International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:2, 2010

# A New Version of Annotation Method with a XML-based Knowledge Base

Mohammad Yasrebi, and Somayeh Khosravi

***Abstract*—**Machine-understandable data when strongly interlinked constitutes the basis for the SemanticWeb. Annotating web documents is one of the major techniques for creating metadata on the Web. Annotating websitexs defines the containing data in a form which is suitable for interpretation by machines. In this paper, we present a better and improved approach than previous [1] to annotate the texts of the websites depends on the knowledge base.

***Keywords*—**Knowledge base, ontology, semantic annotation, XML.

## I. Introduction

SEMANTIC annotation is the process of inserting tags in a document to assign semantics to text fragments allowing creating the documents processable not only by humans but also automated agents [6]. The acquisition of masses of metadata for the web content would allow various Semantic Web applications to emerge and gain wide acceptance. At present there are various Information Extraction (IE) technologies available that allow recognition of named entities within the text, and even the relations, events, and scenarios in which they take part. Thus, metadata could be assigned to the document, presenting part of its information content, suitable for further processing. Such metadata can range from formal reference to the author of the document, to annotations of all the companies and amounts of money referred in the text [7].

By researching about methods and existing semantic annotation platforms we observe that all of these methods are using the source of information which is named knowledge base to define the concepts and semantics of words in texts. The knowledge bases which are used in these tools are defective and unable to define the concepts of some words. So, the idea of using extended knowledge base with more knowledge and information in most domains came to exist and is able to be complete more and more.

In this paper, we present an approach to semantic enrichment website and documents. This system is still semi-automatic, but we perform some changes in various steps, especially in knowledge base in order to 1) increase the rate of search and 2) possibility of managing the knowledge base by advanced and structured methods.

M. Yasrebi is with the Islamic Azad University, Shiraz, Iran (phone: +98917-714-0793; e-mail: mohammadyasrebi@gmail.com).

S. Khosravi is with the Islamic Azad University, Shiraz, Iran (phone: +98917-309-7525).

First of all we discuss about the previous approach generally, and then describe the changes of each step with their reasons.

## II. The Role of Knowledge bases in our Approach

In this approach, two different knowledge bases used as follow:

- Primary knowledge base
- Secondary knowledge base

### *A. Primary Knowledge Base*

The Primary knowledge base is the most important and essential part of knowledge base. In fact, this knowledge base contains information about the concept/instance which is supplied by well-informed users. In the previous approach, the primary knowledge base contains the set of data bases which are related to specific domain, but in this situation when we have lots of data; the rate of search process is very low, in addition storing this amount of data need massive spaces. Therefore, we changed this implementation depends on XML format to solving above problems and possibility of managing the knowledge base by advanced and structured methods.

These XML files which create in each domain are going to become complete as the time passes, and in an ideal situation all words of a specific domain are identified and implemented in the XML file.

### *B. Secondary Knowledge Base*

As its name implies, the secondary knowledge base is used to help the primary knowledge base. The latter the same as previous includes three components as follow:

- basic knowledge source
- data frame library
- lexicons

1. Basic Knowledge Source

WordNet Ontology [8] according to richness of relations between concepts can not use only in order to perform the extraction and induction of data in its data bases and extracted semantic schemas. Because it is defective for some words, and we reduce these defects with other parts such as data frame library and lexicons. For example, the WordNet Ontology can not identify the word "alen" as a person's name, or "222-2222" as a telephone number, or "qwerty@yahoo.com" as an e-mail address, etc. Since WordNet basically consists of information about concepts and their relations (e.g. hyperonyms etc.)

YAGO[1] could be considered as additional BKS, since this ontology incorporates a lot of instanceOf(instance, concept) relations with broad coverage.

