Automatic Entity Recognition and Typing in Massive Text Corpora

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 natural language text data, ranging from news articles, product reviews, advertisements, to a wide range of user-generated content from social media. To turn such massive unstructured text data into actionable knowledge, one of the grand challenges is to gain an understanding of entities and the relationships between them. In this tutorial, we introduce data-driven methods to recognize typed entities of interest in different kinds of text corpora (especially in massive, domain-specific text corpora). These methods can automatically identify token spans as entity mentions in text and label their types (*e.g.*, `people`, `product`, `food`) in a scalable way. We demonstrate on real datasets including news articles and yelp reviews how these typed entities aid in knowledge discovery and management.

Keywords

Entity Recognition and Typing; Massive Text Corpora

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, organizations, and events, 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. Our phrase mining tool, SegPhrase [34], won the grand prize of Yelp Dataset Challenge[1], and our entity recognition and typing tool, ClusType [41], was shipped to facilitate products in Microsoft Bing Ads team.

Example: Entity recognition and typing for social media. In a collection of tweets, entities such as people, restaurants and events are mentioned in text. For example, from the tweet "*Jean Joho (Chef of The Eiffel Tower Restaurant) is on board to present at EC 2010.*", it is desirable to identify "*Jean Joho*" as `person`, and "*The Eiffel Tower Restaurant*" as `restaurant`, and '*EC 2010*" as `event`. However, existing work encounters several challenges when handling such a *domain-specific* text corpus:

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

2. Many entity detection tools are trained on general-domain, regular corpora (*e.g.*, news articles), but they do not work well on domain-specific corpora such as tweets. 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.*, "*Jean Joho*" may refer to different people). Even though 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 surrounding contexts using correlated textual information structured in an information network for holistic analysis.

What will be covered in this tutorial?

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

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

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
DOI: http://dx.doi.org/10.1145/2872518.28910651.

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.

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.

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
         A. Hidden Markov models
         B. Maximum entropy-based models
         C. SVM-based models
         D. Conditional random field-based 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

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
   (a) Combine entity recognition with other related tasks
   (b) Extracting entities from multiple sources
   (c) Integrating multiple NER systems

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.

1. REFERENCES

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

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[3] D. M. Bikel, R. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. *Machine learning*, 34(1-3):211–231, 1999.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT Workshop on Computational Learning Theory*, 1998.

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

[6] 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.

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

[8] M. Collins. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 489–496. Association for Computational Linguistics, 2002.

[9] J. R. Curran and S. Clark. Language independent ner using a maximum entropy tagger. In *HLT-NAACL*, 2003.

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

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

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

[13] 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.

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

[15] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

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

[17] W. Gao, P. Li, and K. Darwish. Joint topic modeling for event summarization across news and social media streams. In *CIKM*, 2012.

[18] 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.

[19] W. Guo, H. Li, Ji, and M. T. Diab. Linking tweets to news: A framework to enrich short text data in social media. In *ACL*, 2013.

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

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

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

[23] H. Ji and R. Grishman. Knowledge base population: Successful approaches and challenges. In *ACL*, 2011.

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

[25] Z. Kozareva, K. Voevodski, and S.-H. Teng. Class label enhancement via related instances. In *EMNLP*, 2011.

[26] 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.

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

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

[29] D. Lin and X. Wu. Phrase clustering for discriminative learning. In *ACL*, 2009.

[30] H. Lin, Y. Jia, Y. Wang, X. Jin, X. Li, and X. Cheng. Populating knowledge base with collective entity mentions: A graph-based approach. In *ASONAM*, 2014.

[31] T. Lin, O. Etzioni, et al. No noun phrase left behind: detecting and typing unlinkable entities. In *EMNLP*, 2012.

[32] W. Lin, R. Yangarber, and R. Grishman. Bootstrapped learning of semantic classes from positive and negative examples. In *ICML Workshop on The Continuum from Labeled to Unlabeled Data*, 2003.

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

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

[35] 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.

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

[37] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[38] N. Nakashole, T. Tylenda, and G. Weikum. Fine-grained semantic typing of emerging entities. In *ACL*, 2013.

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

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

[41] 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.

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

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

[44] W. Shen, J. Wang, P. Luo, and M. Wang. A graph-based approach for ontology population with named entities. In *CIKM*, 2012.

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

[46] P. P. Talukdar, T. Brants, M. Liberman, and F. Pereira. A context pattern induction method for named entity extraction. In *CONLL*, 2006.

[47] P. P. Talukdar and F. Pereira. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *ACL*, 2010.

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

[49] R. Yangarber, W. Lin, and R. Grishman. Unsupervised learning of generalized names. In *COLING*, 2002.