SCIENTIFIC KNOWLEDGE EXTRACTION FROM MASSIVE TEXT DATA

BY

XUAN WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Professor Jiawei Han, Chair and Director of Research
Professor Heng Ji
Professor Chengxiang Zhai
Dr. Zhiyong Lu, National Institutes of Health (NIH)

ABSTRACT

Text mining is promising for advancing human knowledge in many fields, given the rapidly growing volume of text data (e.g., news reports, scientific articles, and medical notes) we are seeing nowadays. Recently, there has been a growing interest in bringing text mining to scientific discovery in various domains, such as mining the biomedical literature and electronic health record for health care and biomedicine, mining the chemistry literature for molecular discovery and synthetic strategy designing, and mining the agriculture literature for agricultural resilience, management, and sustainability. We envision tremendous opportunities in this emerging area of advanced text mining for scientific discovery.

This thesis focuses on developing effective and scalable text mining *algorithms* and *systems* to enable and accelerate scientific discovery. We primarily focus on two research directions: (1) scientific information extraction with weak supervisions, and (2) scientific knowledge discovery applications.

- **Scientific Information Extraction with Weak Supervisions**: 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. To address this challenge, we have developed methods that extract information from text with minimal human supervision. We have contributed a series of algorithms and systems under three weak supervision scenarios: (1) pattern-enhanced weak supervision for scientific information extraction, (2) ontology-guided distant supervision for fine-grained information extraction, and (3) cross-modal supervision between text and graph.

- **Scientific Knowledge Discovery in Real World**: With the advanced text mining methods developed, we future study how to enable and accelerate real-world knowledge discovery. We have been collaborating with experts in various science domains (e.g., biomedicine, chemistry, and health) to achieve this goal. Through the collaborations, we have developed algorithms and systems for two real-world applications: (1) scientific textual evidence discovery and (2) scientific topic contrasting.

Our 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,

UC Davis, UCSD, USC, Purdue, and Iowa State University). Our algorithms and systems can be generally used for any science domain where a knowledge discovery from massive text data is needed. Two examples in the health and chemistry domains are discussed below.

- **Clinical Domain**: We have developed text mining methods to find proteins that are specifically associated with six main categories of heart diseases. Our top-ranked proteins match the knowledge of the clinical researchers very well. Some of our discovered proteins are currently under experimental validation by clinical researchers at the UC Davis Medical Center. This collaboration has a high potential to unveil novel therapeutic targets in patients and repurpose drugs already used in the clinic.

- **Chemistry Domain**: We have also developed text mining methods to support an intelligent molecule discovery process in organic chemistry. We have been collaborating with the researchers in the Chemistry Department at UIUC, finding the most representative catalysts and reaction conditions by comparing different organic reaction types. This collaboration leads to AI-driven systems for automatic chemical/material synthesis plan generation and optimization.

In summary, we tackle a series of technical challenges for automatically extracting a wide range of information from unstructured scientific text. We further address open scientific problems, such as clinical drug discovery and chemical and biological molecule design, based on the rich information we automatically extracted from the scientific text. However, there remain grand challenges for scientific text mining, such as a lack of specialized domain knowledge in a natural language context, multi-modal representations of scientific knowledge, and complex conditions associated with scientific information. In the future, we plan to tackle the above challenges by developing knowledge-enhanced, multi-modal, and condition-aware text mining approaches for scientific discovery.

To my family for their love and support.

ACKNOWLEDGMENTS

First of all, I would like to thank my Ph.D. advisor, Jiawei Han, who is the best mentor I could ever imagine. I appreciate your warm encouragement to me to start my research journey, patient guidance to help me overcome the research obstacles, and your strong support to help me build confidence and grow into a mature researcher. I came from a non-CS background (B.S. in Biology and M.S. in Biochemistry and Statistics). I remember feeling easily confused and unconfident when I first started my Ph.D. study in Computer Science. However, you gave me numerous trust and encouragement to find my own research path. I still remember you told me at the beginning of my Ph.D. that I should confidently work on a research problem as long as I know it is a real problem and will have a significant scientific impact. Then we started working on open information extraction with pattern mining and we see it still has great potential now integrated with new technology (e.g., pre-trained language models). You are a real scientist to me, who always cares the most about real problems and seek real solutions from foundational principles. During my Ph.D. study, you always told me to slow down, break a complex big problem into smaller ones, and solve them one by one in a thorough way. Your research attitude greatly influenced and shaped my research attitude and I hope to bring it to more younger researchers. I also appreciate your tremendous help during my job search, going through my job talk with me so many times, and helping me during every step in this process. I am so grateful for your strong support that enables me to continue my academic career. I owe so much to Jiawei, and the only way to redeem myself is to keep spreading your spirits to younger generations.

I would like to thank all the other thesis committee members, Heng Ji, Chengxiang Zhai, and Zhiyong Lu. I have been collaborating closely with Heng on various biomedical/chemistry text mining projects and received valuable guidance from her. Heng is especially caring and supportive to women students and researchers, serving as a role model to all of us. I met Chengxiang during my first year in PhD study when Chengxiang served on my Program of Study committee. I always remember his advice to me that by the end of your PhD, you should be known as the top expert in a particular research field instead of as someone who published a lot of papers. I met Zhiyong during an international conference in 2018, where we had enlightening discussions on weakly-supervised information extraction. We have had close collaborations since then and I have always been inspired by the real problems Zhiyong raises in the biomedical text mining domain. Moreover, all of them have provided enormous help in my job search process and valuable advice for crafting a successful academic career.

I would also like to thank all our data mining group (DMG) members for not only the exciting research collaborations but also the warm friendship. DMG is a big family to all of us. We support each other during the ups and downs and share happiness and sadness. I am sure that all of our group members will realize your dreams and have a bright future. Alphabetically, they are: Shivam Agarwal, Aabhas Chauhan, Shweta Garg, Xiaotao Gu, Yingjun Guan, Vivian Hu, Jiaxin Huang, Enyi Jiang, Meng Jiang, Minhao Jiang, Bowen Jin, Priyanka Kargupta, Tanay Komarlu, Yuning Mao, Yu Meng, Bangzheng Li, Qi Li, Zoey Li, Liyuan Liu, Weili Liu, Siru Ouyang, Meng Qu, Xiang Ren, Jingbo Shang, Xiangchen Song, Fangbo Tao, Jiaming Shen, Yu Shi, Jinfeng Xiao, Carl Yang, Chao Zhang, Xinyang Zhang, Yu Zhang, Yunyi Zhang, Ming Zhong, Qi Zhu, Wanzheng Zhu.

Lastly, I would like to thank my family and friends, without whom I could have never made it this far to complete my Ph.D. study and continue my academic career. I would like to thank my wonderful parents, Hanxing Wang and Zhongtang Wang, who are always there for me with their warm love and unconditional support. I would like to thank my beloved husband, Aiguo Han, who walked this long way together with me and always back me up to face any challenges. I would also like to thank my precious little one, my son Eric Han, who came into my life at the end of my Ph.D. study and brightened my life ever since then.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1
1.1 Overview 1
1.2 Scientific Information Extraction with Weak Supervisions 2
1.3 Scientific Knowledge Discovery in Real World 7
1.4 Overall Impact 9

CHAPTER 2 PATTERN-ENHANCED WEAK SUPERVISION FOR NESTED BIOMEDICAL NAMED ENTITY RECOGNITION 11
2.1 Introduction 11
2.2 The PeNNER Framework 12
2.3 Experiments 17
2.4 Related Work 24
2.5 Summary 25

CHAPTER 3 ONTOLOGY-GUIDED DISTANT SUPERVISION FOR FINE-GRAINED CHEMISTRY NAMED ENTITY RECOGNITION 27
3.1 Introduction 27
3.2 The ChemNER Framework 29
3.3 Experiments 34
3.4 Related Work 41
3.5 Summary 42

CHAPTER 4 CROSS-MODAL SUPERVISION FOR CHEMICAL REACTANT ENTITY CLASSIFICATION 43
4.1 Introduction 43
4.2 The ReactClass Framework 44
4.3 Experiments 50
4.4 Related Work 54
4.5 Summary 56

CHAPTER 5 SCIENTIFIC TEXTUAL EVIDENCE DISCOVERY 57
5.1 Introduction 57
5.2 The EvidenceMiner Framework 58
5.3 Experiments 67
5.4 Related Work 68
5.5 Summary 68

CHAPTER 6 SCIENTIFIC TOPIC CONTRASTING 69
6.1 Introduction 69
6.2 The SciContrast Framework 70
6.3 Experiments 77
6.4 Related Work 82
6.5 Summary 82

CHAPTER 7 APPLICATIONS AND CONCLUSIONS 83
7.1 Scientific Text Mining: Summary 83
7.2 Applications 85
7.3 Conclusions 90

CHAPTER 8 VISION AND FUTURE DIRECTIONS 92
8.1 Knowledge-Enhanced Scientific Information Comprehension 92
8.2 Multi-Modal Scientific Information Extraction 93
8.3 Multi-Dimensional Scientific Information Analysis 94

REFERENCES 95

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

Text mining is promising for advancing human knowledge in many fields, given the rapidly growing volume of text data (e.g., news reports, scientific articles, and medical notes) we are seeing nowadays. Recently, there has been a growing interest in bringing text mining to scientific discovery in various domains, such as mining the biomedical literature and electronic health record for health care and biomedicine, mining the chemistry literature for molecular discovery and synthetic strategy designing, and mining the agriculture literature for agricultural resilience, management, and sustainability. 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 [1]. We envision tremendous opportunities in this emerging area of advanced text mining for scientific discovery.

Challenges There are several unique challenges for scientific text mining. First, there *lack human annotations* for various science domains (e.g., chemistry and geoscience), especially fine-grained science domains (e.g., organic or inorganic chemistry). Recently, deep learning methods have set up state-of-the-art performance on various text mining tasks. However, deep learning methods rely on massive human-annotated data for model training, which is hard to acquire in science domains due to the limited time and labor of the scientists. We have developed effective text mining methods with minimal human supervision that can be easily applied to various science domains. Second, scientific knowledge usually resides in *multiple modalities*. For example, chemical compounds can be described with both text descriptions and molecule graphs. It is challenging to learn a scientific entity representation with multi-modal information. On the other hand, we see this multi-modal representation as an opportunity since the information in one modality may benefit the tasks in other modalities. We have developed effective chemistry text classification methods with supervision from molecule graph matching. Last, the sentences in scientific writing are usually *long with complex structures*. We have developed text mining methods that specifically deal with the wide-window relation extraction in scientific literature where the two related entities are far apart from each other in the sentence.

This thesis focuses on developing effective and scalable text mining *algorithms* and *systems*

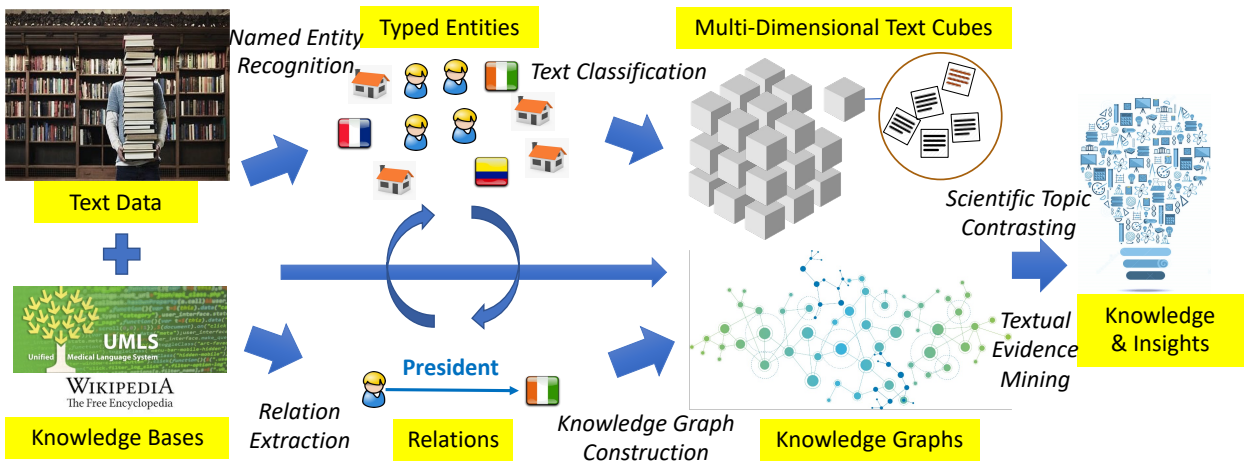


Figure 1.1: This thesis focuses on automatically understanding massive text data to enable and accelerate scientific discovery.

to enable and accelerate scientific discovery (Figure 1.1). Our research has primarily focused on two directions: (1) scientific information extraction with weak supervisions (see Chapters 2, 3, 4) and (2) scientific knowledge discovery in real world (see Chapters 5, 6). The overview and contributions of each research direction are described below.

1.2 SCIENTIFIC INFORMATION EXTRACTION WITH WEAK SUPERVISIONS

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. To address this challenge, we have developed methods that extract information from text with minimal human supervision. We have contributed a series of algorithms and systems under three weak supervision scenarios: (1) pattern-enhanced weak supervision for scientific information extraction, (2) ontology-guided distant supervision for fine-grained information extraction, and (3) cross-modal supervision between text and graph.

1.2.1 Pattern-Enhanced Weak Supervision for Scientific Information 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 (KBs) can be easily accessed, constructed, or integrated, making distant supervision realistic for fine-grained scientific NER tasks. For distant supervision, training labels are automatically generated by matching the mentions in text with the concepts in the KBs. A major challenge of distant supervision is the limited coverage of the dictionaries from the KBs, leading to false-negative errors during the distant training label creation.

To tackle the challenge of *incomplete dictionaries* for distant label generation, we propose several distantly supervised NER methods [2, 3, 4, 5, 6, 7] that effectively deal with noisy distant supervision. One example is PeNNER [2], a BioNER method that relies on massive corpora and unsupervised pattern mining for nested named entity boundary correction. PeNNER takes massive corpora as input, with entities pre-tagged by any flat NER tool. We first perform automatic meta-pattern extraction and take the extracted meta-patterns as candidate outer entity patterns. Then we select two patterns for each entity type as a seed pattern set and perform automatic pattern set expansion. Note that in this step, we only need **very weak supervision** (two user-specified seed patterns instead of a large human-annotated training dataset). The top-ranked meta-patterns in each expanded pattern set are considered correct outer entity patterns for the corresponding entity type. These outer entity patterns are used to correct the boundaries and types of their matched entities in the input corpus. This weakly-supervised pattern method also gives us the advantage to discover new entity types that are not pre-tagged in the input corpus.

Contributions 1.1:

- *Problem*: We study the problem of biomedical named entity recognition with various weak supervision signals (e.g., distant supervision from knowledge bases and weak supervision from seed textual patterns) without requiring human effort for training data annotation.
- *Methodology*: We develop an effective, weakly-superivsed method, PeNNER, for nested biomedical named entity recognition. PeNNER relies on massive corpora and unsupervised pattern mining for nested named entity boundary correction.
- *Effectiveness*: PeNNER outperforms the state-of-the-art supervised biomedical NER methods (e.g., PubTator [8]) while requiring no human supervision. Moreover, PeNNER is also able to accurately extract new types (e.g., biological process and treatment) that are not originally annotated by PubTator in the input corpus.

Impact 1.1:

- Our distantly supervised BioNER methods are future extended into a biomedical named entity annotation system, CORD-NER [6], that automatically annotates 75 fine-grained biomedical entity types in the COVID-19 literature.
- CORD-NER has also been used for downstream applications such as COVID-KG [9]: COVID-19 knowledge graph construction and drug repurposing report generation.
- Our distantly supervised BioNER methods have been taught in a graduate class at University of Illinois at Urbana-Champaign and presented in conference tutorials at IEEE-BigData [10], WWW [11] and SIGKDD [12].

1.2.2 Ontology-Guided Distant Supervision for Fine-Grained Information Extraction

In the chemistry domain, it is important to recognize chemistry entities of diverse and fine-grained types (e.g., "inorganic phophorus compounds", "coupling reactions" and "catalysts") to provide a wide range of information for scientific discovery. Similar to the biomedical domain, we leverage domain-specific ontologies and KBs as distant supervision to develop effective methods for fine-grained chemistry NER. In addition to the aforementioned incomplete dictionary problem, the chemistry domain faces another great challenge of *noisy annotation* where a mention can be erroneously matched due to the potential matching of multiple entity types in the KBs. Previous distantly supervised NER studies largely ignore the noisy annotation problem by simply discarding those multi-labels during the KB-matching process. However, the noisy labels cannot be simply ignored for the chemistry entities because they consist of a large portion of distant training labels. We observe that more than 60% of the entities in the chemistry corpus have multiple labels during KB-matching in the chemistry knowledge bases.

We propose ChemNER [7], an ontology-guided, distantly-supervised NER method for fine-grained chemistry NER. Taking an input corpus, a chemistry type ontology, and associated entity dictionaries collected from the KBs, we propose a flexible KB-matching method with TF-IDF-based majority voting to resolve the incomplete annotation problem. Then we propose an ontology-guided multi-type disambiguation method to resolve the noisy annotation problem. Taking the output from the above two steps as distant supervision, we further train a sequence labeling model to cover additional entities. ChemNER significantly improves the distant label generation for the subsequent NER model training. We also provide an expert-labeled, chemistry NER dataset with 62 fine-grained chemistry types (e.g., chemical compounds and chemical reactions).

Contributions 1.2:

- *Problem*: We study the problem of fine-grained chemistry named entity recognition with distant supervision from domain-specific knowledge bases and ontologies.
- *Methodology*: We develop an ontology-guided, distantly-superivsed method, ChemNER, for fine-grained chemistry named entity recognition. ChemNER leverages the chemistry type ontology structure to provide a global topic constraint for context-aware multi-type disambiguation.
- *Effectiveness*: ChemNER is highly effective, substantially outperforming the state-of-the-art supervised NER methods (i.e., RoBERTa [13] and ChemBERTa [14]), improving the F1 score from 0.2 to 0.45.

Impact 1.2:

- We are further developing a chemistry reaction tracker system, ReactionTracker, that uses ChemNER and information retrieval methods to track chemistry research publications related to user-specified organic chemical reactions. We expect this ReactionTracker system will significantly benefit the query-based tracking of scientific publications.
- ChemNER has further been used in AI-driven systems for automatic chemical/material synthesis plan generation and optimization to support an intelligent molecule discovery.
- ChemNER has been taught in a graduate class at University of Illinois at Urbana-Champaign and presented in conference tutorials at IEEE-BigData [10], WWW [11] and SIGKDD [12].

