Dealing with Semantic Heterogeneity Issues on the Web

The Semantic Web is an extension of the traditional Web in which meaning of information is well defined, thus allowing a better interaction between people and computers. To accomplish its goals, mechanisms are required to make explicit the semantics of Web resources, to be automatically processed by software agents (this semantics being described by means of online ontologies). Nevertheless, issues arise caused by the semantic heterogeneity that naturally happens on the Web, namely redundancy and ambiguity. For tackling these issues, we present an approach to discover and represent, in a non-redundant way, the intended meaning of words in Web applications, while taking into account the (often unstructured) context in which they appear. To that end, we have developed novel ontology matching, clustering, and disambiguation techniques. Our work is intended to help bridge the gap between syntax and semantics for the Semantic Web construction.

Introduction

Although far from being fully deployed, the current Semantic Web constitutes an enormous source of distributed and heterogeneous knowledge, and is very quickly evolving thanks to the new knowledge constantly added by thousands of users. In order to attain the goals of the Semantic Web, the semantics of Web resources should be clearly defined by using ontologies. Nevertheless, it is unrealistic to expect that the volume of manually annotated resources will someday reach the critical mass that the Semantic Web requires to become a reality. For this reason, we think that it can largely benefit from methods that help to automatically determine the semantics of textual resources on the Web.

Availability of online ontologies and semantic content largely benefits interoperability on the Web. However, due to the open nature of the Web, online semantics may be defined by different people, for very different domains, and differing largely in expressiveness, richness, coverage, and quality, thus leading to an increasing semantic heterogeneity. Such heterogeneity originates some issues of which the most remarkable ones are:

1. **Semantic Ambiguity:** many intended meanings are associated with the same word.
2. **Semantic Redundancy:** many semantic descriptions are available to represent the same intended meaning.

These two problems may hamper Semantic Web applications when they need to determine the right meaning of certain textual resources (data to be annotated, keywords to be searched, entities to be extracted, etc.) by means of online ontologies.

In the following, we look at a set of techniques specifically designed for harvesting the Semantic Web, ranging from semantic measures to semantic clustering and disambiguation, which can help to reduce such redundancy and ambiguity issues. In particular, the combined use of these techniques enables the processing of any word on a Web context whose sense we need to discover by retrieving its set of possible senses, expressed as concise ontology terms, and indicating which one is the most adequate for its context.

Semantic Heterogeneity

Let us imagine that we want to interrogate the Semantic Web with a query: “Give me a list, ordered by calories, of recipes containing apple”. The problem has two dimensions: i) semantic querying, i.e., how to clearly specify the semantics of this query to be understandable by a software agent, and ii) semantic annotation, i.e., how semantic content can be added into the Web beforehand to allow satisfying such a kind of semantic queries. We are not entering into the details of semantic querying and annotation here, but we want to point out that both problems need to clearly determine the semantics of the involved terms to operate. In our example, only when the semantics of “apple” is clearly determined in the query, by grounding it into a certain ontology term (e.g., http://dbpedia.org/resource/Apple), the query can be processed to retrieve Web data pointing at the same term (e.g., recipes in a Web page in which the word “apple” has been semantically annotated with http://dbpedia.org/resource/Apple). Nevertheless, the sense selection problem gets more complicated, owing to the fact that “apple” is polysemous and can be interpreted, depending on the context, as “a fruit”, as “a tree”, or as “a company”, for instance. Thus, the intended meaning of “apple” has to be determined to provide a suitable semantic description. This exemplifies the *ambiguity* problem mentioned above. Furthermore, given a particular interpretation of “apple” (e.g., as “a fruit”) it can be annotated with different but semantically equivalent ontology terms in websites or datasets about recipes. The latter illustrates the *redundancy* problem.

