Automatic Entity Recognition and Typing in Massive Text Data

Xiang Ren[†] Ahmed El-Kishky[†] Heng Ji [‡] Jiawei Han[†]
[†] University of Illinois at Urbana-Champaign, Urbana, IL, USA
[‡] Computer Science Department, Rensselaer Polytechnic Institute, USA
[†]{xren7, elkishk2, hanj}@illinois.edu [‡]jih@rpi.edu

ABSTRACT

In today's computerized and information-based society, individuals are constantly presented with vast amounts of text data, ranging from news articles, scientific publications, product reviews, to a wide range of textual information from social media. To extract value from these large, multi-domain pools of text, it is of great importance to gain an understanding of entities and their relationships. In this tutorial, we introduce data-driven methods to recognize typed entities of interest in massive, domain-specific text corpora. These methods can automatically identify token spans as entity mentions in documents and label their fine-grained types (*e.g.*, people, product and food) in a scalable way. Since these methods do not rely on annotated data, predefined typing schema or hand-crafted features, they can be quickly adapted to a new domain, genre and language. We demonstrate on real datasets including various genres (*e.g.*, news articles, discussion forum posts, and tweets), domains (general vs. biomedical domains) and languages (*e.g.*, English, Chinese, Arabic, and even low-resource languages like Hausa and Yoruba) how these typed entities aid in knowledge discovery and management.

1. INTRODUCTION

Recognizing some limitations in organizing large-scale textual data, we motivate the necessity of powerful and scalable methods for entity recognition and typing as database technology is applied to modern-day massive text data.

Motivation

We motivate this tutorial by starting with identifying the key aspects that have led to the success in database technology and suggesting a similar approach to handling the modern day explosion of unstructured big-data.

Entity recognition/typing and structured analysis of massive text corpora. The success of database technology is largely attributed to the efficient and effective management of structured data. The construction of a well-structured database is often the premise of consequent applications. Although the majority of existing data generated in our society is unstructured, big data leads to big opportunities to uncover structures of real-world entities, such as people, products and organizations, from massive amount but interrelated unstructured data. By mining token spans of entity mentions in documents, labeling their structured types and inferring their relations, it is possible to construct semantically rich structures and provide conceptual summarization of such data. The uncovered structures will facilitate browsing information and retrieving knowledge that are otherwise locked in the data.

Several techniques we discussed have already gained some publicity or application within industry. Our phrase mining tool, SegPhrase [25], won the grand prize of Yelp Dataset Challenge[1] and was adopted in TripAdvisor's new feature[2]. Our entity recognition and typing tool, ClusType [31], was shipped to facilitate products in Microsoft Bing Ads team. Our Entity Discovery and Linking system [17] won first place at NIST TAC-KBP Tri-lingual Entity Discovery and Linking Evaluation.

Example: Automatically recognizing and typing entities in Yelp reviews. In a business review corpus like Yelp reviews, entities such as `food`, `restaurant`, `location` and `event` are mentioned in the documents. For example, from the sentences "*The best BBQ I've tasted in Washington! I had the pulled pork sandwich with coleslaw for lunch.*", it is desirable to identify "*BBQ*", "*pulled pork sandwich*", and "*coleslaw*" as `food`, and "*Washington*" as location. Furthermore, the label of "*coleslaw*" can be refined as `salad` (and "*pulled pork sandwich*" as `sandwich`). However, existing work encounters several challenges when handling such a *domain-specific* text corpus.

1. The lack of annotated data for domain-specific corpora presents a major challenge for adapting traditional supervised named-entity recognition techniques. Fortunately, a number of semantically rich knowledge-bases are available, which provides chances for *automatically* recognizing entities by *distant supervision*.

2. The automatically generated training data by distant supervision may introduce noisy labels—labels that are irrelevant to the local context of entity mention. It is necessary to conduct label noise reduction on the training data by leveraging corpus-level statistics.

