The Use of Natural Language Processing to Transform Health Records Information

A. Roberts[1]

[1]Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

Electronic Health Records (EHRs) have opened up the possibility of data re-use, to support both clinical practice, and clinical research. In order to achieve this re-use, we have to address the issue that EHRs contain both structured and unstructured data. By structured data, we mean formally described and organised data, such as that entered via forms, prescriptions, and lab results. By unstructured data, we mean the free text, narrative portion of the record such as patient encounter notes, and correspondence. It is estimated that 80% of the information content of the record is in this unstructured portion. Free text enables the clinician to express complex concepts, events, and uncertainties, in a way that is not possible in the structured record. Unlike structured data, however, free text creates difficulty for re-use and for statistical analysis. These difficulties include the incompleteness of the free text record, the inherent ambiguity of natural language, the noisy nature of the data, dependence on the context in which the record was written, and the author's assumptions about the target audience. Natural Language Processing (NLP), the computerised processing of human language, provides tools and processes to tackle some of these difficulties, and is increasingly used to extract clinically significant information from the textual component of the EHR. We examine ways in which NLP has been used to this end. We review some of the NLP systems and methods used over clinical records, focusing on the techniques of Classification and Information Extraction (IE), and using examples of NLP in use on the South London and Maudsley mental health case register. We will consider some of the NLP tools and resources available to health informaticians, their benefits and their costs.