Research Statement Xuan Wang

My research interests are in **Natural Language Processing** and **Data Mining**, emphasizing **applications to Biological and Health Sciences**. Natural language processing (NLP) and text mining are promising for advancing human knowledge in many fields, given the rapidly growing volume of text data (e.g., scientific articles, medical notes, and news reports) we are seeing nowadays. Recently, there has been a growing interest in bringing NLP and text mining to scientific discovery. For example, materials scientists have demonstrated that unsupervised word embeddings capture complex materials science concepts without explicit chemical knowledge and recommend materials for functional applications several years before their discovery [2]. I envision tremendous opportunities in this emerging area of using modern NLP and text mining techniques to enable and accelerate scientific discovery.

My current research theme is **developing effective and scalable algorithms and systems for automatically understanding massive text data to enable and accelerate biomedical discovery**. My **highly interdisciplinary background** has placed me in a suitable position to conduct research in this emerging area of AI-driven scientific knowledge discovery. I have solid training in biochemistry (**B.S. in biological science** and **M.S. in biochemistry)**, statistics (**M.S. in statistics**), and computer science (**Ph.D. in computer science**). I have published about 20 research/demo papers in both top-tier NLP conferences (e.g., ACL, EMNLP, and NAACL) and biomedical informatics journals (e.g., Bioinformatics) and conferences (e.g., ACM-BCB and IEEE-BIBM). These papers have obtained more than 370 citations. My research has primarily focused on two directions: (1) **information extraction with minimum human supervision** [4-12] and (2) **literature-based knowledge discovery in biomedicine** [3,13-14].

My research benefits from and fosters collaborations with experts in various research areas **within and beyond computer science** from various institutions, including **hospitals** (UC Davis Medical Center), **government** (National Institute of Health and Army Research Lab), **industry** (IBM and Eli Lilly), and **academics** from other universities (Stanford, UCLA, UCSD, USC, Purdue, and Iowa State University). For example, I collaborated with UC Davis Medical School to develop a text mining method that identifies cardiovascular proteins specifically associated with six sub-categories of heart diseases [3]. This method enables a **precision medicine** approach to find new forms of treatment for patients with preserved ejection fraction (HFpEF), which has led to a **recently funded $1.5 million NIH project**. Encouraged by this and other similarly productive collaborations (see §1.1 and §1.2), I plan to **open new research directions** in using NLP and text mining for scientific discovery in the next few years. My short-term goal is to develop **knowledge-enhanced, condition-aware, and multi-modal NLP and text mining approaches for scientific discovery**. My long-term goal is to achieve **AI-driven knowledge discovery with real-world impact** on various applications, such as information systems for health care and biomedicine, AI-driven systems for molecular discovery and synthetic strategy designing, knowledge graph construction, web mining, environmental monitoring, smart city, and cyber-physical systems.

1. Research Contributions

1.1. Information Extraction with Minimum Human Supervision

With the growing volume of text data and the breadth of information, it is inefficient or nearly impossible for humans to manually find, integrate, and digest useful information. A major challenge is to develop methods that automatically understand massive unstructured text data with minimum human effort. To address this challenge, I have been developing methods that extract information from text with minimum human supervision. I have contributed a series of algorithms and systems for two tasks: (1) named entity recognition with distant supervision and (2) open relation extraction.

Named Entity Recognition (NER) aims to locate and classify entity mentions (e.g., "United States") from text into pre-defined categories (e.g., "countries"). Scientific literature analysis needs dozens to hundreds of distinct,

fine-grained entity types (e.g., more than 100 biomedical entity types in the Unified Medical Language System [UMLS] database), making consistent and accurate annotation difficult even for crowds of domain experts. However, domain-specific ontologies and knowledge bases can be easily accessed, constructed, or integrated, making distant supervision realistic for fine-grained scientific NER tasks. In distant supervision, training labels are automatically generated by matching the mentions in text with the concepts in the knowledge bases. A major challenge of distant supervision is the limited coverage of the dictionaries from the knowledge bases, leading to false-negative errors in distant label generation.

