Web-scale Knowledge Collection

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

Xin Luna Dong
Amazon
Seattle, Washington
lunadong@amazon.com

Prashant Shiralkar
Amazon
Seattle, Washington
shiralp@amazon.com

Hannaneh Hajishirzi
University of Washington
Seattle, Washington
hannaneh@washington.edu

ABSTRACT

How do we surface the large amount of information present in HTML documents on the Web, from news articles to scientific papers 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. We cover the key ideas and intuition behind existing approaches to emphasize their applicability and potential in various settings.

CCS CONCEPTS

• Information systems → Data extraction and integration; • Computing methodologies → Information extraction.

KEYWORDS

Information extraction, Web extraction, Semi-structured data, Web mining

ACM Reference Format:

Colin Lockard, Prashant Shiralkar, Xin Luna Dong, and Hannaneh Hajishirzi. 2020. Web-scale Knowledge Collection. In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM '20)*, February 3–7, 2020, Houston, TX, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3336191.3371878>

1 MOTIVATION

The World Wide Web contains vast quantities of textual information in several forms: unstructured text, template-based semi-structured webpages, and tables. Methods for extracting information from these sources and converting it to a structured form have been a target of research from the NLP, data mining, and database communities. Extraction techniques have largely differed based on the form of data, with approaches targeted at unstructured text relying on

learning syntactic or semantic textual patterns, while approaches targeted at semi-structured documents learn to identify structural patterns in the template, and approaches targeting web tables rely heavily on entity linking and type information.

Nevertheless, researchers in these different areas 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.

While these different data modalities have largely been considered separately in the past, recent research has started taking a more holistic view 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 and cross-document understanding rather than simply considering sentences as independent units.

With this in mind, it is worth considering the information extraction problem as a whole to understand how the many modalities of data on the web can be best utilized.

2 OUTLINE OF TOPICS AND SCHEDULE

The tutorial will cover the following topics over 3 hours, including a 15 minute break between Parts 2 and 3.

2.1 Part 1: Introduction and Applications

(15 min) We begin our tutorial by introducing the different modalities in which information is presented on the web, including unstructured text, template-based semi-structured webpages, and relational tables. We then introduce the two main regimes for extraction: Traditional schema-aligned extraction (or “ClosedIE”), where facts are extracted and directly mapped onto a pre-existing ontology of entities and relationships, and “OpenIE”, where the relationships between entities are expressed using strings extracted from the source document with no pre-defined schema.

We then motivate the utility of these extraction approaches by introducing applications in knowledge graph construction, question-answering, personalization, and ontology creation.

2.2 Part 2: IE from Unstructured Text

(1 hr 15 min) Diving deep into extraction work on unstructured text from the natural language processing community, we explore methods for both the ClosedIE and OpenIE approaches introduced

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WSDM '20, February 3–7, 2020, Houston, TX, USA
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6822-3/20/02.
<https://doi.org/10.1145/3336191.3371878>

previously. We organize our survey of these approaches by focusing on the common problems that have motivated researchers across the different extraction communities. One such problem is the difficulty in obtaining sufficient training data, which has motivated work on bootstrapping [6], distant supervision [11, 20], and data programming [30]. Another problem is the question of responding to the variety and scale of knowledge in documents, which motivates cross-sentence [24, 27] and cross-document approaches, as well as multi-task extractors [14, 17, 18].

Along the way, we introduce machine learning models used for extraction including CNNs [7, 22, 31], LSTMs [19, 21], and GCNs [9, 18, 23], and a variety of OpenIE [8, 29] and schema-less approaches [12, 32].

2.3 Part 3: IE from Semi-structured Documents

(1 hr 15 min) Building on the discussion of approaches to extraction from unstructured text, we look at how many of the same problems have motivated researchers working on semi-structured websites and tabular data. We introduce semi-structured extraction models built using both the wrapper induction approach and statistical machine learning methods and then discuss how these have been adapted for dealing with the problem of limited training data, including semi-supervision [10], bootstrapping [28] and distant supervision [15]. We further look at methods that harness cross-document relationships while maintaining computational efficiency [2]. Along the way, we look at related work on identifying relational tables on the web and aligning them to existing knowledge and schemas [3–5].

We then turn toward recent work that has begun to combine methods from both unstructured and semi-structured extraction by taking a more holistic view toward documents, their text, and their structure. These methods use visual elements of documents [1, 13, 16, 25, 26] and attempt to learn generalizable layout features that can be used in conjunction with textual semantics to better inform extraction.

2.4 Conclusion and Open Directions

(15 min) We conclude by discussing how recent trends suggest many open possibilities for building better representations of webpages that will allow for more accurate and scalable extraction technologies. Deep learning methods offer potential for harnessing and unifying the diverse signals from text, layout templates, and visual features of webpages, opening new avenues for mining the wealth of information present on the Web.

