

LANGUAGE AND SEMANTICS: WHAT CAN YOU DO FOR MY SEARCH ENGINE (AND FOR ME)?

Antonio Valderrabanos, CEO, Bitext, Spain

Online Information Conference - London, December 2nd 2009

Keywords

Semantics, semantic search, search technologies, Natural Language Processing, NLP, text analytics, Business Intelligence

500 word abstract

Over the last few years the World Wide Web has become a digital Gutenberg which has unleashed a completely new business and information sharing scenario. Publishers of all types of content have chosen the Web as repository for content previously found in papers or private archives. The Web has even become a medium of publication of native content such as blogs, forums and twitters. Therefore, we can only expect an exponential growth of publisher and user-generated content.

In order to get hold of the explosion of content, searching technologies continue to be the only tool available to individual users. Search itself can be construed as an implementation of dynamic and limitless hyperlinking since every time we do a search we are linking different documents according to the keywords in the search query. And for the time being search remains to be the only technology that can make the web manageable for end users, particularly as a self-service which is simple and intuitive for the average person.

However, search is an old technology which dates back to the sixties and it was not designed to solve the challenge of an increasing number of users and growing complexity in an also increasing number of documents. In fact, for end users search has shifted from being a service provided by librarians to a self-service similar to ATMs. This change generates frustration for users and puts pressure on search engine providers to improve performance and user-friendliness. As a result, the Web community realizes that most of the potential of Web and the knowledge it contains are underexploited or are even unknown.

And here is where Semantics comes to the rescue: the Web community is looking at Semantics as the source of solutions for exploiting all the potential of the Web since Semantics is the science of meaning, and it is the meaning of Web texts the challenge to be addressed. The so-called Semantic Web is the tag under which various research efforts are merging, such as knowledge representation, automatic reasoning, etc. But so far results are falling short of expectations because implementing Semantic Web principles at web level becomes an impossible task even if the task could be handled in an automated fashion, and this becomes a stumbling block to creating semantic knowledge.

That is why Natural Language Processing (NLP) is the solution to automate the knowledge acquisition problem because current NLP technologies provide one of the key ingredients for the Semantic Web to become a reality: text analytics or the ability to extract content from text. This ability can be turned into two highly needed tasks: automatic text tagging of entities, concepts and events; and automatic population of ontologies with selected entities, concepts and facts. In addition, NLP technologies can also provide interfaces capable of natural language understanding which are required by self-service end users.

Since 2007 Bitext is applying this approach to real-life projects in areas such as citizen services and business intelligence.

Key learning points:

1. Search engines are not delivering at the level that users demand. Semantics can help search engines deliver the results users need
2. Semantics is a complex discipline in which language knowledge and world knowledge converge. It is seen as a holy grail for the search problem
3. Semantic search requires the compilation of costly resources, intense in manual work and hard to reuse (ontology population, text tagging). Language technologies (text analytics) are one of the key technologies that can automate these tasks