World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:9, No:02, 2015

Towards a Large Scale Deep Semantically Analyzed Corpus for Arabic: Annotation and Evaluation

Authors : S. Alansary, M. Nagi

Abstract : This paper presents an approach of conducting semantic annotation of Arabic corpus using the Universal Networking Language (UNL) framework. UNL is intended to be a promising strategy for providing a large collection of semantically annotated texts with formal, deep semantics rather than shallow. The result would constitute a semantic resource (semantic graphs) that is editable and that integrates various phenomena, including predicate-argument structure, scope, tense, thematic roles and rhetorical relations, into a single semantic formalism for knowledge representation. The paper will also present the Interactive Analysis tool for automatic semantic annotation (IAN). In addition, the cornerstone of the proposed methodology which are the disambiguation and transformation rules, will be presented. Semantic annotation using UNL has been applied to a corpus of 20,000 Arabic sentences representing the most frequent structures in the Arabic Wikipedia. The representation, at different linguistic levels was illustrated starting from the morphological level passing through the syntactic level till the semantic representation is reached. The output has been evaluated using the F-measure. It is 90% accurate. This demonstrates how powerful the formal environment is, as it enables intelligent text processing and search.

Keywords : semantic analysis, semantic annotation, Arabic, universal networking language
Conference Title : ICEMNLP 2015 : International Conference on Empirical Methods in Natural Language Processing
Conference Location : Kuala Lumpur, Malaysia
Conference Dates : February 12-13, 2015