To tackle the challenge of incomplete dictionaries for distant label generation, I have proposed several distantly supervised NER methods [4-10] that effectively deal with noisy distant supervision. One example is **PatNER** [7] that automatically mines the entity naming principles (e.g., disease entities often contain the words "syndrome" or "disorder") to enhance distant supervision. PatNER **outperforms the state-of-the-art supervised NER method** (i.e., BioBERT [15]) on the NCBI-Disease dataset by a substantial margin (**0.89 versus 0.92 in F1 score**) while requiring no human supervision. I extended PatNER into a comprehensive named entity annotation system, **CORD-NER** [8], that automatically annotates 75 fine-grained biomedical entity types in the COVID-19 literature. CORD-NER was presented at the **ISMB'20** and **has also been used for downstream applications such as COVID-19 knowledge graph construction and drug repurposing report generation** [16].

In my recent paper published in **EMNLP'21**, I proposed **ChemNER** [9], an ontology-guided, distantly supervised method for fine-grained chemistry NER. ChemNER leverages the chemistry type ontology structure to provide a global topic constrain for context-aware multi-type disambiguation. It is highly effective, **substantially outperforming the state-of-the-art supervised NER methods** (i.e., RoBERTa [17] and ChemBERTa [18]), **improving the F1 score from ~0.2 to 0.45**. Collaborating with chemists, we are moving forward to develop a chemistry reaction tracker system that uses ChemNER and information retrieval methods to track chemistry research publications related to particular organic chemical reactions based on user-given queries. We expect this system will significantly benefit the query-based tracking of scientific publications.

Open Relation Extraction requires no pre-specified relation types (e.g., DRUG treat DISEASE) but aims to extract all the relation tuples (e.g., (Denosumab, treat, Osteoporosis)) from a text corpus. Meta-pattern discovery methods utilize data redundancy to derive informative frequent patterns and use the derived patterns as relation types for open relation extraction. Compared to existing open information extraction (OpenIE) methods, meta-pattern discovery produces a more structured relationship that can be used in downstream applications. However, existing meta-pattern discovery methods cannot extract patterns spanning long and complex sentences, which greatly limits their performance in the scientific domains.

I have proposed several meta-pattern-guided open relation extraction methods [11-12] to address the above challenge. For example, I developed **CPIE** [11] that extracts meta-patterns (textual patterns containing entity type tokens, e.g., DRUG treat DISEASE) spanning long and complex sentences in biomedical literature. CPIE breaks down long sentences into shorter yet meaningful segments. It then conducts meta-pattern mining to discover high-quality meta-patterns. Last, it hierarchically groups synonym meta-patterns to better understand and organize patterns. CPIE achieves the highest precision compared to the state-of-the-art OpenIE baselines and keeps the distinctiveness and simplicity of the extracted relation tuples. CPIE is highly effective in extracting rich information from large-scale biomedical literature. **It has been used for downstream applications such as textual evidence discovery in life sciences** [13].

1.2. Literature-based Knowledge Discovery in Biomedicine

This research direction aims to enable and accelerate real-world knowledge discovery with the advanced text mining techniques developed. I have collaborated with experts in various scientific disciplines (e.g., biomedicine,

chemistry, and health) to achieve this goal. Through the collaborations, I have developed systems and algorithms for two tasks: (1) textual evidence discovery and (2) scientific topic contrasting.

