Automatic Entity Recognition and Typing from Massive Text Corpora: A Phrase and Network Mining Approach

Xiang Ren[†], Ahmed El-Kishky[†], Chi Wang[‡], Jiawei Han[†]
[†] University of Illinois at Urbana-Champaign, Urbana, IL, USA
[‡] Microsoft Research, Redmond, WA, USA
[†] {xren7,elkishk2,hanj}@illinois.edu [‡]chiw@microsoft.com

ABSTRACT

In today's computerized and information-based society, we are soaked 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 unlock the value of these unstructured text data from various domains, 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 types (*e.g.*, people, product, food) in a scalable way. We demonstrate on real datasets including news articles and tweets how these typed entities aid in knowledge discovery and management.

Introduction

Motivation: 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 inter-related 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.

Example: Entity recognition and typing in Yelps reviews. In a business review corpus like Yelp reviews[1], entities such as food, locations and restaurants 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. However, existing work encounters several challenges when handling such a *domain-specific* text corpus.

1. The lack of annotated data for domain-specific corpus 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. 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 mention candidates with minimal linguistic assumptions.

3. 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.

Target Audience and prerequisites

Researchers and practitioners in the field of data mining, text mining, information extraction, information retrieval, web search, database systems, 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.

[1] http://www.yelp.com/dataset_challenge

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.
ACM 978-1-4503-3664-2/15/08.
DOI: http://dx.doi.org/10.1145/2783258.2789988.

Tutorial Outline

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

Recognition and typing of entities from massive, unstructured text corpora
We introduce the background of entity recognition and typing problem. We will provide examples on different documentary data collections where entities are explicitly typed and linked with documents, or need to be extracted or inferred from the text data. We demonstrate the growing needs on recognizing and typing entities for a wide range of applications including information extraction and knowledge base population.

Entity mention detection
We discuss different paradigms and methodologies for detecting entity mention in text data, ranging from supervised learning, unsupervised approaches, to distant supervised methods. We introduce efforts to combine corpus statistics from massive document collection with a variety of document-level features, and to leverage distant supervision from structured knowledge bases. We demonstrate the benefits of various phrase mining-based approaches over traditional supervised entity detection methods.

Entity recognition in a single document
We outline different supervised methods for recognizing typed entities in a single, general-domain document. We discuss the limitations of these methods when they are applied to single, domain-specific document such as business review and tweet. We introduce previous work on designing methods that can specifically handle different domains. We further demonstrate efforts towards recognizing fine-grained entity types and adopting distant supervision from large amount of external knowledge sources.

Entity recognition in a massive, domain-specific text corpus
Often times, it is hard to obtain label data for a large, domain-specific and dynamic text corpus. We demonstrate efforts on data-driven entity recognition and typing methods to more effectively model the data redundancy in the massive text corpus. We focus on the problem of entity recognition and typing with distant supervision and demonstrate the benefits of leveraging rich entity information in structured, semantically-rich knowledge bases. We demonstrate how creating information networks that embody the interaction between different text units in text can assist in high-quality entity typing. We argue that such a network mining approach can holistically model the corpus-level information in a principled way.

Evaluation, Case Studies and Recent Progress
We outline different evaluation metrics and benchmark datasets for named-entity recognition (NER), and introduce several popular NER systems and NER shared tasks. We discuss the findings from applying different NER methods on new article domain and Yelp review domain. We outline the recent progress on combining entity recognition with other related text mining tasks, extracting entities from multiple data sources, integrating NER results from multiple systems, and applying deep learning for NER.

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 C. L. and Jane W.-S. Liu Award and Yahoo!-DAIS Research Excellence Gold Award in 2015. He received 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, and network mining. He is the recipient of both the National Science Foundation Graduate Research Fellowship as well as National Defense Science and Engineering Fellowship.

Chi Wang is a researcher in Microsoft Research, Redmond, Washington. He has been researching into discovering knowledge from unstructured and linked data, such as topics, concepts, relations, communities and social influence. His book *Mining Latent Entity Structures* is published by Morgan Claypool Pub. in the series of *Synthesis Lectures on Data Mining and Knowledge Discovery.* He is a winner of Microsoft Research Graduate Research Fellowship.

Jiawei Han is an Abel Bliss Professor of Department of Computer Science at Univ. of Illinois at Urbana-Champaign. His research areas encompass data mining, data warehousing, information network analysis, etc., with over 600 conference and journal publications. He is Fellow of ACM, Fellow of IEEE, the Director of IPAN, supported by Network Science Collaborative Technology Alliance program of the U.S. Army Research Lab, and the Director of KnowEnG: a Knowledge Engine for Genomics, one of the NIH supported Big Data to Knowledge (BD2K) Centers.

Acknowledgments

Research was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1017362, IIS-1320617, and IIS-1354329, HDTRA1-10-1-0120, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC.