TagMe tool was used for mentions detection in the documents. We used the Jena library for the analysis and extraction of the entities in the ontology. We use Jena TDB triple store to operate DBpedia locally.

For space issues, Table 2 shows only the results of the first 25 annotated documents. Column 2 shows the number of words in each document. Column 3 shows the mentions detected in each document. The columns 4 and 5 show the mentions linked in the KIM and DBpedia ontologies, respectively.

Table 2 shows only few mentions linked with KIM knowledge base. This is mainly due to the fact that (i) ontologies and instances are limited and (ii) the entities must have a value in rdfs: label.

Table 3: Precision, recall, F-measure, and accuracy of semantic annotations between context-free and context-based semantic annotation.

Means	Context-free	Context-based
Precision	0.621	0.893
Recall	0.839	0.799
F-measure	0.678	0.815
Accuracy	0.644	0.835

In the first case, if an ontology and knowledge base have a limited scope, a mention in the ontology could not exist. Therefore, ontology with a larger population (as DBpedia) will cover most of the mentions obtained in the documents.

In the second case, the entities must have value in rdfs: label, since this depends on links between the mentions and entities. DBpedia has more mention-entity link since it contains more than 4 million instances.

Table 3 shows the results of the semantic annotation evaluation DBpedia. The standard measures precision, recall, F measure, and accuracy were used for evaluating the annotations obtained. Precision is the rate between the relevant instances of the ontology and the total number of instances retrieved, and recall is the rate between the number of relevant instances retrieved and the total number of relevant instances existing in the ontology:

$$\begin{aligned} \text{Precision} &= \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|}, \\ \text{Recall} &= \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}, \end{aligned} \tag{12}$$

where TP (True Positives) are the set of retrieved instances that are relevant, FP (False Positives) are the set of retrieved instances that are not relevant, and FN (False Negatives) are the set of instances that are wrongly retrieved as nonrelevant.

The results show that our proposed method of context-based semantic annotation improves the results of the context-free annotation method.

Comparison to State of the Art. The results of our similarity calculation approach were compared with different strategies shown of the state of the art. Some approaches only take into account the weight of the edges, the association between each pairwise concept, and the ontology structure. We compared our approach with different methods in the literature that measure document similarity and use the LP50 data set. Among the methods analyzed are Latent Semantic Analysis (LSA) [40], Explicit Semantic Analysis (ESA) [41], Salient Semantic Analysis (SSA) [40], Graph Edit Distance (GED) [42], and ConceptsLearned [43].

The results obtained of comparison of our approach with other methods using LP50 dataset are shown in Table 4. The values of Pearson and Spearmen correlation of our approach were 0.745 and 0.65, respectively. This result was best compared to the results of other approaches. Thus, our approach significantly outperforms, to our knowledge, the most competitive related approaches, although ConceptsLearned has

Table 4: Comparing our approach with other methods using LP50 dataset.

Approach	Pearson correlation	Spearman correlation
LSA	0.59	0.53
ESA	0.68	0.59
SSA	0.71	0.64
GED	0.72	0.64
Our approach	0.745	0.65
ConceptsLearned	0.81	0.75

Table 5: Information content with extrinsic and intrinsic approaches.

Parameter	Metric	Pearson correlation
Depth	Intrinsic	0.743
Descendant	Intrinsic	0.743
Instances	Extrinsic	0.745

a better correlation of Pearson and Spearman (0.81 and 0.75). This is because ConceptsLearned uses 17 more features compared to ours, but the computational cost is high.

Comparison with Other Metrics for Information Content (IC) Calculation. We performed tests with different metrics to calculate the information content and use the extrinsic approach. The information content with the intrinsic approach can be performed using two parameters: (1) the depth of the class and (2) the descendants of a class.

Table 5 shows the slight advantage of considering the ontology instances with the extrinsic information content.

6. Conclusions

In this paper, we have presented a semantic annotation of unstructured documents approach. Which considers concepts similarity in ontology through its semantic relations.

The unstructured documents are represented as graphs, the nodes represent the mentions, and the edges represent the semantics and relationships. Each semantic relationship has a weighting measure assigned. Thus, the significant relationships have a higher weight.

The context extraction was done through the computation of association between pairwise concepts and the weight of entity relations. The sum of the two values is the one that measures the meaning or context of an entity. We also took advantage of instances in the knowledge base to measure the information content classes and relationships.