Textual Evidence Discovery aims to automatically retrieve evidence sentences given a user-input query. Scientists need textual evidence mining to validate and prioritize the scientific hypotheses before expensive experimental validation. Textual evidence discovery is an important but underexplored problem in scientific text mining. Traditional literature search engines (e.g., PubMed for biomedical sciences) are designed for document retrieval and do not allow direct retrieval of specific statements. Some of these statements may serve as textual evidence that is key to hypothesis generation and new finding validation. I have developed a web-based system, **EvidenceMiner** (https://evidenceminer.com/) [13, 14], which incorporates the fine-grained named entity and open relation information to discover textual evidence. EvidenceMiner works on CORD-19 [19], the COVID-19 Open Research Dataset. EvidenceMiner takes a researcher's query (e.g., "UV, kill, Sars-Cov-2") and returns a ranked list of sentences containing the compelling evidence as well as their associated research articles (Fig. 2). It has the following distinctive features: (1) it allows users to query a natural language statement or an inquired relationship at the meta-symbol level (e.g., CHEMICAL and PROTEIN) and automatically retrieves textual evidence from a background corpora of COVID-19; (2) It has been constructed in a completely automated way without requiring any human effort for training data annotation; (3) it achieves the best performance compared with baseline methods such as LitSense [20]. EvidenceMiner has been demonstrated at both **ACL'20** and **ISMB'20** conferences and has attracted great attention. **We have users (including biomedical and clinical researchers) from various universities and institutions**. For example, Dr. David Liem (UC Davis Medical School) used EvidenceMiner to test scientific hypotheses for the relationship between cardiovascular diseases and COVID-19. Dr. Clare Voss (Army Research Lab) used EvidenceMiner to test scientific hypotheses related to the UV inactivation of COVID-19.

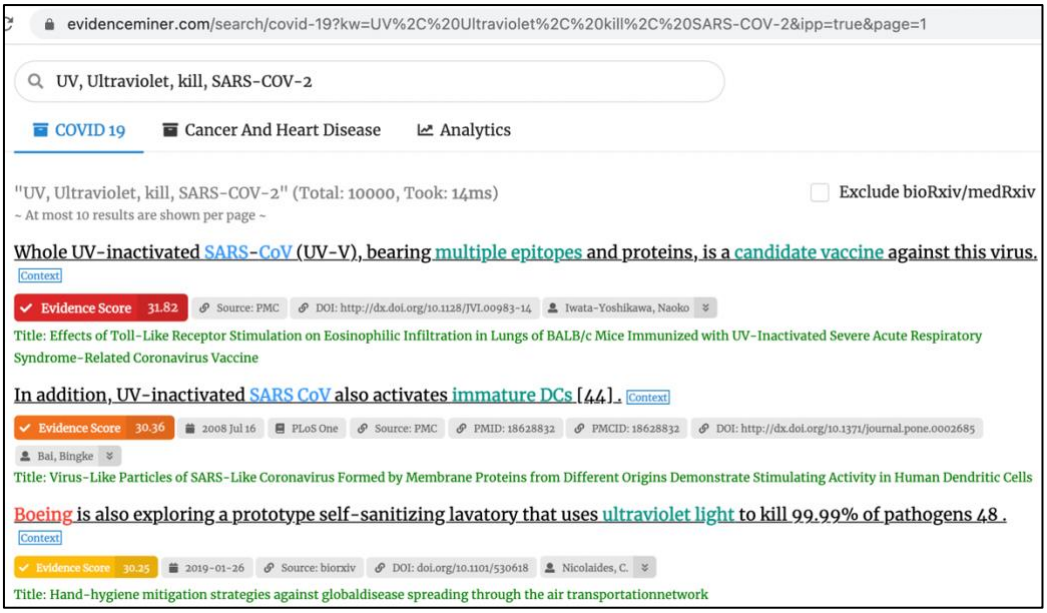

Fig. 2. An example of the evidence sentences retrieved by EvidenceMiner. The top three sentences are highly related to the query (UV, Ultraviolet, kill, SARS-COV-2).

