Big Mechanisms for Processing Big Data in Medical Informatics

Sanda Harabagiu
Professor, Computer Science Department
Director, Human Language Technology Research Institute
The University of Texas at Dallas
(*www.hlt.utdallas.edu*)

Abstract: Unprecedented volumes of clinical data, both in structured and unstructured formats, as well as in multimedia format, give rise to new mechanisms of capitalizing the biomedical data. In order to capture the complex, causal and explanatory models of medical data and to facilitate retrieval of relevant information, big mechanisms operating on this medical big data need to be developed. This talk presents such a big mechanisms facilitated by a MapReduce implementation which generates automatically a Qualified Medical Knowledge Graph (QMKG) which can be used for retrieving patient cohorts with higher precision than state-of-the-art methods.

Clinical data is expressed within the narrative portion of the electronic medical records (EMRs), requiring natural language processing techniques to unlock the medical knowledge referred to by physicians. This knowledge, derived from the practice of medical care, complements medical knowledge already encoded in various structured biomedical ontologies. Moreover, the clinical knowledge derived from EMRs also exhibits relational information between medical concepts, derived from the cohesion property of clinical text, which is an attractive attribute that is currently missing from the vast biomedical knowledge bases.

This talk shall describe an automatic method of generating a graph of clinically related medical concepts by considering the belief values associated with those concepts. The belief value is an expression of the clinician's assertion that the concept is qualified as present, absent, suggested, hypothetical, ongoing, etc. Because the method detailed in this talk takes into account the hedging used by physicians when authoring EMRs, the resulting graph encodes qualified medical knowledge wherein each medical concept has an associated assertion (or belief value) and such qualified medical concepts are spanned by relations of different strengths, derived from the clinical contexts in which concepts are used. The strength of the relations between qualified medical concepts is computed using MapReduce, generating a novel form of big mechanism for big data.

Biography: Professor Harabagiu joined the Computer Science Department at U.T. Dallas in January 2002, coming from U.T. Austin, where she was a faculty member in the Department of Computer Sciences. Her research interests include Natural Language Processing, Information Retrieval, Knowledge Processing, Artificial Intelligence and more recently Medical Informatics. She has been interested for a long time in Textual Question-Answering, reference resolution and textual cohesion and coherence. In 2006 she co-edited a book entitled Advances in Open Domain Question Answering. She has also co-organized several workshops, symposia, tutorials and research tracks focusing on Textual Question-Answering or Reference Resolution. In her research she combines knowledge extracted from the World Wide Web with knowledge coerced from large lexical databases (e.g. WordNet or FrameNet) to be able to model the semantics of language in texts. Semantic interpretations made possible by lexical data inform various textual inference tasks that she is studying. Examples of forms of textual inference that interest her are: Temporal Inference, Causal Inference, Spatial Inference. She is particularly fascinated by the textual inference that can be performed in complex domains, e.g. inference of concepts and relations in Electronic Medical Records, coreference, spatial and temporal grounding in different types of clinical documents.

She is a past recipient of an NSF CAREER Grant for studying reference resolution. She is also a member of AAAI, ACL, ACM SIGIR, IEEE Computer Society. She is the founder and Director of the Human Language Technology Research Institute at University of Texas at Dallas.