According to the state of the art the results obtained with our approach give the best results.

As future work, we are trying to reduce the knowledge base by selecting the entities whose definition is more likely to be used in the corpus. Additionally, Word2vec tool for semantic extraction of terms and documents can be used.

Finally, this approach also has been compared with other proposals available in the literature.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research work has been partially funded by European Commission and CONACYT, through the SmartSDK project. It also has been partially funded by TecNM with the project 6021.17-P.

References

[1] T. Zhang, K. Liu, and J. A. Zhao, "Graph-based similarity measure between wikipedia concepts and its application in entity linking system," *Journal of Chinese Information Processing*, vol. 29, no. 2, pp. 58–67, 2015.

[2] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data—the story so far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, article 122, 2009.

[3] C. Bizer, J. Lehmann, G. Kobilarov et al., "DBpedia—a crystallization point for the Web of Data," *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154–165, 2009.

[4] K. Bollacker, R. Cook, and P. Tufts, "Freebase: a shared database of structured general human knowledge," in *In Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI '07)*, vol. 2, pp. 1962-1963, AAAI Press, British Columbia, Canada, July 2007.

[5] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th International World Wide Web Conference (WWW '07)*, pp. 697–706, Alberta, Canada, May 2007.

[6] L. Bos and K. Donnelly, "The advanced terminology and coding system for ehealth," *Studies in Health Technology and Informatics*, vol. 121, pp. 279–290, 2009.

[7] J. M. Ruiz-Martínez, R. Valencia-García, J. T. Fernández-Breis, F. García-Sánchez, and R. Martínez-Béjar, "Ontology learning from biomedical natural language documents using UMLS," *Expert Systems with Applications*, vol. 38, no. 10, pp. 12365–12378, 2011.

[8] C. Caracciolo, A. Stellato, A. Morshed et al., "The AGROVOC linked dataset," *Journal of Web Semantics*, vol. 4, no. 3, pp. 341–348, 2013.

[9] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[10] R. Berlanga, V. Nebot, and E. Jiménez, "Semantic annotation of biomedical texts through concept retrieval," *Procesamiento del Lenguaje Natural*, vol. 45, pp. 247–250, 2010.

[11] M. Dai, N. Shah, W. Xuan et al., "An efficient solution for mapping free text to ontology terms," in *Proceedings of the American Medical Informatics Association Symposium on Translational BioInformatics (AMIA-TBI '08)*, Washington, DC, USA, November 2008.

[12] R. Navigli and M. Lapata, "An experimental study of graph connectivity for unsupervised word sense disambiguation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 678–692, 2010.

[13] E. Agirre, O. L. de Lacalle, and A. Soroa, "Random walks for knowledge-based word sense disambiguation," *Computational Linguistics*, vol. 40, no. 1, pp. 57–84, 2014.

[14] M. Lee and M. Welsh, "An empirical evaluation of models of text document similarity," in *Proceedings of the 27 Annual Conference of the Cognitive Science Society (CogSci '05)*, pp. 1254–1259, Erlbaum, Stresa, Italy, July 2005.

[15] P. Ferragina and U. Scaiella, "Fast and accurate annotation of short texts with wikipedia pages," *IEEE Software*, vol. 29, no. 1, pp. 70–75, 2012.

[16] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia spotlight: shedding light on the web of documents," in *Proceedings of the 7th International Conference on Semantic Systems (I-SEMANTICS '11)*, pp. 1–8, Graz, Austria, September 2011.

[17] OpenCalais, 2014, http://www.opencalais.com/.

[18] IBM, AlchemiLanguage, 2015, https://alchemy-language-demo.mybluemix.net/.

[19] D. O. C. S. The University of Sheffield, Developing Language Processing Components with GATE, 8 edition, 2017 https://gate.ac.uk/userguide.

[20] M. Laclavík, M. Šeleng, M. Ciglan, and L. Hluchý, "Ontea: platform for pattern based automated semantic annotation," *Computing and Informatics*, vol. 28, no. 4, pp. 555–579, 2009.

[21] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno, "Text processing through web services: calling Whatizit," *Bioinformatics*, vol. 24, no. 2, pp. 296–298, 2008.

[22] C. Tao, D. Song, D. Sharma, and C. G. Chute, "Semantator: semantic annotator for converting biomedical text to linked data," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 882–893, 2013.

