The GMRKB.com Semantic Wiki (2019)

Gabor Melli <gmelli@acm.com>, **Olga Moreira** <olga.moreira@gmail.com>

ABSTRACT

We introduce *GM-RKB*, a linguistically-rich online semantic wiki focusing on concepts and text from the scientific literature on machine learning and related fields of computing, statistics, mathematics, and physics. It systematically describes thousands of concepts for many machine learning-related tasks, systems and algorithms, along with their input/output data type and requirements. Further the text from thousands of scientific publications have their concept mentions semantically annotated and thus linked to concept entries in the wiki. To the best our knowledge, this interlinked ontology-corpus is one of the few scientific linguistically-rich semantic wiki resources freely available to the research community.

Text annotation: Most of the approximately 4,500 publication abstracts and content quotes were annotated using the following two-step process:

1. the SDOI system's mention recognizer (Melli, 2012) automatically pre-annotates abstracts by applying a trained *conditional random field (CRF)*-based chunker;
2. the authors review each abstract to remove remaining editing errors and to add domain-specific repairs (this step takes approximately 1 minute per abstract).

This concept mention interlinking is accomplished using the popular annotation format used in Wikipedia[1] and is a continuation of the work started in (Melli, 2010)..

Semantic Wiki Structure: To support both human and machine readability, the concept pages are represented with the use of a controlled English vocabulary and the structure proposed in (Melli & McQuinn, 2008), where each concept page contains: (1) a unique preferred name; (2) a definitional sentence of the form of "*X is a type of Y that ...*"; (3) words that are commonly synonymous with the concept; (4) a context that contains statements relating the concept to other concepts in the ontology; (5) examples and counter-examples of instances of the concept; (6) a set of related concepts whose relationship has not been formally defined; (7) when possible, references including helpful quoted text from external resources, published research papers, Wikipedia articles and other web-accessible resources such as "*The Encyclopedia of Machine Learning*" (Sammut & Webb, 2011).

Comparison with other semantic resources: When compared to other lexically-rich domain-specific semantic resources such as the Gene Ontology and the MeSH controlled vocabulary, GMRKB contains more mid-level concepts and semantic relationships such as "*textual data*" and "*minimal biclique set cover problem*" but relatively fewer named entity mentions. The main reason is that mentions in these resources are linked to specific entity instances (e.g. proper names, molecules, and organisms) while in GMRKB, when named entities are encountered, they often related to researchers such as "*Gibbs*" or "*Markov*" and usually embedded within a concept-phrase such as "*Gibbs sampling method*" or "*Hidden Markov model*". Table 1 summarizes the key statistics of the wiki.

Concept pages	20,634		
Internal links	122,913		
	Min	Median	Max
Links into a concept	0	3	364
Links out of a concept	3	5	535
Synonyms per concept	0	1	8

Table 1 – Summary statistics of the current wiki

Future Applications: This resource has several possible future applications as it enables direct temporal modeling of topic trends that arise in the field, making it possible for it to be used in the evaluation process of terminology mining systems and the semi-automated annotation of published research (Melli & Ester, 2010). As a long-term goal, we aim to support digital libraries to help researchers navigate scientific literature at a semantic level, and formally align the ontology to existing resources such as the SUMO top-level ontology and the OntoDM domain-specific ontology (Panov et al., 2014).

References

G. Melli. (2012). "Identifying Untyped Relation Mentions in a Corpus Given An Ontology." In: Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing.

G. Melli, and M. Ester. (2010). "Supervised Identification of Concept Mentions and their Linking to an Ontology"

G. Melli. (2010). "Concept Mentions within KDD-2009 Abstracts (kdd09cma1) Linked to a KDD Ontology (kddo1)." In: LREC-2010.

G. Melli, and J. McQuinn. (2008). "Requirements Specification Using Fact-Oriented Modeling: A Case Study and Generalization." In: ORM-2008.

P. Panov, L. Soldatova, and S. Dzeroski. (2014). "Ontology of Core Data Mining Entities." In: Data Mining and Knowledge Discovery Journal, 28(5-6).

C. Sammut, and G. I. Webb (2011). "Encyclopedia of Machine Learning." Springer. ISBN:0387307680

[1] http://wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking