A Hitchhiker's Guide to Ontology

Fabian M. Suchanek

Télécom Paris & Institut Polytechnique de Paris, France

A knowledge base (KB) is a computer-processable collection of knowledge about the world. In its simplest variant, a KB takes the form of a labeled graph, where the nodes are entities (such as people, organizations, and geographical locations), and the edges represent the links between these entities in the real world (such as who was born where, which organization is headed by whom, which city is the capital of which country etc.). The formal definition of the categories and the relations of a KB is called an ontology[1]. Knowledge bases provide the background knowledge for different artificial intelligence applications, ranging from personal assistants to Web search, question answering, and text analysis. In particular, KBs are useful in information retrieval (IR), where they serve for structured search and entity disambiguation. Research has made extraordinary progress in the automated construction of KBs, and today's KBs contain billions of entities [1]. Nevertheless, KBs are still far from perfect. In this keynote talk, I outline several challenges in the construction and maintenance of KBs, and show how our research group approached them.

Construction of KBs. KBs used to be built by hand. In 2008, our YAGO knowledge base [2] was one of the first large knowledge bases that were constructed automatically. While the first version of YAGO fed from Wikipedia, the newest version, YAGO 4 [3], feeds from Wikidata. YAGO 4 cleans up the taxonomy of Wikidata (by replacing it by the one from schema.org), gives entities and relations readable identifiers, and applies schema constraints. This cleans the data, and an OWL DL reasoner can actually run on this KB. We have also ventured beyond Wikidata and Wikipedia, by extracting commercial products from the Web [4]. Our work harvests universal product codes from the Web and builds a shallow KB on top. In such scenarios, one often encounters the problem of entity linking: Given the mention of an entity on a Web page, map it to any of the predefined entities from a catalog. We found that this problem can be solved by a rather lightweight neural architecture [5], which works just as well as heavier solutions such as transformers.

Completion of KBs. KBs are usually highly incomplete. We have worked on this problem along several axes: Our AMIE system [10] can find rules such as *If two people are married, they usually live in the same city* [11]. Such rules can then be used to predict missing facts. We have also developed methods to determine whether a fact is missing in the first place [12]. Another work [13] can determine whether an attribute (such as *hasParent*) appears with all entities of a class (say, *Person*) in the real world – even if it does not in the KB. Finally, we have developed methods to estimate how many entities of a class are missing [14].

Querying KBs. KBs can be pretty large, and thus they are usually loaded into a triple store (database) in order to query them. However, for large KBs, even this loading can take hours, and if we want to launch only a single query, the loading is an overhead. We have developed an approach that transforms a Datalog or SPARQL query into bash commands [15]. These can be executed directly on the files of the KB (in TSV format), thus bypassing the triple store completely. Another work [16] is concerned with querying KBs that can be accessed only by predefined functions. We show that for a certain class of functions and queries, it is decidable whether a query can be answered by an orchestration of these functions.

Applying KBs. We have applied KBs for the purposes of combinatorial creativity [17] and to the digital humanities [18]. With the help of YAGO, we can, e.g., trace the life expectancy over the centuries, drilled down by gender or country of birth. In an attempt to bring semantic understanding to a very different domain, we look into explaining the decisions of a black box machine learning model by help of several decision trees [19].

Extension of KBs. KBs usually contain mainly binary links between entities – a knowledge representation known as *RDF*. In our *NoRDF project* [6], we aim to enrich KBs by beliefs, claims, events, causes, and entire stories. We have so far mainly surveyed the existing literature: how to deal with non-named entities [7], how to deal with vague expressions [8], and how to assess whether transformers can reason on natural language [9].

Conclusion. Knowledge Bases are a fascinating and useful domain of research. Many challenges have been overcome recently, and many new ones are awaiting us.

Acknowledgments

Partially funded by ANR-20-CHIA-0012-01 ("NoRDF").

DESIRES 2021 – 2nd International Conference on Design of Experimental Search & Information REtrieval Systems, September 15–18, 2021, Padua, Italy

suchanek@telecom-paris.fr (F. M. Suchanek)
https://suchanek.name (F. M. Suchanek)

[1] The title uses this word instead of "knowledge base" to rhyme with a book title by Douglas Adams [42].

References

[1] G. Weikum, L. Dong, S. Razniewski, F. M. Suchanek, Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases, in: Foundations and Trends in Databases, 2021.

[2] F. M. Suchanek, G. Kasneci, G. Weikum, Yago - A Core of Semantic Knowledge , in: WWW, 2007.

[3] T. P. Tanon, G. Weikum, F. M. Suchanek, YAGO 4: A Reason-able Knowledge Base , in: ESWC, 2020.

[4] A. Talaika, J. A. Biega, A. Amarilli, F. M. Suchanek, IBEX: Harvesting Entities from the Web Using Unique Identifiers , in: WebDB workshop, 2015.

[5] L. Chen, G. Varoquaux, F. M. Suchanek, A Lightweight Neural Model for Biomedical Entity Linking, in: AAAI, 2021.

[6] F. M. Suchanek, The Need to Move Beyond Triples , in: Text2Story workshop, 2020.

[7] P.-H. Paris, F. M. Suchanek, Non-named entities - the silent majority, in: ESWC short paper track, 2021.

[8] P.-H. Paris, S. E. Aoud, F. M. Suchanek, The Vagueness of Vagueness in Noun Phrases, in: AKBC short paper track, 2021.

[9] C. Helwe, C. Clavel, F. M. Suchanek, Reasoning with Transformer-based Models: Deep Learning, but Shallow Reasoning, in: AKBC short paper track, 2021.

[10] L. Galárraga, C. Teflioudi, K. Hose, F. M. Suchanek, AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases , in: WWW, 2013.

[11] F. M. Suchanek, J. Lajus, A. Boschin, G. Weikum, Knowledge Representation and Rule Mining in Entity-Centric Knowledge Bases , in: RW invited paper, 2019.

[12] L. Galárraga, S. Razniewski, A. Amarilli, F. M. Suchanek, Predicting Completeness in Knowledge Bases , in: WSDM, 2017.

[13] J. Lajus, F. M. Suchanek, Are All People Married? Determining Obligatory Attributes in Knowledge Bases , in: WWW, 2018.

[14] A. Soulet, A. Giacometti, B. Markhoff, F. M. Suchanek, Representativeness of Knowledge Bases with the Generalized Benford's Law, in: ISWC, 2018.

[15] T. Rebele, T. P. Tanon, F. M. Suchanek, Bash Datalog: Answering Datalog Queries with Unix Shell Commands, in: ISWC, 2018.

[16] J. Romero, N. Preda, A. Amarilli, F. M. Suchanek, Equivalent Rewritings on Path Views with Binding Patterns, in: ESWC, 2020.

[17] F. M. Suchanek, C. Menard, M. Bienvenu, C. Chapellier, Can you imagine... a language for combinatorial creativity? , in: ISWC, 2016.

[18] T. Rebele, A. Nekoei, F. M. Suchanek, Using YAGO for the Humanities , in: WHISE workshop, 2017.

[19] N. Radulović, A. Bifet, F. M. Suchanek, Confident Interpretations of Black Box Classifiers, in: IJCNN, 2021.

[42] Douglas Adams, The Hitchhiker's Guide to the Galaxy, 1979.