Semantic Annotation for Semantic Social Networks

Using Community Resources

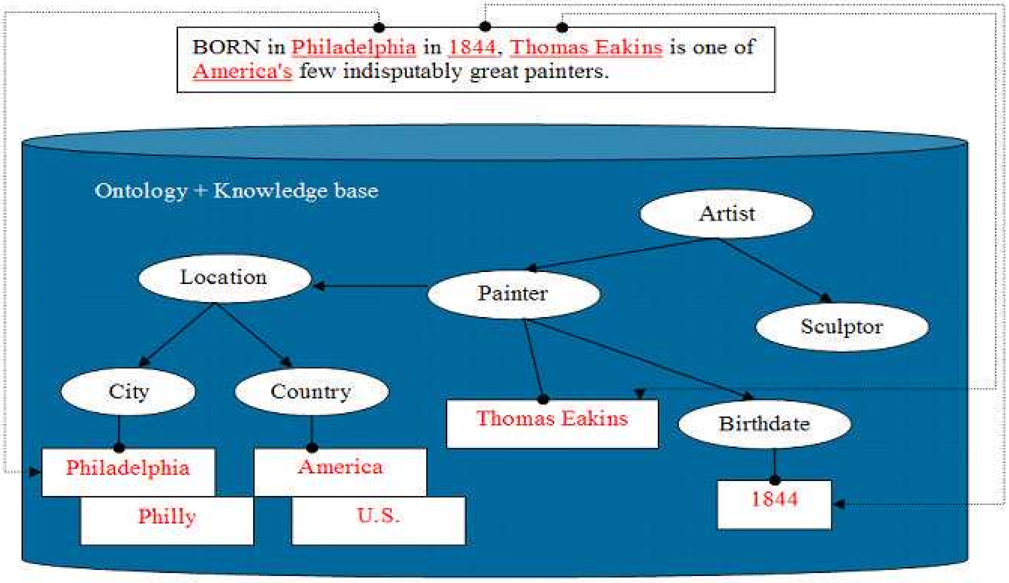
Lawrence Reeve and **Hyoil Han**
College of Information Science and Technology
Drexel University, Philadelphia, PA 19108
lhr24@drexel.edu
hhan@cis.drexel.edu

Reference to this article should be made as follows:
Reeve L., Han H., (2005), Semantic Annotation for Semantic Social Networks
Using Community Resources,
AIS SIGSEMIS Bulletin, Vol. 2, Issue (3&4), 2005, pp: 52-56

Semantic Social Networks (SSN) merge the Semantic Web with social networking so that resources and people related to the resources are linked together (Downes, 2004). An advantage of social and Semantic Web merging is to facilitate information searches among people with related interests. SSN applications environments are to include, among other features, content creation facilities. An ongoing problem with any content creation on the Semantic Web is the semantic annotation (or semantic tagging) of information of the new content. While tagging of content using user profile information is addressed by SSN, semantic tagging of content is not. We propose to provide a semi-automatic semantic tagging facility for new (and existing) Web-based text content using community-based knowledge resources.

Automatic semantic annotation of information content is an open problem, but is crucial to the realization of the Semantic Web. Annotation systems require the initial definition of an ontology and as well as a knowledge base. Both of these resources work together to facilitate markup. The ontology identifies the important concepts in a domain, while the knowledge base provides additional information, such as term synonyms for concepts. For example, the concept {Lung Cancer} can be expressed using at least three different terms: {"Lung Cancer," "Cancer of the Lung," "Carcinoma of the Lung"}. Semi-automatic semantic annotation systems use the synonyms in the knowledge base to find instances in a text source, and then map the instance to an ontological concept (i.e., a concept in the ontology). Figure 1 shows an example of how entities in a text source are mapped into ontological concepts using a knowledge base. Ontological concepts are represented as ovals, while knowledge base entries are indicated by rectangles.

Figure 1: Semantic tagging using an ontology and knowledge base.


BORN in Philadelphia in 1844, Thomas Eakins is one of America's few indisputably great painters.
Ontology + Knowledge base
Artist
Location
Painter
Sculptor
City
Country
Thomas Eakins
Birthdate
Philadelphia
America
Philly
U.S.
1844

There are three classes of semantic annotation systems: manual, semi-automatic, and automatic. Manual annotation provides facilities within a content editing environment allowing a user to select concepts from a predefined knowledge source. An example of this style of annotation is Semantic Word (Tallis, 2003), which provides an environment for authoring as well as marking up documents from within a single interface. The most significant drawbacks to manual annotation are the expense and inconsistency of human annotators. Semi-automatic annotation is currently the most viable approach. Semi-automatic systems perform text analysis to identify instances and then label the text with their corresponding ontological concepts. These systems are not completely automatic, however, due to the problem of disambiguation. For example, in biomedicine applications, there are two concepts for the term {Mass}: a quantitative concept {"how much"} and a finding concept {"found a mass"}. If there are insufficient clues to disambiguate which concept is intended, the system must consult the user to disambiguate the term. If the disambiguation problem can be solved, then automatic systems will be possible. For more descriptions of semantic annotation systems, please see our review paper (Reeve & Han, 2005).

