Taming Unstructured Big Data: Automated Information Extraction from Massive Text

Xuan Wang, Yu Zhang, Qi Li, Jiawei Han

Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
{xwang174, yuz9, qili5, hanj}@illinois.edu

***Abstract*—Text data is a powerful information source that covers almost every aspect of our life. Automated information extraction has attracted considerable attention with various approaches developed to mine *structured* knowledge from *unstructured* text. In this tutorial, we present an organized picture of automated information extraction from massive text to answer the need of a systematic review and comparison of the techniques. We first introduce major tasks of information extraction such as named entity recognition and relation extraction. Then we introduce downstream applications such as heterogeneous information network construction and claim mining that utilize the extracted information. Specifically, we focus on the methods that are scalable, effective, minimum supervised and working on various kinds of text (e.g., news and biomedical science). We also demonstrate on a real-world dataset, PubMed that includes over 29 million biomedical literature, how the heterogeneous information network can be constructed and how the scientific claims can be automatically retrieved based on automated information extraction. The covered topics will be interesting to both advanced researchers and beginners in data mining, text mining, natural language processing and machine learning.**

I. Introduction

Goals and Subtopics. The goal of this tutorial is to give a full picture of automated information extraction from massive text (see tutorial contents and tutorial outline). We will discuss the following key issues: (1) named entity recognition with distant supervision from knowledge bases; (2) open relation extraction without predefined relation types; (3) heterogeneous information network construction from text with minimum supervision; (4) scientific claim retrieval based on automated information extraction.

Relevance to the Attendees. Text data is a powerful information source that covers almost every aspect of our life, including social media and health care. Automated information extraction has attracted considerable attentions with various approaches developed to mine *structured* knowledge from *unstructured* text. Traditionally, studies in NLP, machine learning or text mining communities focus on the methods for information extraction by obtaining a large amount of human annotated data and then train a machine learning model for information extraction. However, this paradigm suffers from the high cost and low efficiency in manual annotation. Recently, the data mining community has demonstrated the tremendous power of the data-driven methods for scalable and effective information extraction from massive text with minimum human supervision. Unfortunately, there lacks a systematic review of the recent progress for automated information extraction from massive text. This tutorial is to bridge this gap and explore the power of data-driven methods for automated information extraction.

II. Intended Audience and Pre-requisites

Targeted Audience. This tutorial is intended for researchers and practitioners in data mining, text mining, natural language processing and machine learning.

Content Level. The content level is 30% beginner, 30% intermediate, and 30% advanced.

Prerequisites. While the audience with a good background in the above areas would benefit most from this tutorial, we believe the material to be presented would give general audience and newcomers a complete picture of the current work, introduce important research topics in this field, and inspire them to learn more.

III. Tutorial Contents

A. Introduction

We will first introduce the big picture of information extraction from massive text. We will then introduce the four technical modules of the tutorial as follows.

B. Named Entity Recognition

Named entity recognition (NER) aims to identify text spans as candidate entities and classifies them into a set of semantic classes, such as person and organization. Existing studies leverage either human-annotated corpus (fully supervised) or knowledge bases (distantly supervised) to train a sequence labeling model. We will cover both directions.

Backgrounds and Supervised Methods. To start from the beginning, we will give an overview of the NER task and state-of-the-art supervised machine learning approaches [7]. Then we will introduce the popularly adopted neural architecture BiLSTM-CRF [10] and its variants (e.g., BiLSTM-CNN-CRF [15]). An application in biomedical NER [24] will also be covered as an example of integrating multiple corpora with different labeled entity types.

Distantly Supervised Methods. To alleviate human effort, dictionaries and knowledge bases have been applied to automatically annotated raw corpus for training. We will cover some most recent studies [8], [19] utilizing heuristic matching

rules and automatic phrase mining techniques when dealing with distant supervision.

C. Relation Extraction

Fully-supervised relation extraction (RE) approaches (e.g., [13], [22], [25]) aims to predict the relationship between a pair of entities in a sentence. These approaches encounter two challenges: (1) They usually require a large number of human-annotated sentences as training data, which are expensive to obtain. (2) The power of the trained relation classifier is confined to a given relation type set, which makes it hard to transfer the model to new relation types or new domains. After giving an overview of the RE task and fully-supervised methods, we will cover three types of minimum supervised approached tackling the above two challenges.

Pattern-based Methods. Pattern-based approaches propose to automatically extract textual patterns *without human supervision*. We will cover approaches utilizing linguistic features (e.g., ReVerb [5] and PATTY [16]) and introduce techniques grouping synonymous patterns using extracted relation tuples (e.g., MetaPAD [9]).

