Linguistic Techniques and Standards for Semantic Mark-up

James Pustejovsky
LingoMotors, 585 Mass. Ave.
Cambridge, MA 02139
jamesp@lingomotors.com
and

Computer Science Department
Brandeis University, Waltham, MA 02254
jamesp@cs.brandeis.edu

1 Semantic Technologies for Mark-up Standards

In this position paper, we present several points regarding current efforts towards the establishment of standards for web content, particularly from the contribution of natural language technologies. The position argued here is that the modeling of natural language meaning is critically important for the robust establishment of content markup languages for web protocols and communication. It is the view here that the largest, and indeed, the most successful, semantic webs are existing linguistic communities. Although this might appear trivial or irrelevant to the concerns of web content mark-up and communication, we shall argue that the efforts recently underway in the natural language technologies community, such as with LingoMotors and other companies, are relevant to the establishment and dissemination of semantic web standards.

The research and development effort at LingoMotors is focused on the automatic identification of semantic content, in the form of digital assets such as web text, for subsequent use by a consuming application. Such applications include information and database retrieval systems, CRM, content and knowledge management systems, as well as categorization and clustering algorithms. From the general perspective of the present workshop, such technologies can be seen as enabling several opportunities relating to the automatic semantic markup, interpretation, and verification of web content and interactions. More specifically, LingoMotors technology is helping to realize a user interface for the web that is mediated through ordinary language interactions.

2 Encoding and Recognizing Content

The semantic technologies being developed at LingoMotors contribute directly to the realization of a robust *Language User Interface* (LUI) for the web, helping to drive the interpretation of business exchanges, commerce, and other communicative transactions. A Language User Interface provides a platform from which users interact with the computer in ordinary language, for most everyday applications. The underlying technology responsible for this is what we call the *Concept Machine* (c-machine). Such an interface effectively hides the complexity of the knowledge and actions behind the words and text.

Web interactions outfitted with semantic technologies can offer distinct capabilities that are impossible to achieve with current search and navigation technologies, for identifying the rich and pertinent structural information about objects in an application area. For example, a "knowledge object" corresponding to the notion of an ecommerce product has `performance`, `form factor`, `price`, and `customer attributes`, each of which can take specific values. The concept machine not only contains these structural descriptions but also the knowledge for recognizing and attaching attributes and their values to the objects. Some of the major areas enabled by semantic technologies include the following:

1. Relation Searching and filtering: Many searches or filtering conditions are interested not in things, but in events and actions; that is, a category or type of company buying or acquiring another company, the announcement of a new alliance or product introduction or price drop, etc. This is possible only with robust recognition of the semantic types of the relations in the texts.

2. Category-based searching and filtering: A significant shortcoming of today's searching and filtering techniques is that the user has to know the name of what he or she is looking for in order to find or exclude it. Semantic technologies allow the user specify search or filtering conditions based upon categories, such as product type, company type, personal title, without having to know the literal names.

3. Categorization and Clustering: Semantic technologies provide recognition of entities and relations between entities. As a result, they enable richer, more meaninful categorization possibilities. Similarly,

clustering algorithms are driven by concepts rather than word tokens, resulting in more informed classifications.

4. Similarity matching: Current similarity searching methods primarily rely upon statistical comparisons of literal strings. Semantic mark-up enables similarity searches and matches to combine the rich description inherent in the concept machine with statistical information metrics to provide more meaningful matches.

We believe that the success and deployment of semantic technologies is essential for the realization of the goals and vision of the semantic web community. The LingoMotors technology is focused directly at addressing these challenges.