Data Sets for Entity Information Extraction from the Web

Vincent W. Zheng
Advanced Digital Sciences Center, Singapore
vincent.zheng@adsc.com.sg

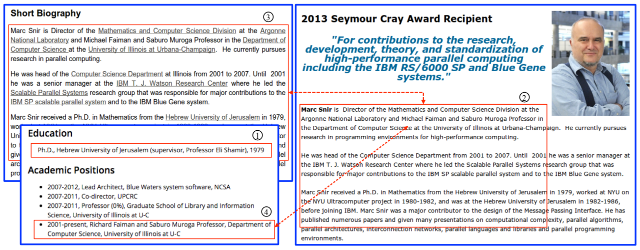
Kevin Chen-Chuan Chang
University of Illinois at Urbana Champaign, USA
kcchang@illinois.edu

ABSTRACT

Entity Information Extraction (EIE) is useful for many downstream applications, such as entity integration and knowledge base consolidation. In this document, we introduce three data sets that we collected for the task of entity information extraction. Reference for these data sets is [1].

1. INTRODUCTION

The task of *Entity Information Extraction* (EIE) aims to extract information about an entity e from a set of its Web pages. EIE is useful for supporting many downstream applications such as entity integration and knowledge base consolidation. In EIE, entities can be researchers, cars, products, etc.; and their set of entity Web pages can be easily collected from many entity information sites (e.g., Edmunds.com, Google Shopping) or search engines with person name disambiguation.



Short Biography

Marc Snir is Director of the Mathematics and Computer Science Division at the Argonne National Laboratory and Michael Faiman and Saburo Muroga Professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. He currently pursues research in parallel computing.

He was head of the Computer Science Department at Illinois from 2001 to 2007. Until 2001 he was a senior manager at the IBM T. J. Watson Research Center where he led the Scalable Parallel Systems research group that was responsible for major contributions to the IBM SP scalable parallel system and to the IBM Blue Gene system.

Marc Snir received a Ph.D. in Mathematics from the Hebrew University of Jerusalem in 1979,

Education

Ph.D., Hebrew University of Jerusalem (supervisor, Professor Eli Shamir), 1979

Academic Positions

- 2007-2012, Lead Architect, Blue Waters system software, NCSA
- 2007-2011, Co-director, UPCRC
- 2007-2011, Professor (0%), Graduate School of Library and Information Science, University of Illinois at U-C
- 2001-present, Richard Faiman and Saburo Muroga Professor, Department of Computer Science, University of Illinois at U-C

2013 Seymour Cray Award Recipient

"For contributions to the research, development, theory, and standardization of high-performance parallel computing including the IBM RS/6000 SP and Blue Gene systems."

Marc Snir is Director of the Mathematics and Computer Science Division at the Argonne National Laboratory and Michael Faiman and Saburo Muroga Professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. He currently pursues research in programming environments for high-performance computing.

He was head of the Computer Science Department from 2001 to 2007. Until 2001 he was a senior manager at the IBM T. J. Watson Research Center where he led the Scalable Parallel Systems research group that was responsible for major contributions to the IBM SP scalable parallel system and to the IBM Blue Gene system.

Marc Snir received a Ph.D. in Mathematics from the Hebrew University of Jerusalem in 1979, worked at NYU on the NYU Ultracomputer project in 1980-1982, and was at the Hebrew University of Jerusalem in 1982-1986, before joining IBM. Marc Snir was a major contributor to the design of the Message Passing Interface. He has published numerous papers and given many presentations on computational complexity, parallel algorithms, parallel architectures, interconnection networks, parallel languages and libraries and parallel programming environments.

Figure 1: An example of EIE from three Web pages.

