Guest Editorial Preface

Special Issue on Web Scale Knowledge Extraction

Aditya Kalyanpur, IBM T.J. Watson Research Center, USA

James Fan, IBM T.J. Watson Research Center, USA

Christopher Welty, IBM T.J. Watson Research Center, USA

Recently, there has been a significant amount of interest in automatically creating large-scale knowledge bases from unstructured text. Workshops such as WEKEX (https://sites.google.com/site/wekex2011), AKBC (http://akbcwekex2012.wordpress.com), and WOLE (http://wole2012.eurecom.fr) organized at top international conferences have brought attention to this issue, and pulled together researchers in the inter-related fields of natural language processing (NLP), knowledge representation and reasoning (KR&R) and data mining to discuss solutions for this interesting problem.

Compared to traditional, manually created representations, auto-generated knowledge bases have the advantage of scale and coverage. They often contain tens of millions of propositions, represented using a variety of encodings, from simple binary assertions to more complicated frame-like structures, and are extracted by parsing and analyzing large Web corpora (for examples, see Etzioni, Fader, Christensen, Soderland, & Mausam, 2011; Carlson, Betteridge, Kisiel, Settles, Hruschka, & Mitchell, 2010). Once built, they can be used in a variety of AI applications such as semantic search and question answering (Fan, Kalyanpur, Gondek, & Ferrucci, 2012).

However, the problem of mining structured knowledge from unstructured text is highly non-trivial. It involves several multi-faceted challenges, such as:

- **Extraction Challenges:** Mining knowledge from large Web corpora and/or multiple sources means having to deal with varying granularity, noise and authenticity of information. Related issues include dealing with ambiguity in text; analyzing semi-structured data (e.g., lists and tables embedded in pages); processing information in multiple languages; parsing poorly written text that still might contain meaningful information; differentiating between opinions, beliefs and facts; and the ubiquitous issue of filtering out spam/faulty information.
- **Representation Challenges:** Even if one had tools to analyze unstructured information, there is the issue of how to model the extracted knowledge appropriately. A variety of KR formalisms can be chosen (e.g., assertions, rules, frames, etc.), encoded in different logics (e.g., RDF, OWL, Horn-clauses, etc.), and figuring out the most feasible representation

model for a given application/use-case can be non-trivial.

- **Retrieval Challenges:** Given the scale of the problem we are dealing with (i.e., Web-scale data), a core challenge is being able to index and retrieve the extracted information efficiently. This issue becomes more critical in applications such as online question answering that require very fast response times. Also, in several use-case scenarios, the input unstructured data is continuously changing (e.g., news feeds analysis) in which case the issue of doing efficient "incremental" updates to the underlying knowledge store becomes crucial.
- **Inference Challenges:** Often there is a need to do non-explicit inference or reasoning over extracted data to deduce new and interesting information (e.g., an application in the medical domain that does diagnosis based on patient symptoms, history etc may need to combine several facts together with domain-rules to arrive at the desired result). Techniques range from using statistical models to do fuzzy 'textual entailment', to using formal logic-based reasoning. Also, it would be interesting to combine the two forms of reasoning (i.e., statistical and logic-based) in a manner that addresses the drawbacks of each approach, however there is very little research in this regard as of yet.

The goal of this special edition is to highlight promising solutions to Web-scale knowledge extraction that address some of the core challenges mentioned above.

In "Towards Large-Scale Unsupervised Relation Extraction from the Web", Min et al. focus on the knowledge extraction problem with the WEBRE framework for large-scale unsupervised entity and relation extraction. The interesting aspect of the work is the combination of distributional-similarity based techniques with extracted hypernymy knowledge. This is followed by the use of clustering models to recognize paraphrases and semantic classes of their arguments. Compared to similar prior approaches to unsupervised relation extraction, the main advantages of WEBRE is its explicit handling of synonymy and polynymy, both of which are core issues in knowledge extraction.

The paper "Enabling Folksonomies for Knowledge Extraction: A Semantic Grounding Approach" also focuses on knowledge extraction though the problem of assigning semantics to tags in folksonomies by grounding them in Wikipedia/DBpedia entities, thereby allowing the vast number of Web 2.0 folksonomies to be used as structured knowledge resources in information extraction. The choice for Wikipedia/DBpedia is a natural one given the large amount of domain-independent encyclopedic knowledge, multilingual content and connections to other linked open data resources. A central task in this work is the disambiguation of user tags, which is realized through standard IR techniques (vector space model) that make use of the explicit disambiguation knowledge in Wikipedia.

In "Elementary: Large-Scale Knowledge-Base Construction via Machine Learning and Statistical Inference", Niu et al. discuss a comprehensive Knowledge-base Construction framework called 'Elementary' to extract knowledge from text and do statistical inference using Markov Logic Networks (MLN). There are several interesting aspects of the work such as the ability to deal with diverse information sources, extract multiple forms of features, encode different machine learning models (along with feature extractions) as MLN rules, and deal with inconsistency in the data. Additionally, the ability to represent the extracted knowledge in relational form is useful as it allows the resulting inference procedures to be Web-scalable.

The final paper, "Predicting Lexical Answer Types in Open Domain QA", Gliozzo and Kalyanpur propose an innovative framework, namely the Generalized Frame Model (GFM), to handle similarity and type abstraction in lexical knowledge bases in a principled way. The framework is defined over the PRISMATIC store, an extremely scalable lexical knowledge base that was used in the IBM Watson Question Answering system. The main contribution of the GFM work is the hybridization of LSA techniques with relational queries done using a lambda calculus. The GFM can be applied to a large variety of tasks, including learning selectional preferences (specifically, inferring types for open-domain questions as is done in IBM Watson) and textual entailment (specifically, lexical substitution and paraphrasing). This paper, co-authored by one of the guest editors, was handled by an external coordinating editor to eliminate any conflict of interest concern.

We would like to thank the reviewers for providing useful constructive feedback to the authors and for helping us decide on the final set of papers for this special edition. We believe that these four papers demonstrate novel, interesting and diverse solutions to the challenging problem of Web Scale Knowledge Extraction. They have sufficient breadth and depth to be excellent introductory papers for newcomers to the field, as well as insightful papers for experts in the area. Given the prominence and relevance of this research topic, we hope to see much more research on automated knowledge extraction in the near future.

Aditya Kalyanpur
James Fan
Christopher Welty
Guest Editors
IJSWIS

REFERENCES

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., Jr., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, Atlanta, GA.

Etzioni, O., Fader, A., Christensen, J., Soderland, S., & Mausam. (2011). Open information extraction: The second generation. In *Proceedings of the International Joint Conferences on Artificial Intelligence 2011-12*, Barcelona, Spain.