2. Data Frame Library

Basically in computer-based sciences, data has poor structure and for describing these data we have to use simple classifications such as "integer", "real", "string", etc. On the other hand, we can not identify concepts with these classifications. Therefore, we have to use a classification with better structure. This classification is presented as data frame library and contains the second part of our secondary knowledge base. One of the ways to extract the concepts such as date, e-mail address, phone number, etc. is to use the regular expressions [9]. In this paper, we name these regular expressions as data frame library.

3. Lexicons

The other part of our secondary knowledge base is lexicons. Lexicons used to enrich WordNet ontology as BKS. Lexicons includes the set of different lists, that each list is the name of various entities such as persons, animals, capitals, etc. However, the lexicon plays an important role for recognizing the instances of the specific concepts and limiting the domain. For example, the WordNet can not identify the concept of the word "alen", but this word exists in the list of the person's name in lexicons and then lexicons can detect this word as the name of person.

## III. Architecture of Knowledge Bases

Fig. 1 shows our knowledge bases architecture briefly. As it is shown, this architecture contains all the knowledge bases which are described in previous sections and their relations.

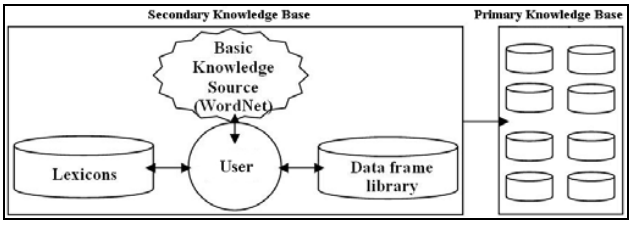


Fig. 1 The architecture of knowledge bases

This architecture is the same as previous architecture generally, but we have changed the implementation of primary knowledge base includes some XML files in each domain. For more information you can see this paper [1].

## IV. The Annotation Method in our Approach

After preparing the needed knowledge base, based on the methods outlined in previous sections, we can discuss on extracting the word and the concepts and also semantic annotation.

[1] http://www.mpi-inf.mpg.de/~suchanek/downloads/yago/

First, it is necessary to describe a general view on the architecture of our approach and then inspect the details of this project. Fig. 2 shows a general view of the architecture of our approach.

As Fig. 2 shows, this process contains 3 separate phases:

1. Determining the text's domain
2. Extracting the words and their concepts
3. Semantic annotation and inserting tag process

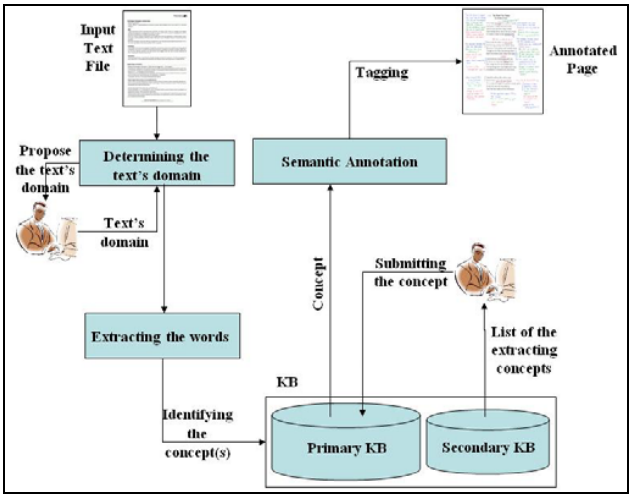


Fig. 2 Architecture of our approach

1. Determining the Text's Domain

We have some changes in this phase. In this approach the system considers one ore more domains for each text. Here we don't need the human intervention to determine the text's domain. In the next steps, system determines the unique text's domain itself. By this mean, this system has two advantages, at first, this step is automatically and we can omit the human intervention, and the second is that in some texts which can we use them in more than one domains, different XML files can completely parallel.

2. Extracting the Words and their Concepts

In this phase, we need to extract words which are concepts or instances of a concept, and also explain a special meaning such as: email address, or name of person, etc.

