ISSN 1870-4069

Semantic Annotation Approach for Information Search

Fernando Pech-May[1], Alicia Martinez-Rebollar[2], Jorge Magaña-Govea[1], Luis A. Lopez-Gomez[1], Edna M. Mil-Chontal[1]

[1] Instituto Tecnológico Superior de los Ríos, Tabasco, Mexico
fpech@tamps.cinvestav.mx, {jgoveaitsr,llopezitsr}@gmail.com, mariled7@hotmail.com

[2] Centro Nacional de Investigación y Desarrollo Tecnológico, CENIDET, Morelos, Mexico
amartinez@cenidet.edu.mx

Abstract. Due to the needs to improve the information search process, new strategies have been created to enhance searches. The semantic search performs the search by means of meaning instead of literals. The semantic search in unstructured documents requires to formalize knowledge through an annotation semantic process. Some annotation proposals use natural language processing tools, ontologies to link document terms; others use the similarity of entities through the weight of the edges, association between pair of concepts or the ontology structure. In this paper we present an alternative for semantic annotation in unstructured documents by semantic context extraction of entities. In the approach we detect the named entities through a data dictionary created from Wikipedia and link the instances in the ontology. The context extraction strategy is based on the concepts similarity; each term is associated with an instance of the ontology and the similarity between relationships explicit is measured by the combination of two types of measures: the association between each pair of concepts and the weight of the relationships. The approach was tested with two ontologies and two datasets in news and business, respectively.

Keywords: semantic annotation, semantic similarity, concept similarity.

1 Introduction

The large amount information stored and shared on the Web in form of unstructured documents [16] has caused difficulties for its search and retrieval. Traditionally two factors are used to classify the results of a search: 1) the relevance that measures the coincidence of the terms, and 2) the documents popularity, which is a complementary factor to documents ranking. Despite this, there are still challenges for searching and information management to reduce effort and search time.

On the other hand, there has been a constant growth in the semantic Web and has opened new opportunities for access and information retrieval and has motivated the development of linked data and knowledge bases for different domains and applications such as DBPedia [2], FreeBase [3], YAGO [32], etc. Also knowledge bases have been developed in specific areas such as Snomed CT [8] and UMLS [27] for medical areas and AGROVOC [4] for the agricultural area. These knowledge bases have become valuable resources for the knowledge extraction. A fundamental component to take advantage of such resources is to formalize knowledge by linking the unstructured text with elements of the knowledge base, called semantic annotation.

Some systems of semantic annotation have been developed in the medical area [20] for the identification of biomedical entities such as proteins, genes, diseases and their relationships. Other approaches have focused on named entities such as people, organizations and places. The first annotation proposals used natural language processing tools for documents analysis; these approaches present problems of: 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 (see Figure 1).

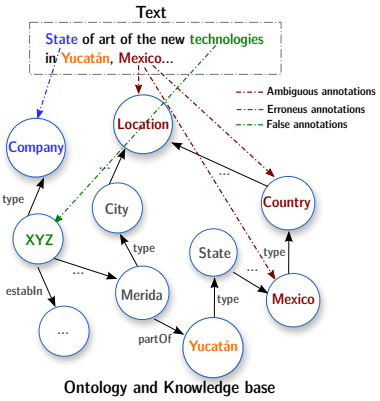

Fig. 1. Problems in semantic annotations: ambiguous, erroneous and false annotations.

This paper presents a semantic annotation approach in unstructured Web documents through its contextual semantic information through the use of ontologies in a limited domain. The proposal of semantic annotation is an improvement presented in [23] to represent unstructured documents by using an ontology, linking the terms or mentions of the document to the entities and to explore the semantic and contextual information by calculating the association of explicit relationships and the weight of the relationships of entities involved.