In our example we need to determine the most suitable meaning of the word “apple” in its context (a Web search about recipes). In a Semantic Web-based scenario we can search online ontologies to get the possible semantic descriptions for “apple”. Nevertheless, we have to deal with the ambiguity of the term and try to select, among all the possible semantic descriptions, the one that represents the intended meaning best. The task of deciding which sense is the correct intended one is relatively easy for humans by simply inspecting the context in which the ambiguous word appears. For example, if “apple” appears as a cooking ingredient in a Web page about recipes, it likely refers to “a fruit”, whereas in a document about electronic equipments, it probably refers to the company Apple Inc. This selection task is however very difficult for a computer program. Word sense disambiguation (WSD) techniques can be applied here to help computers decide the right meaning of the word. Disambiguation techniques try to pick out the most suitable sense of an ambiguous word according to the context (usually its surrounding text) in which it appears.

As it was mentioned before, online ontologies can be accessed to discover possible semantic descriptions for “apple”. However, many redundant terms can be obtained to represent the same meanings. For instance, at the time of writing this, Swoogle Semantic Web search engine (http://swoogle.umbc.edu/) retrieves 445 different ontology terms associated with the label “apple”. Obviously, this number is well above the real polysemy of the word “apple”. This problem may hamper the disambiguation task, as there are a large amount of terms to analyse and to choose among, while a simple inspection confirms that most of the 445 obtained terms fall into one of the three possible interpretations mentioned above. We consider that this redundancy problem can be solved by carrying out a sense clustering process that will group the terms referred to the same meanings. Besides reducing the number of meanings to facilitate disambiguation algorithms, this also allows creating richer integrated semantic descriptions that combine information coming from different sources.

Techniques for Semantic Heterogeneity Reduction

Our goal is to devise suitable techniques to discover and represent the intended meaning of the words as ontology terms in an accurate way, free of ambiguities and redundancies, for their use on the Semantic Web. Particularly, we focus on discovering the meaning of words in unstructured texts (such as search keywords or folksonomy tags), a scenario which maximises the problem owing to the lack of syntactical or structural information in the context that could help us to apply traditional disambiguation methods.

The approach proposed here (see Figure 1) is grounded on a study of semantic measures that numerically compute the degree of similarity and relatedness among different semantic descriptions. Based on such measures, we have developed a set of techniques (ontology matching, sense clustering, and sense disambiguation) with an intention to overcome the above mentioned problems of redundancy and ambiguity on the Semantic Web.

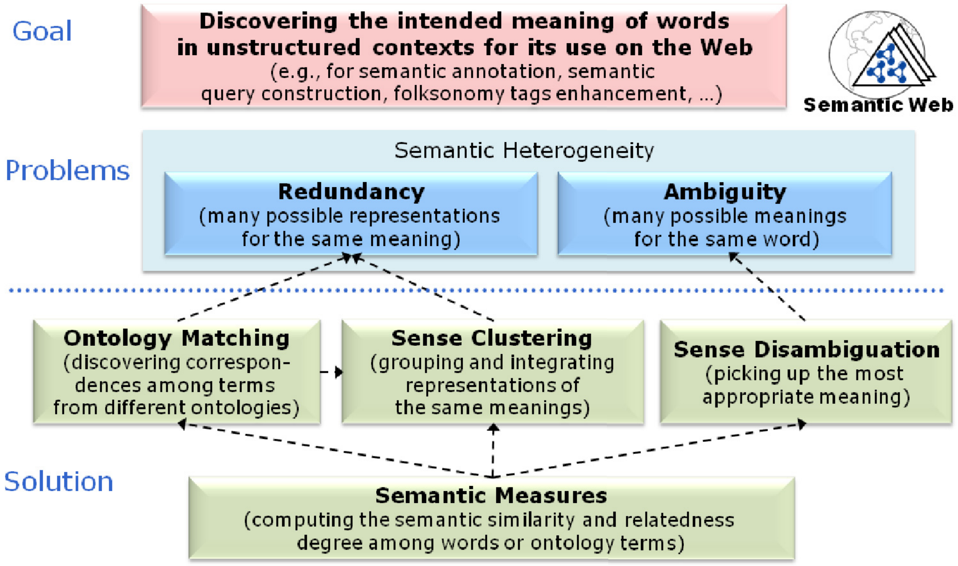
Goal
Discovering the intended meaning of words in unstructured contexts for its use on the Web
(e.g., for semantic annotation, semantic query construction, folksonomy tags enhancement, ...)
Semantic Web
Problems
Semantic Heterogeneity
Redundancy
(many possible representations for the same meaning)
Ambiguity
(many possible meanings for the same word)
Ontology Matching
(discovering correspondences among terms from different ontologies)
Sense Clustering
(grouping and integrating representations of the same meanings)
Sense Disambiguation
(picking up the most appropriate meaning)
Solution
Semantic Measures
(computing the semantic similarity and relatedness degree among words or ontology terms)

