World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:3, No:2, 2009

Semi-Automatic Approach for Semantic Annotation

Mohammad Yasrebi, Mehran Mohsenzadeh

***Abstract*—**The third phase of web means semantic web requires many web pages which are annotated with metadata. Thus, a crucial question is where to acquire these metadata. In this paper we propose our approach, a semi-automatic method to annotate the texts of documents and web pages and employs with a quite comprehensive knowledge base to categorize instances with regard to ontology. The approach is evaluated against the manual annotations and one of the most popular annotation tools which works the same as our tool. The approach is implemented in .net framework and uses the WordNet for knowledge base, an annotation tool for the Semantic Web.

***Keywords*—**Semantic Annotation, Metadata, Information Extraction, Semantic Web, knowledge base.

.

I. INTRODUCTION

ANNOTATING web documents is one of the major techniques for creating metadata on the Web. Annotating websites defines the containing data in a form which is process able and suitable for interpretation not only by humans but also automated agents and machines.

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 [8].

The approach for automatic (versus manual) extraction of metadata is promising scalable, cheap, author-independent and (potentially) user-specific enrichment of the web content. However, at present there is no technology available to provide automatic semantic annotation in conceptually clear, intuitive, scalable, and accurate enough fashion. All existing semantic annotation systems rely on human intervention at hole or some point in the annotation process, Therefore, The annotation process is manual or semi-automatic. In this paper, we present a new approach to semantic enrichment (annotate) websites and documents by taking the annotation process to a conceptual level and by integrating it into an existing knowledge base "WordNet". This approach is semi-automatic system.

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 our developed approach there is no need for manual information extraction. It is not based on learning human-created samples either. The idea of information extraction lies in the concept of knowledge base, including a complete set of words, the collections of grammars, data frames and various lists of entities.

We discuss about the considered knowledge bases and architecture of them in [1].

II. THE PROCESS OF OUR APPROACH

This section discusses the process of our approach. The process consists of four steps (depicted in Fig. 1):

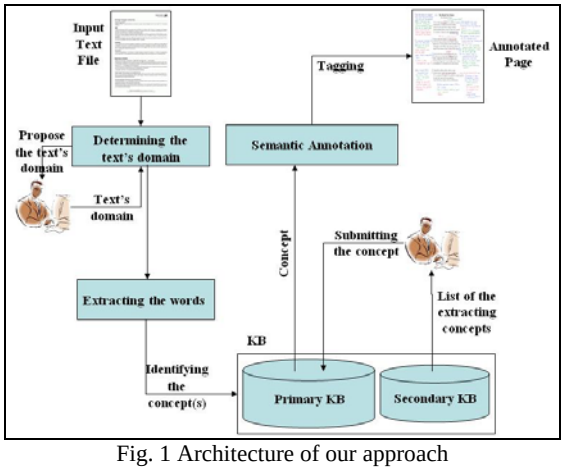

Fig. 1 Architecture of our approach

Input: Text's of a web page.

In our implementation, at first Web pages are cleaned from tags by tag-removers tools such as Emsa[1] and other parts of

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

M. Mohsenzadeh, is with the Islamic Azad University – Science and research branch, Tehran, IRAN (e-mail: m_mohsenzadeh77@yahoo.com).

[1] http://ww.snapfiles.com/get/emsatagremover.html

web page which haven't any relation to the content such as advertisements. Then contents of these web pages in text format are entered to system.
For example: Assume that this is the contents of the web site of one travel agency after performing the tag-removers tools.

Free golf at a beautiful new villa on Florida's sunny gulf coast.
For more information please contact to alen at alen@travel.org or request your offer to GTA.

Step 1: Determining the text's domain
System requests the subject and the text's domain from the user who knows the domain.
This process can be done as an offer. In other words, the various domains are suggested to the user and then he will select one of them or may insert the domain manually.
In above example system propose the domains such as "travel", "location", etc. User annotator who familiar to domain of this web page selects the one of these proposed domains or if there isn't the text's domain in proposed list, entered it manual.

Step 2: Extracting the words
In this phase, system extracts the all of words. Thus, by using a pattern which determines the words such as "/w+" and a loop, we extract the words of the text one by one to the end of the text.
For above example, system extracts these words:
Free, golf, at, a, beautiful, new, villa, on, Florida, sunny, gulf, coast, for, more, information, please, contact, to, alen, alen@travel.org, or, request, your, offer, to, GTA.

