CS410 Tech Review

jnisaac2

Introduction

The creation of the internet was a major advance in the accessibility of knowledge. At no other point in human history has such a wealth of information been publicly available, yet due to its massive size, it will never be fully utilized without the help of tools that can parse and interpret it. Some of these tools are knowledge bases, which extract and store information in complex structures that allow for meaningful connections between data to be identified and saved. Early knowledge bases relied on human effort to organize and structure data, but the inaccuracies and inefficiencies of this strategy have made automatic methods of construction a more appealing solution. This review will discuss three broad approaches to automatic knowledge base construction, exemplified by the knowledge bases YAGO, OLLIE, and NELL. Google Knowledge Vault, a recent project, takes certain aspects from these approaches for its objective of building a probabilistic knowledge base that automatically processes data from the entire internet.

YAGO

YAGO, or Yet Another Great Ontology, is a knowledge base created using Wikipedia for facts and WordNet for semantic relationships. By uniting both resources, YAGO can recognize entities and create relations between them on scale much greater than WordNet alone. It works by using the category pages of Wikipedia, which congregate all pages about entities that belong to a certain category, to intuit the relationships between any two entities [1]. A relationship along with the two entities it describes comprise a fact. In the YAGO model, words, instances of classes, relations, and facts are all treated as entities, which allows for complex relations between any number of entities to form. YAGO, and other similar knowledge bases like DBpedia or Freebase, rely on structured data such as WordNet relationships as sources, which is one of the broad approaches to automatic knowledge base construction.

OLLIE

OLLIE, or Open Language Learning for Information Extraction, is a knowledge base that draws from the entire internet. In contrast to YAGO's structured approach, OLLIE makes use of open information extraction techniques on the unstructured data of the web. A major process of the knowledge base is converting sentences into tuples of argument phrases linked by a relation phrase that determines the relationship between two argument phrases [2]. These tuples are used to gather from the internet OLLIE's training set, sentences that contain all the words that occur in the tuple. Next, OLLIE learns open pattern templates by analyzing the structure of relevant sentences. These templates represent the variety of ways that relations can be encoded in text and can be used to analyze new sentences that haven't been encountered yet. A broad and flexible set of techniques assist OLLIE's and other similar knowledge bases' approach of analyzing unstructured data.

NELL

NELL, or Never-Ending Language Learning, is a knowledge base that performs two tasks, running constantly without stopping. The first task is reading, or extracting information from the web to gather knowledge, and the second task is to ensure its reading ability is strictly increasing compared to the previous day. NELL represents information in terms of categories that noun phrases belong to and relations between noun phrases. Starting with a seed ontology of a small number of categories and relations, it collects information through four main subsystems. First, the *Coupled Pattern Learner* recognizes phrase patterns from which information can be extracted, such as "X was born in Y." Second, the *Coupled SEAL* searches the internet with facts that NELL believes to learn novel ways that the facts can be represented linguistically. Third, the *Coupled Morphological Classifier* uses logistic regression to classify noun phrases by their features, and promotes candidate facts, relations which NELL is unsure of, to beliefs. Fourth, the *Rule Learner* learns probabilistic Horn clauses, rules that can be used to extrapolate new relations from other learned relations [3]. NELL, and other knowledge bases such as PROSPERA and DeepDive, share the approach of extracting information from the internet by starting with fixed ontologies or schema [4].

Google Knowledge Vault

Google Knowledge Vault is one of the largest knowledge bases in the world, constructed automatically from the internet. It stores information in RDF triples, each containing a subject, predicate, and object. Like YAGO and NELL, predicates and entity types are taken from fixed ontologies. It differs from all other knowledge bases, however, in its approach of combining noisy unstructured data from the internet with the structured knowledge of other knowledge bases, which allows it to overcome errors in both sources [4]. There are three main tasks of Knowledge Vault: extracting triples from the web, learning prior probabilities of every triple based on knowledge from another knowledge base, and combining the extracted triples and priors to determine how likely it is that a triple is correct. As with the approaches of OLLIE and NELL, a variety of sources are mined from the internet, including free text, HTML, and human-annotated pages. Knowledge Vault stores its triples in a format irrelevant to their appearance in text, reducing the occurrence of redundant information compared to other knowledge bases. Since the extracted triples may contain information that isn't present in source knowledge base, a local closed world assumption is used, which reduces the reliance on source knowledge. This unique combinatorial approach results in a knowledge base with a number of known facts an order of magnitude greater than the next largest knowledge base.

Conclusion

In conclusion, Google Knowledge Vault takes characteristics from each of three previous approaches to automatic knowledge base construction to create its own unique approach. Like knowledge bases such as YAGO, which utilize structured sources for information, Knowledge

Vault uses organized triples from another knowledge base. As with knowledge bases like OLLIE, which use open information techniques to extract data from unstructured internet text, it processes data from a variety of web sources, stripping away differences in linguistic representation to obtain the singular meaning of relations between entities. Knowledge Vault also takes aspects of knowledge bases similar to NELL, which depend on fixed ontologies to begin learning, by classifying extracted information according to a source knowledge base. Uniquely, Knowledge Vault combines these characteristics to make its own approach, a melding of guidance by structured data and completeness from vast noisy data.

References

[1] F. Suchanek, G. Kasneci, and G. Weikum. YAGO – A Core of Semantic Knowledge. In *WWW*, 2007.

[2] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni. Open language learning for information extraction. In EMNLP, 2012.

[3] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. H. Jr., and T. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[4] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 601-610, 2014.