1.2.3 Cross-Modal Supervision Between Text and Graph

Scientific knowledge can be described on various levels of abstractions: from high-level categorical concepts to low-level concrete entities. For example, the Csp^3-Csp^3 Suzuki cross-coupling reaction is defined by chemists as a process involving a pair of high-level reactant groups (i.e., the M-side reactant group "primary alkyl boronate" and the X-side reactant group "primary alkyl halide"). While in the chemistry literature, this chemical reaction can also be described as a process involving two low-level concrete chemical entities (e.g., "1-bromododecane" and "B-n-octyl-9-BBN"). This gap between high-level and low-level abstractions of scientific knowledge is a common phenomenon in various domains such as biology, chemistry, and physics.

In the general domain, recent work has been done on classifying entities in the text into human-given categories without human annotation. However, in the chemistry domain, the task of reactant entity classification requires more effective methods that take two special characteristics of the chemical molecules into consideration. The first characteristic is that each chemical molecule can be represented in *two modalities*: a chemical name in the text and a molecule structure in the graph. Thus a large amount of high-quality training data for chemical name classification can be automatically created with cross-modal supervision of molecular structure matching. The second characteristic is that there is a *knowledge-aware subword correlation* between the chemical names to be classified and that of the reactant groups as class labels. Thus the interaction between the subwords (e.g., wordpieces in the pre-trained language models) in the chemical names and reactant groups is the most prominent feature of training a reactant entity classification model.

We propose ReactClass [15], a highly effective reactant entity classification method without requiring human effort for training data annotation. ReactClass is designed to take two special characteristics, multi-modal representation and knowledge-aware subword correlation, of the chemical molecules into consideration. First, we propose to use cross-modal supervision to automatically create the training data for chemical name classification in the text via molecular structure matching in the graph. Second, we propose to train a classifier based on the knowledge-aware subword cross-attention map between each chemical name and its corresponding reaction group.

Contributions 1.3:

- *Problem*: We study the problem of reactant entity classification method without requiring human effort for training data annotation.
- *Methodology*: We develop a highly effective reactant entity classification method, ReactClass, without requiring human effort for training data annotation. ReactClass is designed to take two special characteristics, multi-modal representation and knowledge-aware subword correlation, of the chemical molecules into consideration.
- *Effectiveness*: ReactClass is highly effective, achieving state-of-the-art performance in classifying the chemical names into human-defined reactant groups without requiring human effort for training data annotation.

Impact 1.3:

- ReactClass has also been incorporated into the ReactionTracker system for a smart query expansion to enhance the chemical reaction literature tracking.
- ReactClass has further been used in AI-driven systems for automatic chemical/material synthesis plan generation and optimization to support an intelligent molecule discovery.
- ReactClass has been taught in a graduate class at University of Illinois at Urbana-Champaign and presented in conference tutorials at IEEE-BigData [10], WWW [11] and SIGKDD [12].

1.3 SCIENTIFIC KNOWLEDGE DISCOVERY IN REAL WORLD

With the advanced text mining methods developed, we future study how to enable and accelerate real-world knowledge discovery. We have been collaborating with experts in various science domains (e.g., biomedicine, chemistry, and health) to achieve this goal. Through the collaborations, we have developed algorithms and systems for two real-world applications: (1) scientific textual evidence discovery and (2) scientific topic contrasting.

1.3.1 Scientific Textual Evidence Discovery

Scientific 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.

We have developed a web-based system, EvidenceMiner[1] [16, 17], which incorporates the fine-grained named entity and open relation information to discover textual evidence. EvidenceMiner works on CORD-19 [18], 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. EvidenceMiner 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

[1] `https://evidenceminer.com/`

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 [19].

Contributions 1.4:

- *Problem*: We study the problem of textual evidence discovery from scientific literature.
- *Methodology*: We develop a web-based system, EvidenceMiner, which incorporates the fine-grained named entity and open relation information to discover textual evidence.
- *Effectiveness*: EvidenceMiner achieves the best performance compared with baseline methods such as LitSense [19].

Impact 1.4:

- EvidenceMiner has 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.
- We are further extending EvidenceMiner to other scientific domains such as chemistry and material science.

1.3.2 Scientific Topic Contrasting

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. 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.

We have developed a web-based system, SciContrast[2], for scientific topic contrasting based on life science literature. SciContrast enables scientists to select a set of topics of interest,

[2] `https://scicontrast.firebaseapp.com/`

contrasts the representative knowledge across multiple topics, and provides concrete sentences from literature to support such evidence. SciContrast provides a focused list of prioritized candidates (e.g., proteins for each heart disease) for scientists to explore to save time and expensive experimental efforts. SciContrast has the following distinctive features: (1) it addresses the open problem of scientific topic contrasting from biomedical literature; (2) it automatically extracts rich fine-grained knowledge (entity and relation information) from the background corpora; (3) it summarizes the most representative knowledge for each user-given topic using comparative text analysis; and (4) it further provides concrete evidence sentences to support the representative knowledge discovery.

Contributions 1.5:

- *Problem*: We study the problem of scientific topic contrasting in the scientific literature.
- *Methodology*: We develop a web-based system, SciContrast, for scientific topic contrasting in the biomedical literature. SciContrast summarizes and contrasts the most representative knowledge for each user-input topic as well as providing concrete evidence sentences supporting this representative knowledge discovery from the scientific literature.
- *Effectiveness*: SciContrast achieves the best performance compared with baseline methods such as BioBERT [20].

Impact 1.5:

- We have been collaborating with UC Davis Medical School to identify cardiovascular proteins specifically associated with six sub-categories of heart diseases [21]. This collaboration enables a precision medicine approach to find new forms of treatment for patients with preserved ejection fraction (HFpEF). Our collaboration shows the real-world impact of text mining on medical knowledge discovery.
- We are future extending the SciContrast to a multi-omics data mining platform involving scientific literature, electronic health record, and genomic data analysis. We expect this platform will benefit the precision medicine development in various diseases.

1.4 OVERALL IMPACT

Our 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, UC Davis, UCSD, USC, Purdue, and Iowa State University). Our algorithms and systems can be generally used for any science domain where a knowledge discovery from massive text data is needed. In summary, our work has been used in the following settings:

- **Used in real world**:
  - **Clinical Domain**: Our text mining methods have been used to find proteins that are specifically associated with six main categories of heart diseases. Our top-ranked proteins match the knowledge of the clinical researchers very well. Some of our discovered proteins are currently under experimental validation by clinical researchers at the UC Davis Medical School. This collaboration has a high potential to unveil novel therapeutic targets in patients and repurpose drugs already used in the clinic.
  - **Chemistry Domain**: Our text mining methods have been used to support an intelligent molecule discovery process in organic chemistry. We have been collaborating with the researchers in the Chemistry Department at UIUC, finding the most representative catalysts and reaction conditions by comparing different organic reaction types. This collaboration leads to AI-driven systems for automatic chemical/material synthesis plan generation and optimization.
- **Taught in classes and conference tutorials**: Our methods on pattern-enhanced weakly-supervised NER (PeNNER), ontology-guided distantly-supervised NER (ChemNER), and cross-modal supervision between text and graph (ReactClass) are being taught in graduate courses, e.g., University of Illinois at Urbana-Champaign (CS 512), and are introduced as major parts of the conference tutorial in top data mining and database conferences such as SIGKDD, WWW, and IEEE-BigData.
- **Awards**: This thesis work has been awarded YEE fellowship from 2020 to 2021 from the University of Illinois at Urbana-Champaign. It has also impacted an application on COVID-19 knowledge graph construction [9] that has been awarded the Best Demo Paper Award in 2021 from NAACL.

Next, we will discuss how to automatically extract a wide range of fine-grained information from unstructured scientific text. We will further discuss how to address real-world scientific discovery problems, such as scientific textual evidence discovery and scientific topic contrasting, based on the rich information we automatically extracted from scientific text.

CHAPTER 2: PATTERN-ENHANCED WEAK SUPERVISION FOR NESTED BIOMEDICAL NAMED ENTITY RECOGNITION

2.1 INTRODUCTION

Biomedical named entity recognition (BioNER) aims to identify text spans associated with proper names and classify them into a set of semantic classes (e.g., genes, proteins, chemicals, and diseases). BioNER is a fundamental step in the biomedical information extraction pipeline. It facilitates downstream tasks such as relation extraction [22, 23] and knowledge base construction [24, 25, 26, 27].

The common way to approach BioNER is to formulate the task as a sequence labeling problem. Machine learning methods have been proposed for BioNER, from feature-based [28, 29] to neural network methods [3, 30, 31]. However, those flat BioNER methods are unable to handle *nested named entities*. Figure 2.1 shows an example of the nested naming structure: a chemical entity (i.e., "alanine") is nested within a protein entity (i.e., "alanine aminotransferase"). The state-of-the-art flat BioNER system, PubTator [8], recognizes "alanine" as a chemical but misses "alanine aminotransferase" as a protein.

Nested named entities, especially the outermost entities, are important in the BioNER tasks for two reasons. First, nested named entities are common in biomedical literature. For example, 17% of the entities in the GENIA [32] dataset are embedded within another entity. Second, downstream tasks require the BioNER methods to detect the outermost entities as the first step. Failing to recognize the outermost entities may introduce errors to subsequent tasks such as relation extraction and knowledge base construction.

Machine learning methods have been proposed for nested NER [33, 34, 35, 36, 37, 38]. However, those methods are fully supervised, requiring human effort for feature engineering or training data annotations. Feature-based methods [33, 34, 35, 36] rely on handcrafted features carefully designed for each entity type. Neural network methods [37, 38] save efforts for feature engineering, but still require a large amount of human-annotated training data. Therefore, these methods cannot be easily adapted to new entity types. In GENIA, a benchmark dataset for nested BioNER, five types of biomedical entities (i.e., gene/protein, DNA, RNA, cell type, and cell line) are annotated. Despite the success of the supervised nested NER methods on the GENIA dataset, it remains unknown whether those methods perform well at detecting nested naming structures for other important types of biomedical entities such as chemicals and diseases.

In this chapter, we propose PeNNER, a BioNER method that relies on massive corpora and unsupervised pattern mining for nested named entity boundary correction. PeNNER

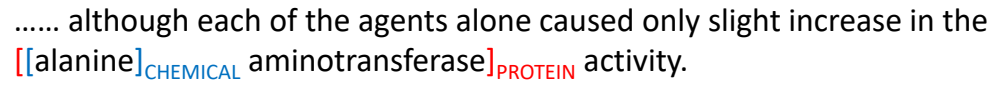


Figure 2.1: An example of biomedical named entities with nested naming structure from PubMed (PMID: 10190572).

takes massive corpora as input, with entities pre-tagged by any flat NER tool. We first perform automatic meta-pattern extraction and take the extracted meta-patterns as candidate outer entity patterns. Then we select two patterns for each entity type as a seed pattern set and perform automatic pattern set expansion. Note that in this step, we only need **very weak supervision** (two user-specified seed patterns instead of a large human-annotated training dataset). The top-ranked meta-patterns in each expanded pattern set are considered correct outer entity patterns for the corresponding entity type. These outer entity patterns are used to correct the boundaries and types of their matched entities in the input corpus. Compared with previous BioNER methods, PENNER greatly enhances nested named entity boundary correction without any human effort for feature engineering or training data annotation. Moreover, our pattern set expansion approach gives us the advantage to discover new entity types that are not pre-tagged in the input corpus. We compare PENNER with the state-of-the-art BioNER system, PubTator, and observed significant improvement in recognizing the outer entities for four types: gene, chemical, disease, and species. PENNER is also able to accurately extract new types (e.g., biological process and treatment) that are not originally annotated by PubTator in the input corpus.

2.2 THE PENNER FRAMEWORK

The PENNER framework is shown in Figure 2.2. The first step is an initial round of entity tagging. Taking an input corpus, we first use PubTator to tag biomedical entities of four types: genes, chemicals, diseases, and species. Then we replace the tagged biomedical entities with their types from PubTator. The second step is meta-pattern extraction. We extract quality sequential patterns containing entity type tokens as *meta-patterns* [39]. From our perspective, a quality meta-pattern is assumed to be frequent, informative, and complete. For example, in Figure 2.2, the green box shows some extracted quality meta-patterns. The third step is pattern expansion. For each entity type to be recognized, we take two user-specified seed patterns as weak supervision and expand the pattern set iteratively. At each round of expansion, we select the meta-patterns sharing the most context similarity with the seed patterns and add them to the seed patterns of the corresponding type. For example, in Figure 2.2, the weak supervision for the GENE type includes two seed patterns:

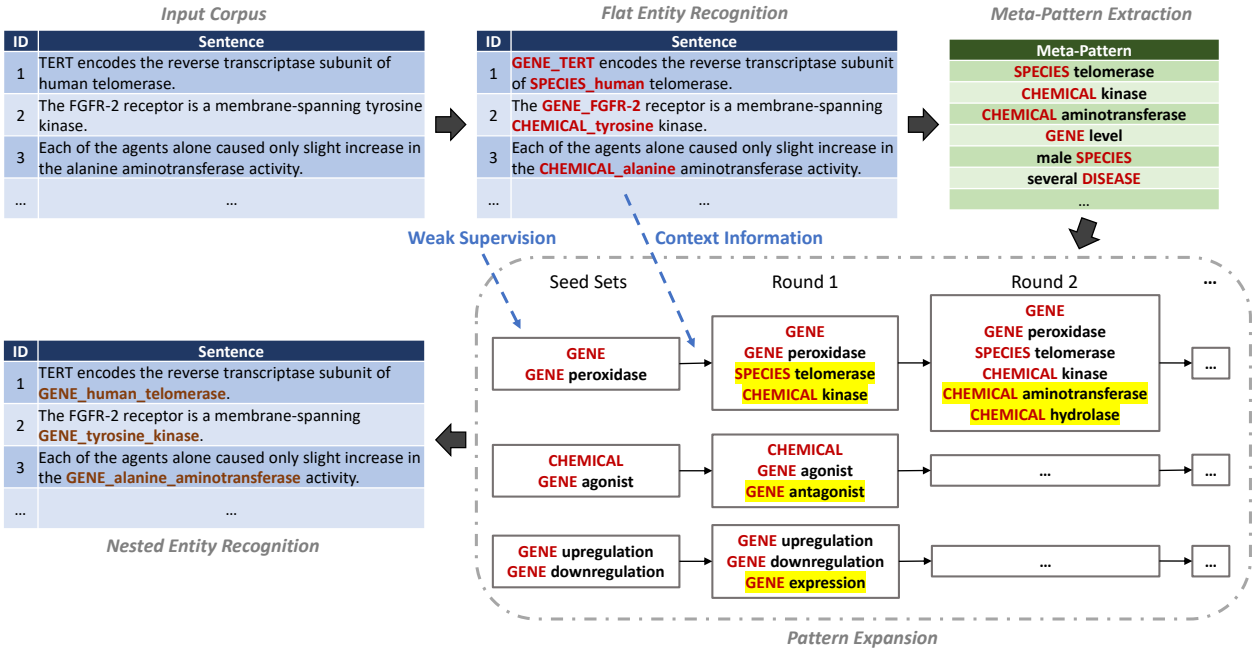


Figure 2.2: Framework overview of PENNER.

"GENE" and "GENE peroxidase". During the first round of pattern expansion, "SPECIES telomerase" and "CHEMICAL kinase" are considered to be similar to the seed patterns of the GENE type. So we add the above two patterns to the pattern set of the GENE type for future expansions. After we finished expanding the pattern sets for each entity type, we take the final expanded pattern sets and match the patterns to concrete entity mentions in the input corpus. Those matched entity mentions are considered as the outer entities and their boundaries and types are corrected by their matched patterns. We discuss the two key components of PENNER, meta-pattern extraction and pattern expansion, in detail below.

2.2.1 Meta-Pattern Extraction

Candidate Meta-Pattern Extraction Taking an input corpus pre-tagged with any flat NER tool, we first replace the tagged entity mentions with their type names. After the replacement, the corpus will be a sequence as a mixture of word tokens and entity-type tokens. Then we conduct frequent sequential pattern mining [40] to extract a big pool of candidate meta-patterns. A meta-pattern is a sub-sequence of the corpus containing at least one entity-type token. For example, "human telomerase" and "mouse telomerase" are two sub-sequences of the original corpus. While they can be represented by the same meta-pattern "SPECIES telomerase" after replacing "human" and "mouse" with their entity-type token "SPECIES".

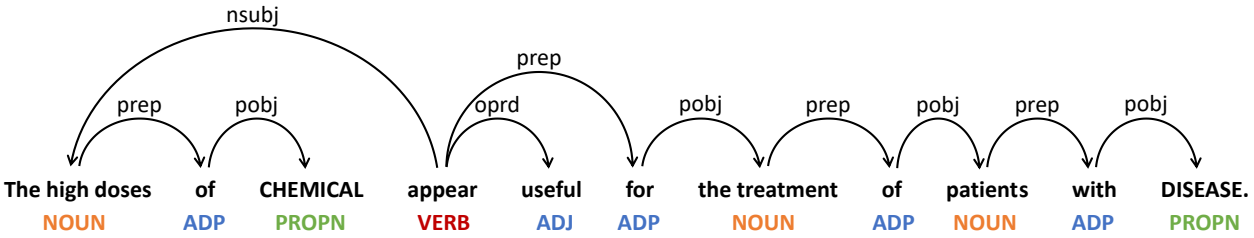


Figure 2.3: An example sentence with dependency parsing tree using SpaCy.

Meta-pattern extraction is important for nested NER for two reasons. First, the nested naming structure is more of a pattern-level phenomenon than an instance-level phenomenon. For example, the meta-pattern "SPECIES telomerase" indicates a nested protein entity no matter whether we are talking about humans or mice as the SPECIES instance. Second, a meta-pattern has aggregated context information of all of its instances, which helps us learn its semantics in a more accurate way.

Quality Meta-Pattern Selection After extracting the big pool of candidate meta-patterns, we further conduct quality meta-pattern selection to remove noisy patterns before the next step of pattern expansion. Quality meta-pattern selection has been studied in MetaPAD [39], TruePIE [41], and BioOpenIE [42]. They propose a set of statistical features (e.g., pattern frequency, IDF score, and co-occurrence) to train a classifier that estimates the quality of the candidate meta-patterns. However, previous methods do not consider a semantic analysis of the meta-patterns based on the sentence structures. In biomedical literature, sentences are usually long with formal language styles. We propose a quality meta-pattern selection method incorporating dependency parsing [43] to utilize the sentence structure features.

We use SpaCy[3] for dependency parsing. The output parsing tree has a set of directed syntactic relations between the words in a sentence. Figure 2.3 shows an example of a sentence parsing tree. The root of the tree is the verb "appear". It is connected to "The high doses" via a subject relation (`nsubj`) and to "for" via a preposition relation (`prep`). According to the parsing tree structures as well as the corpus statistics, we propose the following four criteria to select quality meta-patterns:

- **Frequency**: A quality meta-pattern should occur frequently. In PENNER, we require each meta-pattern candidate to appear more than 10 times in the corpus.
- **Informativeness**: A quality meta-pattern should either be a single entity type (e.g., "DISEASE") or a phrase with one entity type and at least one non-stop-word (e.g., "pa-

[3]https://spacy.io/

tients with DISEASE"). Since we focus on the NER task, meta-patterns with two or more entity mentions (e.g., "CHEMICAL induces DISEASE") will not be considered, but they will be useful in other tasks such as relation extraction.