[23] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov, "KIM—semantic annotation platform," in *Proceedings of the 2nd International Conference on Semantic Web Conference (ISWC '03)*, vol. 2870 of *Lecture Notes in Computer Science*, pp. 834–849, Springer, Sanibel Island, Fla, USA, October 2003.

[24] P. Castells, M. Fernández, and D. Vallet, "An adaptation of the vector-space model for ontology-based information retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 2, pp. 261–272, 2007.

[25] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta, "Semantically enhanced information retrieval: an ontology-based approach," *Journal of Web Semantics*, vol. 9, no. 4, pp. 434–452, 2011.

[26] R. Berlanga, V. Nebot, and M. Pérez, "Tailored semantic annotation for semantic search," *Journal of Web Semantics*, vol. 30, pp. 69–81, 2015.

[27] V. Nebot and R. Berlanga, "Exploiting semantic annotations for open information extraction: an experience in the biomedical domain," *Knowledge and Information Systems*, vol. 38, no. 2, pp. 365–389, 2014.

[28] D. Fuentes-Lorenzo, N. Fernández, J. A. Fisteus, and L. Sánchez, "Improving large-scale search engines with semantic annotations," *Expert Systems with Applications*, vol. 40, no. 6, pp. 2287–2296, 2013.

[29] L. Ding, T. Finin, A. Joshi et al., "Swoogle: a search and metadata engine for the semantic web," in *Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM '04)*, pp. 652–659, Washington, DC, USA, November 2004.

[30] Y. Lei, V. Uren, and E. Motta, "SemSearch: a search engine for the semantic web," in *Managing Knowledge in a World of Networks*, S. Staab and V. Svtek, Eds., vol. 4248 of *Lecture Notes in Computer Science*, pp. 238–245, Springer, Berlin, Germany, 2006.

[31] C. Tao, Z. Yongjuan, Z. Shen, C. Chengcai, and C. Heng, "Building semantic information search platform with extended sesame framework," in *Proceedings of the 8th International Conference on Semantic Systems (ISEMANTICS '12)*, pp. 193–196, New York, NY, USA, September 2012.

[32] S. Saha, A. Sajjanhar, S. Gao, R. Dew, and Y. Zhao, "Delivering categorized news items using RSS feeds and web services," in *Proceedings of the 10th IEEE International Conference on Computer and Information Technology (ScalCom '10)*, pp. 698–702, Bradford, UK, July 2010.

[33] V. Lopez, M. Fernández, E. Motta, and N. Stieler, "PowerAqua: supporting users in querying and exploring the Semantic Web," *Journal of Web Semantics*, vol. 3, no. 3, pp. 249–265, 2012.

[34] A. Singhal, G. Salton, M. Mitra, and C. Buckley, "Document length normalization," *Information Processing & Management*, vol. 32, no. 5, pp. 619–633, 1996.

[35] I. Augenstein, L. Derczynski, and K. Bontcheva, "Generalisation in named entity recognition: a quantitative analysis," *Computer Speech and Language*, vol. 44, pp. 61–83, 2017.

[36] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*, pp. 133–138, ACM, Las Cruces, NM, USA, June 1994.

[37] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI '95)*, vol. 2, pp. 448–453, Morgan Kaufmann Publishers Inc., Quebec, Canada, August 1995.

[38] T. M. Cover and J. A. Thomas, *Elements of Information Theory Wiley Series in Telecommunications and Signal Processing*, John Wiley & Sons, New York, NY, USA, 2007.

[39] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, "Hierarchical interest graph," 2016, http://wiki.knoesis.org/index.php/Hierarchical_Interest_Graph.

[40] S. Hassan and R. Mihalcea, "Semantic relatedness using salient semantic analysis," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pp. 884–889, San Francisco, Claif, USA, August 2011.

[41] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI '07)*, pp. 1606–1611, Hyderabad, India, January 2007.

[42] M. Schuhmacher and S. P. Ponzetto, "Knowledge-based graph document modeling," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*, pp. 543–552, ACM, New York, NY, USA, February 2014.

[43] L. Huang, D. Milne, E. Frank, and I. H. Witten, "Learning a concept-based document similarity measure," *Journal of the Association for Information Science and Technology*, vol. 63, no. 8, pp. 1593–1608, 2012.