Transparent Machine Learning for Information Extraction: State-of-the-Art and the Future

Laura Chiticariu
IBM Research – Almaden
San Jose, CA, USA
chiti@us.ibm.com

Yunyao Li
IBM Research – Almaden
San Jose, CA, USA
yunyaoli@us.ibm.com

Frederick R. Reiss
IBM Research – Almaden
San Jose, CA, USA
frreiss@us.ibm.com

1 Topics and Descriptions

The rise of Big Data analytics over unstructured text has led to renewed interest in information extraction (IE). These applications need effective IE as a first step towards solving end-to-end real world problems (e.g. biology, medicine, finance, media and entertainment, etc). Much recent NLP research has focused on addressing specific IE problems using a pipeline of multiple machine learning techniques. This approach requires an analyst with the expertise to answer questions such as: "What ML techniques should I combine to solve this problem?"; "What features will be useful for the composite pipeline?"; and "Why is my model giving the wrong answer on this document?". The need for this expertise creates problems in real world applications. It is very difficult in practice to find an analyst who both understands the real world problem and has deep knowledge of applied machine learning. As a result, the real impact by current IE research does not match up to the abundant opportunities available.

In this tutorial, we introduce the concept of *transparent machine learning*. A transparent ML technique is one that:

- produces models that a typical real world user can read and understand;
- uses algorithms that a typical real world user can understand; and
- allows a real world user to adapt models to new domains.

The tutorial is aimed at IE researchers in both the academic and industry communities who are interested in developing and applying transparent ML.

Although most recent IE research focuses on techniques that are *opaque* — not transparent — there is a significant minority by various researchers from industry and academia that does not. The main goal of our tutorial is to highlight this line of work and motivate new research directions that will lead to more transparent machine learning in IE, resulting in greater impact in solving real-world problems. We categorize, compare and analyze recent advances in this area, providing a systematic view of the state-of-the-art. We cover transparent machine learning techniques for both learning features useful for information extraction, as well as learning complete extractors. We also share our experience and lessons learned on building an enterprise information extraction system integrated with multiple transparent machine learning techniques.

We start our tutorial by motivating the transparent approach with examples of emerging real-world applications. We outline the practical challenges for information extraction that these applications pose: (1) *accuracy* – the ability to generate extraction results with high precision and recall, (2) *scalability* – the ability to scale with large datasets and complex extraction tasks, (3) *usability* – the ease of building an extractor, and (4) *transparency*. Of these, accuracy and scalability have been traditionally studied in the NLP research community, with lesser emphasis being placed on usability and transparency. However, the need for *usability* and *transparency* is becoming prominent in real world applications where very often, both programmers with NLP background

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 4–6,
Lisboa, Portugal, 17-21 September 2015.

and labeled data are scarce.

We then introduce the concept of *transparent machine learning (ML)*: techniques for generating an IE model that can be easily understood and reasoned about by humans, and whose output can be "explained", thereby enabling maintainability and customizability of the extractor. We explain how transparent ML helps in addressing all four practical challenges, emphasizing on usability and transparency, and give a high-level overview of the requirements involved in enabling transparent machine learning.

Next, we present a summary of the state-of-the-art transparent machine learning techniques for IE based on literature from top NLP conferences published in the past decade. These techniques can be divided into two types: (1) those for learning features for IE (e.g. gazetteers); and (2) those for learning complete extractors. We categorize and discuss both types of techniques based on four key dimensions: (1) *models of representation* – ranging from very simple models such as dictionaries and regular expressions, to more complex models such as sequential patterns or deep parse patterns, certain classes of classifiers that are "explainable", or a combination of the above; (2) *mode of learning* – ranging from unsupervised to semi-supervised and fully supervised; (3) *learning algorithms* – including Inductive Logic Programming, Active Learning, and various other machine-learning algorithms; and (4) *incorporation of domain knowledge* – covering both the type of domain knowledge (such as seed examples or patterns) and the stages in the life cycle of the extractor when this knowledge is incorporated: either offline (at development time), or online (at runtime).

We then present a case study and show how different transparent machine learning techniques can be brought together to enable an information extraction system with end-to-end transparency. Specifically, we emphasize on the need for a standard language that is expressive enough to represent and combine different kinds of transparent models of representation. Such a standard language is a key enabler for transparent ML for multiple reasons. First, NLP researchers can use the standard as the expressivity of the output model and focus on developing and applying transparent machine learning techniques for this target language. Second, it allows the entire IE pipeline to be expressed in a unified framework, thus eliminating the need to understand multiple programming paradigms and simplifying the overall development, maintenance and debugging process. Finally, it enables the development of standard IE runtime engines to automatically optimize and scale the execution of IE programs, which further allow NLP researchers to focus on developing techniques for learning in this target language, rather than dealing with engineering issues inherent when building a pipeline of disparate components.

Finally, we conclude with an outlook of current research challenges, and promising future directions for transparent machine learning for information extraction.

2 Outline

1. Motivating Examples [30 minutes]
   - Examples of real-world applications
   - Practical challenges
     - Accuracy
     - Scalability
     - Transparency
     - Usability
2. Transparent Machine Learning for Information Extraction: Introduction [10 minutes]
3. Transparent Machine Learning for Information Extraction: State of the Art [80 minutes]
   - Learning features for information extraction [40 minutes]

   break

   - Learning extractors [40 minutes]
4. Case Study: Building an End-to-End Transparent Enterprise Information Extraction System [45 minutes]
   - Design considerations
   - Overall architecture
   - Transparent machine learning techniques
   - Live demonstration
5. Conclusion: Research challenges and future directions [15 minutes]

3 Instructors

Laura Chiticariu is a Research Staff Member at IBM Research – Almaden. She received her Ph.D from U.C. Santa Cruz in 2008. Her current research focuses on improving developmental support in information extraction systems.

Yunyao Li is a Research Staff Member and Research Manager at IBM Research – Almaden. She received her Ph.D from the University of Michigan, Ann Arbor in 2007. She is particularly interested in designing, developing and analyzing large scale systems that are usable by a wide spectrum of users. Towards this direction, her current research focuses on enterprise-scale natural language processing.

Frederick Reiss is a Research Staff Member at IBM Research – Almaden. He received his Ph.D. from U.C. Berkeley in 2006. His research focuses on improving the scalability of text analytics in enterprise applications.