- **Syntactic Completeness**: For a quality meta-pattern, all of its tokens in the parsing tree should form a connected subgraph. For example, in Figure 2.3, "CHEMICAL appear useful" is not complete since "CHEMICAL" and "appear" are separated by other nodes. In contrast, "patients with DISEASE" is complete.

- **Semantic Completeness**: Since we focus on the NER task, the extracted pattern should form a complete noun phrase. For example, in Figure 2.3, "of CHEMICAL" is syntactic complete but is not a complete noun phrase. To reduce the noise of incomplete noun phrases, we divide the whole parsing tree into chunks. Starting from the root, we iteratively cut the tree at nouns (i.e., nodes with tags `NOUN` or `PROPN`). The noun serves as the leaf of the current chunk as well as the root of the next chunk. For example, in Figure 2.3, the sentence is divided into four chunks: "the high doses appear useful for the treatment", "the high doses of CHEMICAL", "treatment of patients", and "patients with DISEASE". We require a semantic complete pattern to be a complete chunk in the sentence.

Specifically, for quality meta-pattern selection, we first select all the meta-patterns satisfying our frequency threshold. Then we check each of them by informativeness, syntactic completeness, and semantic completeness. We remove the meta-patterns that do not meet any of the above criteria from our candidate meta-pattern pool.

2.2.2 Pattern Expansion

Pattern Set Expansion Taking the quality meta-patterns as candidates, we further expand the quality patterns into the pattern sets for each entity type we want to recognize. To get rid of the reliance on the entity-type-dependent training corpus, this pattern expansion step needs to be done under **very weak supervision**. Here we adopt the SetExpan framework [44]. SetExpan takes several user-provided seed patterns for each entity type (e.g., "GENE" and "CHEMICAL peroxidase" for the GENE type) and expand the seed patterns with other patterns (e.g., "CHEMICAL aminotransferase", "CHEMICAL hydrolase", and "SPECIES telomerase") belonging to the same entity type. We assume patterns sharing the most context similarities are likely belonging to the same entity type. Specifically, we utilize **skip-grams** as the context features for similarity calculation.

Formally, given a candidate pattern p, one of its skip-grams is "w_{-1} _ w_1" where w_{-1} and w_1 are two context words and p is replaced with a placeholder. For example, in the sentence

"This effect exhibits CHEMICAL peroxidase activity in SPECIES hepatocytes", one skip-gram of the pattern "CHEMICAL peroxidase" is "exhibits _ activity". We can also enlarge the context window size to extract longer skip-grams (e.g., "$w_{-2}w_{-1}$ _ $w_1w_2w_3$"). In our experiments, the maximum context window size is 4. Note that word embedding methods such as Word2Vec [45] also use skip-gram information. One advantage of SetExpan over Word2Vec is that they impose strong positional constraints by using concrete skip-grams.

SetExpan defines the similarity between each pair of pattern p and its context feature c using the TF-IDF transformation [46]:

$$f_{p,c} = \log(1 + X_{p,c})(\log |P| - \log \sum_{p' \in P} X_{p',c}), \tag{2.1}$$

where P is the set of candidate patterns and $X_{p,c}$ is the raw co-occurrence count between p and c in our input corpus. Empirically, SetExpan shows that such weight scaling outperforms other alternatives such as point-wise mutual information (PMI) and BM25. Then the similarity between two patterns p_1 and p_2 under feature set F is defined as

$$sim(p_1, p_2|F) = \frac{\sum_{c \in F} \min(f_{p_1,c}, f_{p_2,c})}{\sum_{c \in F} \max(f_{p_1,c}, f_{p_2,c})}. \tag{2.2}$$

Given the seed pattern set S, we first score each skip-gram feature c based on its accumulated similarity with the seed patterns in S (i.e., $\sum_{p \in S} f_{p,c}$). Then M features with the highest scores will be selected, from which we sample N subsets F_i ($i = 1, 2, ..., N$). Each of the subsets contains M_0 ($M_0 < M$) features. The score of each pattern p in feature set F_i is

$$score(p|F_i) = \frac{1}{|S|} \sum_{p' \in S} sim(p, p'|F_i). \tag{2.3}$$

Therefore, for F_i, we can obtain a ranking list of patterns according to their $score(\cdot|F_i)$. Suppose the rank of p in feature set F_i is $r_{p,i}$, we calculate the mean reciprocal rank of p as

$$MRR(p) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{r_{p,i}}. \tag{2.4}$$

Finally, the patterns with MRR higher than a threshold MRR_{thrs} will be added into the seed set for the next iteration.

Multi-Set Co-Expansion In practice, we need to recognize entities of different types **simultaneously**. For example, we may expand patterns representing genes and chemicals

Algorithm 2.1: MULTISETEXPAN($S_1, ..., S_Q, MRR_{thrs}$)

Data: M seed sets $S_1, ..., S_Q$ representing Q different entity types
Result: Q expanded sets $S_1, ..., S_Q$
while $\exists S_k$ not converged **do**
  for $i = 1$ to Q **do**
    Sample N context feature sets $F_1, ..., F_N$
    for $p \in P \setminus S_i$ **do**
      Calculate $MRR(p)$ for each p
      if $MRR(p) \geq MRR_{thrs}$ and $p \notin \cup_{j \neq i} S_j$ **then**
        $S_i \leftarrow S_i \cup \{p\}$
      end
    end
    if nothing added into S_i in this round **then**
      Mark S_i as converged
    end
  end
end

Table 2.1: Basic statistics of the biomedical literature corpus.

Abstracts	Sentences	Entity Mentions			
		GENE	CHEMICAL	DISEASE	SPECIES
28007	302736	215704	314134	129931	86697

at the same time using two seed pattern sets. In our problem setting, the entity types are assumed to be mutually exclusive (e.g., a disease entity/pattern can hardly be a chemical as well). This property enables different semantic sets to give hints to each other. Therefore, we extend SETEXPAN to the MULTISETEXPAN framework, shown in Algorithm 2.1. Given Q seed sets $S_1, S_2, ..., S_Q$ of different types, MULTISETEXPAN expands patterns for each S_i by turns. If a pattern has already been included in other pattern sets, no matter how large its MRR is, we will not put it in S_i. In our experiments, we find this multi-set co-expansion strategy highly effective in avoiding interference among different seed pattern sets.

2.3 EXPERIMENTS

We aim to answer three questions in the Experiments section. First, at the pattern level, how does PENNER perform in the meta-pattern expansion? Second, at the instance level, how does PENNER perform in nested named entity recognition? Third, after the pattern enhancement, what are the improvements of PENNER over PubTator?

Table 2.2: Pattern expansion results of EMBEDDING on Gene, Chemical, Disease and Species entities. Grey patterns are judged as incorrect.

Seed	{GENE, GENE peroxidase}	{CHEMICAL, GENE agonist}	{DISEASE, cellular DISEASE}	{SPECIES, female SPECIES}
1	unassigned : GENE	CHEMICAL receptor modulator (serm)	DISEASE vera	fischer SPECIES
2	CHEMICAL phosphatase	antagonist of CHEMICAL	potential for DISEASE	SPECIES and adult
3	(CHEMICAL) release	offspring of SPECIES	GENE translocation	exposure to CHEMICAL or
4	SPECIES cardiomyocyte	CHEMICAL oxidase (	SPECIES and adult	SPECIES in vivo
5	potential against DISEASE	DISEASE chemopreventive agent	growth and DISEASE	CHEMICAL protect
6	GENE inducer	GENE receptor activity	a common DISEASE	CHEMICAL interfere
7	effect and mechanism of CHEMICAL	antagonist (CHEMICAL)	rare DISEASE	a cohort of SPECIES
8	inducer of GENE	CHEMICAL blocker	detection of DISEASE	SPECIES albino
9	(GENE) antagonist	CHEMICAL substituent	DISEASE as well as	CHEMICAL exposure ,
10	GENE level and	CHEMICAL vapor	progression and DISEASE	the detrimental effect of CHEMICAL

Table 2.3: Pattern expansion results of SETEXPAN on Gene, Chemical, Disease and Species entities. Grey patterns are judged as incorrect.

Seed	{GENE, GENE peroxidase}	{CHEMICAL, GENE agonist}	{DISEASE, cellular DISEASE}	{SPECIES, female SPECIES}
1	SPECIES telomerase	GENE	hepatic DISEASE	male SPECIES
2	CHEMICAL	DISEASE chemopreventive agent	degradation of GENE	DISEASE
3	DISEASE	DISEASE	dermal DISEASE	CHEMICAL
4	CHEMICAL acetyltransferase	CHEMICAL chelation	clinical DISEASE	DISEASE cell
5	CHEMICAL aminotransferase	SPECIES	GENE phosphorylation	GENE
6	SPECIES	GENE antagonist	-	SPECIES cell
7	CHEMICAL hydrolase	DISEASE cell	-	pregnant SPECIES
8	GENE kinase	underlying mechanism of CHEMICAL	-	adult SPECIES
9	CHEMICAL kinase	CHEMICAL exclusion	-	CHEMICAL channel
10	CHEMICAL influx	10 m CHEMICAL	-	DISEASE cell line

2.3.1 Datasets

PENNER is a general nested named entity boundary correction method that can be applied to any domain. In this study, we evaluate the performance of PENNER and several baseline methods on a biomedical literature corpus. The corpus is constructed from the Comparative Toxicogenomics Database (CTD) [47] that contains a large set of human-curated biomedical entities. Two entities, together with their relation, form a tuple in CTD. We randomly sample 248,064 tuples and extract all the PubMed abstracts associated with these tuples. Table 2.1 shows some basic statistics of our biomedical literature corpus.

2.3.2 Baselines

We demonstrate the effectiveness of PENNER against two baseline methods:

Table 2.4: Pattern expansion results of PeNNER on Gene, Chemical, Disease and Species entities. Grey patterns are judged as incorrect.

Seed	{GENE, GENE peroxidase}	{CHEMICAL, GENE agonist}	{DISEASE, cellular DISEASE}	{SPECIES, female SPECIES}
1	SPECIES telomerase	DISEASE chemopreventive agent	hepatic DISEASE	male SPECIES
2	CHEMICAL aminotransferase	CHEMICAL chelation	degradation of GENE	DISEASE cell
3	GENE promoter	GENE antagonist	dermal DISEASE	pregnant SPECIES
4	CHEMICAL hydrolase	-	clinical DISEASE	adult SPECIES
5	CHEMICAL oxidase	-	GENE phosphorylation	SPECIES hepatocyte
6	CHEMICAL acetyltransferase	-	-	SPECIES embryo
7	GENE kinase	-	-	normal SPECIES
8	CHEMICAL kinase	-	-	juvenile SPECIES
9	CHEMICAL peroxidase	-	-	adult male SPECIES
10	CHEMICAL dismutase	-	-	f334 SPECIES

- **Embedding** [45] adopts Word2Vec to learn the representation vector of each meta-pattern by viewing it as a single token in the corpus. Given the seed patterns, the expanded patterns are ranked by the sum of distances away from the seed patterns and the top-10 expanded patterns are returned as the results.
- **SetExpan** [44] is an ablation of the PeNNER framework, where the seed pattern sets of each entity type are expanded one at a time instead of simultaneously.

For SetExpan and PeNNER, we set M, M_0, N to be 200, 120, and 10, respectively. If the final expanded pattern has more than 10 patterns, we only take the top-10 expanded patterns as the results.

2.3.3 Extracting New Meta-Patterns

We first look at the meta-pattern expansion results. During the initial entity tagging step, we use PubTator to recognize five biomedical entity types: gene/protein, chemical, disease, species, and SNP. We ignore the SNP type in PeNNER because the SNP entities are sparse in our input corpus. For each of the four entity types, we take two seed patterns to see whether PeNNER can find new nested naming structures of the same type. The results are shown in Tables 2.2, 2.3, and 2.4.

At the pattern level, PeNNER consistently achieves better performance than the two baselines. Patterns extracted by Embedding are noisy, and some of them are even syntactically wrong. This is because Embedding only considers semantic similarity while ignoring frequency. For patterns that are not so frequent, their context in the corpus is limited, so the quality of their representations learned by Word2Vec may not be good. In contrast,

PeNNER cares about both semantics and frequency. If the extracted patterns appear very often in the corpus, we have a good reason to trust the quality of its context information.

SetExpan does not exploit the mutual exclusiveness of different seed sets. As we can see, "CHEMICAL" and "DISEASE", as entity types, are far more frequent than other quality meta-patterns in the corpus. Although they may not be semantically similar to "GENE", they will be ranked high if frequency and semantics are considered comprehensively. As a result, "CHEMICAL" and "DISEASE" will be expanded into the gene set after the first round. This may cause severe *semantic drift* problems since other disease- or chemical-related patterns may be excluded in the next few rounds. In contrast, PeNNER never considers "CHEMICAL" or "DISEASE" as candidates for the gene set under the multi-set co-expansion mechanism since the "CHEMICAL" and "DISEASE" patterns have already appeared in the chemical and disease seed sets, respectively. We discuss the four entity types with their expanded pattern sets by PeNNER in detail below.

- **GENE**: PeNNER discovered ten gene/protein meta-patterns that are all correct. One interesting observation is that most of the expanded meta-patterns for the gene type belong to the same fine-grained gene type: enzymes. This fine-grained type enzyme is also the type of the seed meta-pattern "GENE peroxidase" in the gene pattern set. Generally speaking, if entities appear in the same meta-pattern set, they are likely to be similar to each other or belong to the same fine-grained entity type. For example, the chemical instances of the meta-pattern "CHEMICAL aminotransferase" include "alanine", "aspartate", "tyrosine", and "ornithine". All four chemicals above belong to the same fine-grained chemical type: amino acids.

- **CHEMICAL**: PeNNER discovers three chemical meta-patterns, among which "GENE antagonist" is the counterpart of "GENE agonist" in the seed pattern set. The chemical instances of the meta-pattern "CHEMICAL chelation" include "iron", "copper", "zinc", and "EDTA". We observe that "CHEMICAL chelation" as a whole entity is more complete than its partial CHEMICAL entity since metal ions and their chelations are different types of chemicals.

- **DISEASE**: PeNNER discovered five disease meta-patterns, among which three are correct and the other two are biological processes. One correct example meta-pattern is "hepatic DISEASE". The disease instances of the meta-pattern "hepatic DISEASE" include "fibrosis", "inflammation", "tumor", and "toxicity". Similarly, we observe that "hepatic fibrosis" is a more complete entity than the partial entity "fibrosis". PubTator also recognizes "liver fibrosis" or "liver inflammation" as a complete entity (see abstracts

Table 2.5: $NDCG$@10 of different methods on the four types.

Method	GENE	CHEMICAL	DISEASE	SPECIES
EMBEDDING	0.139	0.580	0.073	0.315
SETEXPAN	0.602	0.312	**0.754**	0.417
PeNNER	**1.000**	**1.000**	**0.754**	**0.776**

Table 2.6: Number of instances extracted by different methods on the four entity types.

Method	GENE	CHEMICAL	DISEASE	SPECIES
EMBEDDING	79	139	61	45
SETEXPAN	1734	**458**	**184**	2211
PeNNER	**5254**	**458**	**184**	**3212**

with PMIDs 30079841 and 23813842 as two examples). So our expanded pattern "hepatic DISEASE" for diseases is consistent with the entity recognition principle from PubTator.

- **SPECIES**: PeNNER discovered ten species meta-patterns, among which eight are correct and the other two are cell types. Our expanded meta-patterns incorporate certain attributes (e.g., "male") to the species entities that are beneficial to downstream knowledge extraction tasks. For example, one sentence is "Amphetamine and cocaine decreased susceptibility to myoclonus in young mice and increased susceptibility in mature mice". If the BioNER methods ignore "young" and "mature", the facts extracted from this sentence will be inaccurate and even controversial each other. PeNNER successfully recognized "young mice" and "mature mice" as whole entities, which benefits downstream tasks such as relation extraction from the above sentence.

2.3.4 Recognizing Nested Entities

Taking the expanded meta-patterns for each entity type, we match the meta-patterns to concrete entity mentions in the input corpus. Those matched entity mentions are considered as the outer entities and their boundaries and types are corrected by their matched patterns.

Tables 2.5 and 2.6 show the performance of PeNNER in nested NER. Since the expanded meta-patterns form a ranked list, we use *normalized discounted cumulative gain* (nDCG) [48] to evaluate the rank-aware precision. Besides precision, we also show the number of correct instances extracted by each method on the four entity types. This can be regarded as the instance-level "recall" of nested NER.

From Tables 2.5 and 2.6, we observe that PeNNER consistently outperforms the baselines both in precision and recall. For the genes and chemicals, all the instances extracted by

Table 2.7: Pattern expansion results of PeNNER for Biological Process and Treatment entities. Grey patterns are judged as incorrect.

Seed	{GENE upregulation, GENE downregulation}	{CHEMICAL injection, CHEMICAL inhalation}
1	GENE expression	CHEMICAL treatment
2	GENE phosphorylation	CHEMICAL administration
3	the development of DISEASE	CHEMICAL exposure
4	GENE induction	treatment with CHEMICAL
5	CHEMICAL action	exposure to CHEMICAL
6	identification of GENE	administration of CHEMICAL
7	GENE suppression	pretreatment with CHEMICAL
8	DISEASE reduction	CHEMICAL pretreatment
9	CHEMICAL production	-
10	GENE activity	-

PeNNER are correct. We observe that PeNNER extracted 5,254 nested gene entities from 28,007 PubMed abstracts (i.e., on average, one nested GENE entity in every 5.33 abstracts). It further confirms our motivation that the nested named entities are common in biomedical literature.

2.3.5 Finding New Types of Entities

We further demonstrate an advantage of PeNNER over the fully-supervised methods: finding new types of entities. If one entity type has not been included in the training set, it would be extremely difficult for the supervised methods to recognize entities of this type. In biomedical literature, biomedical processes [49] and treatment entities [50] attract great attention. For example, detecting biological process patterns such as "GENE expression" and "GENE phosphorylation" are useful in connecting gene/protein-disease-drug in the context of gene-variant [51] and protein modification (PTM) [52]. However, these two types, biological process and treatment, are not commonly annotated in the BioNER dataset and have not been included in PubTator. Under this setting, PeNNER shows its power. Similar to the meta-pattern expansion process for the known entity types, we take only two seed meta-patterns for each new entity type. Table 2.7 shows the pattern expansion results on the two new types: biological process and treatment.

- **Biological Process**: Taking "GENE upregulation" and "GENE downregulation" as the seed meta-patterns, PeNNER discovered ten additional meta-patterns from the corpus, among which eight are correct. Similar to the seed meta-patterns, most of the extracted biological process patterns are describing the activities of genes.