SIGMOD'16, June 26-July 01, 2016, San Francisco, CA, USA
 ISBN 978-1-4503-3531-7/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2882903.2912567

[1] http://www.yelp.com/dataset_challenge
[2] http://engineering.tripadvisor.com/using-nlp-to-find-interesting-collections-of-hotels/

3. Many entity detection tools such as noun phrase chunkers are trained on general-domain corpora (*e.g.*, news articles), but they do not work effectively nor efficiently on domain-specific corpora such as Yelp reviews (*e.g.*, "*pulled pork sandwich*" cannot be detected). A domain-agnostic phrase mining algorithm is required to efficiently generate entity mentions with minimal linguistic assumptions.

4. Entity surface names are often ambiguous—multiple entities many share the same surface name (*e.g.*, "*Washington*" may refer to the U.S. government, the capital city or a sport team). Although the contexts surrounding each entity mention provide clues on its types, challenges arise due to the diversity on paraphrasing. With data redundancy in a massive corpus, it is possible to disambiguate entities and resolve synonymous contexts using correlated textual information structured in an information network for holistic analysis.

What will be covered?

Preliminaries: We introduce the audience to the broad subject of entity recognition by providing motivation in the context of *information extraction for knowledge base population* . Within this context introduce entities, types, extracting these entities from within text itself, and examples validating the necessity for entity disambiguation and resolution. We then introduce several different applications to latent entity discovery in data. In particular, we will introduce and explain grouping latent entities by concepts, by topics, and even extracting and understanding relationships between entities.

Entity Mention Detection: In this part of our presentation, we introduce the problem of identifying entity mentions. We formulate the problem and discuss three main schools of thought in tackling the problem.

1. **Supervised Methods:** We begin by introducing IOB, a common representation that transforms entity mention detection into classification. Then, beginning with classical (non-sequential) models, we outline a variety of methods including unigram and higher-order chunkers, SVM's, maximum entropy models, and ensemble methods. We then introduce sequential models for entity mention detection and outline a progressive array of methods from generative to discriminative models.
2. **Unsupervised Methods:** We follow on by introducing two main classes of unsupervised approaches: chunking grammar approaches and methods that draw upon large-text corpora and their relative merits and broad spectrum of applications. We then focus on applications of ToPMine for topical phrase mining and noun-collocation mining.
3. **Distantly / Weakly Supervised Methods:** We focus on a variety of methods including incorporating outside information via dictionary. We mainly emphasize SegPhrase, an approach for extracting high-quality phrases and entity mentions with minimal supervision.

Entity Recognition in Individual Documents:

1. **General Text:** In the context of general text recognition, we discuss introduce many named entity recognition (NER) methods. We discuss entity recognition as sequence labeling as well as the coarse types and manually-annotated corpora these models leverage.
2. **Domain Text:** In the context of domain-specific extraction, we discuss several approaches. We discuss twitter in the context of Tweet segmentation and chunking as well as LabeledLDA based on Freebase. In addition to twitter we discuss entity recognition in product reviews and biomedical text data.

Entity Recognition in Large Domain-Specific Corpora: We contrast single-document cases to the context of large single-domain corpora. Starting with semi-supervised approaches, we present sequence-labeling models and models that combine local and global features. We transition to weakly supervised approaches and their merits -discussing pattern-based bootstrapping methods, SEISA: Set expansion, and a variety of probabilistic modeling methods as well as graph-based label propagation approaches. We then discuss several approaches for distantly supervised entity recognition. These methods include state of the art approaches such as FIGER which performs sequence labeling with automatically annotated data , SemTagger - a contextual classifier that uses seed data , APOLLO which performs label propagation on graphs, and ClusType which employs relation phrase-based clustering for effective entity recognition.

Case Study and Evaluations We conclude our tutorial by demonstrating the capabilities of many of the tools and methods mentioned on a variety of test cases and metrics. We begin by evaluating tools and methodologies for entity recognition. We introduce a variety of evaluation metrics and public datasets, and evaluate a variety of general-domain NER systems including the Stanford Named Entity Recognizer, Illinois Named Entity Tagger, FIGER, and other Named Entity Recognition in NLP toolkits. We then present a few case-studies on two real-world datasets consisting of news articles and tweets. In particular we focus on entity mention detection in these datasets and typing these extracted entity mentions.