Figure 1. Techniques to reduce the semantic heterogeneity issues on the Web.

Semantic Measures

Semantic measures evaluate how semantically related two terms (words, ontology terms, etc.) are. According to [1] measures that take into account all the possible semantic relationships are called *semantic relatedness measures*, while *semantic similarity measures* consider only semantic relationships that involve similitude or likeliness between the two compared terms. On the basis of the previous definitions, one can create many particular semantic relatedness or similarity measures. Particularly, we have been interested in defining semantic measures suitable for their use on the Web. Before introducing our measures, we present here some desirable characteristics that such measures should fulfil in order to make easer their use on the open and dynamic context of the Semantic Web:

1. *Maximum coverage*. In many scenarios, we do not know in advance what the user has in mind when he chooses certain words to interact with Web applications. Therefore, to maximize the chances of taking into account the right user meanings, semantic measures should consider as many interpretations of the measured terms or words as possible, not relying on a particular knowledge source or annotated corpus only, but consulting many of them.

2. *Dynamic knowledge selection*. Expecting that all the accessible knowledge on the Semantic Web can be treated locally is an unrealistic approach. On the contrary, we propose semantic measures with the ability of working among any dynamically discovered ontology term, coming from any pool of online or local ontologies.

3. *Universality*. Semantic measures, in the highly dynamic context of the Web, should be defined independently of their final application.

All these desirable features have motivated the design principles that we adopted to define semantic measures. Our proposal reuses some previous work in the field, however adding interesting novelties, with the use of the Web as a corpus for relatedness computation, or the application of lightweight inference for schema-based similarity computation. Our proposed measures are:

1. *Context and Inference-based Semantic Similarity Measure*. This measure combines different techniques to compare the ontological contexts of two ontology terms (i.e., labels, hyper/hyponyms, domains, roles, etc.). In addition, we enrich these ontological contexts by applying semantic reasoning techniques, in order to give rise to inferred facts that are not present in the asserted ontologies. After extracting the ontological contexts, a comparison is made between them by combining different elementary techniques, as linguistic similarities and vector space modelling [2]. The result of our measure is a value in [0,1] representing how similar the contexts of the compared ontology terms are. A comparison between two entities e_1 and e_2 is as follows:

$$\begin{aligned} & sim(e_1,e_2) = w_{label} \cdot sim_{str}(e_1^{syn}, e_2^{syn}) + \\ & w_{descr} \cdot vsm(e_1^{descr}, e_2^{descr}) + \\ & w_{attr} \cdot vsm(e_1^{attr}, e_2^{attr}) + \\ & w_{\sup} \cdot vsm(e_1^{\sup}, e_2^{\sup}) + \\ & w_{sub} \cdot vsm(e_1^{sub}, e_2^{sub}) \end{aligned} \qquad (1)$$

where $e^{syn}, e^{descr}, e^{attr}, e^{sup}, e^{sub}$ denote the set of synonym labels, textual description, set of attributes characterizing the entity (i.e., set of properties if *e* is a class; domains and ranges if *e* is a property; associated classes and property values if *e* is an individual), and super/subterms of the hierarchical graph respectively. *vsm* is a comparison based on vector space models, sim_{str} is a string-based similarity, and w_i is the weight of the i-th component. Such weights are empirically inferred; in our prototype we experimented with the OAEI benchmark track (see later the experiments section) and choose the ones that lead to the best performance.