- **Treatment**: Taking "CHEMICAL injection" and "CHEMICAL inhalation" as the seed meta-patterns, discovered eight additional meta-patterns from the corpus, among which

six are correct. One interesting mistake PeNNER makes here is about an ambiguous pattern "CHEMICAL exposure". In fact, the pattern "CHEMICAL exposure" indicates different types with different fine-grained chemical types embedded in this pattern. For example, if the chemical is a drug (e.g., "resveratrol", "simvastatin", or "quercetin"), the pattern "CHEMICAL exposure" indicates a treatment entity. On the other hand, if the chemical is a toxic chemical (e.g., mercury, lead, and hydrofluoric acid), the pattern "CHEMICAL exposure" indicates a symptom entity. This observation motivates us to study more fine-grained entity typing as the first step for meta-pattern disambiguation.

2.3.6 Case Study

To further demonstrate the improvements of PeNNER over PubTator, we compare the annotation results on several sentences by PeNNER and PubTator in Table 6.5. In the first two sentences, PubTator can only do flat NER, while PeNNER successfully detects CHEMICAL-GENE and GENE-CHEMICAL nested naming structures. We also observe that recognizing more complete entities benefits downstream applications such as relation extraction. For example, in the second sentence, the inhibition relation happens between "MCP-1" and "Erk1/2 antagonist" instead of its partial entity "Erk1/2". Failing to recognize the whole entity "Erk1/2 antagonist" will lead to an opposite relation with the original sentence because "antagonist" means a suppressor of the protein "Erk1/2". In the third and fourth sentences, PeNNER recognizes entities of the new types: biological process and treatment. Similarly, the GENE-PROCESS nested naming structure leads to an accurate relation extraction because it is "STAT1 phosphorylation" instead of "STAT1" that is being up-regulated in the third sentence.

Despite the impressive results of PeNNER, there is still room for future improvements. For example, the meta-patterns can be utilized in a more general way. PeNNER mainly uses meta-patterns with only one entity type token to deal with the nested naming structures. However, meta-patterns with two or more type tokens may also be useful. We still take the sentences in Table 6.5 as examples. In the first sentence, the abbreviations of the genes (i.e., "SOD" and "Ala-AT") are not recognized. In fact, we do extract a quality meta-pattern "GENE (GENE)". If we already know that the entity outside of the brackets is a gene/protein, we may infer that the inside one is a gene/protein as well. In the third sentence, "STAT3 and STAT5 phosphorylation" is suppressed. However, we only find "STAT5 phosphorylation" and leave "STAT3" alone. It is possible to utilize meta-patterns such as "GENE and GENE phosphorylation" to find a more complete nested naming structure.

Table 2.8: Case study of the NER results. Differences of PubTator and PENNER results are marked in bold. In contrast with PubTator, PENNER is able to detect nested entity structures as well as new types of entities.

	PMID: 15820610
PubTator	The aim of the present study was to determine the effect of HRT on the activities of an antioxidant enzyme [**superoxide**]$_{\text{CHEMICAL}}$ dismutase (SOD) and aminotransferases like [**alanine**]$_{\text{CHEMICAL}}$ aminotransferase (Ala-AT) and [**aspartate**]$_{\text{CHEMICAL}}$ aminotransferase in different age groups ...
PENNER	The aim of the present study was to determine the effect of HRT on the activities of an antioxidant enzyme [[**superoxide**]$_{\text{CHEMICAL}}$ **dismutase**]$_{\text{GENE}}$ (SOD) and aminotransferases like [[**alanine**]$_{\text{CHEMICAL}}$ **aminotransferase**]$_{\text{GENE}}$ (Ala-AT) and [[**aspartate**]$_{\text{CHEMICAL}}$ **aminotransferase**]$_{\text{GENE}}$ in different age groups ...
	PMID: 10919993
PubTator	Mitogen-activated protein (MAP) kinase [**Erk1/2**]$_{\text{GENE}}$ antagonist mainly inhibited the release of [MCP-1]$_{\text{GENE}}$, whereas MAP kinase [**p38**]$_{\text{GENE}}$ antagonist mainly suppressed the release of [IL-8]$_{\text{GENE}}$ and [RANTES]$_{\text{GENE}}$.
PENNER	Mitogen-activated protein (MAP) kinase [[**Erk1/2**]$_{\text{GENE}}$ **antagonist**]$_{\text{CHEMICAL}}$ mainly inhibited the release of [MCP-1]$_{\text{GENE}}$, whereas MAP kinase [[**p38**]$_{\text{GENE}}$ **antagonist**]$_{\text{CHEMICAL}}$ mainly suppressed the release of [IL-8]$_{\text{GENE}}$ and [RANTES]$_{\text{GENE}}$.
	PMID: 21266192
PubTator	... it suppressed [STAT3]$_{\text{GENE}}$ and [**STAT5**]$_{\text{GENE}}$ phosphorylation in HS-578T cells, whereas it up-regulated [**STAT1**]$_{\text{GENE}}$ phosphorylation and down-regulated [**STAT5**]$_{\text{GENE}}$ phosphorylation in MCF-7 cells.
PENNER	... it suppressed [STAT3]$_{\text{GENE}}$ and [[**STAT5**]$_{\text{GENE}}$ **phosphorylation**]$_{\text{PROCESS}}$ in HS-578T cells, whereas it up-regulated [[**STAT1**]$_{\text{GENE}}$ **phosphorylation**]$_{\text{PROCESS}}$ and down-regulated [[**STAT5**]$_{\text{GENE}}$ **phosphorylation**]$_{\text{PROCESS}}$ in MCF-7 cells.
	PMID: 10498651
PubTator	[**COL1A2**]$_{\text{GENE}}$ expression was decreased by [**vitamin E**]$_{\text{CHEMICAL}}$ treatment or transfection with [**manganese superoxide**]$_{\text{CHEMICAL}}$ dismutase, and was further increased after treatment with [**L-buthionine sulfoximine**]$_{\text{CHEMICAL}}$ ...
PENNER	[[**COL1A2**]$_{\text{GENE}}$ **expression**]$_{\text{PROCESS}}$ was decreased by [[**vitamin E**]$_{\text{CHEMICAL}}$ **treatment**]$_{\text{TREATMENT}}$ or transfection with [[**manganese superoxide**]$_{\text{CHEMICAL}}$ **dismutase**]$_{\text{GENE}}$, and was further increased after [**treatment with** [**L-buthionine sulfoximine**]$_{\text{CHEMICAL}}$]$_{\text{TREATMENT}}$ ...

2.4 RELATED WORK

Several methods have been proposed for flat NER. Early supervised methods are based on hidden markov models (HMMs) [53] or conditional random fields (CRFs) [54]. Recently, recurrent neural networks (RNNs) have been widely applied to several sequence labeling tasks. Lample et al. [55] proposed neural models based on long short-term memory networks

(LSTMs) for flat named entity recognition and achieved state-of-the-art performance.

There are fewer approaches address the problem of nested entities. Alex et al. [33] presented several techniques based on CRFs for nested NER in the GENIA dataset. They obtained their best results from a cascaded approach, where they applied CRFs in a specific order on the entity types, such that each CRF utilizes the output derived from previous CRFs. However, their approach could not identify nested entities of the same type. Finkel and Manning [34] proposed a CRF-based constituency parser for nested named entities such that each named entity is a constituent in the parse tree. Their model achieved state-of-the-art results on the GENIA dataset. However, the time complexity of their model is $O(n^3)$, where n is the number of tokens in the sentence, making inference slow. Lu and Roth [35] further proposed a linear time directed hypergraph-based model.

While most previous efforts for nested entity recognition were limited to named entities, Lu and Roth [35] addressed the problem of nested entity mention detection where mentions can either be named, nominal or pronominal. Their hypergraph-based approach is able to represent the potentially exponentially many combinations of nested mentions of different types. They adopted a CRF-like log-linear approach to learn these mention hypergraphs and employed several hand-crafted features defined over the input sentence and the output hypergraph structure. Recently, Muis and Lu [36] introduced the notion of mention separators for nested entity mention detection. In contrast to the hypergraph representation that Lu and Roth [35] adopt, they learn a multigraph representation and are able to perform exact inference on their structure. It is an interesting orthogonal approach for nested entity mention detection.

Neural network models for nested NER are recently proposed as extensions to the state-of-the-art RNN-based models for flat NER. Katiyar and Cardie [37] proposed to learn a hypergraph representation for nested entities using features extracted from a recurrent neural network. Ju et al. [38] proposed to dynamically stack flat NER layers and recognize outer entities with additional information from their inner entities. The neural network models save human effort for feature generation. However, they require a large amount of training data and are not easily adapted to new entity types.

2.5 SUMMARY

In this chapter, we proposed a framework PeNNER that automatically discovers nested naming structures in biomedical literature. Taking a corpus pre-tagged by any existing flat NER tool, PeNNER extracts quality meta-patterns in an unsupervised way and finds meta-patterns associated with each entity type under very weak supervision. Experiments

demonstrate that PeNNER outperforms the baselines by a large margin in finding quality meta-patterns and nested named entities. In addition, PeNNER is also able to find new types of entities with just two user-specified seed patterns. Case studies demonstrate that the PeNNER largely improves the annotation results by PubTator. One interesting mistake PeNNER currently makes is about ambiguous patterns (e.g., "CHEMICAL exposure" can indicate treatment or symptom depending on whether the chemical is a drug or toxic chemical). This observation motivates us to study fine-grained named entity recognition as the first step to benefit meta-pattern disambiguation. In the next chapter, we will introduce a method for fine-grained chemistry named entity recognition under distant supervision.

CHAPTER 3: ONTOLOGY-GUIDED DISTANT SUPERVISION FOR FINE-GRAINED CHEMISTRY NAMED ENTITY RECOGNITION

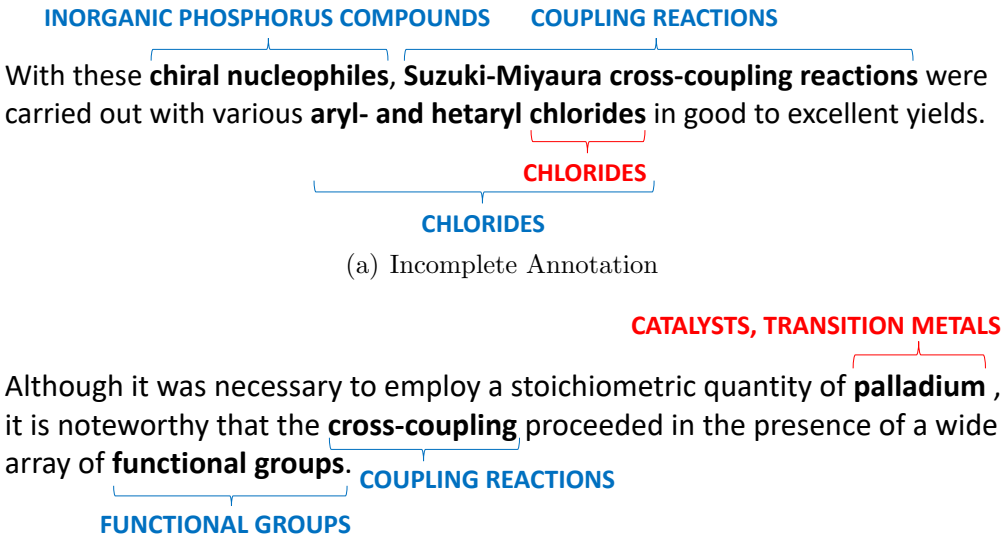
3.1 INTRODUCTION

Named entity recognition (NER) is a fundamental step in scientific literature analysis to build AI-driven systems for molecular discovery, synthetic strategy designing, and manufacturing [24, 25, 26, 27, 56]. The NER task aims to locate and classify entity mentions (e.g., "Suzuki-Miyaura cross-coupling reactions") from unstructured text into pre-defined categories (e.g., "coupling reactions"). In the chemistry domain, previous NER studies are mostly focused on one coarse-grained entity type (i.e., chemicals) [29, 57, 58] and rely on large amounts of manually-annotated data for training deep learning models [3, 13, 55, 59, 60, 61].

In real-world applications, it is important to recognize chemistry entities on diverse and fine-grained types (e.g., "inorganic phophorus compounds", "coupling reactions" and "catalysts") to provide a wide range of information for scientific discovery. It will need dozens to hundreds of distinct types, making consistent and accurate annotation difficult even for domain experts. On the other hand, the domain-specific ontologies and knowledge bases (KBs) can be easily accessed, constructed, or integrated, which makes distant supervision realistic for fine-grained chemistry NER.

Still, challenges exist for correctly recognizing the entity boundaries and accurately typing entities with distant supervision. In distant supervision, training labels are generated by matching the mentions in a document with the concepts in the knowledge bases (KBs). However, this kind of KB-matching suffers from two major challenges: (1) *incomplete annotation* where a mention in a document can be matched only partially or missed completely due to an incomplete coverage of the KBs (Figure 3.1(a)), and (2) *noisy annotation* where a mention can be erroneously matched due to the potential matching of multiple entity types in the KBs (Figure 3.1(b)). Due to the complex name structures (e.g., nested naming structures and long chemical formulas) of chemical entities, these challenges lead to severe low-precision and low-recall for fine-grained chemistry NER with distant supervision.

Several studies have attempted to address the incomplete annotation problem in distantly-supervised NER. For example, AutoNER [62] introduces an "unknown" type that can be skipped during training to reduce the effect of false negative labeling with distant supervision. BOND [63] leverages the power of pre-trained language models and a self-training approach to iteratively incorporate more training labels and improve the NER performance. However, previous methods assume a high precision and reasonable coverage of KB-matching for distant label generation. For example, the KB-matching on the CoNLL03 dataset [63]



(a) Incomplete Annotation

(b) Noisy Annotation

Figure 3.1: Two major challenges of distant supervision for fine-grained chemistry NER: incomplete annotation and noisy annotation. The KB-matching labels are marked in red and the true labels are marked in blue.

reported over 80% on precision and over 60% on recall. These methods do not work well with fine-grained chemistry NER that has severe low precision and low recall with KB-matching. Previous studies also largely ignore the noisy annotation problem by simply discarding those multi-labels during the KB-matching process [63]. However, the noisy labels cannot be simply ignored for the chemistry entities because they consist of a large portion of distant training labels. We observe that more than 60% of the entities have multiple labels during KB-matching in the chemistry domain.

In this chapter, we propose ChemNER, an ontology-guided, distantly-supervised NER method for fine-grained chemistry NER. Taking an input corpus, a chemistry type ontology, and associated entity dictionaries collected from the KBs, we develop a flexible KB-matching method with TF-IDF-based majority voting to resolve the incomplete annotation problem. Then we develop an ontology-guided multi-type disambiguation method to resolve the noisy annotation problem. Taking the output from the above two steps as distant supervision, we further train a sequence labeling model to cover additional entities. ChemNER significantly improves the distant label generation for the subsequent NER model training. We also provide an expert-labeled, chemistry NER dataset with 62 fine-grained chemistry types (e.g., chemical compounds and chemical reactions). Experimental results show that ChemNER is highly effective, achieving substantially better performance (with .25 absolute F1 score improvement) compared with the state-of-the-art NER methods.

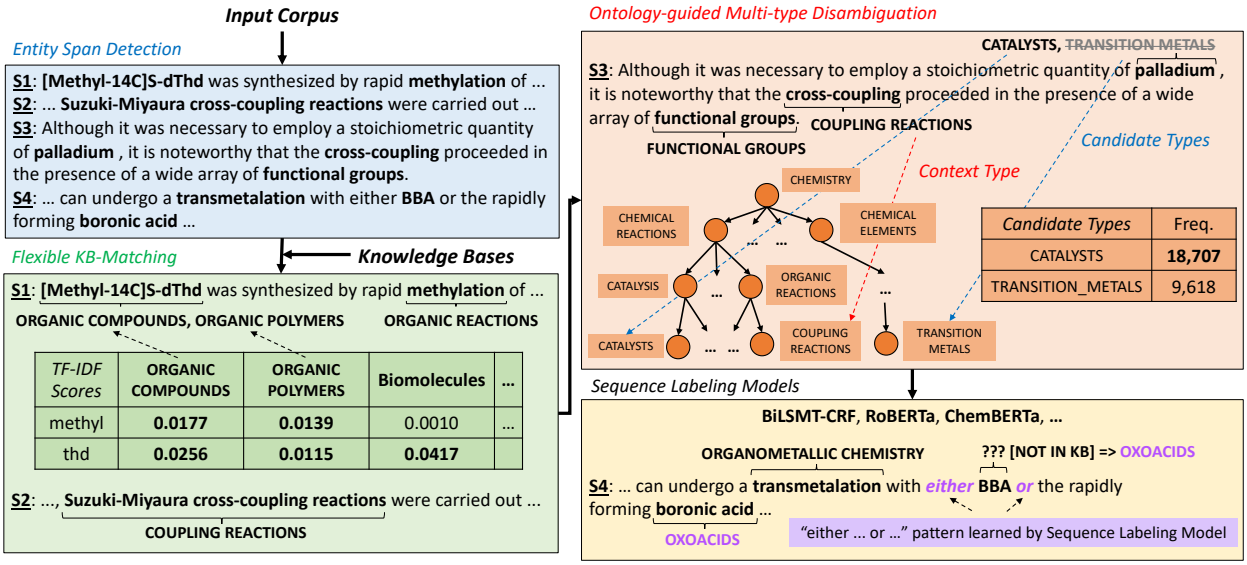


Figure 3.2: The overall framework of ChemNER. It includes a distant label generation (entity span detection, flexible KB-matching, and ontology-guided multi-type disambiguation) and a sequence labeling model training.

3.2 THE CHEMNER FRAMEWORK

The ChemNER framework is shown in Figure 6.1. It includes distant label generation (entity span detection, flexible KB-matching, and ontology-guided multi-type disambiguation) and sequence labeling model training.

3.2.1 Data Preparation

The input to ChemNER includes two parts: (1) a chemistry literature corpus, and (2) a fine-grained chemistry type ontology and associated entity dictionaries for each type.

Corpus Collection For this study, we collected a chemistry literature corpus from PubChem[4]. This corpus contains 4,608 papers, among which 319 papers have the full-text and all have the title and abstract. There are 71,406 sentences in this corpus.