Open-domain Approaches. Open-domain information extraction relies on predicates to extract a subject and an object. We will introduce methods exploiting local context (e.g., ClausIE [3]) and incorporating global cohesiveness (e.g., ReMine [26]). Also, we will cover approaches dealing with high-arity relations [4] and its application in biomedical and clinical text.

Weakly-supervised Methods. Weakly-supervised methods require users to provide a small set of sentences or entity pairs for each relation type. We will introduce traditional pattern-based bootstrapping approaches (e.g., Snowball [1]) and pattern-enhanced embedding learning methods [17].

D. Heterogeneous Information Network Construction

By viewing entities as nodes and relations as edges, a heterogeneous information network (HIN) can be constructed to represent structured knowledge from unstructured corpora. We will cover the following two tasks of heterogeneous information network construction.

Network-based Document Summarization. Using a network to represent concepts and their relationships will bring structures into document summaries. Some attempts have been made to automatically construct concept networks from text. We will review existing studies covering education-related corpora [6], [23] and biomedical literature [20].

Factual Network Construction and Exploration. Automatically constructing structured networks from large amounts of background documents can support efficient exploration of structured factual knowledge. We will introduce Life-iNet [18], a network-based system to explore relationships between genes, drugs and diseases.

E. Claim Mining

Claim mining aims to automatically extract argumentative sentences from the text. There are two categories of claim mining methods based on the input: context-independent and context-dependent claim mining. We will first introduce the task of claim mining and the state-of-the-art supervised methods [11], [14]. Then we will introduce several recent studies that utilize limited human-annotated training data for claim mining [2], [12], [21].

Context-independent claim mining. Context-independent claim mining aims at detecting claims in a document without necessarily resorting to the contextual information [14]. It does not require any pre-defined topics and classifies every sentence to judge if it is a claim.

Context-dependent claim mining Context-dependent claim mining aims at detecting claims that are specifically related to a pre-defined concrete context [11]. There are two basic concepts for context-dependent claim mining: (1) topic: a short query that frames the discussion, and (2) context-dependent claim (CDC): a concise statement retrieved from the text that directly supports or contests the given topic.

Finally, we will conclude our tutorial by demonstrating on a real-world dataset, PubMed that includes over 29 million biomedical literature, how the heterogeneous information network can be constructed and how the scientific claims can be automatically retrieved based on automated information extraction. We will also summarize the techniques and discuss future directions.

IV. Outline of the Tutorial

A detailed outline of the topics that will be covered in the tutorial is presented as follows.

- Introduction
  - Motivations
  - Overview of automatic information extraction from massive text
  - Application of utilizing the extracted information
- Named Entity Recognition
  - What is Named Entity Recognition (NER)?
  - Traditional Supervised Methods
    - CorNLL03 Shared Task
    - Sequence Labeling Framework
    - Conditional Random Fields (CRFs)
    - Handcrafted Features
  - Modern Neural Models
    - Bidirectional Long Short-term Memory (BiLSTM)-based Models
    - Language Model and Contextualized Representations
    - End-to-end Neural Models
  - Distantly Supervised Methods
    - Entity Typing
    - Learning from Domain-Specific Dictionaries
- Relation Extraction
  - What is Relation Extraction (RE)?
  - Supervised Methods

    - Multi-relational Embedding
    - Position-aware Neural Models
    - Dependency-path-based Neural Models
  - Pattern-based Methods
    - Sequential Textual Patterns
    - Patterns with Linguistic Features
    - Synonym Pattern Grouping
  - Open-domain Approaches
    - How to exploit local structure?
    - How to exploit global consistency?
    - High-arity OpenIE
  - Weakly-supervised Methods
    - Bootstrapping Methods
    - Pattern-enhanced Embedding Learning
- Heterogeneous Information Network Construction
  - What is a Heterogenous Information Network (HIN)?
  - Network-based Document Summarization
    - What is a concept map?
    - Constructing Concept Maps from Text
  - Factual Network Construction and Exploration
    - Network Construction
    - Factual Knowledge Exploration
- Claim Mining
  - What is a claim?
  - Comparison between Opinion Mining, Argument Mining and Claim Mining
  - Context-independent claim mining
    - What is context-independent claim mining?
    - Supervised Machine Learning Methods
  - Context-dependent claim mining
    - What is a topic and a context-dependent claim?
    - Supervised Machine Learning Methods
    - Weakly/Distantly-supervised Methods
    - Unsupervised Claim Mining
- System Demos
  - Named Entity Recognition & APIs
  - Open Relation Extraction & APIs
  - Scientific Claim Mining & APIs
  - BioText Mining Demo.
- Summary and Future Directions
  - Summary
    - Principles and Techniques
    - Advantages and Limitations
    - How to choose a method based on your application?
  - Future Directions

V. Duration of the tutorial

The duration of the tutorial will be 2 hours.

