A Lightweight Approach to Semantic Tagging

Nadzeya Kiyavitskaya, Nicola Zeni, Luisa Mich, John Mylopoulus
Department of Information and Communication Technologies, University of Trento
Via Sommarive 14, 38050, Povo, Trento, Italy
{nadzeya.kiyavitskaya, nicola.zeni, luisa.mich}@unitn.it
Department of Computer Science, University of Toronto
jm@cs.toronto.edu

Abstract

Semantic Annotation is a challenging research direction in the area of Semantic Web. Turning the web into a Semantic Web implies widespread semantic annotation of documents. But it is still need to be investigated further in order to make annotation process more efficient, automating it as far as possible. The approach described in this paper aims at semi-automatic semantic tagging by application of linguistic lightweight methods for extraction of relevant concepts and by defining appropriate semantic models.

1. Introduction. Problem Statement

In the Semantic Web (SW) vision [1] of the Web resources are accessible not only to humans, but also to automated processes, e.g., automated "agents" roaming the web performing useful tasks, such as improved search and resource discovery, information brokering and information filtering. The automation of these tasks depends on elevating the status of the web from machine-readable to something we might call machine-understandable. The key idea is to have information on the web defined and linked in such a way that its meaning is explicitly interpretable by software processes rather than just being implicitly interpretable by humans. Within the vision of the SW, ontologies as semantic models have become an increasingly important research topic. To realize this vision, it is proposed to annotate web resources with metadata – data describing their content or functionality. Semantic tagging (or annotation) is now one of the promising methodologies to define semantic structures on the content.

Semantic tagging is the annotation of each content word with a semantic category. In many projects semantic categories are assigned on the basis of a semantic lexicon like WordNet for English [8].

Metadata may carry syntactic or semantic information.

- *Syntactic metadata* reflects the information about the data types and structures at the computer level. It describes non-contextual information about content, focusing on elements such as size, location or date of document creation providing little or no contextual understanding of what the document says or implies.
- *Semantic metadata* captures the information about the contents of the data. It is metadata that describe contextually relevant or domain-specific information about content based on a domain specific metadata model (e.g., industry-specific or enterprise specific) or ontology is known as semantic metadata.

Semantic markup of web documents is usually done manually with web-based knowledge representation languages, such as RDF[1] (Resource Description Framework) and OWL[2] (Ontology Web Language).

The notion of ontology has been adopted from philosophy. In computer science, an ontology[3] is the attempt to formulate an exhaustive and rigorous conceptual schema within a given domain. Typically ontology is represented as a hierarchical data structure containing all the relevant entities and their relationships and rules (theorems, regulations) within that domain. In other words an ontology is "an explicit specification of conceptualization" [3].

Initially in the current project we are focusing on the textual documents. The first thing to observe that most of them still resides in unstructured text archives, lacks metadata and they are only accessible through limited search mechanisms. Conversion of documents to semi-structured semantically annotated documents requires heavy costs in manual operations. Transforming textual data into semantically annotated documents should be largely automated to minimize costly human efforts. With rapidly growing amount of on-line web documents we will need the tools enable to transform web pages to semi-structured semantically annotated representation, effectively selecting to what text or parts of the text are relevant. But if we do not assume that the language used by users has to be a sub-set of natural language (NL), it is

[1] http://www.w3.org/RDF/
[2] http://www.w3.org/TR/owl-features/
[3] http://en.wikipedia.org/wiki/Ontology_(computer_science)

not possible to use ad hoc small natural language processing (NLP) systems, and a successful approach demands a system able to tackle the full range of NL problems [9], [7]. However, such systems are highly expensive to develop, resources and time consuming. Therefore we would like to leave out general-purpose NL understanding methodologies and turn to the lightweight processing. These techniques have advantages of being fast and scalable, and in large variety in the disposal.

We also may distinguish between different input documents:

- structured text,
- semi-structured text,
- unstructured plan text.

For each type of input data different kinds of tools can be applied with distinct level of efficiency.

The main goal of this project is to develop lightweight linguistic techniques to assist semi-automatic semantic tagging task as far as possible and to investigate their performance.

2. Related Works

Along the lines of ontology-based annotation there have been various projects, all of which aim at motivating people to richly annotate electronic documents in order to turn them into a machine-understandable format, and at developing and spreading annotation-aware applications such as content-based information presentation and retrieval. We list some of the current research projects which relates to semantic annotation task.

SHOE Knowledge Annotator [4] is an extension to HTML which provides a way to incorporate machine-readable semantic knowledge in HTML or other WWW documents. The Knowledge Annotator is a Java program that allows users to mark-up web pages with the SHOE ontology. SHOE associates a context with a web page; this context can be used to disambiguate terms and provide background knowledge that might help in interpreting content. It allows representing concepts, their taxonomies, n-ary relations, instances and deduction rules, which are used by its inference engine to obtain new knowledge.