Why a tutorial at SIGMOD 2016

In today's era of 'big data', people are exposed to an explosion of information in the form of documentary data collections, ranging from the scientific knowledge of all humanity, to the daily life of individuals. Most of these collections are unstructured or loosely structured. Effective detection, extraction, and typing of entity structures is key to inducing structure and understanding from messy and scattered raw data. This tutorial will present an organized picture of recent research on entity mention detection, entity typing, and general text extraction algorithms. We will show how exciting and surprising knowledge can be discovered from your own not so well-structured raw data.

Audience and Prerequisites

Researchers and practitioners in the field of database systems, information extraction, data mining, text mining, information retrieval, web search, and information systems. While the audience with a good background in these areas would benefit most from this tutorial, we believe the

material to be presented would give general audience and newcomers an introductory pointer to the current work and important research topics in this field, and inspire them to learn more. Only preliminary knowledge about text mining, information extraction, data mining, algorithms, and their applications are needed.

2. TUTORIAL OUTLINE

This tutorial presents a comprehensive overview of the techniques developed for automatic entity recognition and typing in recent years. We will discuss the following key issues.

1. Preliminaries of Entity Recognition and Typing
   (a) Entities that are explicitly typed and linked externally with documents.
      i. Wikilinks and ClueWeb corpora
   (b) Entities that can be extracted within text.
   (c) Entity disambiguation and resolution.
      i. MENED: Mining evidence outside referent knowledge bases
2. Entity Mention Detection
   (a) Unsupervised Entity Mention Detection
      i. ReVerb: A pattern-based approach for matching entities and relations
      ii. Significance detection of phrases and entities in a corpus
      iii. Jointly identifying significant phrases for entities and relations
   (b) Supervised Entity Mention Detection
      i. Jointly extracting entities and relations
      ii. MaxiEnt markov models for information extraction and segmentation
      iii. Semi-supervised text chunking for entity candidate generation
      iv. Ranking for entity mention
   (c) Distantly / Weakly Supervised Methods
      i. SegPhrase: Weakly supervised phrase extraction and segmentation
      ii. Exploiting dictionaries: Combining semi-markov extraction process with data integration
3. Entity Recognition in Single Text Documents
   (a) Traditional supervised named entity recognition (NER) systems
      i. Entity recognition and typing as a sequence labeling task
      ii. Classic coarse types and manually-annotated corpora
      iii. Sequence labeling models
   (b) Entity recognition in tweets
      i. Tweet segmentation and chunking
      ii. LabeledLDA based on Freebase
      iii. Segment ranking
   (c) Entity recognition in product reviews
   (d) Entity recognition in biomedical text
4. Entity Recognition in A Large, Domain-specific Corpus
   (a) Semi-supervised approaches
      i. Combining local and global features
   (b) Weakly-supervised approaches
      i. Pattern-based bootstrapping methods
      ii. SEISA: A set expansion method
      iii. Probabilistic modeling methods
      iv. Graph-based label propagation
      v. Extracting entities from web tables
   (c) Distantly-supervised approaches
      i. SemTagger: Seed-based contextual classifier for entity typing
      ii. APOLLO: Label propagation on graphs
      iii. ClusType: Effective entity recognition by relation phrase-based clustering
   (d) Fine-grained typing approaches
      i. FIGER: Multi-label classification with automatically annotated data
      ii. HYENA: Hierarchical classification for fine-grained typing
      iii. WSABIE: Embedding method for fine-grained typing
   (e) Label noise reduction in distant supervision
      i. Noisy candidate types in automatically generated training data
      ii. Simple pruning heuristics
      iii. Partial-label learning methods
      iv. Label noise reduction by heterogeneous partial-label embedding
   (f) Liberal entity recognition and typing
      i. Three-level semantic representation
      ii. Joint hierarchical clustering and linking algorithm
5. Evaluations of entity recognition
   (a) Evaluation metrics and public datasets
   (b) Public general-domain NER systems
   (c) Shared tasks on entity recognition
6. Case studies: news articles and tweets.
   (a) Entity recognition in these datasets
      i. Detect entity mentions in these datasets
      ii. Typing entity mentions in these datasets
   (b) Integrating entities in both datasets
7. Recent progress and research problems on entity recognition

Tailor of the Tutorial for Different Durations. The duration of the tutorial is flexible: It is expected to be 3 hours, but it can be compressed into 1.5 hours, based on the need of the conference. The outline presented here is for the full length tutorial. For shorter duration of the tutorial, we plan to cut the detection of entity mentions by half, cut entity recognition in single, general-domain text document in half, and cut the case studies section.

