AutoNet: Automated Network Construction and Exploration System from Domain-Specific Corpora

Jingbo Shang[1], Qi Zhu[1], Jiaming Shen[1], Xuan Wang[1], Xiaotao Gu[1],
Lance Kaplan[2], Timothy Harratty[2], Jiawei Han[1]
[1]Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA
[2]US Army Research Laboratory, MD, USA
[1]{shang7, qiz3, js2, xwang174, xiaotao2, hanj}@illinois.edu, [2]{lance.m.kaplan.civ, timothy.p.hanratty.civ}@mail.mil

ABSTRACT

As a collaborative project funded by US Army Research Lab, our goal is to turn massive unstructured text data into structured heterogeneous information networks (HINs), on which actionable knowledge can be further uncovered flexibly and effectively based on user's instructions. Taking advantage of open knowledge bases, we develop an end-to-end, data-driven system, AutoNet, with no additional human curation and annotation. AutoNet constructs a large-scale HIN from massive (user-provided) domain-specific text corpora (e.g., scientific papers) using our innovative phrase mining, entity typing, and relation extraction methods, and saves these models for later usage. After that, AutoNet supports two real-time functions: (1) *discovery*: given a few user-provided documents, AutoNet will construct a new HIN on the fly and highlight those new nodes (i.e., entities) and/or edges (i.e., relations), which are not in the pre-stored network; and (2) *exploration*: given some user-provided keywords, AutoNet will retrieve a related subnetwork from the large pre-stored HIN. We further design effective visualization tools for both functions. A demo video is available[1].

CCS CONCEPTS

• **Information systems → Information extraction**;

KEYWORDS

Phrase Mining, Entity Recognition, Relation Extraction, Heterogeneous Information Network, Massive Texts

ACM Reference Format:
Jingbo Shang[1], Qi Zhu[1], Jiaming Shen[1], Xuan Wang[1], Xiaotao Gu[1], Lance Kaplan[2], Timothy Harratty[2], Jiawei Han[1]. 2018. AutoNet: Automated Network Construction and Exploration System from Domain-Specific Corpora. In *Proceedings of The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'18).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

INTRODUCTION

The majority of the massive volume of real world data consists of unstructured or loosely structured text, ranging from news to social media, web contents, scientific papers, government documents, and business contracts. The sheer size of such data and the fast pace of new data generation make many existing approaches unscalable and infeasible, due to their reliance on heavy human annotation and curation at the extraction of named entities and their relationships as well as the construction of knowledge graphs. Therefore, automated structure discovery and construction from massive text corpora have become an active research area in the fields of data mining, machine learning, and natural language processing.

We propose a novel and principled *data-driven approach* for *automatic knowledge discovery* in massive, unstructured, and noisy text corpora by constructing *high-quality and structured* heterogeneous information networks (HINs), in a distantly-supervised manner. Note that the HINs that we propose to construct provide stronger typed and structural information than typical designs of knowledge graphs and thus endowing stronger power for mining and inference as shown in [8]. Moreover, our proposed approach is general, extensible to text corpora in *multiple natural languages* and across *multiple domains*. Therefore, instead of common-sense knowledge, the HINs will be directly constructed from the given corpus, and thus can uncover the domain-specific knowledge.

The system will be open-sourced. Consequently, users can build new models using their own data without privacy concerns.

MAIN INNOVATIONS

Fig. 1 shows the workflow for AutoNet depicting its two main innovative features: 1) model learning and network construction, and 2) network exploration and construction on the fly.

Model Learning and Network Construction. AutoNet only requires massive unlabeled texts and existing knowledge bases (KBs), and then learns a series of models, without additional human effort. Such models mine structures (i.e., entities and relations) from text using minimal language- or domain- dependent features. Therefore, the user can easily adapt AutoNet to his/her own domain/language by providing new corpus/KB. Finally, AutoNet constructs a large-scale HIN based on mined structures, builds indices, and stores these models for online exploration and discovery.

Network Exploration and Construction on the Fly. AutoNet will retrieve related nodes and edges, if any, from the large-scale HIN and construct new HINs on the fly from user-provided documents with a similar network construction process guided by the saved models. Also, a user can explore the subnetwork by keywords. The built indices will facilitate efficient selection. An interactive network visualizer enables effective explorations. The node color reflects its type, the node size shows its popularity, and the link thickness means its frequency. As our relation extractor ReMine [9]