2 Background

2.1 Ontology

Ontology is a powerful tool for representation and reasoning of formal knowledge [26,20]. It is used by researchers to represent data of different types and areas and is encoded by OWL ontological languages. It consists of a scheme and instances (see Figure 2) to represent the description of knowledge of their concepts and relationships. An S scheme is defined as $< C, D, P >$, where C is the set of classes $C = c_1, c_2, ..., c_n$, D is the set of data types, and P is the set of properties $P = p_1, p_2, ..., p_n$ which are the relations between the classes. Instances represent knowledge and denote an instantiated class and its relationships. Instances can be defined as a graph $G =< V, E >$, where V is the set of instances and E the set of relations or predicates that join the instances. All classes, properties, data types and instances are explicitly identified by their Uniform Resource Identifier (URI) and are entities of the ontology. Each entity in the ontology is characterized by its textual description declared in the property $rdfs : label$ and it is posible to have lexical variations defined as *rdfs:label* = { *"text1", "text2"*}. Figure 2 shows the fragment of an ontology in the research domain. At the schema level, classes (such as *Laboratory* and *Professor*) and properties (such as *interestedIn*) are defined. At the instance level they indicate the instantiated schemas such as ontologies (instance of the *ResearchGroup* class), *Methodology...* and *Alice Perez* belong to the *Publication* and *Author* classes, respectively. The *Acapulco* instance contains two lexical variations *rdfs:label* = { *"acapulco", "acapulco de juarez"*}.

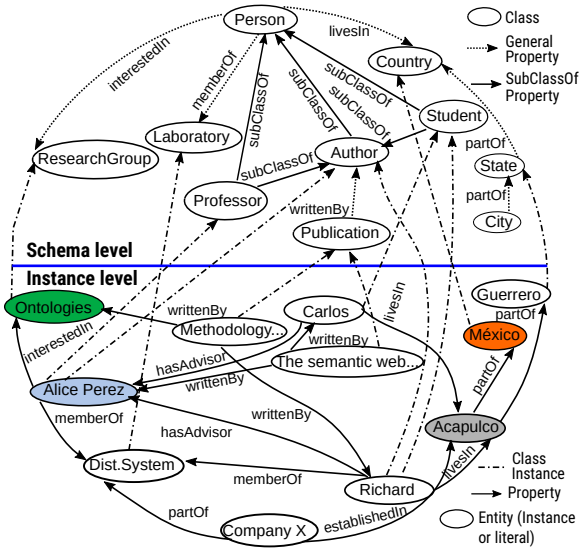

Fig. 2. Ontology example.

2.2 Semantic Annotation

The semantic annotation is the association of some data entity (people, objects, organizations, places, etc. of a text or Web document) to a description or semantic element (ontology) defined in *rdf: label*, in such a way that they are mappings between the fragment of a document term $d = t_1, t_2, \ldots, t_n$ and the concepts that describes the content of the document semantically. The annotation results in metadata that provides information on the classes and instances of entities [34]. It is fundamental in a variety of semantic Web applications such as linked data generation, open information extraction and semantic search. Specifically, semantic search allows users to express their information needs in terms of the knowledge base concepts.

2.3 Related Works

The semantic search involves different processes, which can be divided into: 1) preprocessing, 2) semantic queries translation, 3) semantic annotation, 4) semantic content retrieval, and 5) semantic ranking. In semantic annotation we have classified these works into two categories: 1) general purpose approaches, which help the annotation process and 2) information retrieval based approaches, which use specific domain ontologies and knowledge base.

General-porpuse tools. Let us remark that, AlchemyAPI[3] and OpenCalais [21] 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 [17] 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 RDF formats. Gate [33] 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 [13] 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

[3] http://www.alchemyapi.com

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 and tools are based on a dictionary search strategy. This consists of finding occurrences in text by applying a strict match of terms. They also allow for small variations in the matching of words through translation into regular expressions of the words.

Semantic Annotation Approaches Based on Information Retrieval Techniques. Popov and colleagues [24] presented 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. [5] propose an information retrieval model using ontologies for the annotation classification. This model uses an ontology-based schema for semiautomatic semantic annotation of documents. This research was extended by Fernández et al. [28] to provide natural language queries. Berlanga et al. [1] propose a semantic annotation 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 [1] 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. [9] propose a tool to improve the quality of results of the Web search engines, performing a better classification of the query results.

3 Approach to Context-Based Semantic Annotation

The paper presents a novel proposal of semantic annotation by unstructured documents representation using an ontology to link the document terms/mentions to the ontology entities and to explore the semantic and contextual information. The annotation approach enriches and describes the documents semantic content using the ontology entities similarity by computation two measures: 1) explicit relationships association and 2) the relationships weight of the entities involved. The semantic annotation approach is shown in Figure 3. Below each step described.

