Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web

Xin Luna Dong
Amazon
lunadong@amazon.com

Hannaneh Hajishirzi
University of Washington
Allen Institute for AI
hannaneh@washington.edu

Colin Lockard
University of Washington
lockardc@cs.washington.edu

Prashant Shiralkar
Amazon
shiralp@amazon.com

Abstract

How do we surface the large amount of information present in HTML documents on the Web, from news articles to Rotten Tomatoes pages to tables of sports scores? Such information can enable a variety of applications including knowledge base construction, question answering, recommendation, and more. In this tutorial, we present approaches for information extraction (IE) from Web data that can be differentiated along two key dimensions: 1) the diversity in data modality that is leveraged, e.g. text, visual, XML/HTML, and 2) the thrust to develop scalable approaches with zero to limited human supervision.

1 Description

Motivation: The World Wide Web contains vast quantities of textual information in several forms: unstructured text, template-based semi-structured webpages (which present data in key-value pairs and lists), and tables. Methods for extracting information from these sources and converting it to a structured form have been a target of research from the natural language processing (NLP), data mining, and database communities. While these researchers have largely separated extraction from web data into different problems based on the modality of the data, they have faced similar problems such as learning with limited labeled data, defining (or avoiding defining) ontologies, making use of prior knowledge, and scaling solutions to deal with the size of the Web.

In this tutorial we take a holistic view toward information extraction, exploring the commonalities in the challenges and solutions developed to address these different forms of text. We will explore the approaches targeted at unstructured text that largely rely on learning syntactic or semantic textual patterns, approaches targeted at semi-structured documents that learn to identify structural patterns in the template, and approaches targeting web tables which rely heavily on entity linking and type information.

While these different data modalities have largely been considered separately in the past, recent research has started taking a more inclusive approach toward textual extraction, in which the multiple signals offered by textual, layout, and visual clues are combined into a single extraction model made possible by new deep learning approaches. At the same time, trends within purely textual extraction have shifted toward full-document understanding rather than considering sentences as independent units. With this in mind, it is worth considering the information extraction problem as a whole to motivate solutions that harness textual semantics along with visual and semi-structured layout information. We will discuss these approaches and suggest avenues for future work.

Tutorial Content: We will start by defining unstructured, semi-structured, and tabular text, and discussing the challenges and opportunities that differentiate these data sources, as well as those they have in common. We will then provide introductions to the basic models and learning algorithms used in extraction from unstructured, semi-structured, and tabular text. We will pay special attention to methods that enable extraction to be expanded to the scope of entity and relation types found on the web, such as the distant supervision and data programming paradigms of creating training data, and schema-less "OpenIE" extraction. After introducing the separate approaches targeting these data modalities, we will then explore research that combines signals from textual, visual, and layout information to consider all aspects of a document.

Throughout the tutorial, we will bring together lessons learned from the different communities involved in information extraction research and will

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 23–26
July 5, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

provide insights from industry experiences building a production knowledge graph leveraging both unstructured and semi-structured text. Section 3 contains a full outline of planned content.

Tutorial slides are available at `https://sites.google.com/view/acl-2020-multi-modal-ie`

Relevance to ACL: Information Extraction is a core task in natural language processing, with the web serving as a rich source of information for constructing knowledge bases (KBs). A 2018 NAACL tutorial, "Scalable Construction and Reasoning of Massive Knowlege Bases" (Ren et al., 2018), provided an overview of recent IE and KB research. However, like most NLP research, that tutorial focused on methods that treat text as a simple string of natural language sentences in a `txt` file, while many real-world documents convey information via visual and layout relationships. A separate line of information extraction work has focused on learning to extract from these template-based documents. As interest in multi-modal NLP techniques has grown in recent years, we think the community will be interested in a tutorial that compares and contrasts these approaches and examines recent research that brings together textual, visual, and layout features of documents.

2 Type of the tutorial:

The tutorial will cover **cutting-edge** work in both unstructured and semi-structured information extraction, including visual and GCN-based approaches. However, our coverage of semi-structured and tabular IE will cover **introductory** material since it is likely new to much of the NLP community.