2. *Web-based Semantic Relatedness Measure*. We chose the Normalized Google Distance (NGD), a well-founded existent semantic measure [3], to compute a Web-based relatedness measure between plain words. This measure uses the Web as a knowledge source and is based on counting the co-occurrence of words on Web pages. Relatedness between two words *x* and *y* is given by

$$relWeb(x,y) = e^{-2NWD(x,y)} \qquad (2)$$

where *NWD* is a generalization of NGD to any Web search engine. Based on the latter, we defined a relatedness measure between ontological terms *a* and *b* as follows (see [4] for a detailed discussion):

$$rel_0(a,b) = \frac{\sum_{i,j} relWeb(a_i^{syn}, b_j^{syn})}{|a^{syn}| \cdot |b^{syn}|}$$

$$rel_1(a,b) = \frac{\sum_{i,j} rel_0(a_i^{ctx}, b_j^{ctx})}{|a^{ctx}| \cdot |b^{ctx}|}$$

$$rel(a,b) = w_0 \cdot rel_0(a,b) + w_1 \cdot rel_1(a,b) \qquad (3)$$

where e^{syn} and e^{ctx} are the set of synonym labels and the minimum ontological context of *e* respectively; e^{ctx} contains direct superclasses, domains, or associated classes (if *e* is a class, property, or instance respectively); and w_i are empirically inferred weights ($w_0 = w_1 = 0.5$ in our prototype).

These measures have been designed to fulfil the above mentioned requirements to operate on the Web. In the following, we describe how they are applied to the tasks that we have identified in Figure 1.

Ontology Matching

Ontology matching is the task of determining relationships that hold among terms of two different ontologies. We have developed CIDER (Context and Inference baseD ontology alignER), an ontology matching tool intended to discover semantic equivalence relationships [5]. The inputs to CIDER are ontologies expressed in OWL or RDF; subsequently it carries out the following actions:

1. It extracts the ontological context of the compared terms, enriched by applying a lightweight inference mechanism to add more semantic information that is not explicit in the asserted ontologies.

2. Then, semantic similarities between each pair of terms are computed by using Equation 1. A matrix with all the similarities is obtained.

3. The final alignment is then extracted, finding the highest rated one-to-one relationships among terms, and filtering out the ones that are below a certain threshold.

Besides alignments between whole ontologies, CIDER can also serve individual similarity computations, thus being easily adaptable to other uses such as sense clustering, as we will see in the following section.

Sense Clustering

To tackle the problem of redundancy reduction on the Semantic Web, we define a clustering technique that we apply to the base of ontological terms collected by Watson (a system that crawls the Web and index available semantic resources [6]) and creates groups of ontological terms having similar meanings. Briefly explained, the process is as follows:

1. It starts with an initial grouping of all ontology terms one can find in Watson which are associated to the same keywords (or synonym labels). We call *synonym maps* these sets of ontology terms.

2. Extraction and similarity computation. An iterative algorithm takes each ontology term from a synonym map and computes its similarity degree (Equation 1) with respect to each of the other terms in the synonym map.

3. Integration. When the obtained similarity value is under a given threshold, we consider both as different senses, and the algorithm continues comparing other terms. If, on the contrary, the similarity is high enough, both terms are integrated into a single sense, and the comparison process is reinitiated among the new integrated sense and the rest of terms in the synonym map. In [7] a method for selecting a suitable threshold is discussed.

Steps 2 and 3 constitute an agglomerative clustering algorithm that produces, for each synonym map, a set of integrated senses (called *sense maps*) as output. The clustering process is repeated with the rest of synonym maps, to create eventually a pool of integrated senses which covers all ontology terms in Watson indexes. Each sense map groups the ontology terms that correspond to the same intended meaning.

This is illustrated in Figure 2. The method, applied to a search of “apple”, will return all the ontology terms that refer to the meanings “the fruit”, “the tree” and “the company”, grouped together as three single integrated senses, respectively.