VI. Agreement for the Note Release

We agree to release the notes of this tutorial on the IEEE Big Data 2019 tutorial digital media.

References

[1] E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In *ACM DL'00*, pages 85–94.

[2] K. Al-Khatib, H. Wachsmuth, M. Hagen, J. Köhler, and B. Stein. Cross-domain mining of argumentative text through distant supervision. In *NAACL'16*, pages 1395–1404.

[3] L. Del Corro and R. Gemulla. Clausie: clause-based open information extraction. In *WWW'13*, pages 355–366.

[4] P. Ernst, A. Siu, and G. Weikum. Highlife: Higher-arity fact harvesting. In *WWW'18*, pages 1013–1022.

[5] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP'11*, pages 1535–1545.

[6] T. Falke and I. Gurevych. Bringing structure into summaries: Crowdsourcing a benchmark corpus of concept maps. In *EMNLP'17*, pages 2951–2961.

[7] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL'05*, pages 363–370, 2005.

[8] J. Fries, S. Wu, A. Ratner, and C. Ré. Swellshark: A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*, 2017.

[9] M. Jiang, J. Shang, T. Cassidy, X. Ren, L. M. Kaplan, T. P. Hanratty, and J. Han. Metapad: Meta pattern discovery from massive text corpora. In *KDD'17*, pages 877–886.

[10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *NAACL-HLT'16*, pages 260–270.

[11] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim. Context dependent claim detection. In *COLING'14*, pages 1489–1500.

[12] R. Levy, S. Gretz, B. Sznajder, S. Hummel, R. Aharonov, and N. Slonim. Unsupervised corpus–wide claim detection. In *Proc. Work. Arg. Min.*, pages 79–84, 2017.

[13] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun. Neural relation extraction with selective attention over instances. In *ACL'16*, pages 2124–2133.

[14] M. Lippi and P. Torroni. Context-independent claim detection for argument mining. In *IJCAI'15*, pages 185–191.

[15] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL'16*, pages 1064–1074.

[16] N. Nakashole, G. Weikum, and F. Suchanek. Patty: a taxonomy of relational patterns with semantic types. In *EMNLP'12*, pages 1135–1145.

[17] M. Qu, X. Ren, Y. Zhang, and J. Han. Weakly-supervised relation extraction by pattern-enhanced embedding learning. In *WWW'18*, pages 1257–1266.

[18] X. Ren, J. Shen, M. Qu, X. Wang, Z. Wu, Q. Zhu, M. Jiang, F. Tao, S. Sinha, D. Liem, et al. Life-inet: A structured network-based knowledge exploration and analytics system for life sciences. *ACL'17 Demo*, pages 55–60.

[19] J. Shang, L. Liu, X. Gu, X. Ren, T. Ren, and J. Han. Learning named entity tagger using domain-specific dictionary. In *EMNLP'18*, pages 2054–2064.

[20] J. Shang, Q. Zhu, J. Shen, X. Wang, X. Gu, L. Kaplan, T. Harratty, and J. Han. Autonet: Automated network construction and exploration system from domain-specific corpora. *KDD'18 Demo*, 2018.

[21] E. Shnarch, C. Alzate, L. Dankin, M. Gleize, Y. Hou, L. Choshen, R. Aharonov, and N. Slonim. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *NAACL'18*.

[22] V. Shwartz, Y. Goldberg, and I. Dagan. Improving hypernymy detection with an integrated path-based and distributional method. In *ACL'16*, pages 2389–2398.

[23] C. Tauchmann, T. Arnold, A. Hanselowski, C. M. Meyer, and M. Mieskes. Beyond generic summarization: A multi-faceted hierarchical summarization corpus of large heterogeneous data. In *LREC'18*.

[24] X. Wang, Y. Zhang, X. Ren, M. Zitnik, J. Shang, C. Langlotz, and J. Han. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 2018.

[25] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning. Position-aware attention and supervised data improve slot filling. In *EMNLP'17*, pages 35–45.

[26] Q. Zhu, X. Ren, J. Shang, Y. Zhang, A. El-Kishky, and J. Han. Integrating local context and global cohesiveness for open information extraction. In *WSDM'19*, pages 42–50.