Scientific Topic Contrasting aims to find representative and contrasting knowledge (e.g., entities or relationships) across multiple topics from the scientific literature. For example, clinical researchers want to develop drugs that can precisely treat six main categories of heart diseases (Fig. 3). To find the most representative proteins for each category of heart diseases for drug development, researchers often look into biomedical literature for distinctive associations between proteins and heart diseases before expensive experimental validation. This function is badly needed in scientific research but is under-explored in current literature search and analysis systems. I have developed a web-based system, **SciContrast** (https://scicontrast.firebaseapp.com/), for scientific topic contrasting based on life science literature. SciContrast

enables scientists to select a set of topics of interest, contrasts the representative knowledge across multiple topics, and provides concrete sentences from literature to support such evidence. It provides a focused list of prioritized candidates (e.g., proteins) for scientists to explore to save time and expensive experimental efforts. I collaborated with UC Davis Medical School to identify cardiovascular proteins specifically associated with six sub-categories of heart diseases [3]. **Without SciContrast, clinical researchers would need to read thousands of papers related to the six categories of heart diseases and then experiment with more than 8,000 proteins that have appeared in the papers, which is impossible** with their limited time and experimental funding. **SciContrast enables** this comparative knowledge discovery by advanced methods in text mining. It enables a **precision medicine** approach to find new forms of treatment for patients with preserved ejection fraction (HFpEF), which has led to a **recently funded $1.5 million NIH project**. Our collaboration shows the real-world impact of text mining on medical knowledge discovery.

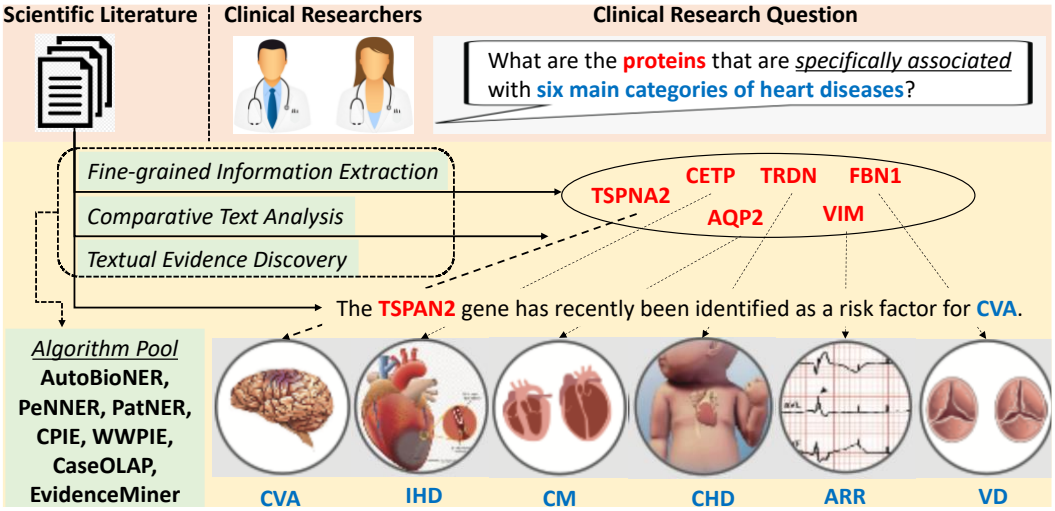

Fig. 3. The overall framework using SciContrast to identify target proteins for six main groups of cardiovascular diseases (CVA, IHD, CM, CHD, ARR, and VD).

2. Future Work

My research tackles a series of technical challenges for extracting a wide range of fine-grained information from unstructured text for biomedical discovery. Based on my prior research and my vision in AI-driven knowledge discovery, I plan to **open new research directions** in using NLP and text mining for scientific discovery in the next few years. There remain grand challenges such as a lack of specialized domain knowledge in a natural language context, complex conditions associated with scientific information, and multi-modal representations of scientific knowledge. To tackle the above challenges, I plan to develop **knowledge-enhanced, condition-aware, and multi-modal NLP and text mining approaches** for scientific discovery. I will collaborate with experts in various research areas **within and beyond computer science**, such as graph mining, natural language processing, computer vision, bioinformatics, computational biology, health informatics, and natural sciences.