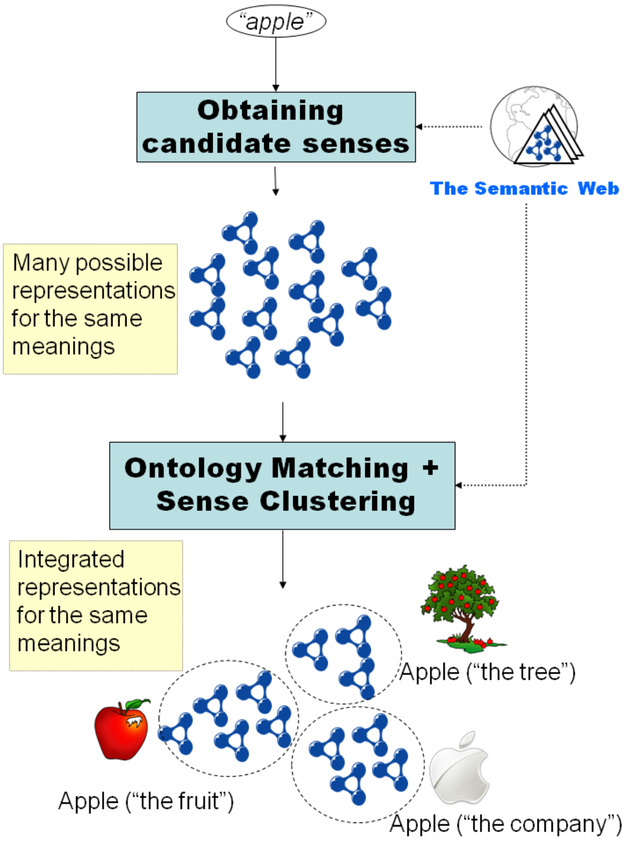
"apple"
Obtaining candidate senses
The Semantic Web
Many possible representations for the same meanings
Ontology Matching + Sense Clustering
Integrated representations for the same meanings
Apple ("the tree")
Apple ("the fruit")
Apple ("the company")

Figure 2: Clustering example.

Sense Disambiguation

Figure 3 illustrates how the ambiguity problem can be solved by applying a disambiguation technique that explores the context words (e.g., "fruit, dessert, pie") and the possible senses of the ambiguous word in order to deduce its right meaning in that context.

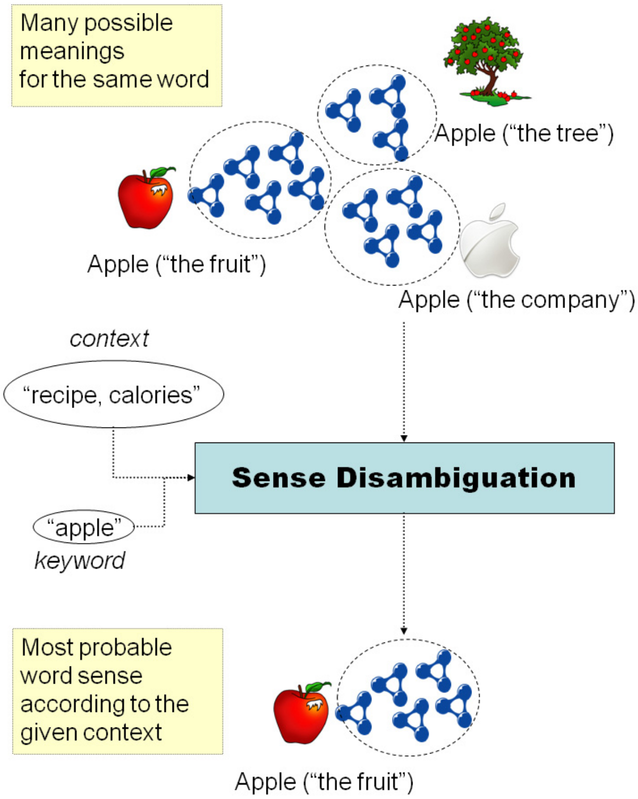

Figure 3: Disambiguation example.