Type Ontology and Dictionary Collection We collected a fine-grained chemistry type ontology from Wikipedia categories rooted under the *Chemistry* category[5]. We treat the Wikipedia category pages as types and the titles of the pages associated with each category as the entity dictionary for each type. We further remove irrelevant types and merge some

[4] https://pubchem.ncbi.nlm.nih.gov/
[5] https://en.wikipedia.org/wiki/Category:Chemistry

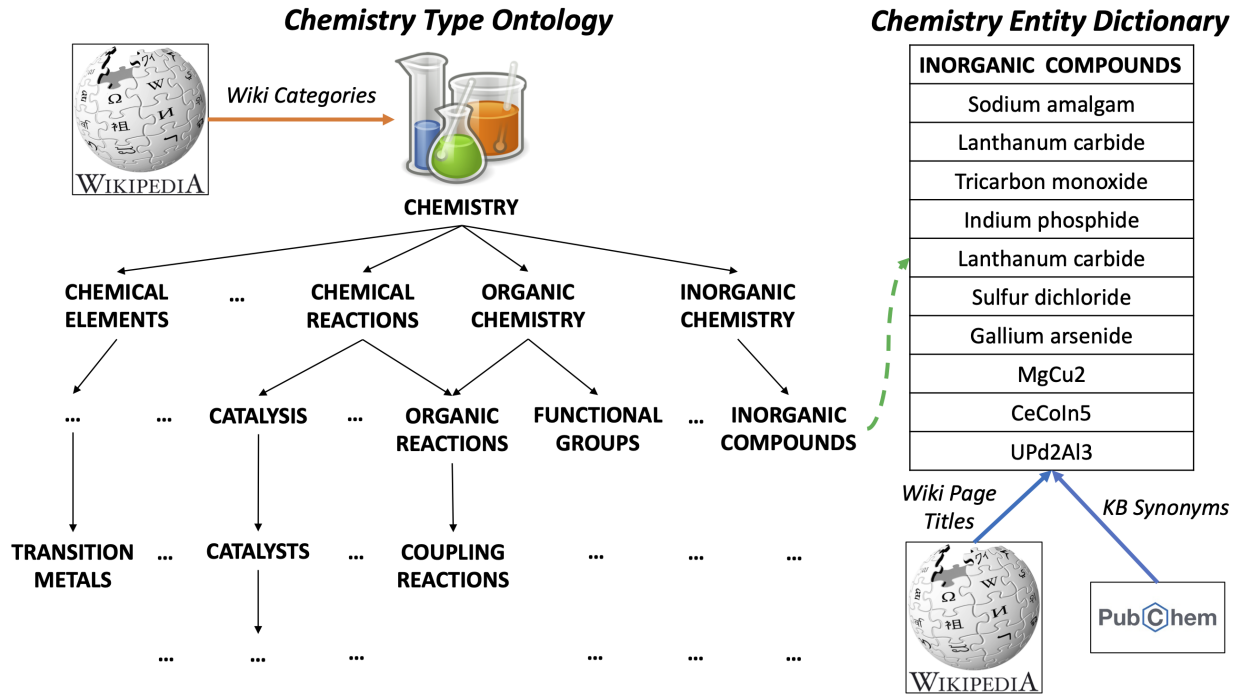


Figure 3.3: Illustration of the chemistry type ontology construction and dictionary collection.

fine-grained types to their coarse-grained parent types based on their term frequencies in the corpus. We also expand the entity dictionaries with synonyms collected from the PubChem knowledge base. Finally, we obtained a fine-grained chemistry entity type ontology with 62 types and its associated dictionaries with 10,551 entities. Figure 3.3 shows a subset of our chemistry type ontology.

3.2.2 Flexible KB-Matching

Taking the input corpus, chemistry type ontology, and associated entity dictionaries collected from the KBs, we first develop a flexible KB-matching method to resolve the *incomplete annotation* problem. Chemistry entities usually have complex naming structures, such as nested naming structures (e.g., "aryl chloride" where "aryl" is a FUCNTIONAL GROUP, "chloride" is a HALIDE, but altogether is an ORGANOHALIDE) and long chemical formulas (e.g., "Methyl 3'-(((Trifluoromethyl)sulfonyl)oxy)-[1,1'-biphenyl]-4-carboxylate"). As a result, the chemical names are quite flexible and cannot be fully covered by the KBs. Simple KB-matching used in previous distantly-supervised NER methods [62, 63] cannot match those complex chemistry entities that do not exist in the KBs, which leads to a severe low precision and low recall for labeling the fine-grained chemistry entities.

We propose to first conduct **entity span detection** with chemistry phrase chunking tools

followed by a flexible KB-matching to resolve the incomplete KB-matching problem. We use two phrase chunking tools, ChemDataExtractor [64] and Genia Tagger [65], to generate candidate entity spans in the input corpus (e.g., in Figure 6.1 sentence S2, the phrase chunking tools find "Suzuki-Miyaura cross-coupling reactions" as a candidate entity span.) Based on the detected candidate entity spans, we develop a flexible KB-matching method with TF-IDF-based majority voting to resolve the incomplete annotation problem.

The flexible KB-matching method can match long and complex chemistry entities (e.g., chemical compounds) that do not exist in the KBs. Specifically, we label each candidate entity span by letting each word token in the entity span vote for several entity types that are most likely to involve this word token. For example, in Figure 6.1 sentence S1, "[Methyl-14C]S-Thd", which is short for "4'-[methyl-14C]thiothymidine" according to the original document, is an author-defined abbreviation that cannot be covered by the existing KBs. However, since "Methyl-" is a common functional group that is usually the prefix of the organic compounds, this word token in "[Methyl-14C]S-Thd" helps vote for the types "ORGANIC COMPOUNDS" and "ORGANIC POLYMERS". Another example is sentence S2, where three ("suzuki", "coupling", "reaction") out of the five word tokens in "Suzuki-Miyaura cross-coupling reactions" help vote for the type "COUPLING REACTIONS".

Formally, let $e = [w_1, w_2, \ldots, w_n], w_i \in \mathcal{V}$, where e denotes each candidate entity span, w_i each word token in the entity span, and $\mathcal{V}$ the vocabulary. Let $\mathcal{T}$ denote the set of fine-grained types and D_t the dictionary of entities for type $t \in \mathcal{T}$. The TF-IDF score of each word token w for each entity type $t \in \mathcal{T}$ is calculated as follows:

$$TF\text{-}IDF(w,t) = TF(w,t) * IDF(w,t), \tag{3.1}$$

$$TF(w,t) = \frac{f(w, D_t)}{\sum_{w' \in \mathcal{V}} f(w', D_t)}, \tag{3.2}$$

$$IDF(w,t) = log\left(\frac{|\mathcal{T}|}{|\{t \mid t \in \mathcal{T}, w \in D_t\}|}\right), \tag{3.3}$$

where $f(w, D_t)$ denotes the frequency of the word token w appearing in the dictionary D_t.

We set a minimum TF-IDF threshold $\theta = 0.02$ to eliminate the common words from voting for the entity types. Then we let each word token vote for several entity types that has the highest TF-IDF scores above the mininum TF-IDF threshold and generate the distant labels by taking the majority voting. Note that this step can generate multi-type labels for the candidate entity spans due to ties in the majority voting. We resolve this problem with an ontology-guided multi-type disambiguation method as the next step.

3.2.3 Ontology-Guided Multi-Type Disambiguation

Based on the output of flexible KB-matching and the chemistry type ontology structure, we develop an ontology-guided multi-type disambiguation method to resolve the *noisy annotation* problem. An intuition of multi-type disambiguation is that the entities in the same sentence, paragraph or document usually follow a focused topic. For example, if a sentence is talking about organic chemistry, the entities in this sentence are more likely to have types related to organic chemistry. Following this intuition and the chemistry type ontology structure (Section 3.2.1), we draw two insights for an automated multi-type disambiguation: (1) the entity types in one sentence are usually confined to one big branch on the chemistry type ontology (e.g., organic or inorganic chemistry), and (2) the type of an entity under local context should be close to the types of the surrounding entities in the same sentence on the chemistry type ontology. For example, in Figure 6.1, sentence S3 contains one entity "palladium" that has two candidate types: "CATALYSTS" that falls under "CHEMICAL REACTIONS" and "TRANSITION METALS" that falls under "CHEMICAL ELEMENTS". By looking at its surrounding entities (e.g., "cross-coupling"), we see that the surrounding entity types (e.g., "COUPLING REACTIONS" for "cross-coupling") fall under the "ORGANIC REACTIONS" branch, which is also under the larger "CHEMICAL REACTIONS" branch, on the type ontology. So the sentence S3 is likely talking about chemical reaction and "palladium" is more suitable to have a type "CATALYSTS" instead of "TRANSITION METALS" based on the local context.

Formally, let $s = [e_1, e_2, \ldots, e_n]$, where s denotes a sentence and e_i ith entity mention in it that has been assigned an initial label set $T_{e_i} = \{t^1_{e_i}, \ldots, t^m_{e_i}\}$ with flexible KB-matching. For an entity e_i with multiple candidate types ($|T_{e_i}| > 1$) to be resolved, we calculate the inverse distance between this candidate type and the distribution of the surrounding types on the type ontology. The disambiguation score for each candidate type $S_d(t^j_{e_i})$ is defined as

$$S_d(t^j_{e_i}) = \frac{\sum_{k\in[1..n],k\neq i,|T_{e_k}|=1} dep(lca(t_{e_k}, t^j_{e_i}))}{n * dep(t^j_{e_i})}, \tag{3.4}$$

where $lca(\cdot,\cdot)$ denotes the lowest common ancestor of two types on the type ontology and $dep(\cdot)$ denotes the depth of the type on the type ontology. A larger $S_d(t^j_{e_i}) \in (0, 1)$ indicates that the candidate type $t^j_{e_i}$ is more likely to be the correct type of entity e_i in sentence s.

If the surrounding types in the sentence still draw ties for the candidate type resolution, we could further enlarge the scope to a few surrounding sentences, the paragraph, the document or the corpus. We introduce a corpus-level global popularity score for each type based on our experimental observations. As shown in Figure 6.1, we calculate the frequency of each

type in our initially labeled corpus with flexible KB-matching. "CATALYSTS" is globally more popular with a frequency of 18,707 compared to "TRANSITION METALS" with a frequency of 9,618. The global popularity score for each candidate type $S_g(t^j_{e_i})$ is defined as

$$S_g(t^j_{e_i}) = \frac{f_c(t_{e_i})}{\sum_{t' \in \mathcal{T}} f_c(t')}, \tag{3.5}$$

where $f_c(\cdot)$ denotes the frequency of the type in the flexible KB-matched corpus. $S_g(t^j_{e_i}) \in (0, 1]$ and a larger score indicates that the candidate type $t^j_{e_i}$ is more likely to be the correct type for the entity e_i globally in the corpus.

The final score $S(t^j_{e_i})$ of the candidate type $t^j_{e_i}$ is a combination of the local disambiguation score $S_d(t^j_{e_i})$ and the global popularity score $S_g(t^j_{e_i})$:

$$S(t^j_{e_i}) = S_d(t^j_{e_i}) * S_g(t^j_{e_i}) \in (0, 1). \tag{3.6}$$

We choose the type $t^j_{e_i}$ for the entity e_i that has a highest score $S(t^j_{e_i})$ for multi-type disambiguation.

3.2.4 Sequence Labeling Models

The flexible KB-matching and multi-type disambiguation still rely on the signals from the KBs and ontologies, which cannot cover all the new entities in the corpus. Taken the output from the above two steps as distant supervision, we further train a sequence labeling model to solve the sparsity labeling problem. For example, in Figure 6.1 sentence 4, "BBA" is a new entity that cannot be labeled by flexible KB-matching since there is no obvious token-level signals. However, there is a "boronic acid" entity with the type "OXOACIDS" in its surrounding context. The sequence labeling models will be able to capture those context patterns such as "either ... or ..." that usually connect entities with similar types. Thus they are likely to recognize "BBA" with the type "OXOACIDS".

Based on the distant labels generated by the flexible KB-matching and multi-type disambiguation, we train a sequence labeling model (e.g., RoBERTa, ChemBERTa) without any constraints on the type of model to use. The loss function is defined as:

$$l =_\theta \sum_i^n loss(h_\theta(x_i), y), \tag{3.7}$$

where $h_\theta(\cdot)$ is the output of the sequence labeling model and y is our generated distant label. This is equivalent to minimizing the cross-entropy error between the outputs of the sequence

labeling model and our generated distant labels.

3.3 EXPERIMENTS

3.3.1 Dataset

We provide a chemistry NER dataset covering 62 fine-grained chemistry types such as chemical compounds and chemical reactions. This dataset can be used to benchmark distantly supervised NER methods for the fine-grained chemistry NER task. The input for training includes two parts: (1) a chemistry literature corpus with 69,806 unlabeled sentences, and (2) a chemistry type ontology with 62 fine-grained chemistry types and associated entity dictionaries for each type (Section 3.2.1). The test set contains 1,600 expert-annotated sentences on the fine-grained chemistry types. We use this test set to compare the performance of different NER methods in our experiments. We report the entity-level micro-precision, micro-recall, and micro-F1 scores[6] of each NER method on the human-annotated test set. We have released all of our data and code for future studies, including the chemistry literature corpus, fine-grained entity type ontology and associated dictionaries collected from Wikipedia-Chemistry, manually-annotated test set for NER performance evaluation, and the code of ChemNER.

Corpus Collection We collected a corpus for Suzuki Coupling reactions in the chemistry domain. Suzuki coupling is an important reaction for carbon–carbon bond formation in organic chemistry. Recent studies have focused on the Suzuki coupling reactions to build AI-driven systems for molecular discovery, synthetic strategy designing, and manufacturing. This corpus contains 4,608 papers that are retrieved from PubChem[7] with the query "Suzuki Coupling", among which 319 papers have the full-text and all have the title and abstract. There are in total 71,406 sentences in this corpus.

Dictionary Collection We collected a fine-grained chemistry entity type ontology from Wikipedia by treating category pages as types and the titles of the pages associated with each category as the entities for each type. We first conducted a depth-first search (DFS) starting from the *Chemistry* category[8] and found that the search did not stop when one million categories had been visited, and it often happened that a category relevant to Chemistry

[6] `https://github.com/chakki-works/seqeval`
[7] `https://pubchem.ncbi.nlm.nih.gov/`
[8] `https://en.wikipedia.org/wiki/Category:Chemistry`

has irrelevant children. Therefore, we decide to use a technical term list to filter out irrelevant categories. We collected a spell-checker dictionary [66] with over 104,000 technical chemistry terms and dropped a category from the search if less than 20% of one-grams in its name and the names of all its direct children were covered by the dictionary. The threshold of 20% was selected empirically. After this step, we obtained a fine-grained chemistry entity type ontology with 3,775 types and 101,415 entities. We future tailor the entity type ontology and their associated entities by removing some irrelevant types and merging some fine-grained types to their coarse-grained parent types based on their frequencies in our chemistry literature corpus. We also expand the entity dictionaries with synonyms collected from the PubChem knowledge base. Finally, we obtained a fine-grained chemistry entity type ontology with 62 types and 10,551 entities.

Test Set Annotation We randomly select 1,600 sentences from the corpus and ask three domain experts to annotate each sentence as our test sets. We leave the remaining sentences (69,806 sentences in the corpus) as the training set for distant supervision. We provide the annotators with an auto-complete drop-down menu consisting of our entity type vocabulary. Each pair of annotators reach a substantial agreement with a Fleiss's κ of 0.72. The conflicts among annotators are resolved by another senior domain expert in the final test set.

3.3.2 Baselines

We compare the performance of ChemNER with several groups of baseline methods.

- **KB-Matching**: This baseline is a simple string matching as [67]. It is a greedy search algorithm that walks through a sentence trying to find the longest strings that match the entities in the dictionaries. For the strings matched with multiple types, we simply discard those multi-labels as [63].
- **KB-Matching (freq)**: This baseline is a simple improvement of KB-Matching. For the strings matched with multiple types, we choose the type that has the highest frequency in the corpus.
- **BiLSTM-CRF**: This baseline is the BiLSTM-CRF model [60] that takes the results of KB-Matching (freq) as distant supervision.
- **AutoNER**: This baseline is the AutoNER model [62] that directly takes the raw corpus and the dictionaries as the input. It has a built-in KB-matching algorithm that maximizes the total number of matched tokens on each sentence to generate distant supervision. For

the strings matched with multiple types, it assigns equal probabilities to each candidate type during training.

- **RoBERTa**: This baseline is the RoBERTa model [13] that takes the results of KB-Matching (freq) as distant supervision.

- **ChemBERTa**: This baseline is the ChemBERTa model [14] that takes the results of KB-Matching (freq) as distant supervision. The ChemBERTa language model is pre-trained on the SMILE strings of the chemical molecule structures instead of the chemistry corpus. To our knowledge, there is no domain-specific pre-trained language model on the chemistry corpus.

- **BOND**: This baseline is the BOND model [63] that takes the results of KB-Matching (freq) as distant supervision. The original distant supervision is our KB-Matching baseline according to the BOND paper. Here we use the improved KB-Matching (freq) baseline to give the BOND baseline an improved performance.

- **ChemNER$_{\mathbf{F}}$**: This is an ablation model of ChemNER with the flexible KB-Matching only. For the strings matched with multiple types, we simply discard those multi-labels.

- **ChemNER$_{\mathbf{FM}}$**: This is an ablation model of ChemNER with the flexible KB-Matching and the ontology-guided multi-type resolution.

- **ChemNER$_{\mathbf{BiLSTM\text{-}CRF}}$**: This is a variation of ChemNER that takes the results of ChemNER$_{\mathrm{FM}}$ as distant supervision and trains a BiLSTM-CRF model for the final prediction.

- **ChemNER$_{\mathbf{RoBERTa}}$**: This is a variation of ChemNER that takes the results of ChemNER$_{\mathrm{FM}}$ as distant supervision and trains a RoBERTa model for the final prediction. It is also the full model of ChemNER that achieves the best performance.

- **ChemNER$_{\mathbf{ChemBERTa}}$**: This is a variation of ChemNER that takes the results of ChemNER$_{\mathrm{FM}}$ as distant supervision and trains a ChemBERTa model for the final prediction.

- **ChemNER$_{\mathbf{BOND}}$**: This is a variation of ChemNER that takes the results of ChemNER$_{\mathrm{FM}}$ as distant supervision and trains a BOND model for the final prediction.

Table 3.1: Runtime and Number of Parameters

Method	Ave. Runtime	# of Parameters
BiLSTM-CRF	6h	2M
AutoNER	20h	8M
RoBERTa	4h	110M
ChemBERTa	3h	110M
BOND	8h	110M

3.3.3 Parameter Settings

Runtime with Parameters We compared all sequence model we adopted during experiments. Our models are trained on a single NVIDIA Titan Xp (12GB) GPU. The details about the average runtime and the number of parameters are given in Table 3.1. All training hyperparameters follow their original implementation.

- **BiLSTM-CRF**: We used the code base of BiLSTM-CRF[9]. The hyperparameters are set to default values. We trained the BiLSTM-CRF on Suzuki Coupling data with 10 epoches with learning rate as 0.001, hidden dimension as 256, drop rate as 0.5 and use word embedding with dimension of 256.

- **AutoNER**: We adopted the code base from AutoNER's original implementation[10]. The hyperparameters are set to default values. We trained AutoNER model on Suzuki Coupling data with 50 epoches and learning rate as 0.05, hidden dimension as 300, drop rate as 0.5 and use pretrained word embedding with dimension of 200.