In order for semantic annotation systems to perform, the knowledge base and ontology must be defined. There is often a considerable amount of work associated with constructing and maintaining these knowledge sources. In addition, the result is usually domain specific. One attempt at large scale automatic semantic tagging is the Seeker platform (Dill et al., 2003). Seeker has tagged 264 million Web pages with 434 million semantic tags. The tagging is done using an application called SemTag. SemTag uses as its knowledge source the TAP knowledge base (TAP KB) from Stanford University (Guha, R., McCool Robert, 2003). TAP KB is shallow but broad, and covers 12 categories with approximately 72,000 tags: Authors, Autos, Baby products, Companies, Consumer electronics, Health, Home Appliances, Movies, Music, Places, Sports and Toys.

SemTag scans text sources, finds tag instances, disambiguates them, and finally annotates the text using TAP tags.

We propose a similar approach where the TAP KB component is replaced with the Wikipedia free encyclopedia (I. Wikimedia Foundation, 2005) as the knowledge source. There are several reasons for choosing Wikipdia as a knowledge source. Wikipedia is an online encyclopedia developed by volunteer authors. The content is subject to consensus, rather than authoritative, approval. In this way, the content reflects the views of the larger community rather than a particular viewpoint. It is this aspect that is useful for using Wikipedia as a knowledge source for tagging on the general Semantic Web. Typically, Semantic Web content is tagged using a domain-specific ontology. It is therefore possible to tag the same content with different ontologies to gain different views of the same content. The use of a community-based, consensus-built knowledge source is one way to bootstrap Semantic Web content. That is, it can provide an initial tagging of content that can later have additional ontology tagging performed to reflect different views. We also find Wikipedia useful because it has an active community, and current event topics are updated or added as they occur. This allows new Semantic Web content to be brought online quickly. Finally, Wikipedia content is licensed using GNU Free Documentation License (Free Software Foundation, Inc., 2002), and is freely downloadable in XML format for machine processing. Wikipedia is available in 200 languages, and has more than 50,000 article entries for each of the ten most active languages, making it a large, multilingual and actively-developed knowledge source.

In order to make use of Wikipedia as knowledge resource for semantic annotation, semantic labels must first be extracted. Since Wikipedia was not designed for semantic annotation, processing must be done to convert the article content into useful tags. We propose converting Wikipedia content into a metathesaurus format to store the concepts and their term representations, similar to the National Library of Medicine's Unified Medical Language System (UMLS) Metathesaurus (United States National Library of Medicine, 2004). The UMLS Metathesaurus is composed of concepts and synonyms. The synonyms are based on medical vocabularies. UMLS also provides a semantic network to organize the concepts.

Each page in Wikipedia describes some topic. The topic name of each page becomes the concept name in the metathesaurus. The implication is that the community has determined that these topics are the most important. The name of the concept (topic page name) also becomes one of the lexical terms for identifying the concept. Figure 2 shows a fragment of markup for the opening text of the Wikipedia topic "John Roberts." We define the opening text as the markup segment bounded by the start of the topic page markup to the first segment divider, which is indicated by the prefix "==" on its own line. The topic name (concept name) is indicated by triple quotes. In example 2, this is '''John Glover Roberts, Jr.'''. Links to other concepts are implemented using opening and closing brackets ([[and]]). These links are used as 'related-to' concepts and are helpful for disambiguating the topic page concept ({John Glover Roberts, Jr.} in this example).

Figure 2: Wikipedia opening text markup for the topic "John Roberts."

'''John Glover Roberts, Jr.''' (born [[January 27]], [[1955]]) is the seventeenth [[Chief Justice of the United States]]. Roberts previously was a judge on the [[United States Court of Appeals for the District of Columbia Circuit]], spent 14 years in [[Law of the United States|private law practice]], and held positions in [[Republican Party (United States)|Republican]] administrations in the [[United States Department of Justice|U.S. Department of Justice]] and [[White House Counsel|Office of the White House Counsel]].

==Personal life, education, and memberships==

From the topic page opening text, "John Roberts" is identified as a concept and also as the preferred term for the concept. In addition, "John Roberts" becomes a lexical term. Additional synonyms are derived from the opening text. Synonyms are usually marked by including them within three single quotes ('''). The three-quote heuristic is one way of identifying synonyms. Synonyms can also be derived from Wikipedia redirect pages. Redirect pages redirect users using one term to the main topic page. For example, the topic page "John Glover Roberts, Jr." redirects to the "John Roberts" topic page. "John Glover Roberts, Jr." is then identified as a synonym for "John Roberts."