We focus on disambiguation in unstructured Web contexts, such as those composed of user keywords or user tags, where traditional WSD techniques have difficulties to operate. In fact, user tags constitute themselves a highly heterogeneous context (usually free text) which hampers disambiguation. Owing to the fact that tags are not in well-formed sentences, syntactic analysis and similar techniques cannot be applied. Furthermore, many tags refer to subjective impressions of users (e.g., "my favourite", "amazing") or technical details (e.g. "Nikon", "photo"), which leads to contexts where many words are useless (or even harmful) for disambiguation.

The overall disambiguation process, used in combination with the sense clustering technique, is described in Figure 4. It receives a keyword and its context words as input and gives its most suitable sense as output. It consists of the following steps:

1. Context selection. We consider the hypothesis that the most significant words in the disambiguation context are the most highly related to the word to disambiguate. Based on that, we compute the Equation 2 between each context word and the keyword to disambiguate, filtering out the context words that score below a certain threshold. Such a threshold is empirically inferred and depends on the search engine used to compute Equation 2 (we use a 0.22 threshold for Yahoo! in our current prototype). The resultant set is called *active context*.

2. After context selection, online and local resources are accessed to provide a set of candidate senses for the keyword. The output of this process is a set of candidate senses that describe the possible meanings of the keyword to disambiguate. Each sense corresponds to an ontology term or to the integration of various ontology terms, as result of applying the above described sense clustering process.

3. Finally, a disambiguation algorithm is run and the senses are weighted according to their likeliness of being the most suitable one for the given context [8]. This is performed by exploring the semantic relatedness among the keyword senses and the words in the context, the overlap between the words that appear in the context and the words that appear in the semantic definition of the sense [9] and, finally, considering the frequency of usage of senses (if available).

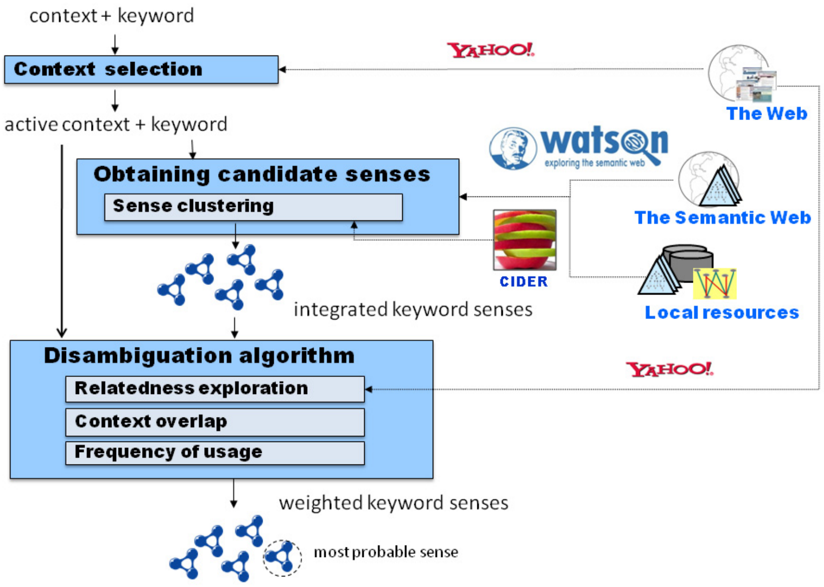

Figure 4: Disambiguation scheme.

Experiments

Our techniques have been tested in order to assess their suitability for the tasks they were created for. For instance, CIDER system participated in the Ontology Alignment Evaluation Initiative (OAEI) [5], where it showed a good behaviour in the benchmark track (97% precision, 62% recall; quite above the 43% precision, 59% recall of the string matching-based baseline), while the results in the directory track were the second best in the competition that year (60% precision, 38% recall).