OntoAnnotate [11] can be regarded as a workbench for semantic annotation of documents using domain-specific ontologies and this enriching HTML pages with semantics that a software agent is capable to automatically process the content of the page and reason about it. The interface dynamically adapts to the given ontology. The tool makes the relationship between particular ontologies and their parts, i.e. concepts and properties, explicit. The core of OntoAnnotate is used for viewing web pages and actually providing annotations. The approach uses shallow text processor for German.

Melita [2] is a semi-automatic annotation tool that has an Adaptive Information Extraction engine (Amilcare) integrated in it. Melita aims to support the user in the process of annotation. The system takes the initiative to do any pre-processing which will be used in the future. The novelty of Melita is the possibility of tuning the Adaptive Information Extraction system altering precision and recall, so as to provide the desired level of pro-activity and intrusiveness. Melita sorts documents after every annotation in order to find the document that best covers the unexplored areas of the domain. Documents with the least number of tags are taken to cover unexplored areas of the domain where new rules can be learned if they are annotated. This approach has led to a quicker convergence of the learning algorithm whilst overcoming the problem of data sparseness.

SemanticWord [12] is a semantic annotation tool, an environment based in MS Word that integrates content and markup authoring. It aims to reduce the effort to create semantically annotated documents by including the features for generation content and annotations simultaneously.

DIAsDEM system [14] was developed for semi-automated semantic tagging of domain-specific text documents. It includes knowledge discovery process. This process groups structural text units based on similarity of their content. Then acceptable clusters are labeled with default semantic names, which are refined by experts. Clusters labels serve as semantic tags for the corresponding text units. The system derives a flat, unstructured DTD that semantically describes an archive of XML documents.

The approach towards SW Information Extraction (IE) presented in [13] is implemented in KIM – a platform for semantic indexing, annotation, and retrieval. It combines IE based on the text engineering platform GATE with SW-compliant knowledge representation and management. The cornerstone is automatic generation of named-entity (NE) annotations with class and instance references to a semantic repository. The semantic annotation offered here is a specific metadata generation and usage schema targeted to enable new information access methods and extend existing ones. It is based on the hypothesis that the named entities mentioned in the documents constitute important part of their semantics.

Among the examples of multilingual semantic annotation let us refer to MUCHMORE project [10] where an XML annotation format and tool are developed on the basis of English-German corpus. The annotation scheme was designed specifically for the purposes of Cross-Lingual Information Retrieval in the medical domain so as to allow both efficient and flexible access to layers of information. Parallel English-German corpus of medical abstracts is used and annotated with linguistic information (tokenisation, part-of-speech tagging,

lemmatisation and decomposition, phrase recognition, grammatical functions) as well as semantic information from various sources. The annotation of medical terms (concepts), semantic types and semantic relations is based on the Unified Medical Language System (UMLS). Additionally, EuroWordNet is used as a general-language resource in annotating word senses and to compare domain-specific and general language use. A major aim of the project was also to complement existing ontological resources by extracting new terms and new semantic relations. The annotation scheme presented conceptually relates to stand-off annotation.

As we see from the variety of projects usually the solutions realize the following approaches. Large part of the semantic annotation intelligent tools that provide user with semi-automatic tagging facilities use knowledge learning techniques. The next well-known method uses regular expression based rules in conjunction with various semantic techniques to extract ontology-driven metadata from structured and semi-structured content. Often solutions are not general-purpose, but adapted to one particular semantic model and allow extraction of concepts from the restricted predefined range.

3. Our approach

In our project we are addressing the problem of automating semantic annotation.

We develop our approach basing on two main assumptions. First assumption is that in order to perform semantic tagging, this process needs to be sufficiently supported by lightweight NLP instruments, rather than relying on the heavyweight natural language processing approaches of building general purpose text understanding system.

Starting from the ontology-based tagging approach in current project, the markups for semantic annotation are supposed to be taken from the predefined semantic model, i.e. ontology. It reflects the user's vision of the domain. Ontologies allow one to define what is relevant to a particular problem and what should be ignored. However, as we may observe from the real life situations, what is relevant in data source for one user is not necessarily relevant for the other user. The approach of defining the unique domain ontology for a single way of annotation cannot always cover all the variety of possible concepts for any probable future user. New user can have own vision on the domain, which was not reflected in the model before. Distinct conceptual models can be applied to the same documents collection. Then, there are plenty of ontologies that also change over the time to reflect changes in the world.

From this observation we naturally derive our second initial assumption that the documents should not be annotated with some fixed ontology once and forever. The same text could be annotated with different semantic metadata by different users depending on the user's current vision of problem and on his/her goals in current search and consequently the information he/she is interested to extract.

The most difficult and tricky issue doing annotation is that the algorithm has to decide the degree of irrelevance of the text for it not to be included as related to the concept node of semantic model. The problem can be formulated as follows: given the text unit and the concept definition we have to decide if the input relates to this concept, eventually defining the measure of relevance. It would be the first answer when performing semantic, instead of syntactic tagging.

In this, also it is important to make sure that irrelevant concepts will not be associated and matched and that relevant concepts will not be discarded. In other words, it is important to insure that high precision and high recall will be preserved during concept selection for document unit. For this purpose the processing environment, having in disposal the range of shallow tools, may also posses the tuning facilities to vary the level of performance.