Mentions detection. The documents are analyzed to detect terms or phrases that may be names people, organizations, places, expressions of time, quantities, etc; these terms are known as mentions or named entities. Mentions detected may correspond to entities in the knowledge base [31]. For mentions identification process , the Tagme tool has been used to analyzing the n-grams in documents

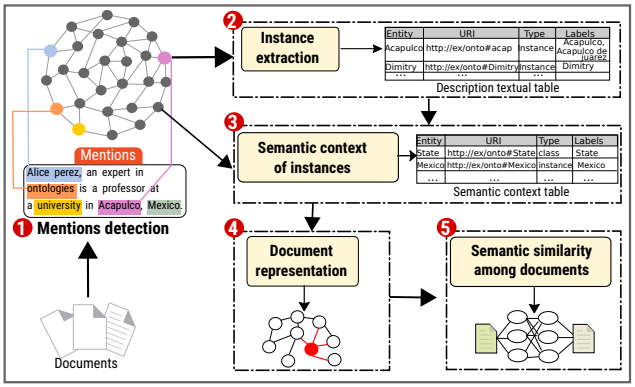

Fig. 3. Approach to context-based semantic annotation.

text by a entities dictionary (see Figure 4). A entities dictionary is built with Wikipedia, taking into account four sources: 1) anchor texts of Wikipedia articles, 2) redirect pages, 3) Wikipedia page titles, and 4) titles variants.

In dictionary construction, the mentions of a single character or with little occurrence are discarded and the further filtering is performed in the words that have low link probability (for example, less than 0.001). Link probability is defined as:

$$Lprobability(m) = P(link|m) = \frac{link(m)}{freq(m)}, \tag{1}$$

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

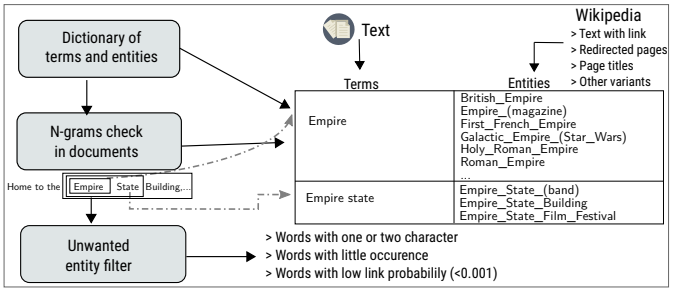

Fig. 4. Process for mentions detection.

The texts of input documents are analyzed to extract terms that may be possible mentions. All n-grams of the input text (up to n = 6), are compared with the entities dictionary. If a n-gram n_1 is contained by another one (that

is to say, that is substring), the shorter n-gram is discarded, if it has lower link probability than the longer one.

Instances Extraction in Knowledge Base. The detected mentions are searched in the ontology by means $rdfs:label$ to find their coincidence in some entity or instance. All values contained in $rdfs:label$ (lexical variations) are considered as labels. Figure 5 shows a code fragment of entity Mexico. In the source code, line 1 shows that the entity is an instance or individual; line 2 entity name; in line 3 the class to which it belongs and 4 its textual description ($rdfs:label$) with two lexical variations: "Mexico"and "Estados Unidos Mexicanos".

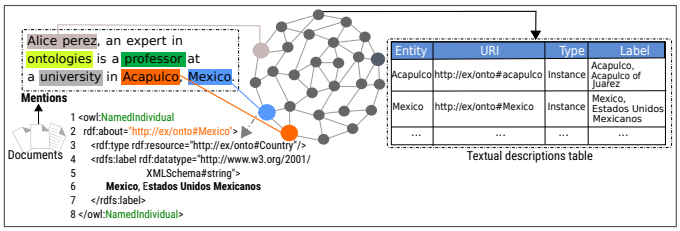

Fig. 5. Instance extraction process.

Semantic Context Extraction of Instances. In this stage, the entities semantic context detected previously is extracted. The explicit relationships in *URI* are also analyzed. The strategy to extract the semantic context is based on measuring the weight of the properties. To measure the association strength among each pair of entities, we have taken into account the entities characteristics and relationships in the knowledge base and it is calculated as a combination of two types of measures: 1) association between each concept pair and 2) the relationships weight.