We have also studied our proposed large scale method for ontology terms [7]. Our intention was to confirm empirically whether the method scales up well and is feasible when applied to the large scale of terms indexed by Watson. We answered positively to this question after applying our technique to a pool of 73,169 different ontology terms, obtaining a strong linear dependence of the time response with respect to the size of the keyword maps (linear correlation coefficient R=0.97).

Additionally, we explored the behaviour of our disambiguation algorithm to determine the sense of a set of ambiguous words associated to 350 pictures in Flickr, comparing them to human opinion [8]. The resultant 58% accuracy beat both the random and the “most frequent sense” (MFS) baselines in this experiment (20% and 43% accuracy respectively). This is a remarkable achievement, because the state of the art indicates that non-supervised disambiguation techniques rarely score above MFS baseline.

Therefore, our semantic measures behave well for the different tasks that we have considered. Besides that, the principles in which their design is based allow us to use them with any ontology no matter the domain, and no matter whether the ontologies were pre-defined or discovered runtime, thus being suitable for their use on such a dynamic and heterogeneous environment as the Web. In http://sid.cps.unizar.es/SEMANTICWEB/EXPERIMENTS/ more details about the data of these experiments can be found.

Related Work

A Co-reference Resolution Service (CRS) [10] has been proposed in the context of Linked Data [11] which tackles the problem of redundancy. This service aims to determine equivalent Web identifiers (URIs) referring to the same concept or entity. The system is targeted to storing, manipulating, and reusing co-reference information. Equivalent identifiers are stored in *bundles*, similarly as we do in our *sense maps.* CRS has the advantage, also shared with our clustering approach, that the knowledge regarding co-reference and equivalence are treated separately from the semantic data sources, thus avoiding the overuse of the `owl:sameAs` clause to identify duplicate entities. Nevertheless, CRS does not propose specific algorithms to identify semantically equivalent groups of ontology terms, as we do. Furthermore, our approach deals with ontology terms while CRS deals with resource identifiers (URIs) no matter whether they constitute semantic descriptions or not.

PowerAqua [12], a Semantic Web-based question answering system, shares our target of discovering the semantics of words in a Web-based context. It receives a simple question posed in natural language as input and obtains the data that satisfy this question as output. It performs first a linguistic analysis of the input question, transforming it into a set of possible query triples. Then, Watson is accessed to retrieve online semantic documents describing the involved entities. A mapping algorithm is applied to find the most suitable correspondence between the query triples and the information in the candidate ontologies, deriving the searched information finally. Both our approach and PowerAqua exploit knowledge from *dynamically selected* online ontologies. Their respective inputs are different, though: while PowerAqua processes well formed sentences, we deal with keywords in unstructured contexts, thus not relying specifically on linguistic analysis.

Conclusions

We have tried to recall the attention of the reader about the interest and necessity of dealing with semantic heterogeneity issues on the Web, for the sake of semantic interoperability and more precise Web data recovering. Although treated locally in particular domains and systems, there is a lack of overall strategies that solve these issues (namely, redundancy and ambiguity) when dynamically harvesting the Semantic Web. Our approach is a step in that direction, intended to allow expressing in a concise way the meaning of terms that appear in unstructured contexts on the Web. Discovering the meaning of keywords can assist semantic query construction, semantic annotation of Web pages, semantic classification of tagged resources, etc. Precisely, our future steps will focus on creating and enriching such type of systems, thus facilitating a practical realization of the Semantic Web.

Acknowledgments

This work has been supported by the CICYT projects TIN2010-21387-C02-02 and TIN2010-17550.

References

1. A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," Computational Linguistics, vol. 32, no. 1, pp. 13-47, Mar. 2006.

2. V. V. Raghavan and M. S. K. Wong. "A Critical Analysis of Vector Space Model for Information Retrieval," Journal of the American Society for Information Science, vol. 37, no. 5, pages 279-287, 1986.

3. R. L. Cilibrasi and P. M. B. Vitányi, "The Google similarity distance," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3, pp. 370-383, Mar. 2007.