- **RoBERTa**: We use the HuggingFace [11] Transformers Python Interface to train the RoBERTa model on the Suzuki Coupling data using the *roberta-base* model with 10 epochs and a batch size of 32. The other hyperparameters are set as default.

- **ChemBERTa**: For ChemBERTa also, we use the HuggingFace Transformers to train the BERT model on the Suzuki Coupling data using the *seyonec/ChemBERTa-zinc-base-v1* model with 10 epochs and a batch size of 32. The other hyperparameters are set as default.

- **BOND**: To train our Suzuki Coupling data using BOND, we use their publicly available code[12] that also uses the HuggingFace Transformers *roberta-base* model as the base model for training. We train the model for 20 epochs with a learning rate of 2e-5. The other hyperparameters are set as default.

[9] `https://github.com/Gxzzz/BiLSTM-CRF`
[10] `https://github.com/shangjingbo1226/AutoNER`
[11] `https://github.com/huggingface/transformers`
[12] `https://github.com/cliang1453/BOND`

Table 3.2: Overall results (%) on the test set.

Method	Precision	Recall	F1
KB-Matching	32.26	4.95	8.58
KB-Matching (freq)	20.51	11.88	15.05
BiLSTM-CRF	21.88	10.40	14.09
AutoNER	20.51	3.96	6.64
RoBERTa	23.55	17.74	20.24
ChemBERTa	17.54	12.28	14.45
BOND	18.84	12.87	15.29
ChemNER	**69.47**	**34.34**	**45.96**

Table 3.3: Results (%) of ChemNER ablation models.

Method	Precision	Recall	F1
ChemNER	69.47	**34.34**	**45.96**
$\text{ChemNER}_{\text{F}}$	**74.76**	29.06	41.85
$\text{ChemNER}_{\text{FM}}$	71.90	32.83	45.08
$\text{ChemNER}_{\text{BiLSTM-CRF}}$	48.65	17.82	26.09
$\text{ChemNER}_{\text{RoBERTa}}$	69.47	**34.34**	**45.96**
$\text{ChemNER}_{\text{ChemBERTa}}$	58.78	29.06	38.89
$\text{ChemNER}_{\text{BOND}}$	52.21	26.79	35.41

3.3.4 Overall Performance

Table 3.2 shows the overall results on the test set of our fine-grained chemistry NER dataset. ChemNER achieves .25 absolute F1 score improvement over the best performing baseline model *RoBERTa*. As we have discussed, the KB-Matching method suffers from severe low precision (32%) and low recall (5%) for labeling the fine-grained chemistry entities, which greatly limits the performance of the baseline NER methods that use KB-Matching for distant supervision.

3.3.5 Ablation Study

Table 3.3 shows the results of ablation studies on the test set of our fine-grained chemistry NER dataset. We compared our ChemNER full model with several ablations and variations. Our ablation model ChemNER_{F} significantly improves the precision and recall over *KB-matching* and ChemNER_{FM} further improves the recall. These two ablations show the effectiveness of our proposed methods, flexible KB-matching and ontology-guided multi-type resolution, for fine-grained chemistry NER under distant supervision. The four full model variations further shows that *RoBERTa* is the best sequence labeling model that takes the output of ChemNER_{FM} as distant supervision.

Table 3.4: Results (%) with different minimum TF-IDF threshold θ for the flexible KB-Matching.

ChemNER$_{\mathbf{F}}$	**Precision**	**Recall**	**F1**
$\theta = 0.005$	66.67	24.15	35.46
$\theta = 0.02$	**74.76**	**29.06**	**41.85**
$\theta = 0.05$	71.19	28.81	41.43

Table 3.5: Results (%) with different enlarged scopes for the ontology-guided multi-type resolution.

ChemNER$_{\mathbf{FM}}$	**Precision**	**Recall**	**F1**
Sentence Only	73.64	30.57	43.20
Sentence+Document	**74.04**	29.06	41.73
Sentence+Corpus	71.90	**32.83**	**45.08**
Sentence+Document+Corpus	70.83	32.07	44.15

3.3.6 Parameter Study

Table 3.4 shows the effect of different mininum TF-IDF threshold θ on the performance of ChemNER$_F$. This threshold θ is used to eliminate common word tokens from voting for the candidate entity types during the flexible KB-Matching. We observe that $\theta = 0.02$ gives the best performance of of ChemNER$_F$.

Table 3.5 shows the effect of different enlarged scopes on the performance of ChemNER$_{FM}$. This enlarged scope is used to control the performance of ontology-guided multi-type disambiguation. We observe that when the context types in one sentence still draw ties for multi-type disambiguation, it is more effective to directly go to the corpus-level to look at the popularity scores for each type instead of extending the ontology-guided multi-type disambiguation mechanism to the document level.

3.3.7 Case Study

Table 3.6 shows some example sentences from our test set. We compare the prediction results of ChemNER with two baseline methods: *KB-Matching* and *RoBERTa*. We also show the prediction results of our ablation models, ChemNER$_F$ and ChemNER$_{FM}$, to demonstrate the contribution of each component and how the ChemNER full model achieves the best performance step by step.

KB-Matching can only match entities that exactly appear in the KB dictionaries, which often leads to incomplete or missing annotations. Based on the results of *KB-Matching*, *RoBERTa* learns to give one context-specific label for each entity. For example, in Sentence

Table 3.6: Examples showing how ChemNER improves the fine-grained chemistry NER performance. The ground truth labels and correct model predictions are in blue and the wrong model predictions are in red. The correct labels are in *italics*.

Sentence # 1	... two **aryl chlorides** *ORGANOHALIDES* can be coupled to one another without the isolation of the intermediate **boronic acid** *OXOACIDS* ...
KB-Matching	... two **aryl** AROMATIC COMPOUNDS, SUBSTITUENTS, FUNCTIONAL GROUPS **chlorides** CHLORIDES can be coupled to one another without the isolation of the intermediate **boronic acid** *OXOACIDS* ...
RoBERTa	... two **aryl** FUNCTIONAL GROUPS **chlorides** CHLORIDES can be coupled to one another without the isolation of the intermediate **boronic acid** *OXOACIDS* ...
ChemNER$_F$	... two **aryl chlorides** CHLORIDES, ORGANOHALIDES can be coupled to one another without the isolation of the intermediate **boronic acid** *OXOACIDS* ...
ChemNER$_{FM}$	... two **aryl chlorides** CHLORIDES can be coupled to one another without the isolation of the intermediate **boronic acid** *OXOACIDS* ...
ChemNER	... two **aryl chlorides** *ORGANOHALIDES* can be coupled to one another without the isolation of the intermediate **boronic acid** *OXOACIDS* ...
Sentence # 2	The total synthesis of **narciclasine** *ALKALOIDS* is accomplished by the late-stage, amide-directed **C–H hydroxylation** *ORGANIC REDOX REACTIONS* ...
KB-Matching	The total synthesis of **narciclasine** FREE RADICALS, ALKALOIDS, BIOMOLECULES is accomplished by the late-stage, amide-directed **C–H hydroxylation** *ORGANIC REDOX REACTIONS* ...
RoBERTa	The total synthesis of **narciclasine** BIOMOLECULES is accomplished by the late-stage, amide-directed **C–H hydroxylation** *ORGANIC REDOX REACTIONS* ...
ChemNER$_F$	The total synthesis of **narciclasine** ALKALOIDS, BIOMOLECULES is accomplished by the late-stage, amide-directed **C–H hydroxylation** *ORGANIC REDOX REACTIONS* ...
ChemNER$_{FM}$	The total synthesis of **narciclasine** *ALKALOIDS* is accomplished by the late-stage, amide-directed **C–H hydroxylation** *ORGANIC REDOX REACTIONS* ...
ChemNER	The total synthesis of **narciclasine** *ALKALOIDS* is accomplished by the late-stage, amide-directed **C–H hydroxylation** *ORGANIC REDOX REACTIONS* ...

1, *KB-Matching* failed to recognize "aryl chlorides" as a whole unit, yet it does match "aryl" to three types (i.e., "AROMATIC COMPOUNDS", "SUBSTITUENTS', and "FUNCTIONAL GROUPS"). *RoBERTa* learns the best label (i.e., "FUNCTIONAL GROUPS") for the multi-type entity (i.e., "aryl") based on the context. Although "FUNCTIONAL GROUPS" is indeed the best type for "aryl" if we look at the word individually, *RoBERTa* still achieves imperfect performance due to the incomplete boundaries inherited from *KB-Matching*.

With flexible KB-Matching, ChemNER$_F$ detects the complete boundaries and assigns much more suitable types in most cases. Based on the results of ChemNER$_F$, using ontology-guided multi-type resolution, ChemNER$_{FM}$ determines the context-specific label

that fits the best. For example, in Sentence # 2, ChemNER$_F$ matches "narciclasine" to two types (i.e., "ALKALOIDS" and " BIOMOLECULES"). Here "ALKALOIDS" is a more suitable type and can be detected by ChemNER$_{FM}$ because "ALKALOIDS" and the context type "ORGANIC REDOX REACTIONS" are both under the ontology branch "ORGANIC CHEMISTRY". However, there are also a few cases that the ontology-guided multi-type resolutions are imperfect. For example, in Sentence # 1, ChemNER$_{FM}$ choose the type "CHLORIDES" over "ORGANOHALIDES" for "aryl chlorides" because "CHLORIDES" and the context type "OXOACIDS" are both under the ontology branch "INORGANIC COMPOUNDS", whereas the ground truth label is just the opposite. This issue could further be resolved by the sequence labeling model trained on top of ChemNER$_{FM}$. For example, in Sentence # 1, ChemNER finally chooses "ORGANOHALIDES" over "OXOACIDS" instead probably because the sequence labeling model captures the pattern on the co-occurrence of "ORGANOHALIDES" and "OXOACIDS". Interestingly, from the perspective of chemistry, organohalides and organoboron species (a sector of oxoacids) are the exact two couplers of the Suzuki Coupling reaction.

3.4 RELATED WORK

Distantly-Supervised NER Aiming to reduce expensive manual annotation, distant supervision has been used to generate training labels automatically by utilizing the entity information from existing KBs. The major research efforts lie in dealing with the incomplete annotation problem caused by an incomplete coverage of the KBs [4, 5, 6, 62, 63, 67].

AutoNER [62] proposes a "tie-or-break" tagging scheme to leverage distant supervision from entity dictionaries. Compared with the traditional "BIOES" tagging scheme, the "tie-or-break" tagging scheme introduces an "unknown" type that can be skipped during training to reduce the effect of false negative labeling brought by the incomplete KB-matching. However, AutoPhrase often misses low-frequency phrases for the "unknown" entity generation using a phrase mining method AutoPhrase [68]. Positive and unlabeled learning (PU-learning) is used in distantly-supervised NER to provide an unbiased and consistent estimator of the objective function [67]. However, there are two limitations in using PU-learning for distantly-supervised NER. First, PU-learning uses the prior distribution for each entity type, a parameter that is estimated from an existing human-annotated test set that is not always available for new entity types. Second, the performance of PU-learning is highly sensitive to the class-imbalance rate for each entity type, a parameter that is heuristically determined. It is difficult to apply PU-learning to distantly-supervised NER tasks on new entity types in new domains due to the above two limitations. BOND [63] leverages the power

of pre-trained language models (e.g., BERT and RoBERTa) and a self-training approach to iteratively incorporate more training labels and improve the NER performance. However, they do not work well with fine-grained chemistry entities that have a severe low-precision and low-recall problem with KB-matching. They also largely ignore the noisy annotation problem by simply discarding those multi-labels during the KB-matching process.

Other Related Tasks One similar task to fine-grained NER is entity linking [69, 70, 71, 72] that maps a candidate entity in the text to a concept identifier in the knowledge bases. However, entity linking cannot deal with new entities that do not exist in the background knowledge bases. Another similar task is fine-grained entity typing (FET) [73, 74, 75, 76, 77, 78] that has been extensively studied in the general domain. FET aims at classifying an entity mention into a wide range of entity types by disambiguating the pre-identified entity mentions into a set of candidate entity types. It is formulated as a multi-class, multi-label classification problem and does not assume type exclusiveness. The fine-grained NER task targets both entity boundary detection and entity type recognition and assumes each entity to be tagged with only one type in a given context.

3.5 SUMMARY

In this chapter, we proposed ChemNER, an ontology-guided, distantly-supervised method for fine-grained chemistry NER. It leverages the chemistry type ontology structure to generate distant labels with methods of flexible KB-matching and ontology-guided multi-type disambiguation. We also provide an expert labeled, chemistry NER dataset with 62 fine-grained chemistry types (e.g., chemical compounds and chemical reactions). Experimental results show that ChemNER is highly effective, outperforming substantially the state-of-the-art NER methods on fine-grained chemistry NER. Although achieving great performance, there is still large room for improvement of ChemNER. In the future, we plan to further refine and enrich the type ontology and incorporate more information in the dictionaries (e.g., chemical structures in the KBs) for a better NER performance. We also plan to apply our fine-grained NER method to other scientific domains such as biomedicine and geoscience.

CHAPTER 4: CROSS-MODAL SUPERVISION FOR CHEMICAL REACTANT ENTITY CLASSIFICATION

4.1 INTRODUCTION

Scientific knowledge can be described on various levels of abstractions: from high-level categorical concepts to low-level concrete entities. For example, in Figure 4.1, the Csp^3-Csp^3 Suzuki cross-coupling reaction is defined by chemists as a process involving a pair of high-level reactant groups (i.e., the M-side reactant group "primary alkyl boronate" and the X-side reactant group "primary alkyl halide"). While in the chemistry literature, this chemical reaction can also be described as a process involving two low-level concrete chemical entities (e.g., "1-bromododecane" and "B-n-octyl-9-BBN"). This gap between high-level and low-level abstractions of scientific knowledge is a common phenomenon in various domains such as biology, chemistry, and physics.

In the general domain, recent work has been done on classifying entities in the text into human-given categories without human annotation [79, 80, 81, 82, 83]. However, in the chemistry domain, the task of reactant entity classification requires more effective methods that take two special characteristics of the chemical molecules into consideration. The first characteristic is that each chemical molecule can be represented in two modalities: a chemical name in the text and a molecule structure in the graph. Thus a large amount of high-quality training data for chemical name classification can be automatically created with cross-modal supervision of molecular structure matching. The second characteristic is that there is a knowledge-aware subword correlation between the chemical names to be classified and that of the reactant groups as class labels. Thus the interaction between the subwords (e.g., wordpieces in the pre-trained language models) in the chemical names and reactant groups is the most prominent feature of training a reactant entity classification model.

In this chapter, we propose ReactClass, a highly effective reactant entity classification method without requiring human effort for training data annotation. For example, in Figure 4.1, ReactClass automatically classifies "1-bromododecane" into "primary alkyl halide" and "B-n-octyl-9-BBN" into "primary alkyl boronate", respectively. ReactClass benefits various downstream applications, such as chemistry knowledge base completion [84], chemistry information retrieval [85, 86, 87], and prediction of chemical reactions, products, and properties [88, 89, 90, 91]. Specifically, ReactClass is designed to take the two special characteristics of the chemical molecules into consideration. First, we propose to use cross-modal supervision to automatically create the training data for chemical name classification in the text via molecular structure matching in the graph. Specifically, we first

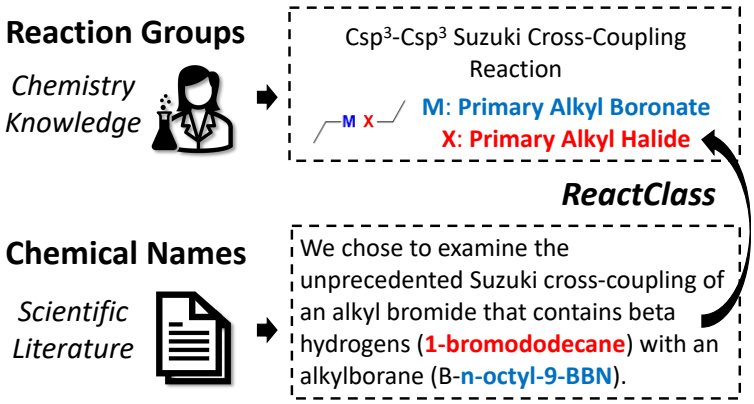


Figure 4.1: ReactClass automatically classifies the concrete chemical entities in the text into the high-level reactant groups defined by the chemical scientists.

convert both the chemical entities and the reactant groups into graph representations and then conduct a subgraph matching. By definition from chemistry knowledge, the training data for each reactant group can be automatically created by finding the chemical names with the graph representations that match the graph representation of the reactant group. Second, we propose to train a classifier based on the subword cross-attention map between each chemical name and its corresponding reaction group. Specifically, we first construct the subword cross-attention map between each chemical name and its corresponding reactant group using their subword embeddings generated from a Transformer-based neural language model. Then we take this 2-D subword cross-attention map as the input feature and encode it with a Convolutional Neural Network (CNN), transforming the text classification task into an image classification task. ReactClass is highly effective, achieving state-of-the-art performance on classifying the chemical names into human-defined reactant groups without requiring human effort for training data annotation.

4.2 THE REACTCLASS FRAMEWORK

The framework of ReactClass consists of two steps: (1) cross-modal supervision of molecule structure matching (Figure 4.2), and (2) subword cross-attention-guided chemical name classification (Figure 4.3). Specifically, ReactClass is designed to take the two special characteristics of the chemical molecules into consideration. First, we propose to use cross-modal supervision to automatically create the training data for chemical name classification in the text via molecular structure matching in the graph. Second, we propose to train a classifier based on the subword cross-attention map between each chemical name and the corresponding reactant group.

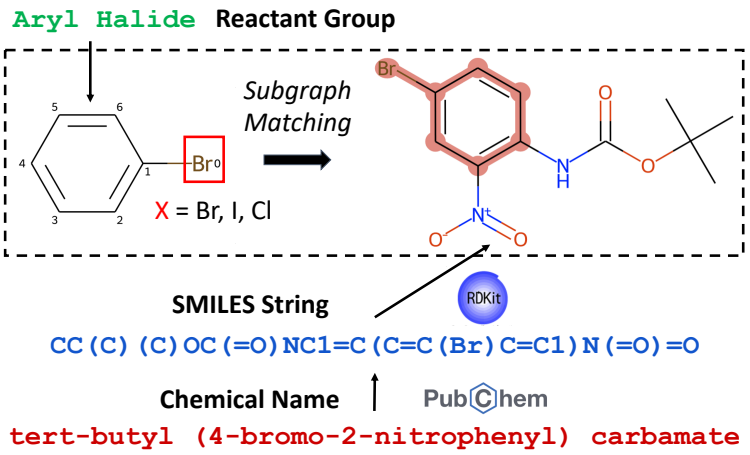


Figure 4.2: Illustration of cross-modal supervision of molecular structure matching.

