Scalable Construction and Reasoning of Massive Knowledge Bases

— Proposal for a Tutorial at NAACL 2018 —

Xiang Ren[1] **Nanyun Peng**[2] **William Yang Wang** [3]
[1] University of Southern California, Department of Computer Science
[2] University of Southern California, Information Sciences Institute
[3] University of California, Santa Barbara, Department of Computer Science
xiangren@usc.edu, npeng@isi.edu, william@cs.ucsb.edu

Abstract

In today's information-based society, there is abundant knowledge out there carried in the form of natural language texts (e.g., news articles, social media posts, scientific publications), which spans across various domains (*e.g.*, corporate documents, advertisements, legal acts, medical reports), and grows at an astonishing rate. How to turn such massive and unstructured text data into structured, actionable knowledge for computational machines, and furthermore, how to teach machines learn to reason and complete the extracted knowledge is a grand challenge to the research community.

Traditional IE systems assume abundant human annotations for training high quality machine learning models, which is impractical when trying to deploy IE systems to a broad range of domains, settings and languages.

In the first part of the tutorial, we introduce how to extract structured facts (*i.e.*, entities and their relations of different types) from text corpora to construct knowledge bases, with a focus on methods that are minimally-supervised and domain-independent for timely knowledge base construction across various application domains.

In the second part, we introduce how to leverage other knowledge, such as the distributional statistics of characters and words, the annotations for other tasks and other domains, and the linguistics and problem structures, to combat the problem of inadequate supervision, and conduct low-resource information extraction.

In the third part, we describe recent advances in knowledge base reasoning. We start with the gentle introduction to the literature, focusing on path-based and embedding based methods. We then describe DeepPath, a recent attempt of using deep reinforcement learning to combine the best of both worlds for knowledge base reasoning.

1 Introduction

Motivation. The success of data mining and artificial intelligence technology is largely attributed to the efficient and effective analysis of structured data. The construction of a well-structured, machine-actionable knowledge base (KB) from raw (unstructured or loosely-structured) data sources 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 (*e.g.*, `person`, `product`), attributes (*e.g.*, `age`, `weight`), relations (*e.g.*, `employee_of`, `manufacture`) from massive text corpora. By integrating these semantic structures, one can construct a powerful KB as a conceptual abstraction of the original corpus. The constructed knowledge base will facilitate browsing information and inferring knowledge that are otherwise widely scattered in the text corpora. Computational machines can effectively perform algorithmic analysis at a large scale over these KBs, and apply the new insights to improve human productivity in various downstream tasks.

Proceedings of NAACL-HLT 2018: Tutorial Abstracts, pages 10–16
New Orleans, Louisiana, June 1, 2018.

Our Focus. In this tutorial, we focus our discussion on two tightly related problems: automatic construction of knowledge bases from text, and knowledge reasoning for knowledge base completion. While traditional information extraction techniques have heavy reliance on human-annotated data, our tutorial will devote more time on introducing methods that can reduce human efforts in the process, by leveraging external knowledge sources (e.g., distant supervision) and exploiting rich data redundancy in massive text corpora (e.g., weak supervision). We also discuss how data sources from various domains and languages could opens up tremendous opportunities to leverage and transfer existing knowledge about domains, tasks and language, and help knowledge extraction in low-resource settings with minimal supervision. In the reasoning part, we aim to leverage the existing background knowledge and design various algorithms to fill in the missing link between entities in the KB, given the extracted KBs are likely incomplete. More specifically, this part will introduce two lines of research for KB reasoning: path-based and embedding-based methods.

Topics to be covered in this tutorial. The first 2/3 of this tutorial presents a comprehensive overview of the information extraction techniques developed in recent years for constructing knowledge bases (see also Section 2 for a more detailed outline). We will discuss the following key issues: (1) data-driven approaches for mining quality phrases from massive, unstructured text corpora; (2) entity recognition and typing: preliminaries, challenges, and methodologies; and (3) relation extraction: previous efforts, limitations, recent progress, and a joint entity and relation extraction method using distant supervision; (4) multi-task and multi-domain learning for low-resource information extraction; (5) distill linguistic knowledge into neural models to help low-resource information extraction. The second half of the tutorial presents a comprehensive overview of KB reasoning techniques. For path-based methods, we will first describe the Path-Ranking Algorithm (PRA) (Lao et al., 2011) and briefly describe extensions such as ProPPR (Wang et al., 2013). Our tutorial will also cover the recent integration of PRA with recurrent neural networks. For the embedding based method, we will briefly describe RESCAL (Nickel et al., 2011) and TransE (Bordes et al., 2013). Finally, we discuss DeepPath (Xiong et al., 2017), a novel deep reinforcement learning model that combines the embedding and path-based approaches for the learning to reason problem.

