Università di Salerno

Dottorato di ricerca in
Ingegneria dell'informazione

Probabilistic Lightweight Ontology for the extraction and representation of Semantics

Abstract

Author:
Fabio Clarizia

Tutor:
Prof. Massimo De Santo

Coordinator:
Prof. Angelo Marcelli

Anno Accademico 2009-2010, IX ciclo - Nuova Serie

The Semantic Web and Knowledge Engineering communities are both confronted with the endeavor to design and build ontologies by means of different tools and languages, which in turn raises an "ontology management problem" related to the peculiar tasks of representing, maintaining, merging, mapping, versioning and translating. These mentioned above are well known concerns animating the debate in the ontology field. However, we argue that the utilization of different tools and languages is mainly due to a personal view of the problem of knowledge representation, which in turn raises a not uniform perspective.

Most important each ontology scientist may rely, deliberately or implicitly, on a different definition of the role of ontology as mean for semantics representation [San06]. Therefore we argue that a special effort should be devoted to better explain and clarify the theory of semantic knowledge and how we should correctly model the latter for being properly represented and used on a machine. A simple process to convey meaning through language can be summarized as follows:

$$\textit{meaning} \rightarrow \text{encode} \rightarrow \text{language} \rightarrow \text{decode} \rightarrow \textit{meaning'},$$

where, since encoding/decoding processes are noisy, *meaning'* is the estimation of the original *meaning*. In order to understand why those processes are noisy we assume that a communication act through language is in the form of writing/reading a book. Here, the origin of the communicative act is a meaning that resides wholly with the author, and that the author wants to express in a permanent text. This meaning is a-historical, immutable, and pre-linguistic and is encoded on the left-hand side of the process; it must be wholly dependent on an act of the author, without the possibility of participation of the reader in an exchange that creates, rather than simply register, meaning. The author translates such creation into the shared code of language, then, by opening a communication, he sends it to the reader at the encoding stage. It is well known that, due to the accidental imperfections of human languages, such translation process may be imperfect, which in turn means that such a process is corrupted by "noise". Once the translated meaning is delivered to reader, a process for decoding it starts. Such process (maybe also corrupted by some more noise) obtains a reasonable approximation of the original meaning as intended by the

author. As a consequence meaning is never fully present in a sign, but it is scattered through the whole chain of signifiers: it is deferred, through the process that Derrida [Der97] indicates with the neologism differànce, a dynamic process that takes plane on the syntagmatic plane of the text [Eco79].

In the light of this discussion we argue that, as pointed out by Steyvers and his colleagues [TLG07], the semantic knowledge can be thought of as knowledge about relations among several types of elements, including *words*, *concepts*,*actions* and *percepts*. According to such definition the following relations must be taken into account:

1. *Concept – concept* relations. For example: knowledge that dogs are a kind of animal, that dogs have tails and can bark, or that animals have bodies and can move;

2. *Concept – action* relations: Knowledge about how to pet a dog or operate a toaster.

3. *Concept – percept* : Knowledge about what dogs look like, how a dog can be distinguished from a cat;

4. *Word – concept* relations: Knowledge that the word dog refers to the concept "dog," the word animal refers to the concept "animal," or the word toaster refers to the concept toaster;

5. *Word – word* relations: Knowledge that the word dog tends to be associated with or co-occur with words such as tail, bone.

Obviously these different aspects of semantic knowledge are not necessarily independent, rather those can influence behavior in different ways and seem to be best captured by different kinds of formal representations. As a consequence result, different approaches to modeling semantic knowledge tend to focus on different aspects of this knowledge, specifically we can distinguish two main approaches:

I The focus is on the structure of associative relations between words in natural language use and relations between words and concepts, along with the contextual dependence of these relations [EK95, Kin88]. This approach is related to points 4 and 5, which can be defined as *light semantics*;

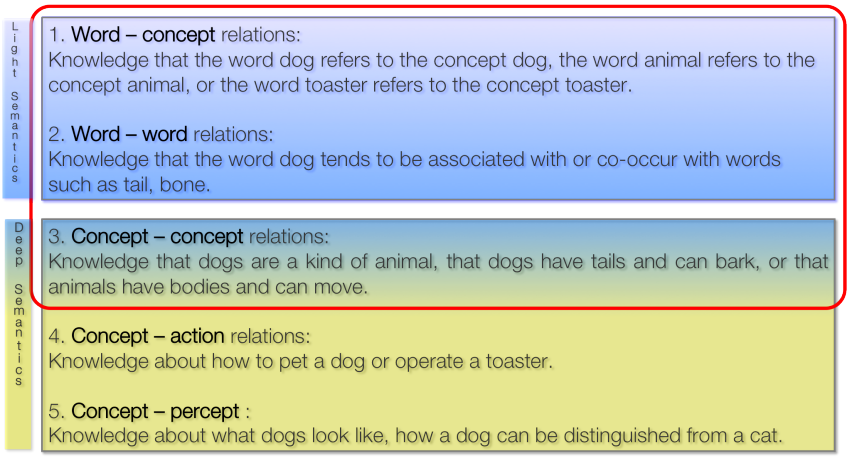

Figura 1 Levels of representation of Semantics: the computational model proposed in this work refers to the reports included in the red frame.