3 Outline

1. **(30 mins) Introduction and Applications**
   - Knowledge Base Population
     - Intro to knowledge graphs
     - Applications
     - Industry examples
     - Importance of the long tail
   - Unstructured, Semi-structured, and Tabular text
     - Unstructured Text
     - HTML and DOM trees
     - Webtables
     - Template learning vs. generalization
   - Schema-aligned extraction vs. OpenIE
   - Common challenges, opportunities, and key intuitions
2. **(45 mins) IE from unstructured text:**
   - Tasks
     - Named Entity Recognition
     - Co-reference Resolution
     - Relation Extraction
     - Event Extraction
   - Featurization and Modeling
     - OpenTag (Zheng et al., 2018)
     - DyGIE (Luan et al., 2019)
   - Limited Training Data
     - Distant Supervision (Mintz et al., 2009)
     - Data Programming (Ratner et al., 2017)
   - OpenIE
3. **(45 mins) IE from semi-structured documents**
   - Supervised Wrapper Induction
     - Vertex (Gulhane et al., 2011)
   - Distantly Supervised approaches
     - LODIE (Ciravegna et al., 2012)
     - DIADEM (Furche et al., 2012)
     - Ceres (Lockard et al., 2018)
   - OpenIE / Schema-less approaches
     - WEIR (Bronzi et al., 2013)
     - OpenCeres (Lockard et al., 2019)
4. **(15 mins) IE from tables**
   - WebTables (Cafarella et al., 2018)
   - Subject detection (Venetis et al., 2011)
   - Joint approaches (LimayeGirija et al., 2010)
5. **(30 mins) Multi-modal extraction**
   - Benefits of multi-modal extraction
     - Connecting tables and text (Ibrahim et al., 2019)
     - Visual signals for keyphrase extraction (Xiong et al., 2019)
     - Documents as images (Katti et al., 2018)
     - GCN-based encoders (Qian et al., 2019; Liu et al., 2019)
   - Multi-modal signals for creating training data (Wu et al., 2018)

- Multi-modal OpenIE

6. **(15 mins) Conclusion and Open Directions**

4 Prerequisites

The tutorial should be accessible to anyone with a background in natural language processing. It would be helpful to have a basic understanding of classification algorithms, preferably with some knowledge of neural network approaches, as well as unsupervised clustering algorithms.

5 Reading list

- "Web-Scale Information Extraction With Vertex", Gulhane et al. (2011)
- "Ten Years of WebTables", Cafarella et al. (2018)
- "Fonduer: Knowledge Base Construction from Richly Formatted Data", Wu et al. (2018)
- "Document-level N-ary Relation Extraction with Multi-Scale Representation Learning", Jia et al. (2019)
- "Extraction and Integration of Partially Overlapping Web Sources" Bronzi et al. (2013)
- "Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion", Dong et al. (2014)
- "A General Framework for Information Extraction Using Dynamic Span Graphs", Luan et al. (2019)
- "OpenCeres: When Open Information Extraction Meets the Semi-Structured Web", Lockard et al. (2019)
- "GraphIE: A Graph-Based Framework for Information Extraction", Qian et al. (2019)

6 Presenters

In alphabetical order,

Xin Luna Dong is a Principal Scientist at Amazon, leading the efforts of constructing Amazon Product Knowledge Graph. She was one of the major contributors to the Google Knowledge Vault project, and has led the Knowledge-based Trust project, which is called the "Google Truth Machine" by the Washington Post. She co-authored the book "Big Data Integration", was awarded ACM Distinguished Member, VLDB Early Career Research Contribution Award for "advancing the state of the art of knowledge fusion", and Best Demo award in Sigmod 2005. She serves on the VLDB endowment and PVLDB advisory committee, and was a PC co-chair for VLDB 2021, ICDE Industry 2019, VLDB Tutorial 2019, Sigmod 2018 and WAIM 2015. She has given multiple tutorials on data integration, graph mining, and knowledge management.
Email: lunadong@amazon.com
Homepage: http://lunadong.com/.

