Yet Another Great Ontology (YAGO)

by Ole Berg & Lennard Feuerbach

on the 5th of November, 2024

Table of Contents

1.1 Ontologies (Recap):

- Formal Representation: Structured definitions of concepts and relationships
- Shared Vocabulary: Common language for effective data exchange
- Knowledge Integration: Unifies diverse data sources for a cohesive view
- Reasoning and Inference: Enables logical reasoning to derive new knowledge

1.2 Motivation and history

- Researchers at the Max Planck Institute for Informatics in Saarbrücken saw potential of knowledge bases
- Already had various application fields
- Criticized that applications only utilized a single source of knowledge, often either WordNet or Wikipedia
- Laid out what features and traits a heavily improved application should have:
  - High accuracy (close to 100 %)
  - Combination of concepts and entities
  - Extensible, easily re-usable, and application independent

1.3 Yet Another Great Ontology (YAGO):

- Knowledge Base: Contains well over millions of entities and facts from real world
- Entities and Relations: Includes entities and their relationships
  - e.g. who played in which movie, which city is located in which country
- Class Taxonomy: Organizes entities in different hierarchical classes
  - e.g. class of cities is a subclass of the class of populated places
- Defined Relations: Specifies relations between entities, forming the ontology
  - e.g. birthPlace can only be between person and a place

1.3 Yet Another Great Ontology (YAGO):

- Wikidata's vast facts are combined with schema.org's ontology
- All identifiers are human-readable and comparable easy to understand
- Logical constraints maintain data quality and enables reasoning over the data
- Simpler version of Wikidata, offering a cleaner knowledge base

2.1 YAGO 1

- First version of YAGO, created in 2007
- Unifies facts from Wikipedia with clean taxonomy from WordNet
- Categories from Wikipedia are brought together with nouns from WordNet
- 2 million entities and 20 million facts

2.2 YAGO2

- Improvement in 2012
- "anchoring [...] along both the spatial and the temporal dimensions"
- *When and where was a fact true?*
- Introduction of *SPOTL(X)* for representing and querying "spatio-temporally enhanced facts"
- Time (or time frame) for entities and facts
- Single geo–coordinate pair for *geo-entities*
- 9.8 million entities and 447 million facts
- Coverage: Time 47 %, location 30 %

2.3 YAGO2s

- Refactoring of YAGO2 in 2013
- Transparent and modular architecture
- Compliance of YAGO syntax with RDF
- Usage of Turtle format
- Introduction of domains from WordNet

2.4 YAGO3

- Improvement in 2015
- Expansion into Wikipedias in other languages
- Use of Wikidata to identify same entities in different-language Wikipedias
- Fact extraction from infoboxes in Wikipedia
- 17 million entities and more than 150 million facts

2.5 YAGO4

- Complete overhaul in 2020
- WordNet is abandoned, replaced by Schema.org
- Focus on Wikidata instead of Wikipedia infoboxes
- Browser and SPARQL endpoint are available
- Upper classes from Schema.org and selected lower classes from Wikidata
- 67 million entities and 343 million facts (Full YAGO)

2.6 YAGO4.5

- Addresses issues of YAGO4
- Paper has been published this year (2024)
- Important principles:
  - Upper taxonomy to define properties
  - Lower taxonomy to convey information in human-readable form
- 49 million entities and 132 million facts

3. Applications

3.1 Various examples

- NER: Enhances entity recognition with structured data
- Semantic Web: Builds knowledge graphs for data linking
- AI Chatbots: Improves chatbot responses with factual data
- Knowledge Graphs: Maps complex entity relationships
- NLP: Supports semantic analysis and word disambiguation

3.2 IBM Watson

- The most known application of YAGO is IBM Watson
- IBM Watson is a system capable of answering questions
- Developed by at IBM, team led by David Ferucci, released in 2011
- Designed to answer questions from the TV show Jeopardy
- IBM Watson actually beat the real champions and won