II The emphasis is on abstract conceptual structure, focusing on relations among concepts and relations between concepts and percepts or actions [CQ69]. This approach is related to points 1, 2 and 3, which can be defined as *deep semantics*.

Once a computational model for each of the two components of semantics has been formulated, the very aim of this research project is to investigate the interaction between them and how such interaction can be modeled through probabilistic methods.

We argue that probabilistic inference is a natural way to address problems of reasoning under uncertainty, and uncertainty is plentiful when retrieving and processing linguistic stimuli. In this direction, it has been demonstrated that language possesses rich statistical structure that could be captured through probabilistic models of language based on recent techniques from machine learning, statistics, information retrieval, and computational linguistics.

Specifically, the description of both *Word – Word* and *Word – Concept* relations, namely *light semantics*, is based on an extension of the computational model, namely the topic model, introduced by Steyvers in [TLG07], where statistic dependence among words is assumed. Topic model is based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model

is a generative model for documents: it specifes a simple probabilistic procedure by which documents can be generated.

The deep semantics is traditionally represented in terms of systems of abstract proposition [CQ69]. Models in this tradition have focused on explaining phenomena such as the development of conceptual hierarchies that support propositional knowledge, reaction time to verify conceptual propositions in normal adults, and the decay of propositional knowledge with aging or brain damage.

Once introduced the general model, we have focused the attention on some of the aspects discussed above so that the core of our proposal is the definition of a type of informal knowledge (see figure 1), that we named *informal Lightweight Ontology*(*iLO*), and that can be derived automatically from documents. Precisely, the *vector of features*, that we call mixed *Graph of Terms*, can be automatically extracted from a set of documents $\mathcal{D}$ using a *global* method for *term extraction* based on a supervised *Term Clustering* technique [Seb02] weighted by the *Latent Dirichlet Allocation* [BNJ03] implemented as the *Probabilistic Topic Model*. The graph is composed of a directed and an a-directed subgraph (or levels). We have the lowest level, namely the *word level*, that is obtained by grouping terms with a high degree of pairwise semantic relatedness; so there are several groups (clusters), each of them represented by a cloud of *words* connected to their respective centroids (directed edges), also called *concepts*. Further, we have the second level, namely the *conceptual level*, obtained by inferring semantic relatedness between centroids, and so between *concepts* (undirected edges).

We have experimented the research proposal setting up two different scenarios, specifically:

I) The first exploits the potential of conceptual categorization of the *iLO* on large collections of textual data, such as repository of web pages;

II) the second is more focused in the area of *User Satisfaction* where the aim is the use of the proposed technique to retrieve, from a large repository of web pages, documents that are as close as possible to the user intentions.

In the first environment, we prove that the accuracy of a text retrieval system can be improved if we employ a query expansion method

based on a mixed *Graph of Terms* instead of a method based on a simple list of words. The graph is composed of a directed and an a-directed subgraph and can be automatically extracted from a set of documents using a method for *term extraction* based on the *probabilistic Topic Model*.

In the second phase of testing, we have shown how the performance of a classic web search engine (in this case, we have used a customized version of Google - *Google Custom Engine*), in terms of quality of results that have been retrieved by performing informational querying tasks, can be improved through the use of an innovative informal lightweight ontology based search technique. As a consequence, the proposed method is such that the retrieved pages are closer to user intentions and thus it improves the overall level of user satisfaction.

In both cases, the results confirmed that the proposed technique certainly increases the performance in terms of relevance, confirming that an informal structure made of concepts and links between them, is capable of providing a greater specialization of the intention and so reducing the inherent problems ambiguity of language.

Bibliografia

[BNJ03] D. M Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(993–1022), 2003.

[CQ69] A. M. Collins and M. R. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, (8):240–247, 1969.

[Der97] Jaques Derrida. De la grammatologie. *Paris:Minuit*, 1997.

[Eco79] Umberto Eco. A theory of semiotics. *Bloomington:Undiana University Press.*, 1979.

[EK95] K. A. Ericsson and W. Kintsch. Long-term working memory. *Psychological Review.*, 102:211–245, 1995.

[Kin88] W. Kintsch. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95:163–182, 1988.

[Pot93] M. C. Potter. Very short term conceptual memory. *Memory & Cognition*, (21):156–161, 1993.

[San06] Simone Santini. Summa contra ontologiam. *Current Trends in Database Technology*, 2006.

[Seb02] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47, March 2002.

[TLG07] J. B. Tenenbaum T. L. Griffiths, M. Steyvers. Topics in semantic representation. *Psychological Review*, 114(2):211–244, 2007.