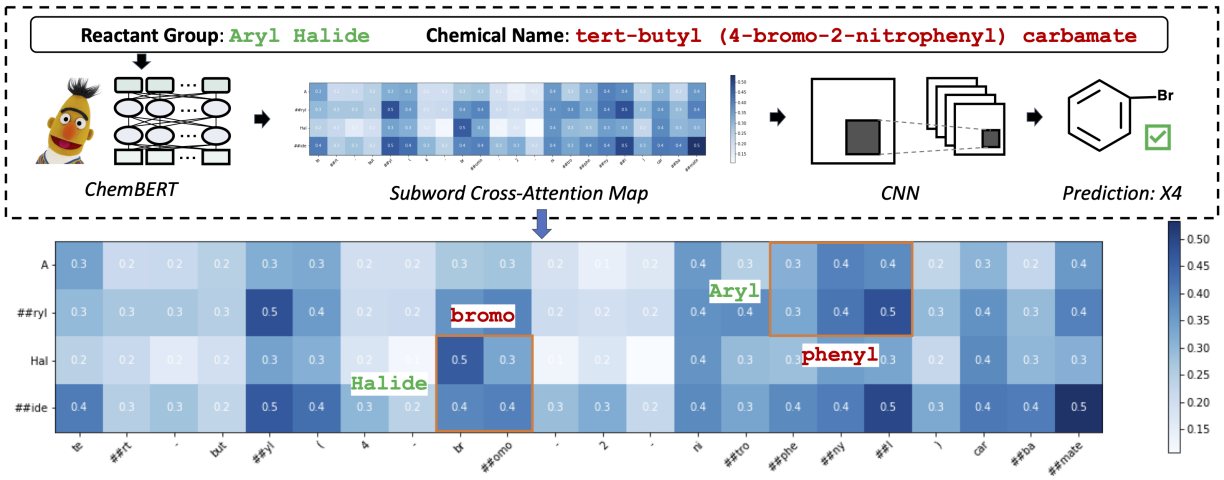


Figure 4.3: Illustration of subword cross-attention-guided name classification.

4.2.1 Cross-Modal Supervision of Molecular Structure Matching

In this section, we describe the detailed method of automatically creating the training data for chemical name classification in the text with cross-modal supervision of molecular structure matching in the graph. Specifically, we first convert both the chemical entities and the reactant groups into graph representations and then conduct a subgraph matching. By definition in chemistry knowledge, the training data for each reactant group can be automatically created by finding the chemical names with graph representations that match the graph representation of the reactant group.

To convert the chemical names into their graph representations, we first collect a large number of candidate chemical names [e.g., "tert-butyl (4-bromo-2-nitrophenyl) carbamate"

in Figure 4.2] from both the chemical reaction knowledge base (Reaxys[13] [92]) and the chemical named entity recognition (ChemNER [7]) results in the chemistry literature. Then we convert the chemical names into their corresponding SMILES strings [e.g., "CC(C)(C)OC (=O)NC1=C(C=C(Br)C=C1)N(=O)=O" in Figure 4.2], a character-based sequence representation of the chemical molecules. This chemical name to SMILES string conversion is automatically done by linking the chemical names to a chemistry knowledge base (PubChem[14] [93]) where we can directly find their corresponding SMILES strings. Finally, the SMILE strings can be converted into molecular structures for the next step of subgraph matching, using an open-source cheminformatics software RDKit[15].

To convert the human-defined reactant groups into their graph representations, we first get the ten reactant groups (e.g., "Aryl Halide" in Figure 4.2) for Suzuki cross-coupling reactions from chemists. The reactant groups cannot be directly converted into molecular structures by knowledge base linking because the reactant groups do not correspond to any specific chemical molecules. However, a reactant group can be converted into a subgraph regular expression. For example, in Figure 4.2, the reactant group "Aryl Halide" can be converted into a subgraph regular expression of a benzene ring with a Br connected to carbon #1, where this Br can be replaced with either I or Cl. By definition in chemistry knowledge, a candidate chemical name belongs to a reactant group if any subgraphs in its molecular structure match the subgraph regular expression of that reactant group. So any chemical names (e.g., "tert-butyl (4-bromo-2-nitrophenyl) carbamate") with a molecular structure that can match the graph representation of "Aryl Halide" can be classified into the "Aryl Halide" reactant group. The subgraph regular expressions of the ten reactant groups are also defined by chemists.

After we get the graph representations of both the chemical names and the reactant groups, we use the RDKit software to conduct the subgraph matching in the molecular structures. Specifically, the RDKit function *HasSubstructMatch()* uses the VF2 subgraph isomorphism algorithm to search for substructures in molecules. After subgraph matching, we have created a large amount of quality training data for the human-defined ten reactant groups. We further create the training data for an additional group "Other" that contains all the chemical names with molecular structures that cannot match the graph representations of any of the ten given reactant groups. This training data creation has perfect accuracy since it strictly follows the chemistry knowledge of how human defines a chemical molecule belonging to a reactant group.

[13] `https://www.reaxys.com/#/search/quick`
[14] `https://pubchem.ncbi.nlm.nih.gov/`
[15] `https://www.rdkit.org/`

However, there are two remaining problems with this automatic training data creation. One problem is that some chemical names can be mapped to multiple reactant groups by the subgraph matching. We ignored all the multi-labeled chemical names in our experiments and put this multi-label disambiguation in our discussions. The other problem is that not all the chemical names in the text can be converted into molecular structures due to two reasons. First, the state-of-the-art chemical linking tools (e.g., PubChem and OPSIN [94]) cannot link all the chemical names in the text to the chemistry knowledge bases perfectly. For example, those chemical names with plural forms or near-miss spellings can be easily missed by the chemical linking tools. Second, and more importantly, not every chemical name in the text has a corresponding molecular structure in theory. For example, the chemical name "2-aryl 5-(4-oxo-3-phenethyl-2-thioxothiazolidin-ylidenemethyl) furan", although looks like a concrete molecule, refers to a group of molecules thus do not have a corresponding molecular structure. We observe that less than 10% of all the chemical names in the text can be directly converted into molecular structures in our experiments. This motivates us to conduct the next step of classification model training for a generalized chemical name classification that can deal with any chemical name that has appeared in the text.

4.2.2 Subword Cross-Attention-Guided Chemical Classification

Based on the training data obtained from the previous step of subgraph matching, we observe a knowledge-aware subword correlation between the chemical names to be classified and the reactant groups as class labels. For example, in the bottom part of Figure 4.3, we see a subword cross-attention map between the chemical name "tert-butyl (4-bromo-2-nitrophenyl) carbamate" and its corresponding reactant group "Aryl Halide". The subword cross-attention map is constructed by first extracting the subword representations from a pre-trained language model in the chemistry domain and then calculating the cosine similarities between the subword representations of the chemical names and the reactant groups. Looking at this subword cross-attention map, we observe that the subword string "phenyl" in the chemical name is highly correlated with the subword string "Aryl" in the reactant group. This is well-aligned with the chemistry knowledge: "Aryl" means any species created by removing a hydrogen atom from an aromatic hydrocarbon and "phenyl" is a specific aryl radical, which is created by removing a hydrogen atom from a benzene ring. Similarly, we can observe that the subword string "bromo" in the chemical name is highly correlated with the subword string "halide" in the reactant group. This observation indicates that rich chemistry domain knowledge is captured by the subword cross-attention map that is created based on the pre-trained language model in the chemistry domain.

This observation of knowledge-aware subword correlation also motivates us to train a classifier based on the subword cross-attention maps between the chemical names and the reaction groups. The general idea of our proposed method for subword cross-attention-guided chemical name classification is shown in the top part of Figure 4.3. First, we take ChemBERT [95], a Transformer-based language model pre-trained on massive chemistry literature, as our base model. We obtain the subword representations from the ChemBERT model and then construct the subword cross-attention map by calculating the cosine similarity between the subword representations of the chemical name and the reactant group. Then we take this 2-D map of subword cross-attention as input and encode it with a CNN, converting the text classification task into an image classification task. We demonstrate the effectiveness of our proposed method by comparing it with baseline methods that directly use the output states of ChemBERT plus a linear layer for prediction in the experiments.

We formally describe our method of subword cross-attention-guided chemical name classification as follows. We first describe how we construct the cross-attention map. Taken each chemical name $e_i = \langle w_1, w_2, ..., w_{|e_i|} \rangle$ and reactant group $g_j = \langle w_1, w_2, ..., w_{|g_j|} \rangle$ as a sequence of subword tokens w_k, we first extract the representation for each subword token from last hidden states of the fine-tuned ChemBERT model. Then we calculate the cosine similarities between the representations of subword tokens in the chemical names and the reactant groups. Specifically, we obtain the cross-attention matrix as follows.

$$q = \mathbf{x}_{\text{group}}\mathbf{W}_q, k = \mathbf{x}\mathbf{W}_k, v = \mathbf{x}\mathbf{W}_v \tag{4.1}$$

$$\mathbf{A} = \text{softmax}(qk^T/\sqrt{C/h}) \tag{4.2}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{C\times(C/h)}, C, h$ are the embedding dimension and the number of heads, $\mathbf{x}$ is representation of the input chemical name, and $\mathbf{x}_{\text{group}}$ is the representation of the input reactant group. The attention matrix $\mathbf{A}$ will be used as our input feature for the classification model training.

Taken the attention matrix $\mathbf{A}$ constructed above, we then describe our design of the classification model. We consider each cross-attention map as a single-channel image and encode it with a three-layer CNN, transforming the text classification task into an image classification task for the final prediction. For each chemical name, we first create positive training data by constructing cross-attention matrix between the chemical name and its corresponding reaction group from subgraph matching. We then create negative training data by constructing cross-attention matrix between the chemical name and other group names. Our learning tasks is a binary classification task. For the learning objectives, we adopt binary-class cross-entropy loss for simplicity with our created training data. Thus,

Table 4.1: Dataset Statistics

Dataset	Suzuki Coupling
# Training Samples	30,488
# Validation Samples	3,855
# Testing Samples	3,858
# Groups	11

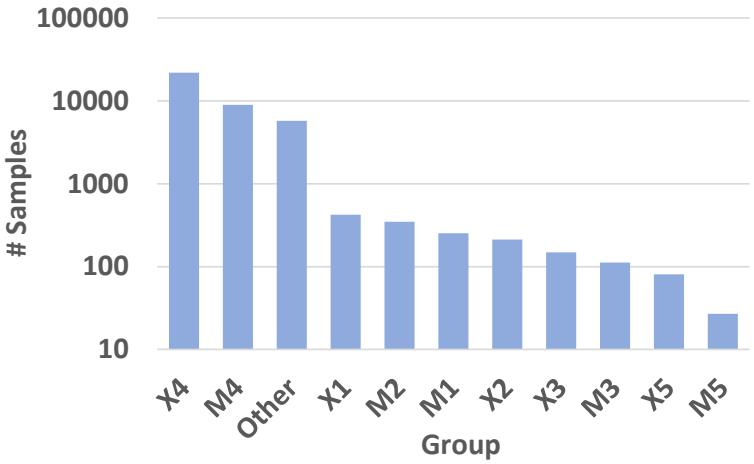


Figure 4.4: Class Distribution

the training loss function for the classification model can be formulated as

$$\mathcal{L} = -\frac{1}{|\mathbb{D}|} \sum_{x_i, y_i \in \mathbb{D}} y_i \log(f(x_i)) + (1 - y_i) \log(1 - f(x_i)) \tag{4.3}$$

where $\mathbb{D} = \{(x_i, y_i)\}$ is the training dataset, x_i is the attention matrix, and $y_i \in \{+1, -1\}$.

During inference, we compute the scores of the attention matrices between each chemical name and all the ten reactant groups to find the reactant group with the highest probability. Since different molecule names are tokenized into subword sequences with different lengths, it results in different dimensions for each cross-attention map. Some extremely long chemical names even contain a lot of repeated information. So we consider padding and truncating the matrices to $K \times K$ dimensions, where K is a hyperparameter during training. The overall efficiency now mainly depends on K, the size of the attention map. We quantitatively compared the performance with different K in our experiments. Another possible hyperparameter is the number of transformer layers to use for subword representation extraction. Existing studies have observed considerable redundancy in the outputs of different Transformer layers, including attention distributions [96, 97]. We also compare the attention matrix constructed from different layers of ChemBERT in our experiments.

4.3 EXPERIMENTS

4.3.1 Dataset

We create a dataset for our task of chemical name classification. We first get ten reactant groups (i.e., M1, M2, M3, M4, M5, X1, X2, X3, X4, and X5 in Figure 4.4) from chemistry experts to serve as our class labels. Each reactant group has a corresponding reactant group name (e.g., "M1" is "Primary Boronate") that can be used in our experiments. Then we collect the chemical names to be classified from both the reaction database (Reaxys [92]) and the named entity recognition (ChemNER [7]) results in chemistry literature. Last, following the training data creation process described in Section 4.2.1, we automatically create around 38K training data for the ten reactant groups plus an "Other" class with cross-modal supervision of molecular structure matching. We split the 38K training data into training/validation/test sets with a ratio of 8:1:1. The dataset details can be found in Table 4.1.

In Figure 4.4, we observe that the class distribution is imbalanced in our automatically created training data. Some reactant groups (e.g., X4 and M4) have more than 10K training samples, while some reactant groups (e.g., X5 and M5) have less than 100 training samples. This class imbalance issue makes it hard for direct training to learn enough characteristics for the reactant groups with few training samples. We leverage an oversampling strategy, weighted bootstrapping, to ensure that the models receive about the same number of data in each class during training. This oversampling strategy is proved to be highly effective in dealing with the class imbalance issue in our experiments.

4.3.2 Baselines

We compare the performance of ReactClass with several baseline methods.

- **BERT/BioBERT/ChemBERT + Softmax**: This is a simple baseline method that directly uses the output states of a pre-trained language model plus a linear layer for prediction. We explored various pre-trained language models in different domains (e.g., BERT [61] in the general domain, BioBERT [20] in the biomedical domain, and ChemBERT [95] in the chemistry domain).
- **ChemBERT + Triplet Loss**: To tackle the class imbalance problem in our training data, we tried the triplet loss that is less sensitive to the imbalanced training data compared to the softmax loss.

- **ChemBERT + Softmax (Oversampling)**: To further tackle the class imbalance problem in our training data, we leverage the weighted bootstrapping strategy to ensure that the models receive about the same number of data in each class during training.
- **Subword + CNN + Softmax**: This is our proposed method that takes the subword cross-attention map between each chemical name and the corresponding reaction group as the input feature and then encodes it with a 3-layer CNN for the final prediction.
- **Subword + CNN + Softmax (Oversampling)**: This is our final model that has the same architecture as Subword + CNN + Softmax, only adding the weighted bootstrapping strategy to further tackle the class imbalance problem in our training data.

We use the micro-F1 and macro-F1 scores[16] as the evaluation metrics for our performance comparison.

4.3.3 Overall Performance

Table 4.2 shows the main results on the test set of our chemical name classification dataset. Comparing different pre-trained language models (BERT/BioBERT/ChemBERT + Softmax), the domain-specific pre-trained language model achieves better performance than that is trained in the general domain. Comparing ChemBERT + Softmax, ChemBERT + Triplet Loss, and ChemBERT + Softmax (Oversampling), both the triplet loss and the oversampling strategy are effective in dealing with the class imbalance problem in our automatically created training data. The oversampling strategy is the most effective one that brings the most performance improvement. Our final model (ReactClass + Oversampling) achieves 98.56% micro-F1 and 90.76% macro-F1 scores with significant performance improvements compared with all the baseline methods. It demonstrates the effectiveness of our proposed method that takes the subword cross-attention maps between the chemical names and the reaction groups as the input feature for classification.

4.3.4 Parameter Study

We perform several experiments on the hyperparameters in our framework to study the efficacy of our classification model. One important hyperparameter is the dimension of our cross-attention matrix K during training and inference. Due to a large amount of computation in calculating cross-attention maps, large dimensions result in much longer computation

[16] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html`

Table 4.2: Main Results (F1 Scores in %)

Method	Micro F1	Macro F1
BERT + Softmax	97.72	82.83
BioBERT + Softmax	97.85	85.00
ChemBERT + Softmax	97.95	84.20
ChemBERT + Triplet Loss	98.16	88.25
ChemBERT + Softmax (Oversampling)	98.16	89.46
Subword + CNN + Softmax (**ReactClass**)	98.28	83.44
Subword + CNN + Softmax (**ReactClass** + Oversampling)	**98.56**	**90.76**

time while small dimensions may cause information loss. We conducted experiments on a dimension size K from 10 to 50. In Figure 4.5, we observe that a larger dimension K will lead to better performance. We use $K = 50$ in all our experiments.

Another important hyperparameter is the number of transformer layers to use for subword representation extraction. In our experiments, we observe that different transformer layers produce redundant information in the constructed subword cross-attention maps. For this reason, as the default setting of ReactClass, we only preserve the attention map from the last layer in ChemBERT. As ChemBERT has 15 layers in total, this saves more than 90% of computational resources.

4.3.5 Discussions

There are still challenges for completely resolving this chemical name classification problem. For example, one chemical name can be matched to multiple reactant groups via subgraph matching. We currently ignored all the multi-labeled entities during our model training, but they can be further disambiguated based on the chemical reactions they are involved in the original text. Also, there is positional information in the molecule structures that are not captured by our current subword cross-attention maps. We discuss each of the two challenges in detail below.

Figure 4.6 illustrates the challenge of multi-label disambiguation based on the chemical reactions. The chemical name "3-diethylboranylpyridine" has a molecular structure that can match the subgraph representations of both the reactant group "Primary Alkyl Boronate" (green circle) and the reactant group "Aryl Boronate" (blue circle). To determine the correct reactant group for "3-diethylboranylpyridine", we need to go back to the original paper and

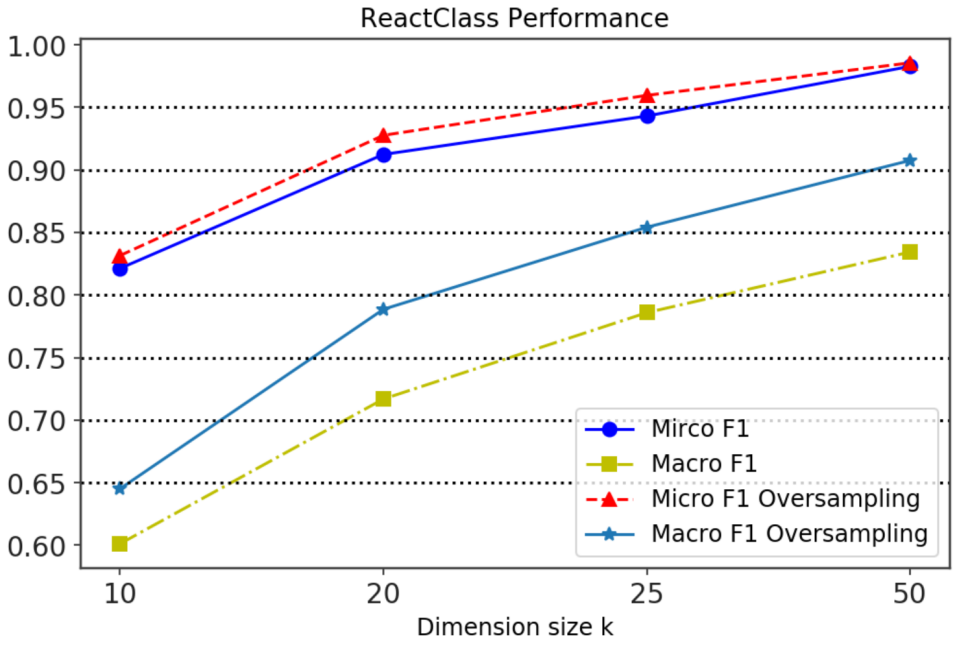


Figure 4.5: Parameter Studies on Dimension K

find the concrete chemical reaction where "3-diethylboranylpyridine" has participated in. From the original paper, we can see that "3-diethylboranylpyridine" mainly function as an "Aryl Boronate" (blue circle) in the concrete chemical reactions (shown in the bottom part of Figure 4.6). So the correct reactant group for "3-diethylboranylpyridine" should be "Aryl Boronate". We currently ignored all the multi-labeled entities in our experiments. However, this reaction-based multi-label disambiguation can be added to further improve the performance of ReactClass.