We give an example of how to do EIE for researcher entities. In Figure 1, we are extracting a set of predefined attributes (e.g., BIO, EMPLOYMENT and so on) about a researcher "Marc Snir from University of Illinois" from three of his Web pages. We formulate this EIE task as a multi-class classification problem: suppose we can segment each Web page into a set of text snippets $X = \{x_k \in \mathcal{X} | k = 1, ..., n\}$; for each text snippet x_k, we wish to assign a label $y_k \in \mathcal{Y}$ for extraction, where $\mathcal{X}$ and $\mathcal{Y}$ are the possible text snippet space and attribute space respectively. We emphasize how to segment the Web page is not our focus; in practice, we find that defining each text snippet as a contiguous text block in HTML (thus consistent in content) suffices to serve our purpose.

Table 2: Data statistics.

	$\|E\|$	$\|W\|$	$\|X\|$	$\|E_L\|$	$\|E_U\|$	$\|E_T\|$	$\|\mathcal{Y}\|$
People	1003	3002	48.2K	100	800	103	11
Product	84	460	4.9K	1	49	34	6
Car	143	1286	25.3K	10	115	18	8

2. DATA COLLECTION

We collected real-world data sets from three domains: *Researcher*, *Car* and *Product*. For each domain, we prepared a list of entities. For each entity, we collected a set of Web pages, each of which is further parsed into text snippets as described in Section 1. In Table 1, we explain what the entities are and how we collected the pages. For each domain, we randomly chose a subset of entities, and asked two human judges to label their data. The judges achieved an agreement of 84% for Researcher domain, 79% for Car domain and 93% for Product domain. The judges resolved the disagreements through discussion.

In Table 2, we summarize the data statistics for each domain. We denote $E = E_L \cup E_U \cup E_T$ as the whole set of entities in a domain, W as the whole set of Web pages collected for E, and $X = X_L \cup X_U \cup X_T$ as the whole set of snippets parsed from W. For example, in people domain, we totally had $|E| = 1003$ entities. We crawled $|W| = 3002$ Web pages for all these entities. We further parsed these pages into $|X| = 48.2K$ snippets. We randomly picked $|E_L| + |E_T| = 203$ entities, and labeled all their snippets. We left the other $|E_U| = 800$ entities' snippets as unlabeled. We evaluated our model on this data set for five times so as to get an average performance. Each time, we randomly selected $|E_L| = 100$ entities to generate the labeled training data, and the other $|E_T| = 103$ entities to generate the test data. For fair comparison, we always compared our model with all the baselines on the same set of data each time. In people domain, we considered $|\mathcal{Y}| = 11$ attributes, as listed in Table 1.

3. REFERENCES

[1] V. W. Zheng and K. C.-C. Chang. Semi-supervised structured classification with conditional probabilistic constraints. In *Proc. of The 25th ACM International Conference on Information and Knowledge Management*, CIKM '16, 2016.

CIKM '16, October 24–28, 2016, Indianpolis, IN, USA.
© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...$15.00
DOI: http://dx.doi.org/XXXX.XXXX

Table 1: Data generation procedures for three domains.

	Description
People	The entities are researchers randomly picked from DBLP's Most Prolific Author List[1] and authors' colleagues. For each entity, we used the name and her affiliation as a query to search in Google, and collected the top three pages in the search result. The attributes are *Bio*, *Presentation*, *Award*, *Research*, *Education*, *Employment*, *Phone*, *Fax*, *Email*, *Address* and *Others*.
Product	The entities are tennis racquets randomly picked from Google Shopping[2]. For each entity, we got all its seller pages in Google Shopping. The attributes are *HeadSize*, *UnstrungWeight*, *StrungWeight*, *BeamWidth*, *Length* and *Others*.
Car	The entities are cars released in 2009. For each entity, we used its name and a keyword "reviews" as a query to search in Google, and collected the top ten pages in the search result. Besides, for each entity, we also collected its specifications from Edmunds[3], which we used to construct constraints. The attributes are *Verdict*, *Interior*, *Exterior*, *DriveRide*, *Reliability*, *Safety*, *Price* and *Others*.