3. ABOUT THE INSTRUCTORS

Xiang Ren is a Ph.D. candidate of Department of Computer Science at Univ. of Illinois at Urbana-Champaign. His research focuses on knowledge acquisition from text data and mining linked data. He is the recipient of the 2016 Google PhD Fellowship in Data Management and Databases, was the recipient of C. L. and Jane W.-S. Liu Award and Yahoo!-DAIS Research Excellence Award in 2015, and received the Microsoft Young Fellowship from Microsoft Research Asia in 2012.

Ahmed El-Kishky is a Ph.D. candidate at Univ. of Illinois at Urbana-Champaign. His research interests include mining large unstructured data, text mining, learning to rank, and network mining. He is the recipient of both the National Science Foundation Graduate Research Fellowship and National Defense Science and Engineering Fellowship.

Jiawei Han is an Abel Bliss Professor at the Department of Computer Science, UIUC. His research areas encompass data mining, data warehousing, database systems, and information networks, with over 700 publications. He is Fellow of ACM, Fellow of IEEE, Director of IPAN (2009-2016), supported by Network Science Collaborative Technology Alliance program of the U.S. Army Research Lab, and the co-Director of KnowEnG: a Knowledge Engine for Genomics, one of the NIH supported Big Data to Knowledge (BD2K) Centers.

Heng Ji is an Edward P. Hamilton Development Chair Associate Professor of Computer Science Department of Rensselaer Polytechnic Institute. Her research interests focus on Natural Language Processing and its connections with Data Mining and Vision. She received "AI's 10 to Watch" Award by IEEE Intelligent Systems in 2013 and NSF CAREER award in 2009. She coordinated the NIST TAC Knowledge Base Population task in 2010, 2011, 2014, 2015 and 2016.

4. RELATED TUTORIALS GIVEN BY THE AUTHORS

1. **Conference tutorial:** Xiang Ren, Ahmed El-Kishky, Chi Wang and Jiawei Han. "Automatic Entity Recognition and Typing from Massive Text Corpora: A Phrase and Network Mining Approach." KDD 2015.
2. **Conference tutorial:** Jiawei Han, Heng Ji and Yizhou Sun. "Successful Data Mining Methods for NLP." ACL 2015.
3. **Conference tutorial:** Han, Jiawei, Chi Wang, and Ahmed El-Kishky. "Bringing structure to text: Mining phrases, entities, topics, and hierarchies" KDD 2014.
4. **Conference tutorial:** Dan Roth, Heng Ji, Ming-Wei Chang and Taylor Cassidy. "Wikification and Beyond: The Challenges of Entity and Concept Grounding." ACL 2014.
5. **Conference tutorial:** Jiawei Han and ChiWang, "Mining Latent Entity Structures from Massive Unstructured and Interconnected Data", SIGMOD 2014.
6. **Conference tutorial:** Tim Weninger, and Jiawei Han, "Information Network Analysis and Extraction on the World Wide Web", WWW 2013.
7. **Conference tutorial:** Tim Weninger and Jiawei Han, "Exploring Structure and Content on the Web: Extraction and Integration of the Semi-Structured Web", WSDM 2013.
8. **Conference tutorial:** Yizhou Sun, Jiawei Han, Xifeng Yan, and Philip S. Yu, "Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach", VLDB 2012.

The above are the related tutorials given by the authors. Several of our early tutorials were on mining heterogeneous information networks. However, the power of such mining comes from structures of such networks (entities and links). Recent advances on information extraction and typing from massive unstructured text makes it possible to garner the informative and illuminating structures lying beneath the raw data. This tutorial presents this new line of research on starting with the shallow information extraction and ending up with deep network analysis, by mining latent entity structures from text and constructing a structured database for knowledge discovery.

References

[1] R. K. Ando and T. Zhang. A high-performance semi-supervised learning method for text chunking. In *ACL*, 2005.

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.

[3] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, and T. M. Mitchell. Coupled semi-supervised learning for information extraction. In *WSDM*, 2010.