In order to provide the motivating example for possible applications consider the following use case scenario. Usually large organizations have a huge collection of textual data (internal and external documents, web resources, etc.) on the domain. The problem that often rises up when mining this information is extraction of data relevant to the goals of current particular request. The ontology-based models can be fruitfully deployed to facilitate information selection requests. Ontology-based approach demonstrates its power over keyword-based search techniques by providing many different levels of abstraction in a flexible manner with greater accuracy. The semantic structure provides index terms (concepts) that can be also used to match with user queries.

Keeping in mind all the described above aspects of the problem statement, the logical architecture of the semantic annotation system to be developed in this project foresees three main structural modules:

- *Semantic modelling module*: its functionality will deals with domain semantic model (re)definition and processing;
- *Language processing module*: its purpose is to provide variety of linguistic instruments to support different kinds of text processing and to regulate the desirable granularity of semantic annotation;
- *Output representation module*: to assist analysts to handle both plain textual, as well as user-friendly graphical views and storing of documents processing results.

At the conceptual level with respect to semantic annotation task we need to answer the following questions:

- How to adapt the given general semantic model to particular domain;
- How to annotate text at the finely grained level of semantic model classification.

The main problems to be addressed in this work at the design level are:

- The choice of the knowledge representation language to develop semantic models;
- The choice of the linguistic tools for efficient tagging.

In order to start experiments with lightweight tools, first we developed our prototype, which used pattern-based domain dependent method to extract the relevant data. Extraction processing part was implemented in TXL programming language[4].

4. Conclusions

The presented work considers the problem of automating semantic annotation. With rapidly growing amount of on-line web documents transforming textual data into semantically annotated documents should be largely automated to decrease costly and time-consuming human efforts.

Approach suggested here concentrates on the application of shallow text processing techniques for the associating semantic metadata to the documents' content on the text units level in order to find the parts, which particularly relates to the concepts. The pool of concepts is given by the semantic model and provides the set of appropriate tags for annotation.

The goal of this project is to experiment with different kinds of lightweight linguistic tools of various types and to investigate the different levels of semantic annotation accuracy.

5. References

[1] T. Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web", *Scientific American*, May 2001.

[2] F. Ciravegna, A. Dingli, D. Petrelli, Y. Wilks, "User-System Cooperation in Document Annotation based on Information Extraction", *Proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management*, Siguenza, Spain, 1-4 October 2002.

[3] T. R. Gruber. A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220, 1993.

[4] J. Heflin, J. Hendler, S. Luke, "SHOE: A Knowledge Representation Language for Internet Applications", *Technical Report CS-TR-4078* (UMIACS TR-99-71), Dept. of Computer Science, University of Maryland at College Park. 1999.

[5] J. Hendler, B. Parsia, "XML and the semantic web", *XML Journal*, Oct 2002.

[6] D.R. Karger, B. Katz, J. Lin, D. Quan "Sticky notes for the semantic web", *In Proceedings of the 8th international conference on Intelligent user interfaces table of contents*, Miami, Florida, USA, 2003, p. 254-256.

[7] N. Kiyavitskaya, N. Zeni, L. Mich, J. Mylopoulos, "Experimenting with Linguistic Tools for Conceptual Modelling: Quality of the models and critical features". *Submitted to NLDB2004*, UK, 2004.

[8] G.A. Miller, "WordNet: A Lexical Database for English". *Communications of ACM 11*, 1995

[9] R. Morgan, R. Garigliano, P. Callaghan, S. Poria, M. Smith, A. Urbanowicz, R. Collingham, M. Costantino, C. Cooper. Description of the LOLITA System as used for MUC-6. *In Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Francisco, 1995, p. 71-86.

[10] B. Sacaleanu, V. Martin, P. Buitelaar, "A Cross-Language Document Retrieval System Based on Semantic Annotation", *In Proceedings of EACL 2003 Demo Session*, Budapest, Hungary, April 2003.

[11] S. Staab, A. Maedche, S. Handschuh, "An Annotation Framework for the Semantic Web", *In the First International Workshop on Multimedia Annotation*, Tokyo, Japan, 2001.

[12] M. Tallis, "Semantic Word Processing for Content Authors", *In Workshop Notes of the Knowledge Markup and Semantic Annotation Workshop (SEMANNOT 2003), Second International Conference on Knowledge Capture (K-CAP 2003)*, Sanibel, Florida, USA, October 26, 2003.

[13] M. Uschold, R. Jasper, "A Framework for Understanding and Classifying Ontology Applications", *In Proceedings of the IJCAI99 Workshop on Ontologies and Problem-Solving Methods*, Stockholm, Sweden, August 1999.

[4] http://www.txl.ca/

[14] K. Winkler, M. Spiliopoulou, “Extraction of semantic XML DTDs from texts using data mining techniques”, *In Proceedings of the K-CAP 2001 Workshop on Knowledge Markup and Semantic Annotation*, Canada, October 2001, p.59-68.