Research Impact. Our phrase mining tool, SegPhrase (Liu et al., 2015), won the grand prize of Yelp Dataset Challenge[1] and was used by TripAdvisor in productions[2]. Our entity recognition and typing system, ClusType (Ren et al., 2015), was shipped as part of the products in Microsoft Bing and U.S. Army Research Lab. We built the first named entity recognizer on Chinese social media (Peng and Dredze, 2015, 2016) and closed the gap between NER on English and Chinese social media. The same technique was applied to build the first relation extractor for cross-sentence, n-ary relation extraction between drug, gene, and mutation (Peng et al., 2017).

Duration and Sessions. The duration of the tutorial is flexible: It is expected to be 3 hours, but it can be extended into 6 hours, based on the need of the conference. The outline presented here is for the 3-hour tutorial. For longer duration of the tutorial, we plan to extend entity and relation extraction parts, and add in more case studies and applications.

Relevance to ACL. Machine "reading" and "reasoning" of large text corpora have long been the interests to CL and NLP communities, especially when people now are exposed to an explosion of information in the form of free text. Extracting structured information is key to understanding messy and scattered raw data, and effective reasoning tools are critical for the use of KBs in downstream tasks like QA. This tutorial will present an organized picture of recent research on knowledge base construction and reasoning. We will show how exciting and surprising knowledge can be discovered from your own not so well-structured raw corpora, and such incomplete KBs can be further used to derive new insights and more complex knowledge with reasoning techniques.

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

[2] http://engineering.tripadvisor.com/mining-text-review-snippets/

2 Outline

This tutorial presents a comprehensive overview of techniques for automatic knowledge base construction from text data (especially from a large, domain-specific text corpora), and techniques for reasoning over large-scale knowledge bases. We will discuss the following key issues:

1. Overview
   (a) Knowledge base: A little history
   (b) Knowledge base preliminaries
   (c) Knowledge base construction: An overview
      i. From phrases to entities and relations

2. Phrase Mining from Massive Text Corpora
   (a) Preliminaries
      i. Criteria of Quality Phrases
      ii. The Origin of Phrase Mining
         A. Automatic Term Recognition
         B. Supervised Noun Phrase Chunking
         C. Dependency Parser-based Methods
   (b) Data-Driven Phrase Mining in A Large Text Corpus
      i. Unsupervised Frequency-based Methods
      ii. Weakly Supervised Method: SegPhrase
      iii. Automated Quality Phrase
         A. No Extra Human Effort
         B. Support Multiple Languages
         C. High Performance

3. Automated Entity Recognition and Typing
   (a) Preliminaries
      i. Entities that are explicitly typed and linked externally with documents.
         A. Wikilinks and ClueWeb corpora
         B. Probase: A Probabilistic Taxonomy
         C. MENED: Mining evidence outside referent knowledge bases
      ii. Entities that can be extracted within text.
      iii. Traditional named entity recognition (NER) systems
         A. Entity extraction as a sequence labeling task
         B. Classic coarse types and manually-annotated corpora
         C. Sequence labeling models
   (b) Entity Recognition and Typing in A Large, Domain-specific Corpus
      i. Semi-supervised approaches
         A. Combining local and global features
      ii. Weakly-supervised approaches
         A. Pattern-based bootstrapping methods
         B. SEISA: A set expansion method
         C. Extracting entities from web tables
      iii. Distantly-supervised approaches
         A. SemTagger: Seed-based contextual classifier for entity typing
         B. ClusType: Effective entity recognition by relation phrase-based clustering
      iv. Fine-grained entity typing approaches
         A. FIGER: Multi-label classification with automatically annotated data
         B. Embedding methods for entity typing: AFET and WSABIE
      v. Label noise reduction in distant supervision
         A. Noisy type issue in distant supervision
         B. Simple pruning heuristics
         C. Partial-label learning methods
         D. Label noise reduction by heterogeneous partial-label embedding

4. Automated Extraction of Structured Entity Relationships
   (a) Preliminaries of relation extraction (RE)
      i. Basic concepts: relation instance, relation mention
      ii. Explicit relation vs. implicit relation
      iii. Downstream applications
         A. Knowledge base completion
         B. Question answering systems
   (b) Traditional supervised RE systems
      i. Supervised RE methods
         A. Supervised models
         B. Features for relation extraction
         C. Training data
         D. Evaluation of RE task
      ii. Systems from Stanford and IBM
   (c) Extracting typed relations from A Massive Corpus
      i. Weak supervision methods

