Mining Structures from Massive Text Data: A Data-Driven Approach

Jiawei Han
Abel Bliss Professor, Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
hanj@illinois.edu

Abstract

The real-world big data are largely unstructured, interconnected, and in the form of natural language text. One of the grand challenges is to mine structures from such massive unstructured data, and transform such big data into structured networks and actionable knowledge. We propose a text mining approach that requires only distant supervision or minimal supervision but relies on massive data. We show that quality phrases can be mined from such massive text data, types can be extracted from massive text data with distant supervision, and entity-attribute-value triples can be extracted from meta-patterns discovered from such data. Finally, we propose a data-to-network-to-knowledge paradigm, that is, first turn data into relatively structured information networks, and then mine such text-rich and structure-rich networks to generate useful knowledge. We show such a paradigm represents a promising direction at turning massive text data into structured networks and useful knowledge.

1 Introduction

The success of data mining technology is largely attributed to the efficient and effective analysis of structured data. The construction of a well-structured, machine-actionable database 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`, `company`, `product`), attributes (*e.g.*, `age`, `weight`), relations (*e.g.*, `employee_of`, `manufacture`) from massive text corpora. By integrating these semantic-rich structures with other inter-related structured data (*e.g.*, `product specification`, `user transaction log`), one can construct a powerful StructDB as a conceptual abstraction of the original text corpora. The uncovered StructDBs will facilitate browsing information and inferring knowledge that are otherwise locked in the text corpora. Computers can effectively perform algorithmic analysis at a large scale over these StructDBs and apply the new insights and knowledge to improve human productivity in various downstream tasks. Our phrase mining tool, SegPhrase (Jialu 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 (Xiang Ren, et al., 2015), was shipped as part of the products in Microsoft Bing and U.S. Army Research Lab.

The remaining of the paper is organized as follows. Section 2 introduces our recent work on automated mining of quality phrases from massive corpora. Section 3 introduces our recent studies on automated recognition and typing of entities and relations with distant supervision. Section 4 presents our initial study on meta-pattern discovery and its application to information extraction. We conclude our study in Section 5 by pointing out some future research topics on turning massive unstructured data into structured knowledge

2 Automated Quality Phrase Mining

Concepts are words and phrases that represent terms or ideas that people are interested in. A lot of concepts, especially scientific concepts, are in the form of phrases and are not restricted to noun phrases (*e.g.*, "*NP Complete*" and "*Learning to Rank*"). Concepts are also often arranged in hi-

[1] `http://www.yelp.com/dataset_challenge`
[2] `http://engineering.tripadvisor.com/mining-text-review-snippets/`

erarchies where each node is a topic represented by a ranked list of concepts (*e.g.*, {'*social network analysis*', '*mining information networks*', ...}, is a child node of a general topic node: {'*knowledge discovery*', '*data mining*', ...}). Such hierarchical organization of concepts allows exploration of corpus at varied granularity, and has applications like visualization, search and summarization.

The NLP community has conducted extensive studies on automatic extraction of quality phrases, but mostly rely on many kinds of linguistic processing (*e.g.*, *chunking, dependency parsing*), domain-dependent language rules, and a large amount of labeled data (*e.g.*, *treebanks*).

In our recent research, we have developed several interesting automated phrase mining methods. The general philosophy is that instead of relying on explicit training, we explore statistical redundancy in document collections by frequent-pattern mining and semi-supervised learning. Such data-driven approaches leverage statistical or heuristic measures derived from corpus and achieve impressive results. Our newly developed phrase mining approach consists of three methods: (1) *unsupervised approach* (*i.e.*, requiring neither expert explicitly labeled training data nor knowledge-base), represented by ToPMine (Ahmed El-Kishky, et al., 2014), (2) *weakly supervised approach* (*i.e.*, requiring a small set of human labeled training data on the quality of phrases), represented by SegPhrase (Jialu Liu, et al., 2015), and (3) *distantly-supervised approach* (*i.e.*, requiring only distantly labeled knowledge-bases, such as Wikipedia), represented by AutoPhrase (Jialu Liu, et al., 2017; Jingbo Shang, et al., 2017).

