Abstract

The ever-increasing effect of information overload requires humans to be extremely selective in what they read and store as knowledge. Much more can be made human accessible if such textual content can be encoded into forms which lend themselves more readily to inferencing. This gap of converting natural language textual content to machine-processable form requires the extraction of surface knowledge and induction of ontological knowledge.

In this research, semi-automated approaches are proposed to first extract surface knowledge from open domain Sri Lankan English news content, and then to develop a technique to induce ontological knowledge from these through automatic means based on the exploitation of a novel combination of redundancy computations, syntactic collocation, term clustering and intra-cluster permutations.

The approach introduced for surface knowledge extraction needs only a single pass over the corpus, and is thus, unique, and fast. From the open domain corpus of 54,201 sentences the surface knowledge extraction algorithm was able to return 45,258 (83.5%) meaningful knowledge triples. Moreover, the algorithm generalized to an independently compiled BBC news corpus by returning 16,838 meaningful triples from 18,184 sentences (92.6%).

The proposed ontological knowledge generation model first converts these surface knowledge triples into Abstract Semantic Patterns (candidate ontological knowledge) by employing regular semantic abstraction, named entity recognition based abstraction and hypernym based abstraction. Based on these patterns, valid ontological knowledge was induced using novel mechanisms. The Joined Directed Syntactic Collocation method inferred 95,491 ontological knowledge facts of which 56% were estimated to be effective. The Cluster Permutations method inferred 127,874 ontological knowledge of which 66% were estimated to be effective.

The main contribution of this research includes novel approaches for automatic knowledge extraction from open domain unstructured text. Thus, it facilitates to enhance the research domain of knowledge extraction. On a practical level, this research contributes towards extracting knowledge embedded in news to be converted to machine interpretable knowledge forms that would benefit mankind.