1. **Concept pairwise Association.** It is used to calculate the relevance degree of a property for entities connected. We compare each pairwise (concepts c_1 and c_2) by calculating similarity. Figure 2 shows the *Acapulco* entity with five explicitly related concepts (*Carlos*, *Guerrero*, *Mexico*, *Richard*, and *CompanyX*). 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 [7]. We have adopted the Resnik approach [25] to measure the similarity between two concepts c_1 and c_2 according to the information content, using the formula:

$$IC = -log_2 \frac{I(D(c))}{I(C)}, \tag{2}$$

where $I(D(c))$ denotes the number instances of the concept c and $I(C)$ represents the number of instances on the ontology.
If we consider that the ontology of Figure 2 contains 1000 resources in *Person*, *Publication* and *ResearchGroup* classes; of which 600 people are interested in a research group (*ResearchGroup*) and 100 people (*Author*) wrote a publication (*Publication*). The information content in *interestedIn* and *writtenBy* is obtained:

$$IC(interestedIn(Person, ResearchGroup)) = -log_2 pr(interestedIn(Person, ResearchGroup)) = -log_2 \frac{600}{1000} = -log_2 0.6 \approx 0.73,$$

$$IC(writtenBy(Publication, Author)) = -log_2 pr(writtenBy(Publication, Author)) = -log_2 \frac{100}{1000} = log_2 0.1 \approx 3.32.$$

Although the information content in a property represents the property discrimination strength, may not be sufficient to determine the entity meaning and extract the semantic context of instances. We propose to measure the weight of each property linked to a concept c.

2. **Relationships Weight.** Based on information theory, the amount of information contained in a random variable over another variable is measured by mutual information (MI). This strategy has been proposed by Cover [6] and we have adapted it to measure the relationship strength of pairwise c_1 and c_2:

$$MI(p(d, r)) = \sum \sum pr(c_1, c_2) \cdot log_2 \frac{pr(c_1, c_2)}{pr(c_1) \cdot pr(c_2)}, \quad (3)$$

where $pr(c_1, c_2)$ is the probability of relationship e belonging to a set of properties of c_1 and c_2. $pr(c_1)$ is the probability of relationship belonging to set of properties of c_1 , whereas $pr(c_2)$ is the probability of relationship e belonging to set of properties c_2. Figure 6 shows the relationships *writtenBy*, *memberOf*, *hasAdvisor*, and *livesIn* belonging to *Richard* entity in the ontology. The instances of these relationships are shown in Figure 2.

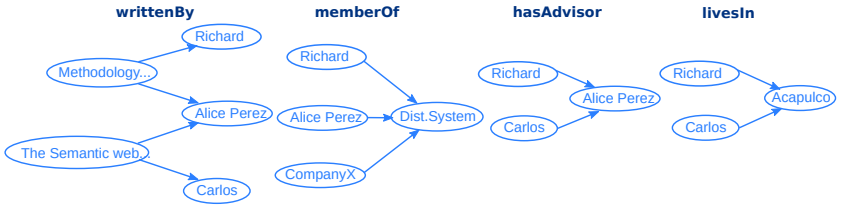

Fig. 6. Examples of *writtenBy*, *memberOf*, *hasAdvisor*, and *livesIn* property.

As an example, and without generality loss, suppose we want to calculate the relationship weight between *Richard* and *Methodology...*, (which is *writ-*

tenBy), is calculated as follows:

$$\begin{aligned}
&\mathbf{MI}(\mathbf{writtenBy}(\mathbf{Publication}, \mathbf{Author})) = pr(Methodology, Richard) \cdot log_2 \\
&\left(\frac{pr(Methodology, Richard)}{pr(Methodology) \cdot pr(Richard)}\right) + pr(Methodology, AlicePerez) \cdot log_2 \\
&\left(\frac{pr(Methodology, AlicePerez)}{pr(Methodology) \cdot pr(AlicePerez)}\right) + pr(TheSemanticWeb, AlicePerez) \cdot log_2 \\
&\left(\frac{pr(TheSemanticWeb, AlicePerez)}{pr(TheSemanticWeb) \cdot pr(AlicePerez)}\right) + pr(TheSemanticWeb, Carlos) \cdot log_2 \\
&\left(\frac{pr(TheSemanticWeb, Carlos)}{pr(TheSemanticWeb) \cdot pr(Carlos)}\right) \\
&= \frac{1}{4} \cdot log_2 \left(\frac{\frac{1}{4}}{\frac{1}{2} \cdot \frac{1}{4}}\right) + \frac{1}{4} \cdot log_2 \left(\frac{\frac{1}{4}}{\frac{1}{2} \cdot \frac{1}{2}}\right) + \frac{1}{4} \cdot log_2 \left(\frac{\frac{1}{4}}{\frac{1}{2} \cdot \frac{1}{2}}\right) + \frac{1}{4} \cdot log_2 \left(\frac{\frac{1}{4}}{\frac{1}{2} \cdot \frac{1}{4}}\right) = 0.5.
\end{aligned} \tag{4}$$

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 approximate mutual information as stated in:

$$MI(e) \approx log_2 \left(\frac{\frac{1}{[I(e)]}}{\frac{1}{I(c_1)} \cdot \frac{1}{I(c_2)}}\right), \tag{5}$$

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 $c2$ (object).

Combining Association and Relationship Weights. 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 equation 2) and the weights of the relationships (see equation 3), we calculate the final weight to obtain the entities context, as stated in:

$$W(p(c_i, c_j)) = \alpha \cdot Sim(c_1, c_2) + \beta \cdot MI(p(c_1, c_2)), \tag{6}$$

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:

$$\frac{Sim - min_{p \in P} Sim}{max_{p \in P} Sim - min_{p \in P} Sim} \text{ and } \frac{MI - min_{p \in P} MI}{max_{p \in P} MI - min_{p \in P} MI}.$$

3.1 Document Representation

Each document is represented as a contextual graph. The contextual graph is constructed by means of extracted instances in each document and the extraction

of its semantic context by calculating the association between the concepts and the weight of relationship. It can expressed as: Given a document corpus $C = d_1, d_2, ..., d_n$ and a knowledge base, a contextual graph GC_t is constructed to a document d_n; we consider the entities set $E = e_1, e_2, ..., e_m$ that occur in the entire contextual graph.

3.2 Semantic Similarity among Documents

Because the proposal is limited to the semantic annotation process, we have used the strategy of Paul et al. [22], they considers that two documents are similar if many annotations of a document are related to at least one annotation in another document (see Figure 7). The figure shows that the entities of document A are compared with the entities of document B. The edges $e = (v, w)$ with greater similarity are selected to calculate the similarity between both documents by means of the following formula:

$$SimDoc(docA, docB) = \frac{\sum_{a_{1i} \in A_1} (sim_{ent}(a_{1i}, matched(a_{1i})))}{|A_1| + |A_2|}. \quad (7)$$

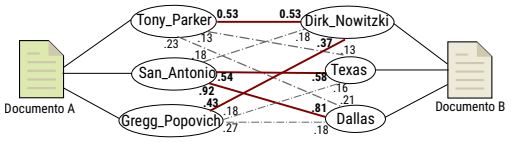

Fig. 7. Similarity approach between pairwise of Paul et al. [22].

4 Evaluation

For the tests, the following resources were used:

- *Ontology and knowledge base.* Two resources were used: DBpedia and KIM platform [24]. The KIM ontology is about politics news, finance and sports. It consists of more than 250 classes, 100 relationships and attributes. The knowledge base consists of 200,000 instances; 50,000 locations 130,000 organizations 6,000 people, etc. DBpedia uses a large multi-domain ontology. It contains 685 classes and 2795 properties and the knowledge base is over 4 million instances.
- *Hierarchy for DBpedia categories.* According to Lam et al. [14] category systems Wikipedia has greater coverage of entities for DBpedia. However, it has two problems: 1) it has no tree structure and 2) it contains cycles.

Kapanipathi et al. [12] created a system of categories[4] using Wikipedia and covering the aforementioned problems. In our approach we have used this classification system for DBpedia and it has been converted into triples for its use.