2.1 Knowledge-Enhanced Information Extraction for Knowledge Discovery

One major challenge for comprehending the fine-grained scientific information in the text is the urgent need for domain-specific background knowledge. Domain knowledge (e.g., from knowledge bases and ontologies) can be naturally expressed in **logical rules** or **symbolic patterns**. Recently, deep learning-based approaches are dominating NLP research and routinely produce strong benchmark accuracy results. However, these approaches lack explainability to human experts, and it is hard to incorporate domain knowledge into these learning-based models. On the other hand, even though methods based on matching symbolic patterns are less accurate on standard test splits than the deep-learning methods, they still offer significant practical advantages. The pattern-based approaches are more transparent to human experts and support human examination of intermediate representations and reasoning steps. They are amenable to having a human in the loop through

intervention, manipulation, and incorporation of domain knowledge. I propose developing approaches combining deep learning and symbolic patterns to better understand the scientific text.

2.2 Condition-Aware Information Extraction for Knowledge Discovery

Another major challenge for scientific information extraction is that scientific knowledge can only be valid under certain **conditions**. For example, a drug may only be considered effective to a disease with a certain dosage or to certain patient groups (e.g., age, gender, or comorbidity with other diseases). I propose to organize massive text into multi-dimensional text cube structure and synthesize knowledge in **multi-dimensional space**. Specifically, I propose developing multi-dimensional information extraction approaches to extract entities, relationship triplets, and knowledge graphs by considering user-specified dimensions or aspects. This multi-dimensional information extraction and knowledge organization facilitate complex real-world applications, such as knowledge summarization of distinct entities, relationships, and networks under each dimension and knowledge discovery through cross-dimensional comparison and inference.

2.3 Multi-modal Information Extraction for Knowledge Discovery

In addition to the text information, scientific knowledge is usually embedded in multi-modal formats (e.g., text and graphics) in the scientific literature. Scientific data also exist in multi-omics formats in some domains (e.g., genomics and proteomics data in the biology domain). A multi-modal information extraction system significantly benefits literature-based scientific discovery in various domains: (1) **biomedicine (text + image + table)**, (2) **chemistry (text + molecular graph)**, and (3) **health (electronic health record + genomics data)**. However, it is nontrivial to integrate information from the text and other modalities. Based on my expertise in text mining, I propose to bridge the gap of multi-modal scientific information extraction by collaborating with experts in various research areas, such as graph mining, natural language processing, computer vision, image processing, bioinformatics, and natural sciences. Specifically, I propose developing cross-media semantic representation learning and information extraction approaches to support complex real-world applications such as multi-modal knowledge base curation and completion.

2.4 NLP and Text Mining for Other Domains

My long-term goal is to achieve **AI-driven knowledge discovery with real-world impact** on various applications, such as information systems for health care and biomedicine, AI-driven systems for molecular discovery and synthetic strategy designing, knowledge graph construction, web mining, environmental monitoring, smart city, and cyber-physical systems. Although the techniques are initially designed for the biomedical domain, they can be generally applicable to other domains. I plan to collaborate with experts in various domains, such as biology, chemistry, medicine, physical sciences, and various engineering fields, to identify real-world problems and propose new NLP and text mining solutions. **I strongly believe NLP and text mining can have a much broader impact on real-world knowledge discovery in the future.**

References

1. M. AlQuraishi. "AlphaFold at CASP13." Bioinformatics 35, no. 22 (2019): 4862-4865.
2. V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain. "Unsupervised word embeddings capture latent knowledge from materials science literature." Nature 571, no. 7763 (2019): 95-98.
3. D. Liem, Alexandre, S. Murali, D. Sigdel, Y. Shi, **X. Wang**, J. Shen, H. Choi, *et al*. "Phrase Mining of Textual Data to Analyze Extracellular Matrix Protein Patterns Across Cardiovascular Disease", American Journal of Physiology-Heart and Circulatory Physiology, 1;315(4):H910-H924, 2018

