

"Water, water, Everywhere...": Data-driven Computational Linguistics Research

By

Dr Kumaran

Head, Multilingual Systems Research Group
Microsoft Research India, Bangalore

Abstract of the Talk

The phrase by the Ancient Mariner captures the essence of data needed for Computational Linguistics research aptly. Thanks to the Internet, natural language data are readily available in the media & Internet, but in a highly unstructured form, with a widely varying quality. In this talk, we outline the current statistical and machine learning based approaches in Computational Linguistics research, and underscore the critical need for data to support such research. We'll discuss some simple problems to show the effectiveness of such approaches. However, these approaches critically depend on large amounts of appropriate linguistics data, making such research possible only in a handful of languages of the world. While the natural language data are available in large quantities, they need to be "structured" to glean useful information from them.

We present two broad approaches in structuring the data, namely, mining the web and involving internet users to provide data. Mining the web data addresses mining multilingual news stories that get produced by news organizations on the same timeline for useful information. Engaging the Internet population in providing data may be useful in specific tasks and domains, paralleling the successes of community contributed efforts like Wikipedia. We discuss our experiences in implementing such approaches for creating linguistic data.

Profile of the Speaker:

Dr Kumaran is currently the head of the Multilingual Systems Research group at Microsoft Research India. He has a Bachelors degree from Anna Univ., Chennai, and a Masters from Rutgers Univ., New Jersey, USA.

After his Masters, he worked for Bell Communications Research, US and Oracle Corporation, US/India for 15 years, before moving back to a research-oriented career. He joined Indian Institute of Science, Bangalore to pursue doctoral research in the area of Multilingual Information Retrieval Architectures.

After PhD in 2005, he joined Microsoft Research India, where he heads a research group engaged in research on multi-lingual and cross-language systems. His current interests are resource creation for research, Machine Translation/Transliteration and crosslingual IR systems; he is also interested in methodologies for collecting linguistic data to enable research in resource-poor languages

Date & Time of Talk: 19th Dec 2009 at 13.45 hrs

Invited Talk on