A. Pattern-based boostrapping methods
B. Seed examples selection
C. DIPRE system
D. KnowItAll system
E. Snowball system

ii. Distant supervision (DS) methods
A. Distant supervision for RE: A typical workflow
B. Challenges of DS: noisy candidate labels
C. Noise-robust DS models

iii. Joint extraction of entities and relations
A. Supervised methods: linear programming and sequence models
B. CoType: A distantly-supervised method

5. Transfer Knowledge for Low Resource Information Extraction
   (a) Multi-task and multi-domain learning for named entity recognition
   (b) Cross-lingual entity extraction
   (c) Distilling linguistics knowledge into relation extraction system

6. Knowledge Base Reasoning: Background and State-of-the-Arts
   (a) Preliminaries
      i. KB Reasoning and Information Extraction
         A. Difference with IE
      ii. Challenges of KB Reasoning
         A. Noisy Background Knowledge
         B. Combinatorial explosion and huge search space
         C. Scalability
   (b) Path-Based Approaches
      i. The Path-Finding Algorithm
      ii. ProPPR
      iii. Combining PRA and Recurrent Neural Networks
   (c) Embedding-Based Approaches
      i. RESCAL
      ii. TransE
      iii. Other Recent Studies
   (d) DeepPath: Reinforcement Learning for KB Reasoning
      i. Problem Formulation
      ii. The DeepPath Algorithm
      iii. Imitation Learning
      iv. Experimental Results

7. Research Frontier

3 Organizers

Xiang Ren, Assistant Professor, Department of Computer Science, University of Southern California. His research focuses on creating computational tools for better understanding and exploring massive text data. He has published over 25 papers in major conferences. He received Google PhD Fellowship, KDD Rising Star by Microsoft, Yahoo!-DAIS Research Excellence Award, C. W. Gear Outstanding Graduate Student Award by UIUC and Yelp Dataset Challenge Award. Mr. Ren has rich experiences in delivering tutorials in major conferences, including SIGKDD 2015, SIGMOD 2016 and WWW 2017. Homepage: http://xren7.web.engr.illinois.edu/.

Nanyun Peng is a Research Assistant Professor at the Department of Computer Science, and a Computer Scientist at the Information Sciences Institute, University of Southern California. She is broadly interested in Natural Language Processing, Machine Learning, and Information Extraction. Her research focuses on low-resource information extraction, creative language generation, and phonology/morphology modeling. Nanyun is the recipient of the Johns Hopkins University 2016 Fred Jelinek Fellowship. She has a background in computational linguistics and economics and holds BAs in both. Home page: http://www.vnpeng.net.

William Wang is an Assistant Professor at the Department of Computer Science, University of California, Santa Barbara. He received his PhD from Carnegie Mellon University, where he worked on scalable probabilistic reasoning language ProPPR with William Cohen. He focuses on information extraction and he is the faculty author of DeepPath—the first deep reinforcement learning system for multi-hop knowledge reasoning. He has published more than 40 papers at leading conferences and journals including *ACL, EMNLP, NAACL, COLING, IJCAI, CIKM, SIGDIAL, IJCNLP, INTERSPEECH,*

ICASSP, ASRU, SLT, Machine Learning, and Computer Speech & Language, and he has received paper awards and honors from CIKM, ASRU, and EMNLP. Website: http://www.cs.ucsb.edu/~william/.

4 Previous Editions and Related Tutorials

A list of tutorials on the most related topics:

1. **Conference tutorial:** X. Ren, Y. Su, X. Yan, "Construction and Querying of Large-scale Knowledge Bases" (CIKM'17). http://xren7.web.engr.illinois.edu/tutorial-cikm17.html.
2. **Conference tutorial:** J. Pujara, S. Singh, B. Dalvi, "Knowledge Graph Construction From Text" (AAAI'17). https://kgtutorial.github.io/.
3. **Conference tutorial:** X. Ren, M. Jiang, J. Shang and J. Han, "Constructing Structured Information Networks from Massive Text Corpora" (WWW'17). http://xren7.web.engr.illinois.edu/www17tutorial.html.
4. **Conference tutorial:** W. Y. Wang, W. Cohen "Scalable Probabilistic Logics" (IJCAI'16). http://www.cs.cmu.edu/~yww/tutorials.html.
5. **Conference tutorial:** W. Y. Wang, W. Cohen "Statistical Relational Learning for NLP" (NAACL'16). http://www.aclweb.org/anthology/N16-4005.
6. **Conference tutorial:** E. Gabrilovichn, N. Usunier, "Constructing and Mining Web-scale Knowledge Graphs" (SIGIR'16). http://dl.acm.org/citation.cfm?id=2914807/.
7. **Conference tutorial:** X. Ren, A. El-Kishky, C. Wang and J. Han, "Automatic Entity Recognition and Typing in Massive Text Corpora" (WWW'16). http://web.engr.illinois.edu/~elkishk2/www2016/.
8. **Conference tutorial:** X. Ren, A. El-Kishky, C. Wang and J. Han, "Automatic Entity Recognition and Typing from Massive Text Corpora: A Phrase and Network Mining Approach" (SIGKDD'15). http://research.microsoft.com/en-us/people/chiw/kdd15tutorial.aspx.