[4] W. W. Cohen and S. Sarawagi. Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods. In *SIGKDD*, 2004.

[5] M. Collins. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *ACL*, 2002.

[6] B. B. Dalvi, W. W. Cohen, and J. Callan. Websets: Extracting sets of entities from the web using unsupervised information extraction. In *WSDM*, 2012.

[7] L. Dong, F. Wei, H. Sun, M. Zhou, and K. Xu. A hybrid neural model for type classification of entity mentions. In *IJCAI*, 2015.

[8] X. L. Dong, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, 2014.

[9] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable topical phrase mining from text corpora. *VLDB*, 2015.

[10] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in knowitall: (preliminary results). In *WWW*, 2004.

[11] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134, 2005.

[12] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, 2011.

[13] V. Ganti, A. C. König, and R. Vernica. Entity categorization over large document collections. In *SIGKDD*, 2008.

[14] A. Gattani, D. S. Lamba, N. Garera, M. Tiwari, X. Chai, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *VLDB*, 6(11):1126–1137, 2013.

[15] S. Gupta and C. D. Manning. Improved pattern learning for bootstrapped entity extraction. In *CONLL*, 2014.

[16] Y. He and D. Xin. Seisa: set expansion by iterative similarity aggregation. In *WWW*, 2011.

[17] Y. Hong, D. Lu, D. Yu, X. Pan, X. Wang, Y. Chen, L. Huang, and H. Ji. Rpi_blender tac-kbp2015 system description. In *Proc. Text Analysis Conference (TAC2015)*, 2015.

[18] R. Huang and E. Riloff. Inducing domain-specific semantic class taggers from (almost) nothing. In *ACL*, 2010.

[19] D. S. Kim, K. Verma, and P. Z. Yeh. Joint extraction and labeling via graph propagation for dictionary construction. In *AAAI*, 2013.

[20] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: named entity recognition in targeted twitter stream. In *SIGIR*, 2012.

[21] Q. Li and H. Ji. Incremental joint extraction of entity mentions and relations. In *ACL*, 2014.

[22] Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan. Mining evidences for named entity disambiguation. In *SIGKDD*, 2013.

[23] G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and searching web tables using entities, types and relationships. *VLDB*, 3(1-2):1338–1347, 2010.

[24] X. Ling and D. S. Weld. Fine-grained entity recognition. In *AAAI*, 2012.

[25] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining quality phrases from massive text corpora. In *SIGMOD*, 2015.

[26] A. McCallum, D. Freitag, and F. C. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML*, volume 17, pages 591–598, 2000.

[27] P. McNamee and J. Mayfield. Entity extraction without language-specific resources. In *COLING*, 2002.

[28] N. Nguyen and R. Caruana. Classification with partial labels. In *SIGKDD*, 2008.

[29] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *CIKM*, 2000.

[30] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *ACL*, 2009.

[31] X. Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, and J. Han. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *SIGKDD*, 2015.

[32] A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In *EMNLP*, 2011.

[33] W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *TKDE*, 27(99):1–20, 2014.

[34] Y. Sun and J. Han. Mining heterogeneous information networks: a structural analysis approach. *SIGKDD Explorations*, 14(2):20–28, 2013.

[35] J. Tang, M. Qu, and Q. Mei. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *SIGKDD*, 2015.

[36] J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In *ACL*, 2010.

[37] P. Venetis, A. Halevy, J. Madhavan, M. Paşca, W. Shen, F. Wu, G. Miao, and C. Wu. Recovering semantics of tables on the web. *VLDB*, 4(9):528–538, 2011.

[38] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In *SIGMOD*, 2012.

[39] M. Yakout, K. Ganjam, K. Chakrabarti, and S. Chaudhuri. Infogather: entity augmentation and attribute discovery by holistic matching with web tables. In *SIGMOD*, 2012.

[40] D. Yogatama, D. Gillick, and N. Lazic. Embedding methods for fine grained entity type classification. In *ACL*, 2015.

[41] M. A. Yosef, S. Bauer, J. Hoffart, M. Spaniol, and G. Weikum. Hyena: Hierarchical type classification for entity names. In *COLING*, 2012.