A Reflection on “YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia”, by Suchanek et al.

About: This paper claims to create a useful ontology that is both orders of magnitude larger than hand-curated ontologies such as WordNet and is significantly more accurate than other automatically curated ontologies. As such, it’s contributions can generally be thought of as consisting of two parts - the development of an ontological system that is more expressive and decidable than previous ontologies, and the method for populating that ontological system with facts. The ontological system is based on the (argument, relation, argument) triple, with reification of instantiations of such a triple as its own entity. The author freely admits that it is similar to RDFS, although it possesses some additional expressivity and semantics via rewrite rules which add to the ontology, and claims that this is actually an advantage. The author also compares the ontology to OWL, and claims it is almost as expressive while being decidable. The population of the ontology is done by examining and doing basic processing on Wikipedia categories of individual pages, creating triples from the combinations of categories and pages, and guiding this automatic process from knowledge already in WordNet.

My Thoughts: While I am aware of the significance of this paper’s contributions and am sure that some real-world applications exist, I feel very strongly that calling this ontology “a core of semantic knowledge” is extremely misleading. There are only 14 relationship types in the ontology, and these include “HasWonPrize”, “BornInYear”, “DiedInYear,” and other extremely specific details. I would argue that knowing that Einstein won the Nobel prize in 1921, while useful to an application such as google’s search engine, is far from being a “core of semantic knowledge.” For instance, in the entire paper itself, apart from the examples given to illustrate the ontology, and perhaps biographical information about authors in the bibliography a very small percentage of readers might care about, there was not a single instance where I needed to know that someone was born in a certain year to understand the paper. In contrast, consider a randomly chosen sentence from the paper: “Preprocessing ensures that words in the query are considered in all their possible meanings.” To truly understand this sentence, I’d need to understand that preprocessing is an algorithmic task that may be one of the contributions of the paper, and that when words are “considered in all their possible meanings,” the assumed meaning will be an element or subset of the set of meanings consisting of “all their possible meanings.” I’d be willing to call SUMO or perhaps even ConceptNet an attempt at “a core of semantic knowledge”, but giving this title to a list of dates seems somewhat absurd.

In addition to this major objection, I had a few other questions regarding the paper. One considered the accuracy claims. The authors claim an increase from 90% accuracy to 97% accuracy over other automatically constructed ontologies. This is certainly impressive and worthy of publication; however, it is very difficult to see what exactly it means for the end user without some idea of the relationship between accuracy of ontologies and accuracy of an end task it was designed for. Is there system 7% more effective in a certain task, or 2^1.07 more effective? Thus, it would have been nice if they chose a real-world task to demonstrate the dominance of their method (although I know this is too much to realistically expect of a single paper). Finally, I wonder how their (argument, relation, argument) would handle ditransitive verbs.

Overall, this paper was well written, and clearly advanced the landscape of ontologies based on information extraction. My only main concern, that such tasks represent a very small step towards commonsense reasoning, does not diminish the fact that, for Google search engine or other end-user tasks, this paper seems to improve the state-of-the-art.