Our experiments on large text corpora show ToPMine and SegPhrase, with minor adaptation, generate quality phrases in large corpora of multiple languages (*e.g.*, English, Arabic, Chinese and Spanish) since both methods rely mainly on statistical analysis instead of language parsing and linguistic features. For AutoPhrase, it demonstrates additional power over Segphrase on four aspects: (i) *minimized human effort*, using a robust positive-only distant training method which estimates the phrase quality by leveraging existing general knowledge bases; (ii) *supporting multiple languages* including English, Spanish, and Chinese, where the language in the input will be automatically detected, (iii) *high accuracy*, using a POS-guided phrasal segmentation model incorporating POS tags when POS tagger is available, and moreover, the new framework is able to extract single-word quality phrases; and (iv) *high efficiency*, due to a better indexing method and an almost lock-free parallelization, which lead to both running time speedup and memory saving.

3 Distantly Supervised Entity/Relation Recognition and Typing

Extracting entities and relations for types of interest from text is important for understanding massive text corpora. Traditionally, systems of entity relation extraction have been relying on human-annotated corpora for training and adopted an incremental pipeline. Such systems require additional human expertise to be ported to a new domain and are vulnerable to errors cascading down the pipeline.

Recently, we have investigated a distantly supervised approach for extraction and typing of entities and relations and developed several interesting methods to reduce human effort and enhance the performance. These include (1) ClusType (Xiang Ren, et al., 2015), which explores an integrated, entity typing and relation-phrase clustering approach, (2) PLE (Xiang Ren, et al., 2016) for refined entity typing, and (3) Co-Type (Xiang Ren, et al., 2017) for jointly embedding and typing entities and relations in a mutually enhanced framework.

ClusType (Xiang Ren, et al., 2015) explores data-driven phrase mining to generate entity mention candidates and relation phrases, and enforces the principle that relation phrases should be softly clustered when propagating type information between their argument entities. Then the method predicts the type of each entity mention based on the type signatures of its co-occurring relation phrases and the type indicators of its surface name, as computed over the corpus. The two tasks, type propagation with relation phrases and multi-view relation phrase clustering, are put in a joint optimization framework and achieves high performance.

For extraction and typing of fine-grained entity types in conjunction with existing knowledge bases, a major difficulty is that the type labels obtained from knowledge bases are often noisy (*i.e.*, incorrect for the entity mentions' local context). We proposed a framework, called PLE (Xiang Ren, et al., 2016), which conducts Label

Noise Reduction in Entity Typing (LNR), to automatically identify correct type labels (type-paths) for training examples, given the set of candidate type labels obtained by distant supervision with a given type hierarchy. PLE jointly embeds entity mentions, text features and entity types into the same low-dimensional space where objects whose types are semantically close have similar representations. Then we estimate the type-path for each training example in a top-down manner using the learned embeddings. We formulate a global objective for learning the embeddings from text corpora and knowledge bases, which adopts a novel margin-based loss that is robust to noisy labels and faithfully models type correlation derived from knowledge bases.

To Further enhance the overall performance for entity and relation extraction and typing, We propose a novel domain-independent framework, called Co-Type (Xiang Ren, et al., 2017), that runs a data-driven text segmentation algorithm to extract entity mentions, and jointly embeds entity mentions, relation mentions, text features and type labels into two low-dimensional spaces (for entity and relation mentions respectively), where, in each space, objects whose types are close will also have similar representations. COTYPE, then using these learned embeddings, estimates the types of test (unlinkable) mentions. We formulate a joint optimization problem to learn embeddings from text corpora and knowledge bases, adopting a novel partial-label loss function for noisy labeled data and introducing an object "translation" function to capture the cross-constraints of entities and relations on each other and achieved high performance over existing embedding-based methods.

4 Meta-Pattern Guided Information Extraction

Mining textual patterns in news, tweets, papers, and many other kinds of text corpora may facilitate effective information extraction from massive text corpora. Previous studies adopt a dependency parsing-based pattern discovery approach. However, the parsing results lose rich context around entities in the patterns, and the process is costly for a corpus of large scale. Recently, we have proposed a typed textual pattern structure, called *meta pattern*, to represent a general form of frequent, informative, and precise subsequence patterns in certain context. We propose an efficient framework, called *MetaPAD* (Meng Jiang, et al., 2017), which discovers meta patterns from massive corpora with three techniques: (1) it develops a context-aware segmentation method to carefully determine the boundaries of patterns with a learned pattern quality assessment function, which avoids costly dependency parsing and generates high-quality patterns; (2) it identifies and groups synonymous meta patterns from multiple facets—their types, contexts, and extractions; and (3) it examines type distributions of entities in the instances extracted by each group of patterns, and looks for appropriate type levels to make discovered patterns precise.