- *Document corpus.* Two corpus were used, one for each ontology. The corpus used for KIM ontology consists of 100 HTML documents of news in politics and international business and politics in the United Kingdom. For the tests with DBPedia, a corpus compiled by Lee et al. [16] named LP50[5]. It consists of 50 general purpose news documents with lengths between 50 and 126 words.
- *Evaluation metrics.* The Pearson correlation was used to evaluate our similarity results [15,18]. This metric is used to measure the approximation of our context with human judgment. Spearman correlation is a measure between two continuous random variables.

For space issues, Table 1 shows only the results of the first 20 annotated documents of LP50; column 3 shows the mentions detected in each document, in column 4 the entities and their semantic context detected in KIM and column 5 the entities and their semantic context detected by DBpedia. The greater reach of DBpedia ontology and knowledge base with respect to KIM is evident. The tests carried out with ontology and knowledge base KIM, were not very satisfactory; This is due to two factors: 1) the ontology and the instances are limited. If an ontology has a limited scope, there may not be a mention in the ontology and therefore its neighboring entities can not be extracted. On the other hand, an ontology with a larger population is more likely to cover a large part of the mentions obtained in the documents, and 2) the entities must have value in $rdfs : label$, on this depends the link between the mention and the entity of the ontology. Therefore, if an entity lacks the value in $rdfs : label$, it will not be taken into account.The tests performed with DBPedia were more satisfactory, this is because the ontology is greater and the knowledge base contains more than 4 million instances, so its scope is superior.

Table 2 shows the results of the semantic annotation evaluation DBpedia. The 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:

$$Presicion = \frac{|TP|}{|TP| + |FP|}, Recall = \frac{|TP|}{|TP| + |FN|}, \tag{8}$$

where TP are the set of retrieved instances that are relevant, FP the set of retrieved instances that are not relevant, and FN are the set of instances that are wrongly retrieved as nonrelevant.

[4] https://github.com/pavan046/higdataset
[5] https://webfiles.uci.edu/mdlee/LeePincombeWelsh.zip

Table 1. 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	35
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	34

Comparision to state of art. 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) [19], Explicit Semantic Analysis (ESA) [10], Salient Semantic Analysis (SSA) [29], Graph Edit Distance (GED) [30], and ConceptsLearned [11]. The results are shown in Table 3. 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 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.

Table 2. 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

Table 3. Our approach with other methods using LP50 dataset.

Approach	Person correlation	Spearman correlation
LAS	0.59	0.60
ESA	0.68	0.727
GED	0.72	0.63
Our approach	0.745	0.65
ConceptsLearned	0.81	0.75

Comparison with Other Metrics for Information Content (IC) Calculation. We performed tests with different metrics. 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 4 shows the slight advantage of considering the ontology instances with the extrinsic information content.

Table 4. Information content with others metrics.

Parameters	Pearson correlation
Common ancestor [7]	0.548
Intrinsic IC [30]	0.744
Extrinsic IC (used in our approach) [10]	0.0.745

5 Conclusions

In this paper, we have presented a semantic annotation of unstructured documents approach. By using ontologies of a specific domain ontology. 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.

