Semantic Annotation and Human Language Technology

Kalina Bontcheva, Hamish Cunningham, Atanas Kiryakov and Valentin Tablan

Abstract

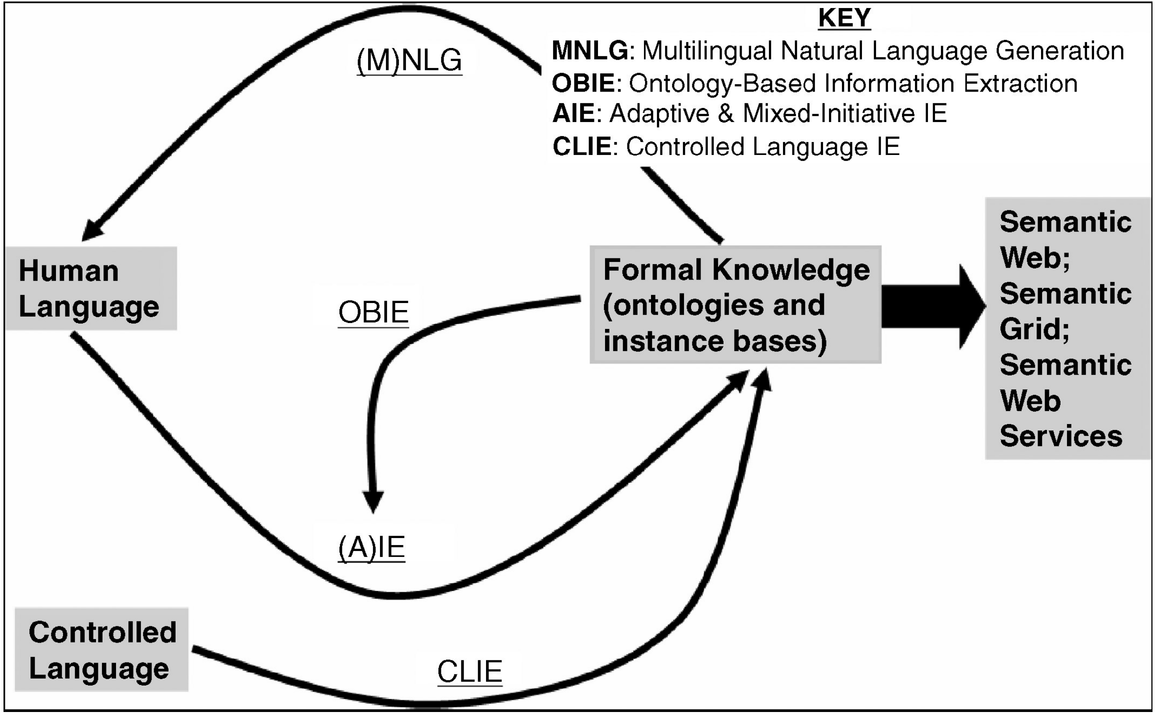
Gartner reported in 2002 that for at least the next decade more than 95% of human-to-computer information input will involve textual language. They also report that by 2012, taxonomic and hierarchical knowledge mapping and indexing will be prevalent in almost all information-rich applications. There is a tension here: between the increasingly rich semantic models in IT systems on the one hand, and the continuing prevalence of human language materials on the other. The process of tying semantic models and natural language together is referred to as Semantic Annotation. This process may be characterised as the dynamic creation of inter-relationships between ontologies (shared conceptualisations of domains) and documents of all shapes and sizes in a bidirectional manner covering creation, evolution, population and documentation of ontological models. Work in the Semantic Web (Berners- Lee, 1999; Davies et al., 2002; Fensel et al., 2002) has supplied a standardised, web-based suite of languages (e.g., Dean et al., 2004) and tools for the representation of ontologies and the performance of inferences over them. It is probable that these facilities will become an important part of next-generation IT applications, representing a step up from the taxonomic modelling that is now used in much leading-edge IT software. Information Extraction (IE), a form of natural language analysis, is becoming a central technology to link Semantic Web models with documents as part of the process of Metadata Extraction.

The Semantic Web aims to add a machine tractable, repurposeable layer to complement the existing web of natural language hypertext. In order to realise this vision, the creation of semantic annotation, the linking of web pages to ontologies and the creation, evolution and interrelation of ontologies must become automatic or semi-automatic processes.

In the context of new work on distributed computation, Semantic Web Services (SWSs) go beyond current services by adding ontologies and formal knowledge to support description, discovery, negotiation, mediation and composition. This formal knowledge is often strongly related to informal materials. For example, a service for multimedia content delivery over broadband networks might incorporate conceptual indices of the content, so that a smart VCR (such as next generation TiVO) can reason about programmes to suggest to its owner. Alternatively, a service for B2B catalogue publication has to translate between existing semistructured catalogues and the more formal catalogues required for SWS purposes. To make these types of services cost-effective, we need automatic knowledge harvesting from all forms of content that contain natural language text or spoken data. Other services do not have this close connection with informal content, or will be created from scratch using Semantic Web authoring tools. For example, printing or compute cycle or storage services. In these cases the opposite need is present: to document services for the human reader using natural language generation. An important aspect of the world wide web revolution is that it has been based largely on human language materials, and in making the shift

to the next generation knowledge-based web, human language will remain key. Human Language Technology (HLT) involves the analysis, mining and production of natural language. HLT has matured over the last decade to a point at which robust and scaleable applications are possible in a variety of areas, and new projects like SEKT in the Semantic Web area are now poised to exploit this development. Figure below illustrates the way in which Human Language Technology can be used to bring together the natural language upon which the current web is mainly based and the formal knowledge at the basis of the Semantic Web.

Figure: HLT and Semantic Web


KEY
MNLG: Multilingual Natural Language Generation
OBIE: Ontology-Based Information Extraction
AIE: Adaptive & Mixed-Initiative IE
CLIE: Controlled Language IE
(M)NLG
Human Language
OBIE
Formal Knowledge (ontologies and instance bases)
Semantic Web; Semantic Grid; Semantic Web Services
(A)IE
Controlled Language
CLIE