Natural language processing: challenges and applications

Prof. Osvaldo N. Oliveira Jr.
Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Física de São Carlos, USP, Brazil
chu@ifsc.usp.br

Some of the innovations brought about by information technology, such as search machines for the Internet and voice-controlled equipment, bring natural language processing (NLP) to the forefront of science and technology areas with great prospects of increasing impact in the near future. The ultimate NLP goal of a fully-fledged communication with machines remains elusive, however, particularly because – as it is now widely accepted – representing knowledge to deal with language subtleties is a very complex task. In this lecture, an overview will be presented of the main challenges facing scientists and technologists to improve the performance of NLP tasks, with emphasis on machine translation and speech recognition systems. The paradigms for NLP, namely the symbolic, the connectionist and the hybrid approaches, will be discussed in terms of their strengths and limitations. Regardless of the paradigm used, any NLP application involves a multi-step process requiring various computational linguistic resources, e.g. dictionaries, part-of-speech taggers, knowledge bases, signal processing units for speech recognition, which need to be integrated. Suggestions will be made of possible efforts to provide open access resources that could be reusable to generate high performance NLP applications.

Dealing with complexity in natural language processing tasks

Prof. Osvaldo N. Oliveira Jr.
Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Física de São Carlos, USP, Brazil
chu@ifsc.usp.br

The discovery that a number of artificial as well as natural systems exhibit the topology of scale-free networks has boosted the area of complex networks, through which concepts of graph theory and statistical physics are merged. Among such systems are included texts represented by a weighted adjacency matrix where the lemmatized words are the network nodes and the edges are created by co-occurrence of words in the text. The scale-free nature of the networks representing text has been exploited, for example, in the investigation of semantic networks and thesauri. In this lecture, the fundamental concepts of complex networks will be presented, to be followed by applications in linguistics in general, with particular attention to natural language processing (NLP). These applications include the use of complex network metrics to assess the quality of written essays by high school students and of machine translation systems, in addition to serving for establishing summarizing strategies and identifying authorship. It will be shown that important features of the text can be captured by combining various metrics of the resulting complex network, which is promising for embedding into NLP systems based on the connectionist approach.