Our extensive experiments demonstrate that our proposed framework discovers high-quality typed textual patterns efficiently from different genres of massive corpora and facilitates information extraction. For example, from an Associate Press and Reuter dataset (APR 2015), one can discover meta-patterns for *country* and *president* and extract *country- president* pairs even for rarely mentioned pairs, like `Burkina Faso-Blaise Compaoré`, and find which bacteria are resistant to which antibiotics from the PubMed abstracts.

5 Conclusions and Future work

Mining structures from massive text copora is an important task for turning big text data into big structured knowledge. Traditional approaches relying on extensive human labeling or annotation of a nontrivial sample set of documents in specific application domain are not scalable. A new direction is to develop effective weakly or distantly supervised methods to explore existing domain-agnostic labels and massive existing text corpora to achieve high performance on phrase mining, entity and relation extraction and typing, and information extraction.

Our recent development of phrase mining methods, such as ToPMine, SegPhrase and AutoPhrase, entity/relation recognition and typing methods such as ClusType, PLE and CoType, as well as pattern-based discovery with massive text corpora, such as MetaPAD, contribute to this direction.

There are a lot of future research problems along this direction. Besides further consolidating these distantly supervised methods, an important direction is to study automated multi-faceted taxonomy direction from massive text to turn extracted concepts (*e.g.*, phrases) into organized

structures as well as identifying trusted claims and comparative and succinct summaries, and build up structured, multi-dimensional text-cubes and information networks, from massive data. We have been working along these lines and developing some new methods, such as SetExpan (Jiaming Shen, et al., 2017), REHession (Liyuan Liu, et al., 2017) and indirect supervision for relation extraction using question-answer pairs (JZeqiu Wu, et al., 2018). Still, this is a huge and promising area, with a vast unexplored territory waiting to be explored.

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 16-18481, IIS 17-04532, and IIS-17-41317, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. Scalable topical phrase mining from text corpora. *PVLDB*, 8(3):305–316, 2014.

Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance Kaplan, Timothy Hanratty, and Jiawei Han. MetaPAD: Meta patten discovery from massive text corpora. In *Proc. 2017 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'17)*, Halifax, Nova Scotia, Canada, Aug. 2017.

Jialu Liu, Jingbo Shang, and Jiawei Han. *Phrase Mining from Massive Text and Its Applications*. Morgan & Claypool Publishers, 2017.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. Mining quality phrases from massive text corpora. In *Proc. 2015 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'15)*, Melbourne, Australia, May 2015.

Liyuan Liu, Xiang Ren, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji, and Jiawei Han. Heterogeneous supervision for relation extraction: A representation learning approach. In *Proc. of 2017 Conf. on Empirical Methods in Natural Language Processing EMNLP'17*, pages 46–56, Copenhagen, Denmark, Sept. 2017.

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, Heng Ji, and Jiawei Han. ClusType: Effective entity recognition and typing by relation phrase-based clustering. In *Proc. 2015 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'15)*, Sydney, Australia, Aug. 2015.

Xiang Ren, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, and Jiawei Han. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *Proc. of 2016 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1825–1834, 2016.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare Voss, Heng Ji, Tarek Abdelzaher, and Jiawei Han. CoType: Joint extraction of typed entities and relations with knowledge bases. In *Proc. 2017 World-Wide Web Conf. (WWW'17)*, Perth, Australia, Apr. 2017.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. Automated phrase mining from massive text corpora. *CoRR*, abs/1702.04457, 2017.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Jingbo Shang, Xiang Ren, and Jiawei Han. SetExpan: Corpus-based set expansion via context feature selection and rank ensemble. In *Proc. 2017 European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD'17)*, Skopje, Macedonia, Sept. 2017.

Zeqiu Wu, Xiang Ren, Frank F. Xu, Ji Li, and Jiawei Han. Indirect supervision for relation extraction using question-answer pairs. In *Proc. of 2018 ACM Int. Conf. on Web Search and Data Mining (WSDM'18)*, Los Angeles, CA, Feb. 2018.