Step 3: Identifying the consept(s)
In this phase, we need to inspect words which are concepts or instances of a concept, and also explain a special meaning such as: email address, or name of person, etc. So, after analyzing the text to words, we have to send the words 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 data base which contains the words related to the domain. If the word exists in the inspected data base, the concept will be returned. In above example we assumed that the primary knowledge base just finding the words *"gulf"* and *"coast"* in domain *"travel"* and returned their concepts such as *"Ocean"* and *"Shore"*.

Then, for other identifying the concepts of other words the secondary knowledge base will help the primary knowledge base and determine its concept. The first choice for determining the concept of current word is the WordNet [4] as BKS. In this part, we have to inspect the word as a noun, verb, adjective or adverb. If the word is a noun the concepts will be extracted. So, we can get count of senses which are related to current word in WordNet. Just three modes may occur:

1. No sense exists for being noun.
2. Existing sense(s) for being noun and also other types(verb, adjective, adverb)
3. Existing sense(s) just for noun and no sense for other types.

For the first mode, we do not have to inspect the current word and then extract the concepts for this word, because the current word is not a noun at all. For the second mode, we have to compare the count of sense(s) related to the noun with the other sense(s) which are related to the each type such as verb, adjective or adverb. If the counts of the sense(s) which are related to the noun are more than the other types, it is obvious that this word can be a noun. Otherwise, we do not have to inspect the current word and then extract the concepts for this word. For the third mode, it is obvious that the current word is certainly a noun and we have to extract its concepts. After we recognized that the word is a noun, we search the concepts in WordNet. A list of the extracted concepts is shown to the user and the user will choose the related concept of the word from the list, or if the user's concept is not in the list, he has to insert it manually. In example above WordNet shows below list for user:
golf: outdoor game, athletic game, sport, activity, event
at: chemical element, substance, physical entity, halogen, group
a: metric linear unit, linear unit, linear measure, measure
villa: revolutionist, radical, person, organism, living thing
Florida: American state, state, administrative district, district, region, location
information: message, communication, collection, group
contact: interaction, action, act, event, connection
or: American state, state, administrative district, district, region, location

After chooses the one of concepts for each word which related to domain by user and needed to annotate too:
golf [Sport]
florida [State]
alen [Person_Name]
villa [Building]

User eliminates some words such as:
at, a, information, contact, or.

After the user submits this process that word will be inserted with its concept into the data base which is related to the text's domain, and as a result the primary knowledge base is updated and completed more and more.

The above cases happen when WordNet can identify the concept of the word, otherwise, data frame library or lexicons will help the WordNet.

If the word is the same as the one of the existing patterns (regular expressions) in data frame library, the concept is determined. For example, it specifies that this word is an email address, or a phone number, or IP address, etc. So, data frame library can detect *"alen@travel.org"* as an *"E_Mail"*.

Otherwise we have to search in different lists of lexicons and if the same case is found the concept will be determined. For example, it specifies:

"golf" as *"Sport"* , *"florida"* as *"State"* and *"alen"* as *"Person_Name"*.

If all of these knowledge bases could not find the concept(s) of one word, the user who knows the text's domain has to insert the concept manually. For example: the wotd *"GTA"* has the same condition and he inserted *"Travel_Agency"* as its concept manually.

The user removes the probable inconsistency among concept titles in basic knowledge base, lexicon, and data frame library (If the different parts of the secondary knowledge base have the different outputs for one word, the user can eliminate the inconsistency of these concepts and select the main concept of the current word).

After determining the concept of the current word, we have to go to the next word and we continue this process to the end of the text. To prevent doing this process twice for the words which are repeated more than once, we recognize these repeated words, and the process of extracting the concept for these words just operates once.

Step 4: Inserting the semantic tags

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, system insert and add tags which contain the concept of the words into the text. For example:

Free <Sport> golf </Sport> at a beautiful new <Building> villa </Building> on <State> Florida </State>'s sunny <Ocean> gulf </Ocean> <Shore> coast </Shore>.

For more information please contact to <Person_Name> alen </Person_Name> at <E_mail> alen@travel.org </E_mail> or request your offer to <Travel_agency> GTA </Travel_Agency>.

At the end of this phase, the first text that is considered as an input file is annotated with semantic tags. The performed tagging is only for presentation, and RDF format would be considered at the moment.