Hannaneh Hajishirzi is an Assistant Professor at the Paul G. Allen School of Computer Science & Engineering at the University of Washington. She works on NLP, AI, and machine learning, particularly designing algorithms for semantic understanding, reasoning, question answering, and information extraction from multimodal data. She has earned numerous awards for her research, including an Allen Distinguished Investigator Award, a Google Faculty Research Award, a Bloomberg Data Science Award, an Amazon Research Award, and a SIGDIAL Best Paper Award.
Email: hannaneh@washington.edu
Homepage:
https://homes.cs.washington.edu/ hannaneh/

Colin Lockard is a PhD student at the Paul G. Allen School of Computer Science & Engineering at the University of Washington, where he has published papers on knowledge extraction from both unstructured and semi-structured text.
Email: lockardc@cs.washington.edu
Homepage:
https://homes.cs.washington.edu/ lockardc/

Prashant Shiralkar is an Applied Scientist in the Product Graph team at Amazon. He currently works on knowledge extraction from semi-structured data. Previously, he received a Ph.D. from Indiana University Bloomington where his dissertation work focused on devising computational approaches for fact checking by mining knowledge graphs. His research interests include machine learning, data mining, information extraction and NLP, and Semantic Web technologies.
Email: shiralp@amazon.com
Homepage:
https://sites.google.com/site/shiralkarprashant/

References

Mirko Bronzi, Valter Crescenzi, Paolo Merialdo, and Paolo Papotti. 2013. Extraction and integration of partially overlapping web sources. *VLDB*, 6(10):805–816.

Michael Cafarella, Alon Halevy, Hongrae Lee, Jayant Madhavan, Cong Yu, Daisy Zhe Wang, and Eugene Wu. 2018. Ten years of webtables. *VLDB*, 11(12):2140–2149.

Fabio Ciravegna, Anna Lisa Gentile, and Ziqi Zhang. 2012. LODIE: Linked open data for web-scale information extraction. *SWAIE*, 925:11–22.

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. 2014. From data fusion to knowledge fusion. *VLDB*, 7(10):881–892.

Tim Furche, Georg Gottlob, Giovanni Grasso, Omer Gunes, Xiaoanan Guo, Andrey Kravchenko, Giorgio Orsi, Christian Schallhart, Andrew Sellers, and Cheng Wang. 2012. Diadem: domain-centric, intelligent, automated data extraction methodology. In *WWW*, pages 267–270. ACM.

Pankaj Gulhane, Amit Madaan, Rupesh Mehta, Jeyashankher Ramamirtham, Rajeev Rastogi, Sandeep Satpal, Srinivasan H Sengamedu, Ashwin Tengli, and Charu Tiwari. 2011. Web-scale information extraction with vertex. In *ICDE*, pages 1209–1220. IEEE.

Yusra Ibrahim, Mirek Riedewald, Gerhard Weikum, and Demetrios Zeinalipour-Yazti. 2019. Bridging quantities in tables and text. *ICDE*, pages 1010–1021.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *NAACL-HLT*, pages 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. *EMNLP*.

LimayeGirija, SarawagiSunita, and Chakrabarti-Soumen. 2010. Annotating and searching web tables using entities, types and relationships. In *VLDB 2010*.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. In *NAACL-HLT*.

Colin Lockard, Xin Luna Dong, Arash Einolghozati, and Prashant Shiralkar. 2018. Ceres: distantly supervised relation extraction from the semi-structured web. *VLDB*, 11(10):1084–1096.

Colin Lockard, Prashant Shiralkar, and Xin Luna Dong. 2019. OpenCeres: When open information extraction meets the semi-structured web. In *NAACL-HLT*, pages 3047–3056.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *NAACL-HLT*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL/IJCNLP*.

Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 2019. Graphie: A graph-based framework for information extraction. In *NAACL-HLT*.

Alexander J. Ratner, Stephen H. Bach, Henry R. Ehrenberg, and Christopher Ré. 2017. Snorkel: Fast training set generation for information extraction. In *SIGMOD*.

Xiang Ren, Nanyun Peng, and William Yang Wang. 2018. Scalable construction and reasoning of massive knowledge bases. In *NAACL-HLT, Tutorial Abstracts*, pages 10–16.

Petros Venetis, Alon Y. Halevy, Jayant Madhavan, Marius Pasca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. 2011. Recovering semantics of tables on the web. *PVLDB*, 4:528–538.

Sen Wu, Luke Hsiao, Xiao Cheng, Braden Hancock, Theodoros Rekatsinas, Philip Levis, and Christopher Ré. 2018. Fonduer: Knowledge base construction from richly formatted data. *SIGMOD*, 2018:1301–1316.

Lee Xiong, Chuan Hu, Chenyan Xiong, Daniel Fernando Campos, and Arnold Overwijk. 2019. Open domain web keyphrase extraction beyond language modeling. In *EMNLP/IJCNLP*.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. OpenTag: Open attribute value extraction from product profiles. In *KDD*, pages 1049–1058. ACM.