3.2 IBM Watson

- Created as a question-answering system using NLP, IR and other technologies
- Written in languages including Java, C++ and Prolog, and runs on Linux servers
- The system Watson uses has 2,880 processor threads and 16 TB of RAM
- The data Watson uses includes encyclopaedias, dictionaries, articles and more
- Also databases, taxonomies and ontologies as DBPedia, WordNet and **YAGO**

4. YAGO and SPARQL

```
PREFIX schema: <http://schema.org/>
PREFIX yago: <http://yago-knowledge.org/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?movie ?duration WHERE {
  ?movie rdf:type schema:Movie .
  ?movie schema:duration ?duration .
  FILTER (xsd:decimal(?duration) >= 100)
}
LIMIT 3
```

	movie	duration
1	yago:The_Way_of_the_Dragon	"100"^^xsd:decimal
2	yago:Carrie__u0028_2013_film_u0029_	"100"^^xsd:decimal
3	yago:The_Profession_of_Arms__u0028_2001_film_u0029_	"100"^^xsd:decimal

Figure 1: Result set of the SPARQL query

Try it for yourself: *https://yago-knowledge.org/sparql*

5. Comparison

- YAGO vs. DBpedia: Fixed schema, non-redundant relations, 50m vs. 4m instances
- YAGO vs. ConceptNet: Focus on specific instances, not common sense knowledge
- YAGO vs. BabelNet: Structured taxonomy and defined schema
- YAGO vs. Freebase: Actively maintained with ongoing updates
- YAGO vs. Wikidata: Readable identifiers, clean taxonomy, logical constraints

6. Conclusion

- **Combines Strengths**: Merges Wikidata's facts with schema.org's ontology
- **User-Friendly**: Has human-readable identifiers and well-organized classifications
- **Streamlined Data**: Includes only well-populated, essential classes and properties
- **Logical Constraints**: Ensures data quality and supports advanced reasoning

7. References

- F. M. Suchanek, G. Kasneci, und G. Weikum, “YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia”, 2007.
- J. Hoffart, F. M. Suchanek, K. Berberich, und G. Weikum, “YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia”, Artificial Intelligence, Bd. 194, S. 28-61, Jan. 2013, doi: 10.1016/j.artint.2012.06.001.
- J. Biega, E. Kuzey, und F. M. Suchanek, “Inside YAGO2s: a transparent information extraction architecture”, in Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro Brazil: ACM, Mai 2013, S. 325-328. doi: 10.1145/2487788.2487935.
- F. Mahdisoltani, J. Biega, und F. M. Suchanek, “YAGO3: A Knowledge Base from Multilingual Wikipedias”.

7. References (cont.)

- T. Pellissier Tanon, G. Weikum, und F. Suchanek, “YAGO 4: A Reason-able Knowledge Base”, in The Semantic Web, Bd. 12123, A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, und M. Cochez, Hrsg., in Lecture Notes in Computer Science, vol. 12123. , Cham: Springer International Publishing, 2020, S. 583-596. doi: 10.1007/978-3-030-49461-2_34.
- F. M. Suchanek, M. Alam, T. Bonald, L. Chen, P.-H. Paris, und J. Soria, “YAGO 4.5: A Large and Clean Knowledge Base with a Rich Taxonomy”, in Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington DC USA: ACM, Juli 2024, S. 131-140. doi: 10.1145/3626772.3657876.

7. References (cont.)

- Ferrucci, David & Brown, Eric & Chu-Carroll, Jennifer & Fan, James & Gondek, David & Kalyanpur, Aditya & Lally, Adam & Murdock, J William & Nyberg, Eric & Prager, John & Schlaefer, Nico & Welty, Christopher. (2010). Building Watson: An Overview of the DeepQA Project. AI Magazine. 31. 59-79. 10.1609/aimag.v31i3.2303.
- "YAGO 3.0". Zugriff am 01.11.2024 [Online]. Link: https://github.com/yago-naga/yago3
- "YAGO 4.5". Zugriff am 01.11.2024 [Online]. Link: https://github.com/yago-naga/yago-4.5
- "Getting Started". Zugriff am 01.11.2024 [Online]. Link: https://yago-knowledge.org/getting-started