4. J. Gracia and E. Mena, "Web-based measure of semantic relatedness," in Proc. of 9th International Conference on Web Information Systems Engineering (WISE 2008), Auckland, New Zealand, vol. 5175. Springer Verlag LNCS, pp. 136-150, Sep. 2008.

5. J. Gracia and E. Mena, "Ontology matching with CIDER: Evaluation report for the OAEI 2008," in Proc. of 3rd Ontology Matching Workshop (OM'08), at ISWC'08, Karlsruhe, Germany, vol. 431. CEUR-WS, pp. 140-146, Oct. 2008.

6. M. d'Aquin, C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, and E. Motta, "Characterizing knowledge on the semantic Web with Watson," in 5th International EON Workshop, at ISWC'07, Busan, Korea, Nov. 2007.

7. J. Gracia, M. d'Aquin, and E. Mena, "Large scale integration of senses for the semantic web," in WWW '09: Proceedings of the 18th international conference on World Wide Web, Madrid, Spain, ACM, pp. 611-620, 2009.

8. J. Gracia and E. Mena, "Multiontology semantic disambiguation in unstructured web contexts," in Proc. of Workshop on Collective Knowledge Capturing and Representation (CKCaR'09) at K-CAP'09, Redondo Beach, California (USA), Sep. 2009.

9. S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness." in Proc. of the 18th International Joint Conference on Artificial Intelligence, pp. 805-810, Aug. 2003.

10. H. Glaser, A. Jaffri, and I. Millard, "Managing co-reference on the semantic web," in WWW2009 Workshop: Linked Data on the Web (LDOW2009), Apr. 2009.

11. C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," International Journal on Semantic Web and Information Systems (IJSWIS), vol. 5, no. 3, pp. 1-22, 2009.

12. V. López, M. Fernández, E. Motta, N. Stieler, "PowerAqua: supporting users in querying and exploring the Semantic Web content," Semantic Web journal, to appear, 2011.

ABSTRACT

The Semantic Web is an extension of the traditional Web in which meaning of information is well defined, thus allowing a better interaction between people and computers. To accomplish its goals, mechanisms are required to make explicit the semantics of Web resources, to be automatically processed by software agents (this semantics being described by means of online ontologies). Nevertheless, issues arise caused by the semantic heterogeneity that naturally happens on the Web, namely redundancy and ambiguity. For tackling these issues, we present an approach to discover and represent, in a non-redundant way, the intended meaning of words in Web applications, while taking into account the (often unstructured) context in which they appear. To that end, we have developed novel ontology matching, clustering, and disambiguation techniques. Our work is intended to help bridge the gap between syntax and semantics for the Semantic Web construction.

KEYWORDS

Semantic Web, semantic heterogeneity, ontology matching, disambiguation

CONTACT INFORMATION

Jorge Gracia
Email: jgracia@fi.upm.es,
Affiliation: Ontology Engineering Group, Universidad Politécnica de Madrid
Address: Campus de Montegancedo sn, Boadilla del Monte 28660 Madrid, Spain
Phone: +34 913363672
Fax: +34 913524819

Eduardo Mena
Email: emena@unizar.es
Affiliation: IIS Department, University of Zaragoza
Address: Edificio Ada Byron, María de Luna 1, 50018 Zaragoza, Spain
Phone: +34 976762340
Fax: +34 976761914

SHORT BIOS

Jorge Gracia is a postdoctoral researcher in the Artificial Intelligence Department of Universidad Politécnica de Madrid, Spain. His research interests include ontology matching, semantic disambiguation, semantic measures, and multilingualism in the field of the Semantic Web. He obtained his PhD in computer science at University of Zaragoza. Contact him at jgracia@fi.upm.es

Eduardo Mena is an Associate Professor at the University of Zaragoza, Spain. He leads the Distributed Information Systems research group at his university. His research interest areas include interoperable, heterogeneous and distributed information systems, Semantic Web, and mobile computing. He received his Ph.D. degree in Computer Science from the University of Zaragoza. Contact him at emena@unizar.es