A basic semantic network to classify related concepts can be generated from Wikipedia topic categories. A Wikipedia category is a special page provided within Wikipedia to organize topic pages. A Wikipedia topic page can belong to multiple categories. For example, the topic page (concept) "John Roberts" belongs to nine categories ("Chief Justices of the U.S.," "Judges of the U.S. Court of Appeals for the DC Circuit," "American lawyers," "Harvard alumni," "Harvard Law School graduates," "Roman Catholic jurists," "Ambidextrous people," "People from New York," "1955 births.") Subcategories are also used within Wikipedia, forming a hierarchical tree of topics. For example, "Chief Justices of the U.S." is a subcategory of "United States Supreme Court." Building a semantic network from topic categories is somewhat problematic because multiple categorization schemes can exist at the same time (Wikimedia Foundation, 2005).

An important part of any semantic tagging application is the disambiguation step. It is possible for a term to map to multiple concepts (Wikipedia topics), and automated tagging is often unable to distinguish which concept is the intended concept. In these cases, the machine defers to a user to disambiguate among many candidate concepts. Wikipedia provides disambiguation pages to let users know a topic may also refer to other topic pages with similar topic page names. Using the example from above, the term "John Roberts" can map to "John Roberts" the Chief Justice of the United States, "John Roberts" the television journalist, or "John C. Roberts" the founder of an Australian construction company, among others. In our observation, the most important information is contained in the opening text. To perform automatic disambiguation, the concepts in the opening text are extracted as 'related-to' instances of the main concept. In Figure 2, these are indicated with the markup contained in opening and closing brackets ([[and]]). The related-to concepts are then used in the disambiguation step in the following manner. First, the source text is processed to identify all ambiguous concepts and unambiguous concepts. Each unambiguous concept has its corresponding unique concept and is annotated (or tagged) with the unique concept. Ambiguous concepts have more than one candidate concept. For each candidate concept of an ambiguous concept, the unambiguous concepts in the source text already tagged are used to find matching related-to concepts for each candidate concept. The candidate concept having the highest frequency count of related-to concepts becomes the disambiguated concept, and the source text is tagged with this concept. If all candidate concepts of an ambiguous concept do not have any related-to concepts, or a frequency count tie occurs, the machine is unable to successfully complete disambiguation and the user must be consulted for a final determination of the semantic label. This approach to disambiguation does not require any prior training, as is the case with other disambiguation approaches, such as SemTag (Dill et al., 2003). This approach also allows the re-use of concepts already manually tagged by the topic editors of Wikipedia.

Social semantic networks (SSN) integrate the existing technologies of social networks and the Semantic Web. We believe that by extending this integration to include community-based knowledge sources and applying these resources to semantic annotation, Semantic Web content for SSNs will appear more rapidly and be more valuable to their users.

References

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., & Jhingran, A. et al. (2003). SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. *Twelfth International World Wide Web Conference,* Budapest, Hungary, 178-186.

Downes, S. (2004). *The Semantic Social Network.* Retrieved October 9, 2005 from www.downes.ca/cgi-bin/website/view.cgi?dbs=Article&key=1076791198.

Free Software Foundation, Inc. (2002). *GNU Free Documentation License.* Retrieved October 9, 2002 from http://en.wikipedia.org/wiki/Wikipedia:Text_of_the_GNU_Free_Documentation_License.

Guha, R. and McCool, R. (2003). TAP: A Semantic Web Platform. *Computer Networks: The International Journal of Computer and Telecommunications Networking. Special Issue: The Semantic Web: An Evolution for a Revolution, 42*(5), 557-577.

Reeve, L., & Han, H. (2005). Survey of Semantic Annotation Platforms. *Proceedings of the 20th Annual ACM Symposium on Applied Computing, Web Technologies and Applications track,* Santa Fe, New Mexico.

Tallis, M. (2003). Semantic Word Processing for Content Authors. *Second International Conference on Knowledge Capture,* Sanibel, Florida, USA.

United States National Library of Medicine. (2004). *UMLS Metathesaurus Fact Sheet.* Retrieved July 31, 2005 from http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html.

Wikimedia Foundation. (2005). *Wikipedia Category.* Retrieved October 9, 2005 from http://en.wikipedia.org/wiki/Wikipedia:Category.

Wikimedia Foundation, I. (2005). *Wikipedia.* Retrieved October 9, 2005 from http://en.wikipedia.org/wiki/Main_Page.