[1]http://dmserv4.cs.illinois.edu/AutoNet-Demo/

KDD'18, August, 2018, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

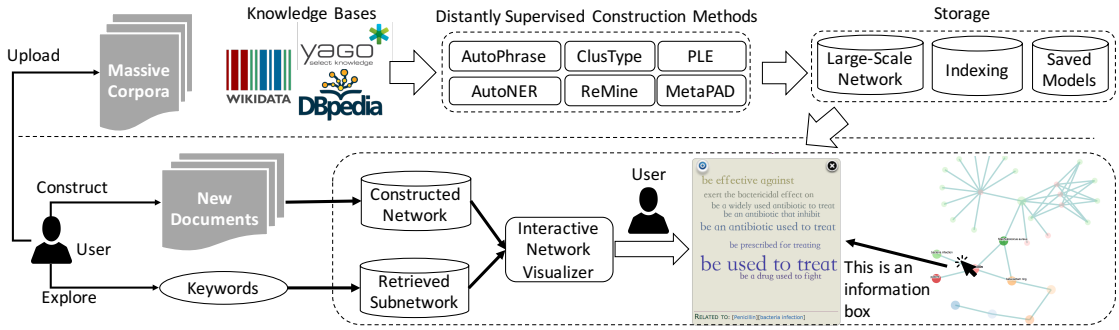

Figure 1: An overview of the AutoNet **system. 1) Top: Offline Construction and 2) Bottom: Online Discovery and Exploration.**

provides relational phrases, we summarize every relation between two entities using a word cloud of all its relational phrases weighted by their frequencies. Moreover, for each relation, AutoNet presents its grounded documents to the user for further investigation.

To our best knowledge, AutoNet is the first system that can construct HINs in the user-specified domain and language. Specifically, it has the following three innovative components.

1. Phrase Mining. We have successfully developed two novel phrase mining methods, SegPhrase [2] and AutoPhrase [7]. They can automatically extract high-quality phrases from domain-specific text corpora written in different languages under light supervision or distant supervision. It's worth mentioning that our phrase mining tools have received Yelp Dataset Challenge Grand Prize[2] and reported by TripAdvisor in their business usage[3].

2. Entity Recognition. At the corpus-level, we have recently developed ClusType [4] and PLE [5], two distantly-supervised models for coarse-grained and fine-grained typings at the corpus-level. At the sentence-level, our LM-LSTM-CRF model has achieved the state-of-the-art on benchmark datasets under the supervised setting without any other external resources [3]. Going beyond, we have developed another distantly supervised sentence-level entity recognition model, AutoNER, which is under review now.

3. Relation Extraction and Attribute Discovery. ReMine [9] is a novel distantly supervised open-domain information extraction (Open IE) method. It can extract high-confidence relational phrases from domain-specific texts in an end-to-end manner, therefore, it receives WWW'18 best poster award honorable mentioning. Built upon phrase mining methods, we have developed MetaPAD [1] to extract attribute names and values with light effort.

Previous Efforts and Limitations. Domain-specific search engines, such as PubMed, using keywords and Medical Subject Headings (MeSH) terms might not work well on capturing cross-document entity relations or identifying publications related to these relations. Life-iNet is a recently proposed network-based knowledge exploration system [6], however, it cannot support online network construction and also can only explore pre-defined relations.

DEMONSTRATION & REPOSITORIES

Some Statistics. Within a few hours, AutoNet can construct a HIN of more than 64 million nodes and 186 million edges based on 2.93 million Cancer-related PubMed papers and the MeSH database, or a HIN of more than 40 thousand nodes and 110 thousand edges based on 2.77 million computer science paper abstracts and Wikipedia.

Stable System. The stable version of the system will be open-sourced on the author's GitHub later.

Public Repos. Our key methods used in this system have received **over 858 stars** on GitHub as follows.

- AutoPhrase: https://github.com/shangjingbo1226/AutoPhrase
- SegPhrase: https://github.com/shangjingbo1226/SegPhrase
- LM-LSTM-CRF: https://github.com/LiyuanLucasLiu/LM-LSTM-CRF
- ClusType: https://github.com/shanzhenren/ClusType
- PLE: https://github.com/shanzhenren/PLE
- ReMine: https://github.com/GentleZhu/ReMine
- MetaPAD: https://github.com/mjiang89/MetaPAD

ACKNOWLEDGEMENTS

Research was sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreement No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HDTRA11810026, Google PhD Fellowship, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). Any opinions, findings, and conclusions or recommendations expressed in this document are those of the author(s) and should not be interpreted as the views of any U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] 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 *SIGKDD*, 2017.
[2] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining quality phrases from massive text corpora. In *SIGMOD*, 2015.
[3] L. Liu, J. Shang, F. Xu, X. Ren, H. Gui, J. Peng, and J. Han. Empower sequence labeling with task-aware neural language model. *AAAI*, 2017.
[4] 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.
[5] X. Ren, W. He, M. Qu, C. R. Voss, H. Ji, and J. Han. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *SIGKDD*, 2016.
[6] 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, System Demonstrations*, 2017.
[7] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han. Automated phrase mining from massive text corpora. *TKDE*, 2018.
[8] Y. Sun and J. Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2012.
[9] Q. Zhu, X. Ren, J. Shang, Y. Zhang, F. F. Xu, and J. Han. Open information extraction with global structure constraints. In *WWW, poster*, 2018.

[2]https://www.yelp.com/dataset/challenge/winners
[3]http://engineering.tripadvisor.com/using-nlp-to-find-interesting-collections-of-hotels/