4. **X. Wang**, Y. Zhang, Q. Li, C. Wu, and J. Han, "PENNER: Pattern-enhanced Nested Named Entity Recognition in Biomedical Literature", in Proc. 2018 Int. Conf. on Bioinformatics and Biomedicine (BIBM'18), 2018
5. **X. Wang**, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, and J. Han, "Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning", Bioinformatics 35(10): 1745-1752, 2019
6. **X. Wang**, Y. Zhang, Q. Li, X. Ren, J. Shang, and J. Han, "Distantly Supervised Biomedical Named Entity Recognition with Dictionary Expansion", in Proc. 2019 IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM'19), 2019
7. **X. Wang**, Y. Guan, Y. Zhang, Q. Li, and J. Han, "Pattern-enhanced Named Entity Recognition with Distant Supervision", in Proc. 2020 IEEE Int. Conf. on Big Data (BigData'20), 2020
8. **X. Wang**, X. Song, B. Li, K. Zhou, Q. Li, and J. Han, "Fine-Grained Named Entity Recognition with Distant Supervision in COVID-19 Literature", in Proc. 2020 IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM'20), 2020
9. **X. Wang**, V. Hu, X. Song, S. Garg, J. Xiao, and J. Han, "ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-guided Distant Supervision", in Proc. 2021 Conf. on Empirical Methods in Natural Language Processing (EMNLP'21), 2021
10. Y. Meng, Y. Zhang, J. Huang, **X. Wang**, Y. Zhang, H.Ji and J. Han, "Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training", in Proc. 2021 Conf. on Empirical Methods in Natural Language Processing (EMNLP'21), 2021
11. **X. Wang**, Y. Zhang, Q. Li, Y. Chen and J. Han, "Open Information Extraction with Meta-pattern Discovery in Biomedical Literature", in Proc. of 2018 ACM Conf. on Bioinformatics, Computational Biology, and Health Informatics (BCB'18), 2018
12. Q. Li, **X. Wang**, Y. Zhang, F. Ling, C. Wu, and J. Han, "Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature", in Proc. 2018 Int. Conf. on Bioinformatics and Biomedicine (BIBM'18), 2018
13. **X. Wang**, Y. Guan, W. Liu, A. Chauhan, E. Jiang, Q. Li, D. Liem, D. Sigdel, J. Caufield, P. Ping and J. Han, "EVIDENCEMINER: Textual Evidence Discovery for Life Sciences", in Proc. 2020 Annual Conf. of the Association for Computational Linguistics (ACL'20) (System demo), 2020
14. **X. Wang**, Y. Zhang, A. Chauhan, Q. Li, and J. Han, "Textual Evidence Mining via Spherical Heterogeneous Information Network Embedding", in Proc. 2020 IEEE Int. Conf. on Big Data (BigData'20), 2020
15. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36, no. 4 (2020): 1234-1240.
16. Q. Wang, M. Li, **X. Wang**, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, *et al*., "COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation", in Proc. of 2021 Annual Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL'21) (System demo), Best Demo Paper Award, 2021
17. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
18. S. Chithrananda, G. Grand, and B. Ramsundar. "Chemberta: Large-scale self-supervised pretraining for molecular property prediction." arXiv preprint arXiv:2010.09885 (2020).
19. L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk *et al*. "Cord-19: The covid-19 open research dataset." ArXiv (2020).
20. A. Allot, Q. Chen, S. Kim, R. V.Alvarez, D. C. Comeau, W. J. Wilbur, and Z. Lu. "LitSense: making sense of biomedical literature at sentence level." Nucleic acids research 47, no. W1 (2019): W594-W599.