REFERENCES

- [1] Yonatan Aumann, Ronen Feldman, Yair Liberzon, Binyamin Rosenfeld, and Jonathan Schler. 2006. Visual information extraction. *Knowledge and Information Systems* 10 (2006), 1–15.
- [2] Mirko Bronzi, Valter Crescenzi, Paolo Merialdo, and Paolo Papotti. 2013. Extraction and integration of partially overlapping web sources. *Proceedings of the VLDB Endowment* 6, 10 (2013), 805–816.
- [3] Michael Cafarella, Alon Halevy, Hongrae Lee, Jayant Madhavan, Cong Yu, Daisy Zhe Wang, and Eugene Wu. 2018. Ten years of webtables. *Proceedings of the VLDB Endowment* 11, 12 (2018), 2140–2149.
- [4] Michael J Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. Webtables: exploring the power of tables on the web. *Proceedings of the VLDB Endowment* 1, 1 (2008), 538–549.
- [5] Michael J Cafarella, Alon Y Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu. 2008. Uncovering the Relational Web. In *WebDB*.
- [6] Andrew Carlson and Charles Schafer. 2008. Bootstrapping information extraction from semi-structured web pages. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 195–210.
- [7] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. In *ACL*.
- [8] Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *WWW*. ACM, 355–366.
- [9] Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction. In *ACL*.
- [10] Pankaj Gulhane, Amit Madaan, Rupesh Mehta, Jeyashankar Ramamirtham, Rajeev Rastogi, Sandeep Satpal, Srinivasan H Sengamedu, Ashwin Tengli, and Charu Tiwari. 2011. Web-scale information extraction with vertex. In *ICDE*. IEEE, 1209–1220.
- [11] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based Weak Supervision for Information Extraction of Overlapping Relations. In *ACL (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 541–550. <http://dl.acm.org/citation.cfm?id=2002472.2002541>
- [12] Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R. Voss, Jiawei Han, and Avirup Sil. 2016. Liberal Event Extraction and Event Schema Induction. In *ACL*.
- [13] Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. *arXiv preprint arXiv:1809.08799* (2018).
- [14] Qi Li, Heng Ji, and Liang Huang. 2013. Joint Event Extraction via Structured Prediction with Global Features. In *ACL*.
- [15] Colin Lockard, Xin Luna Dong, Arash Einolghozati, and Prashant Shiralkar. 2018. CERES: distantly supervised relation extraction from the semi-structured web. *Proceedings of the VLDB Endowment* 11, 10 (2018), 1084–1096.
- [16] Colin Lockard, Prashant Shiralkar, and Xin Luna Dong. 2019. OpenCeres: When Open Information Extraction Meets the Semi-Structured Web. In *NAACL-HLT*. 3047–3056.
- [17] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *EMNLP*.
- [18] 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*.
- [19] Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *ArXiv abs/1603.01354* (2016).
- [20] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL/IJCNLP*.
- [21] Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint Event Extraction via Recurrent Neural Networks. In *HLT-NAACL*.
- [22] Thien Huu Nguyen and Ralph Grishman. 2015. Relation Extraction: Perspective from Convolutional Neural Networks. In *VS@HLT-NAACL*.
- [23] Thien Huu Nguyen and Ralph Grishman. 2018. Graph Convolutional Networks With Argument-Aware Pooling for Event Detection. In *AAAI*.
- [24] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen tau Yih. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *TACL* 5 (2017), 101–115.
- [25] Giuseppe Della Penna and Sergio Orefice. 2016. Supporting Information Extraction from Visual Documents.
- [26] Yujie Qian, Enrico Santus, Zhijiang Jin, Jiang Guo, and Regina Barzilay. 2018. GraphIE: A Graph-Based Framework for Information Extraction. In *NAACL-HLT*.
- [27] Chris Quirk and Hoifung Poon. 2016. Distant Supervision for Relation Extraction beyond the Sentence Boundary. In *EACL*.
- [28] Martin Rezk, Laura Alonso Alemany, Lasguido Nio, and Ted Zhang. 2019. Accurate Product Attribute Extraction on the Field. In *ICDE*. IEEE, 1862–1873.
- [29] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *EMNLP-CoNLL*. Association for Computational Linguistics, 523–534.
- [30] 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 (2018), 1301–1316.
- [31] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In *EMNLP*.
- [32] Congle Zhang, Stephen Soderland, and Daniel S. Weld. 2015. Exploiting Parallel News Streams for Unsupervised Event Extraction. *TACL* 3 (2015), 117–129.