References

1. Berlanga, R., Nebot, V., Jimenez, E.: Semantic annotation of biomedical texts through concept retrieval. Procesamiento del lenguaje natural, 45, 247–250 (2010)
2. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 7, 154–165 (2009)
3. Bollacker, K., Cook, R., Patrick, T.: Freebase: A shared database of structured general human knowledge. In: Proceedings of the 22 National Conference on Artificial Intelligence, AAAI'7, pp. 1962–1963, AAAI Press (2007)
4. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J.: The agrovoc linked dataset. Semantic Web, 4, 341–348 (2013)
5. Castells, P., Fernandez, M., Vallet, D.: An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. Journal of IEEE Transactions on Knowledge and Data Engineering, 19, 261–272 (2009)
6. Cover, T., Joy, T.: Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience (2006)
7. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the american society for information science, 41, 391–407 (1990)
8. Donnelly, K.: SNOMED-CT: The advanced terminology and coding system for eHealth. Journal of Studies in health technology and informatics, 121, 279–90 (2009)
9. Fuentes-Lorenzo, D., Fernandez, N., Fisteus, J., Sanchez, L.: Improving large-scale search engines with semantic annotations. Expert Systems with Application, 40, 2287–2296 (2019)
10. Gabrilovich, E., Markovitch, S.: 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. Morgan Kaufmann Publishers Inc. (2007)
11. Huang, L., Milne, D., Frank, E., Witten, I.: Learning a Concept-based Document Similarity Measure. Journal of American Society Information science Technology, 63, 1593–1608 (2012)
12. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: Hierarchical interest graph. http://wiki.knoesis.org/index.php/Hierarchical_Interest_Graph.
13. Laclavik, M., Hluchy, L., Seleng, M., Ciglan, M.: Ontea: Platform for Pattern Based Automated Semantic Annotation. Computing and Informatics, 28, 555–579 (2009)
14. Lam, S., Hayes, C., Galway, N., Dangan, L.: Using the Structure of DBpedia for Exploratory Search. In: KDD 13: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM.(2015)
15. Leal, J., Rodrigues, V., Queiros, R.: Computing Semantic Relatedness using DBPedia. In: Simes, A., Queiros, R., Cruz, da. (eds.). OASIcs-OpenAccess Series in Informatics, volume 21 of OASICS, pp. 133–147. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2014)
16. Lee, J., Kim, K.-S., Kwon, Y., Ogawa, H.: Understanding human perceptual experience in unstructured data on the web. In: Proceedings of the International Conference on Web Intelligence (eds.) WI'17, pp. 491–498. ACM (2017)
17. Mendes, P., Jakob, M., Garcia-Silva, A., Bizer, C.: 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. ACM (2011)

18. Lee, M.D., Welsh, M.: An empirical evaluation of models of text document similarity. In: In CogSci2005, pp. 1254–1259. Erlbaum (2006)
19. Nakov, P., Popova, A., Mateev, P.: Weight functions impact on LSA performance. In: Proceedings of the EuroConference Recent Advances in Natural Language Processing (RANLP'01), pp. 187–193 (2001)
20. Nebot, V., Berlanga, R.: Exploiting semantic annotations for open information extraction: an experience in the biomedical domain. Knowledge and Information Systems, 38, 365–389 (2014)
21. OpenCalais http://www.opencalais.com/, 2014, last accessed: 2017-04-02.
22. Paul, C., Rettinger, A., Mogadala, A., Knoblock, C., Szekely, P.: Efficient graph-based document similarity. In: Proceedings of the 13th International Conference on The Semantic Web. Latest Advances and New Domains, Volume 9678, pp. 187–193 (2016)
23. Pech, F., Martinez, A., Estrada, H., Hernandez, Y.: Semantic Annotation of Unstructured Documents Using Concepts Similarity. Scientific Programming, 2017, 1–10 (2017)
24. Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov M.: KIM: Semantic Annotation Platform. In: Proceedings of the Second International Conference on Semantic Web Conference, pp. 834–849. Springer Verlag (2004)
25. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence-Volume 1, IJCAI'95, pp. 448–453 Morgan Kaufmann Publishers Inc. (2004)
26. Ristoski, P., Paulheim, H.: Semantic Web in data mining and knowledge discovery: A comprehensive survey. Web Semantics: Science, Services and Agents on the World Wide Web, 36, 1–22 (2016)
27. Martinez, J., Valencia R., Fernandez J., Garcia F., Martinez R.: Ontology learning from biomedical natural language documents using umls. Expert Systems with Applications, 38, 12365–12378 (2011)
28. Saha, G.: Web ontology language (owl) and semantic web. Ubiquity 2007, 1:1-1:1 (2007)
29. Samer, H., Rada, M.: Semantic relatedness using salient semantic analysis. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, pp. 884–889 AAAI Press (2004)
30. Schuhmacher, M., Ponzetto, S.: 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 (2014)
31. Shaalan, K.: A survey of arabic named entity recognition and classification. Computational Linguistics, 40, 469–510 (2014)
32. Suchanek, F., Kasneci, G., Weikum, G.: Yago: A core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, pp. 697–706 ACM (2007)
33. Department of Computer Science The University of Sheffield: Developing Language Processing Components with GATE. 8 edition, https://gate.ac.uk/userguide (2017)
34. Wei, w., Barnaghi, P., Bargiela A.: Rational research model for ranking semantic entities. Journal of Information Sciences, 181, 2823–2840 (2013)