Most of the previous tutorials focused exclusively on the knowledge base construction aspect. In the proposed tutorial, we will give a systematic discussion on the problem of knowledge base reasoning, for which extensive studies have been conducted recently but systematic tutorials are lacking. This tutorial also presents recent advances in applying distant and weak supervision to the extraction of structured facts in knowledge base construction, in addition to the traditional supervised techniques and rule-based approaches.

Target audience and prerequisites. Researchers and practitioners in the field of natural language processing, computational linguistic, text mining, information retrieval, semantic web and machine learning. 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 NLP, algorithms and their applications are needed. We expect there will be around 70 people interested in our tutorial.

Tutorial material and equipment. We will provide attendees a website and upload our tutorial materials (slides, references, softwares). There is no copyright issue. Standard equipment will be enough for our tutorial.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *ACM conference on Digital libraries*. pages 85–94.

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II* .

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*. pages 2787–2795.

Andrew Carlson, Justin Betteridge, Richard C Wang, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Coupled semi-supervised learning for information extraction. In *WSDM*.

Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*. ACM, pages 100–110.

Venkatesh Ganti, Arnd C König, and Rares Vernica. 2008. Entity categorization over large document collections. In *SIGKDD*.

Sonal Gupta and Christopher D. Manning. 2014. Improved pattern learning for bootstrapped entity extraction. In *CONLL*.

Yeye He and Dong Xin. 2011. Seisa: set expansion by iterative similarity aggregation. In *WWW*.

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*.

Ruihong Huang and Ellen Riloff. 2010. Inducing domain-specific semantic class taggers from (almost) nothing. In *ACL*.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. *ACL-HLT* .

Ni Lao, Tom Mitchell, and William W Cohen. 2011. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 529–539.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *ACL*.

Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan. 2013. Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 1070–1078.

Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. 2010. Annotating and searching web tables using entities, types and relationships. *VLDB* 3(1-2):1338–1347.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *AAAI*.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *SIGMOD*.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *EMNLP*.

Paul McNamee and James Mayfield. 2002. Entity extraction without language-specific resources. In *proceedings of the 6th conference on Natural language learning-Volume 20*. Association for Computational Linguistics, pages 1–4.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *ICML*. volume 11, pages 809–816.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisboa, Portugal.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media via learning segmentation representations .

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics* 5:101–115.

Vasin Punyakanok and Dan Roth. 2001. The use of classifiers in sequential inference. In *NIPS*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *ACL*.

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, and Jiawei Han. 2015. ClusType: Effective entity recognition and typing by relation phrase-based clustering. In *KDD*.

Xiang Ren, Wenqi He, Meng Qu, Lifu Huang, Heng Ji, and Jiawei Han. 2016a. AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding. In *EMNLP*.

Xiang Ren, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, and Jiawei Han. 2016b. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *KDD*.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. 2017. CoType: Joint extraction of typed entities and relations with knowledge bases. In *arXiv:1610.08763*.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *TKDE* (99):1–20.

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gur, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for qa evaluation. In *EMNLP*. pages 562–572.

Yizhou Sun and Jiawei Han. 2013. Mining heterogeneous information networks: a structural analysis approach. *SIGKDD Explorations* 14(2):20–28.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 1165–1174.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *ACL*.

William Yang Wang, Kathryn Mazaitis, and William W Cohen. 2013. Programming with personalized pagerank: a locally groundable first-order probabilistic logic. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, pages 2129–2138.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, pages 481–492.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. *EMNLP* .

Endong Xun, Changning Huang, and Ming Zhou. 2000. A unified statistical model for the identification of english basenp. In *ACL*.

Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. 2012. Infogather: entity augmentation and attribute discovery by holistic matching with web tables. In *SIGMOD*.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *ACL*.