Thus, by using a pattern which determines the words and a loop, we extract the words of the text one by one to the end of the text. So, after analyzing the text to words, we have to send the word one by one to knowledge base for determining their concepts.

At first, we send the word to the primary knowledge base and the primary knowledge base by identifying the determined text's domain will search the word in the XML files which contains the words related to the domain. If the word exists, the concept will be returned; otherwise, the secondary knowledge base will help the primary knowledge base and determine its concept. This process is the same as previous.

3. Semantic Annotation and Inserting Tag Process

In this last phase, the extracted words in the text with their concept are accessible. Thus, by identifying the location of the words in the text, we insert and add tags which contain the concept of the words into the text.

However, according to exist XML files, we can use RDF method and OWL language for annotating simply. This phase is under construction.

## V. Conclusion

The Semantic Web requires the widespread availability of document annotations in order to be realized. Benefits of adding meaning to the Web include: query processing using concept-searching rather than keyword-searching [2]; custom web page generation for the visually-impaired [5]; using information in different contexts, depending on the needs and viewpoint of 48 the user [3]; and question-answering [4].

In this system, concepts are extracted based on a quite comprehensive knowledge base. This knowledge base includes a Basic Knowledge Base including a quite complete set of words, the sets of grammars and data frames, and various lists of different entities' names. The performed procedure in our system has been done under the control of a user familiar with the text domain, and therefore annotation process is performed semi-automatically. The superiority of our system to other similar ones is illustrated through a comparative study. Our future endeavor is enhancing the used algorithm, enriching the primary and secondary knowledge base, and also increasing the system's capability in identifying numerical concepts in unstructured web-pages. Other future work would be further evaluation on our suggested method considering other aspects. We hope to evaluate the system on higher number of pages, numerous domains, and pages with various contents including words, numbers, and figures.

## References

[1] M. Yasrebi, M. Mohsenzadeh, M. Abbasi-Dezfuli, “A new approach the text’s of the websites and documents with a quite comprehensive knowledge base,” in *International conference of WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY VOLUME 35, Laval, France, pp. 280-284.*

[2] T. Berners-Lee, J. Hendler., O. Lassila, “The Semantic Web,” Scientific American, 2001, pp. 34-43.

[3] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, “SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation,” in *12th International World Wide Web Conf.*, Budapest, Hungary, 2003, pp. 178-186.

[4] P. Kogut, W. Holmes, “AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages,” in *Proc. Workshop on Knowledge Markup and Semantic Annotation at the First International Conference on Knowledge Capture* (K-CAP 2001), Victoria, BC, 2001.

[5] Y. Yesilada, S. Harper, C. Goble, R. Stevens, “Ontology Based Semantic Annotation for Visually Impaired Web Travellers,” in *Proc. 4th International Conference on Web Engineering (ICWE 2004),* Munich, Germany,2004, pp. 445-458.

[6] N. Kiyavitskaya, N. Zeni1, J.R. Cordy, L. Mich, J. Mylopoulos, “Semi-Automatic Semantic Annotations for Web Documents,” 2005.

[7] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, M. Goranov, “KIM – Semantic Annotation Platform,” in *2nd International Semantic Web Conf. (ISWC2003),* Florida, USA, 2003, pp. 834-849.

[8] G. Miller, “WordNet: An On-line Lexical Database,” *Special Issue, International Journal of Lexicography*, vol. 3, 1990. WordNet: http://wordnet.princeton.edu/

[9] M. Laclavik, M. Seleng, E. Gatial, Z. Balogh, L. Hluchy, “Ontology based Text Annotation – OnTeA,” *Information Modelling and Knowledge Bases XVIII. IOS Press, Amsterdam, Marie Duzi, Hannu Jaakkola, Yasushi Kiyoki, Hannu Kangassalo (Eds.), Frontiers in Artificial Intelligence and Applications*, vol. 154, February 2007, pp.311-315.