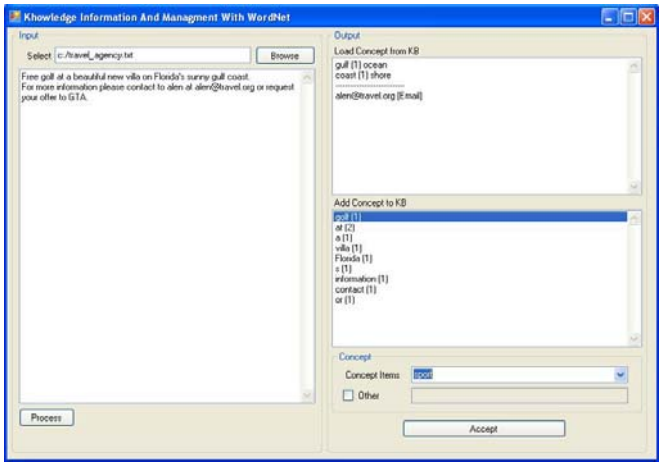

Fig. 2 The user interface of our system

The screenshot in Fig. 2 shows the user interface of our system. In the right side of the screenshot you can see the progress dialog for the primary knowledge base and secondary knowledge base queries. Upper side concepts returned from primary KB and lower side concepts returned from secondary KB. In lower right corner user can choose the related concept of the words from the list, or insert it manually.

III. Evaluations

In this section we deal with the performance and achievement of our system. To do so, the evaluation process is carried out in two phases. First, the system output was compared with manual output of a human annotator. It was thought that manual annotation is done under an ideal, highly accurate condition. Such evaluation, however, would be time-consuming and awkward especially when it involves a great number of documents and web-pages. As such, relying on software even with a margin of error would be reasonable. In the second phase of evaluation, the system output was compared with one of the existing annotation tools, called Ontea [3]. We selected this tool since it was noticeably compatible with our system. Ontea employs regular expressions and patterns as well as knowledge base to perform annotation process. In this evaluation, 50 html web-pages on business job offer were delivered to both systems and both systems' outputs were compared. To cope with the task, standard parameters (Recall and Precision) and F-measure (the harmonic mean of recall and precision) were taken into account.

After achieving the outputs, the relevant parameters were calculated. The results are shown in Table I:

TABLE I
Comparison of our system with Ontea

	Recall	Precision	F-measure
Our approach	90%	75%	81%
Ontea	83%	64%	71%
Human	High	High	High

As shown in the Table I, the measure of recall indicates that only %10 of the required correct annotation is not performed by this system. In other words, in %90 of cases our system has managed to map the instances existing in the text to the appropriate concepts of the ontology, and the result is statistically satisfying. Needless to say, the amount of recall is likely to reach %100 if the structure of pages are improved.

The measure of precision parameter indicates that %25 of annotation performed by the system is incorrect, or an instance is mapped to a wrong concept. The high rate of this figure, i.e. %25 is due to the polysemy of words in different pages. Even sometimes one word may have two totally different concepts in two different documents with one similar domain. In such special case, our system inputs the concept in the second document as it was done in the former one. It would be wrong, however, a user familiar with the domain is able to resolve the trouble. The F-measure also shows the general status of the system. In sum, the results of performance of our system imply its efficiency.

The main reason of our system's better performance is our more comprehensive knowledge base. As Ontea works only with patterns, it is more useful in pages which follow explicit, pre-defined structures. For example, if the name of a company that offers a job is as follows, Ontea would be able to identify it:

Company: Logitech

Therefore, it would be an appropriate tool to identify such pages. But, on pages which lack a clear-cut structure, Ontea fails to identify the existing entities of the text. The knowledge base of our system is a database including a quite complete lexicon as well as a comprehensive grammar and regular expressions, and also lists of various entities. It is not only a much better knowledge base that can identify the entities on explicit structures, but also it is able to identify the entities on unstructured pages.

In general, our system performs successfully on pages which make use of numerous words and concepts. When the pages include a great number of figures, however, our system loses its efficiency. This problem arises because of our basic knowledge base, i.e. WordNet. The drawback could be overcome by structuring such pages using regular expressions.

IV. 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 [6]; using information in different contexts, depending on the needs and viewpoint of 48 the user [5]; and question-answering [7].

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 Dezfulli, "A new approach to annotate the text's of the webpages and documents with a quite comprehensive knowledge base," in The International Conference on Computer, Electrical, and Systems Science, and Engineering CESSE, France, 2008.

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

[3] M. Ciglan, M. Laclavik, M. Seleng, L. Hluchy, "Document indexing for automatic semantic annotation support," 2007.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.