

Research Notes

Nicholas Monath, Niklas Schulze, Klim Zaporozjets

March 5, 2014

Please be sure to provide links to the sources you take notes from either in the form of a link to the webpage or as a Bibtex citation

Possible Future Points of Research/Papers to Read:

- *Semantic Measures for the Comparison of Units of Language, Concepts or Instances from Text and Knowledge Representation Analysis* by Harispe, Ranwez, Janaqi, Montmain
- *Introduction to Information Retrieval* by Christopher Manning
- *Semantic distance in WordNet: An experimental, application-oriented evaluation of ?ve measures* by Budanitsky and Hirst

1 Semantic Similarity

- **Semantic measures:** mathematical tools used to estimate the strength of the semantic relationship between units of language, concepts or instances, through a numerical description obtained according to the comparison of information formally or implicitly supporting their meaning or describing their nature
- **Semantic similarity:** measures the likeness of terms, words, documents (or any objects which can be characterized through semantics). The likeness of compared objects is based on their meaning or semantic content, as opposed to similarity which can be estimated regarding their syntactical representation (e.g. their string format).
- An **ontology** formally represents knowledge as a set of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts
- **Semantic similarity** can be estimated for instance by defining a topological similarity, by using **ontologies** to define a distance between terms/concepts
 - A naive metric for the comparison of concepts ordered in a partially ordered set and represented as nodes of a directed acyclic graph (e.g., a taxonomy), would be the minimal distance in terms of edges composing the shortest-path linking the two concept nodes. Based on text analyses, semantic relatedness/distance between units of language (e.g., words, sentences) can also be estimated using statistical means such as a vector space model to correlate words and textual contexts from a suitable text corpus (co-occurrence).
- Note the difference between semantic *similarity* and semantic *antonymy* (how *unrelated* things are) and semantic **meronymy**
 - A **meronym** denotes a constituent part of, or a member of something. For example, “finger” is a meronym of “hand” because a finger is part of a hand. Similarly, “wheels” is a meronym of “automobile”.

1.1 Measures

- Two main approaches to measuring the similarity of ontological concepts: **edge-based** and **node-based**
 - Edge-based: which use the edges and their types as the data source
 - Node-based: in which the main data sources are the nodes and their properties.
- Other measures calculate the similarity between *ontological instances*:
 - Pairwise: measure functional similarity between two instances by combining the semantic similarities of the concepts they represent
 - Groupwise: calculate the similarity directly not combining the semantic similarities of the concepts they represent
- There are also a number of statical similarity approaches such as: Latent semantic analysis, Pointwise mutual information, etc. (see article for more information)

2 Information Retrieval by Christopher Manning

- **Information retrieval** (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
- **Unstructured data:** refers to data which does not have clear, semantically overt, easy-for-a-computer structure¹

¹Note how most text we would want to search is *semistructured*, that is it has tags such as a title or headings etc