Title
Hierarchical semantic structures for medical NLP.

Authors
Taira, Ricky K
Arnold, Corey W

Publication Date
2013

DOI
10.3233/978-1-61499-289-9-1194

Peer reviewed

MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-1194

Hierarchical Semantic Structures for Medical NLP

Ricky K. Taira and Corey W. Arnold

Department of Radiological Sciences
University of California, Los Angeles, USA

Abstract and Objective

We present a framework for building a medical natural language processing (NLP) system capable of deep understanding of clinical text reports. The framework helps developers understand how various NLP-related efforts and knowledge sources can be integrated. The aspects considered include: 1) computational issues dealing with defining layers of intermediate semantic structures to reduce the dimensionality of the NLP problem; 2) algorithmic issues in which we survey the NLP literature and discuss state-of-the-art procedures used to map between various levels of the hierarchy; and 3) implementation issues to software developers with available resources. The objective of this poster is to educate readers to the various levels of semantic representation (e.g., word level concepts, ontological concepts, logical relations, logical frames, discourse structures, etc.). The poster presents an architecture for which diverse efforts and resources in medical NLP can be integrated in a principled way.

Keywords:

Natural language processing, knowledge representation, structured reporting

Introduction

Medical natural language processing (NLP) implies the development of mathematical models of languages and the study of the transformations that map clinical texts into well-formed logical representations that preserve meaning. A direct mapping from text to logical interpretation is unreasonable given the large variability of clinical text reports and the large state space associated with the universe of all possible logical interpretations. To deal with these difficulties, it is typical to introduce layers of intermediate structure to be imposed onto the input free text data. This significantly reduces the dimensionality of the mapping problem. Thus a more efficient means of encoding the meaning of medical free-text is to organize meaning in a semantic hierarchy, and in particular in the form of a generative model.

Methods

This work presents efforts at critically reviewing the medical and general NLP literature and proposing a framework for understanding how diverse theories, resources, and algorithms fit into the overall NLP problem of translating an input character string into a logical interpretation. We define various transformational levels of abstraction including: 1) surface words to functional words (e.g., collocations and idioms); 2) functional words to syntactic and word-level semantic equivalence classes; 3) word feature sequences to ontologic sequences (i.e., concept coding); 4) ontologic sequences to a set of ontologic propositions; 5) ontologic propositions to ontologic objects/events; 6) ontological objects onto a spatial layout; 7) ontologic events onto a logical timeline; 8) ontologic objects/events to association graphs; 9) ontologic objects/events to discourse templates (e.g., patient discharge information).

For each level of mappings, we list associated NLP subproblems such as tokenization, hyphenation rules, part-of-speech tagging, semantic chunking, ontologic coding, syntactic parsing, semantic selection restriction, thematic role assignment, coreference resolution, and logical entailment.

For subproblems within each semantic transformational layer we present references to various knowledge sources and models that can assist in the mapping task (e.g., specific lexicons, word sequence models, topic models, semantic patterns, domain specific grammars, object appearance models, etc. General computational and representational methods are also referenced for each transformational layer (e.g., spelling correction algorithms, context-free grammars, clustering algorithms for word sense disambiguation, Hidden Markov models, stochastic entity-attribute models; stochastic block models, recursive logical frames, object-oriented models, temporal evolutionary object models, etc.). Various theories and global optimization algorithms imposed on any combination of semantic levels are presented to demonstrate how such optimization across semantic levels can be achieved. For example, topic priming of higher conceptual representations can assist in disambiguation at lower local levels. Optimization algorithms such as concept spreading, multi-pass sieves, and back propagation algorithms are referenced. Finally, system architectures such as the UIMA pipeline are presented for system integration and efficient processing.

Results

This poster summarizes how to view the medical NLP problem in terms of manageable sub-problems. We identify intermediate semantic constituents and methods, resources, and models for instantiating the transformation between hierarchical semantic layers for a given input text.

Conclusion

The medical NLP community can benefit from a theoretical blueprint for coordinating its diverse and mostly fragmented efforts. This work presents an architecture for addressing the medical NLP problem which may potentially accelerate the development of highly accurate NLP tools that sufficiently scale and which have satisfactory processing speed.