Figure 4.7 illustrates the challenge of distinguishing the primary and secondary carbons with positional information in the chemical names. For example, based on subgraph matching, we know that "1-dodecylbromide" should belong to the reactant group "Primary Alkyl Bromide" rather than the reactant group "Secondary Alkyl Bromide". However, when we look at the surface names of "1-dodecylbromide" and "Primary Alkyl Bromide", we cannot observe a clear subword correlation between "primary" and "1-dodecylbromide". This indicates that using the subword cross-attention map as the input feature may lose certain information that is originally contained in the molecular structures. The reason we know from the surface names that "1-dodecylbromide" contains a "primary" rather than a "secondary" carbon is that the subwords "1-" is in front of the subword "dodecyl". This positional information needs to be captured to further improve the performance of ReactClass.

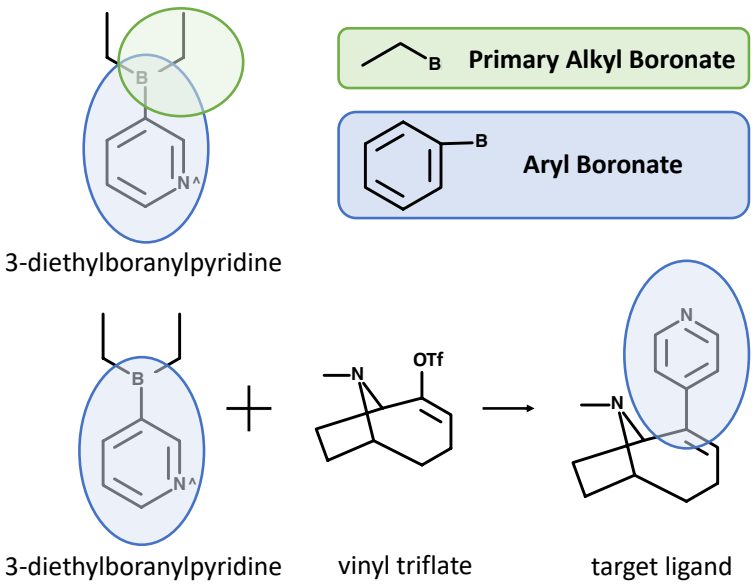


Figure 4.6: Case study of the challenge of multi-label disambiguation based on the chemical reactions.

4.4 RELATED WORK

Chemical Molecule Representation Chemical molecules can be represented in two modalities: a chemical name in the text and a molecular structure in the graph. For the chemical names in the text, representation learning based on chemical IUPAC strings has shown noticeable progress in the recent studies [95]. ChemBERT [95] is a transformer-based language model pre-trained on vast amounts of unlabeled chemistry literature, which is effective for two chemistry literature understanding tasks: chemical product extraction and reaction role labeling. For the molecular structures in the graph, representation learning based on molecular structures has long been studied. Traditional fingerprinting methods [98, 99] have long been used for molecule structural representations without learning from the data. Inspired by recent advances in word embedding methods, new approaches have been developed for molecule representation learning [14, 91, 98, 100, 101]. For example, Mol2Vec [100] treats each molecule as a sentence and its substructures as the words in the sentence and then applies Word2Vec [45] to generate the molecule representation. Additionally, MolBERT [101] and ChemBERTa [14] use SMILES strings (a character-based sequence representation of chemical molecules) as inputs and then apply BERT [61] to generate a transformer-based molecule representation. In ReactClass, we use ChemBERT [95] as our base model to obtain the subword representations for chemical name classification.

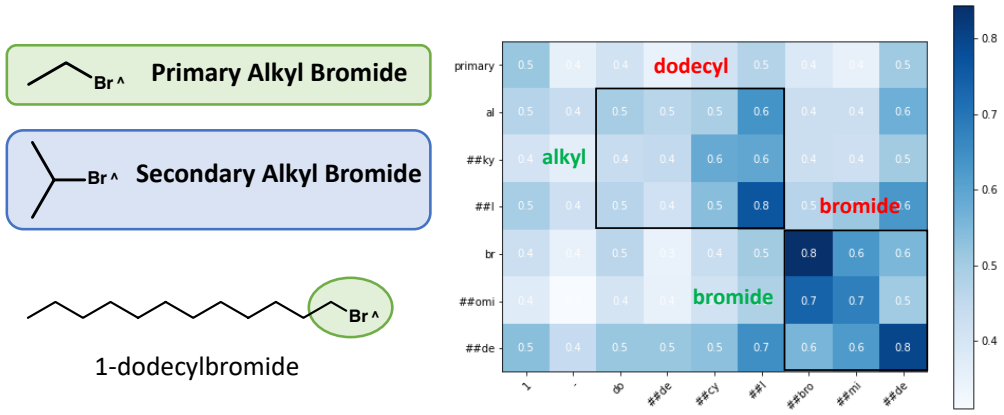


Figure 4.7: Case study of the challenge of distinguishing the primary and secondary carbons with positional information.

Cross-Modal Supervised Learning. There has been a growing interest in using cross-modal supervision for various applications, such as speaker detection and identification [102, 103, 104], action recognition [105, 106], lip-reading [104, 107], touch representation learning [108], radar object detection [109], and multi-modal event extraction [110]. The existing work has mostly focused on speech recognition and computer vision. Some work has been done for cross-modal retrieval between text and graphs [86, 87]. Zhou et al. [86] replace chemical entities in the text with a unique canonical key in a database to perform query expansion by including additional molecules with similar structures in the database. In contrast, Edwards et al. [87] perform a direct semantic cross-modal retrieval by constructing a paired dataset of molecules and their corresponding text descriptions and then learning an aligned common semantic embedding space for retrieval. However, little work has been done for cross-modal supervision in the text or graphs. Due to the special characteristics of chemical molecules (i.e., a chemical molecule can be represented both as a chemical name in the text and a molecular structure in the graph), we propose to automatically create the training data for chemical name classification in the text with cross-modal supervision of molecular structure matching in the graph.

Attention Map Representation Recent studies have focused on using the attention maps from pre-trained transformer-based language models [13, 61] as the input features to capture inter-relation information of tokens in the sentences [97, 111, 112]. Linzen et al. [111] showed that a sufficient amount of linguistic knowledge, such as noun determiners and objects of verbs and prepositions, are captured by the attention maps of BERT [61]. Moreover, using only attention maps as the input features, a model can be trained to perform high-quality dependency parsing [111] and constituency tree construction [112]. UCPhrase

[97] utilizes only the attention maps as the input feature to identify quality phrases in the text. Compared with directly using the output states of RoBERTa [13] as the input feature, Gu et al. [97] showed that the model using the attention maps is less likely to overfit and has a more robust generalization. In this study, we observed a strong subword correlation between the chemical names to be classified and the reactant group names as labels. This subword correlation is highly indicative of chemical name classification. So we propose to train a classifier that takes the subword cross-attention map between each chemical name and the corresponding reaction group as the input feature.

4.5 SUMMARY

In this chapter, we proposed a highly effective method, ReactClass, for reactant entity classification without requiring human effort for training data annotation. ReactClass is designed to take two special characteristics of the chemical molecules into consideration. First, we propose to automatically create the training data for chemical name classification in the text with cross-modal supervision of molecular structure matching in the graph. Second, we propose to train a classifier that based on the subword cross-attention map between each chemical name and the corresponding reaction group. Our method achieves state-of-the-art performance in classifying the chemical names into ten Suzuki cross-coupling reactant groups. Future improvements include multi-label disambiguation based on concrete chemical reactions and adding positional information to better reveal the hidden structural information in the chemical surface names.

CHAPTER 5: SCIENTIFIC TEXTUAL EVIDENCE DISCOVERY

5.1 INTRODUCTION

Search engines on scientific literature have been widely used by life scientists for discoveries based on prior knowledge. Each day, millions of users query PubMed[17] and PubMed Central[18] (PMC) for their information needs in biomedicine [19]. However, traditional search engines for life sciences (e.g., PubMed) are designed for document retrieval and do not allow direct retrieval of specific statements [113, 114, 115]. With the results from those search engines, scientists still need to read a large number of retrieved documents to find specific statements as textual evidence to validate the input query. This textual evidence is key to tasks such as developing new hypotheses, designing informative experiments, or comparing and validating new findings against previous knowledge.

While the last several years have witnessed substantial growth in interests and efforts in evidence mining [19, 116, 117, 118, 119, 120, 121], little work has been done for evidence mining system development in the scientific literature. A significant difference between evidence in the scientific literature and evidence in other corpora (e.g., the online debate corpus) is that scientific evidence usually does not have a strong sentiment (i.e., positive, negative or neutral) in the opinion it holds. Most scientific evidence sentences are objective statements reflecting how strongly they support a query statement. Therefore, if scientists are interested in finding textual evidence for "*melanoma is treated with nivolumab*", they may expect a ranked list of statements with the top ones like "bicytopenia in primary lung melanoma treated with nivolumab" as the textual evidence that supports the input query.

In this chapter, we propose EvidenceMiner, a web-based system for textual evidence discovery for life sciences (Figure 5.1). Given a query as a natural language statement, EvidenceMiner automatically retrieves sentence-level textual evidence from a background corpora of biomedical literature. EvidenceMiner is constructed in a completely automated way without any human effort for training data annotation. It is supported by novel data-driven methods for distantly supervised named entity recognition and open information extraction. EvidenceMiner relies on external knowledge bases to provide distant supervision for named entity recognition (NER) [2, 4, 62]. Based on the entity annotation results, it automatically extracts informative meta-patterns (textual patterns containing entity types, e.g., CHEMICAL inhibit DISEASE) from sentences in the background corpora

[17] `https://www.ncbi.nlm.nih.gov/pubmed/`
[18] `https://www.ncbi.nlm.nih.gov/pmc/`

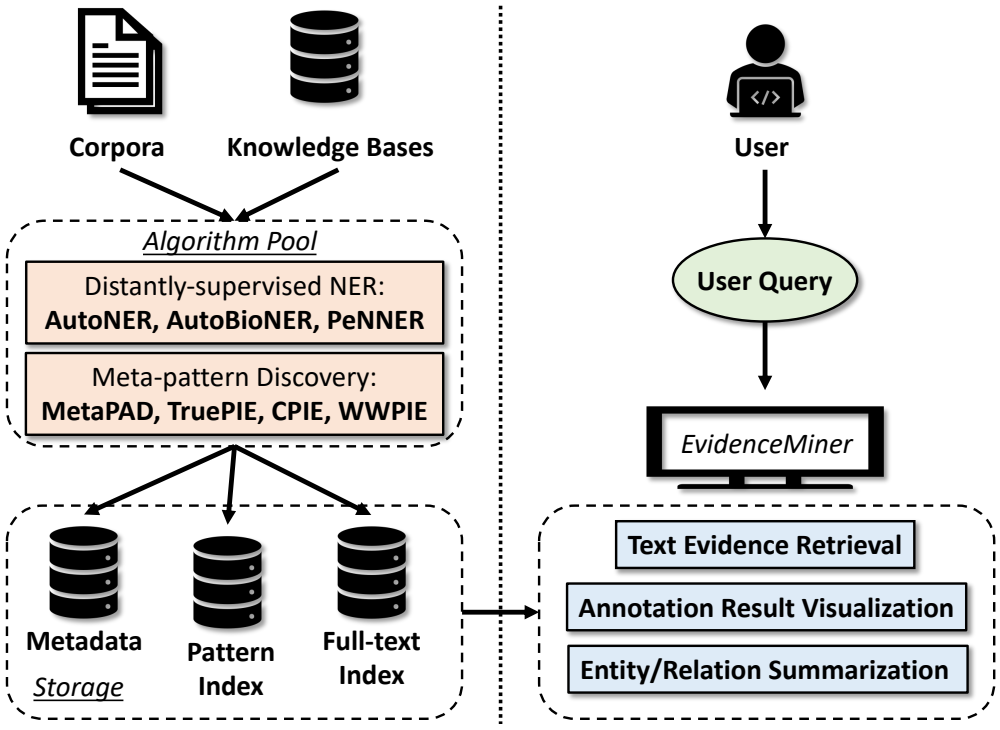


Figure 5.1: System architecture of EvidenceMiner.

[39, 41, 42, 122]. Sentences with meta-patterns that better match the query statement is more likely to be textual evidence. The entities and patterns are pre-computed and indexed offline to support fast online evidence retrieval. The annotation results are also highlighted in the original document for better visualization. EvidenceMiner also includes analytic functionalities such as the most frequent entity and relation summarization.

5.2 THE EVIDENCEMINER FRAMEWORK

EvidenceMiner consists of two major components: an open information extraction pipeline and a textual evidence retrieval and analysis pipeline. The open information extraction pipeline includes two functional modules: (1) distantly supervised NER, and (2) meta-pattern-based open information extraction; whereas the textual evidence retrieval and analysis pipeline includes three functional modules: (1) textual evidence search, (2) annotation result visualization in the original document, and (3) the most frequent entity and relation summarization. Figure 5.1 shows the system architecture of EvidenceMiner. The functional modules are introduced in the following sections.

Table 5.1: Basic statistics of background corpora. It includes PubMed abstracts and PMC full-text papers related to cancers and heart diseases published in 2019.

Background Corpora	Cancers	Heart Diseases
# of PubMed abstracts	48,201	11,766
# of PMC full-text papers	7,130	1,151
# of Sentences in total	1,466,091	246,106
# of Entity instances	3,315,092	400,327
# of Relation instances	29,160	9,576

5.2.1 Open Information Extraction

The open information extraction pipeline extracts entities with distant supervision from knowledge bases and relations with automatic meta-pattern discovery methods. In particular, to extract high-quality entities and relations, we design noise-robust neural models for distantly supervised named entity recognition [4, 62] and wide-window meta-pattern discovery methods to deal with the long and complex sentences in biomedical literature [42, 122].

Corpus Collection To obtain the background corpora for EvidenceMiner, we collect the titles and abstracts of 26M papers from the entire PubMed[19] dump, and the full-text contents of 2.2M papers from PubMed Central[20] (PMC). For the demonstration purpose, we select a subset of documents published in 2019 that are specifically related to two important diseases (cancers and heart diseases) to form the background corpora. The subset of documents are selected by concept matching on MeSH[21], a biomedical concept ontology with the concepts related to cancers (Neoplasms) and heart diseases (Cardiovascular Diseases). Table 6.1 summarizes the statistics of the background corpora.

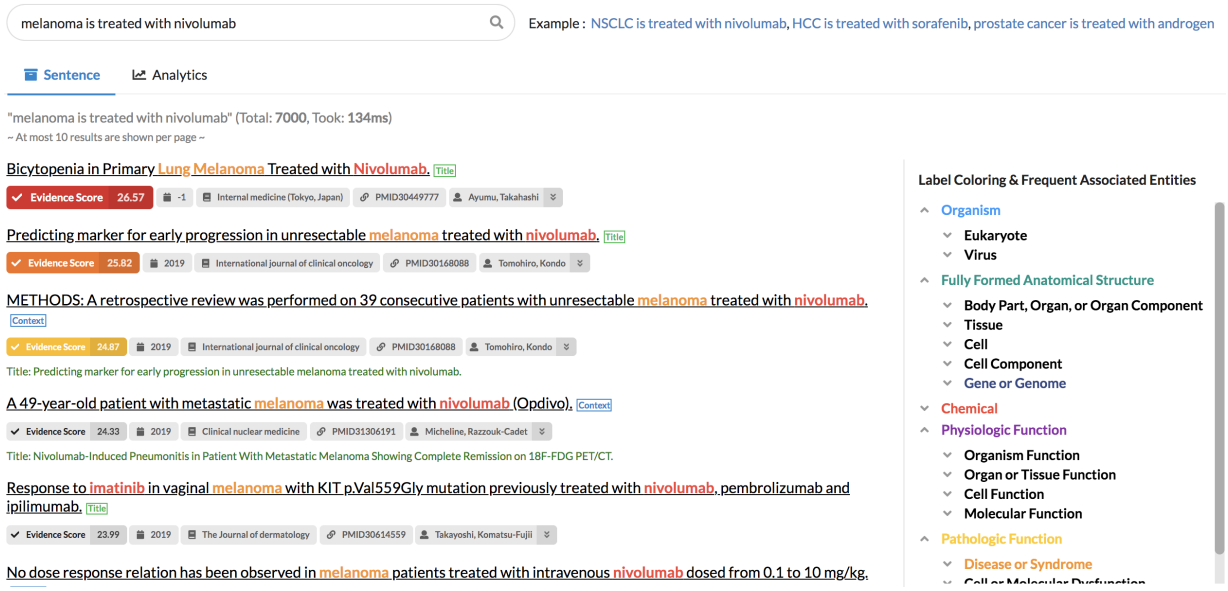
Distantly Supervised Named Entity Recognition Taking the corpus as input, EvidenceMiner relies on UMLS[22], a comprehensive biomedical knowledge base to provide distant supervision for named entity recognition. We select 5 major biomedical entity types (Organism, Fully Formed Anatomical Structure, Chemical, Physiologic Function, and Pathologic Function) including 17 fine-grained entity types (Archaeon, Bacterium, Eukaryote, Virus, Body Part/Organ/Organ Component, Tissue, Cell, Cell Component, Gene or Genome, Chemical, Organism Function, Organ or Tissue Function, Cell Function, Molecular Function, Disease or Syndrome, Cell or Molecular Dysfunction, Experimental Model of

[19] `https://pubmed.gov/pubmed`
[20] `https://pubmed.gov/pmc`
[21] `https://www.nlm.nih.gov/mesh/`
[22] `https://www.nlm.nih.gov/research/umls/index.html`

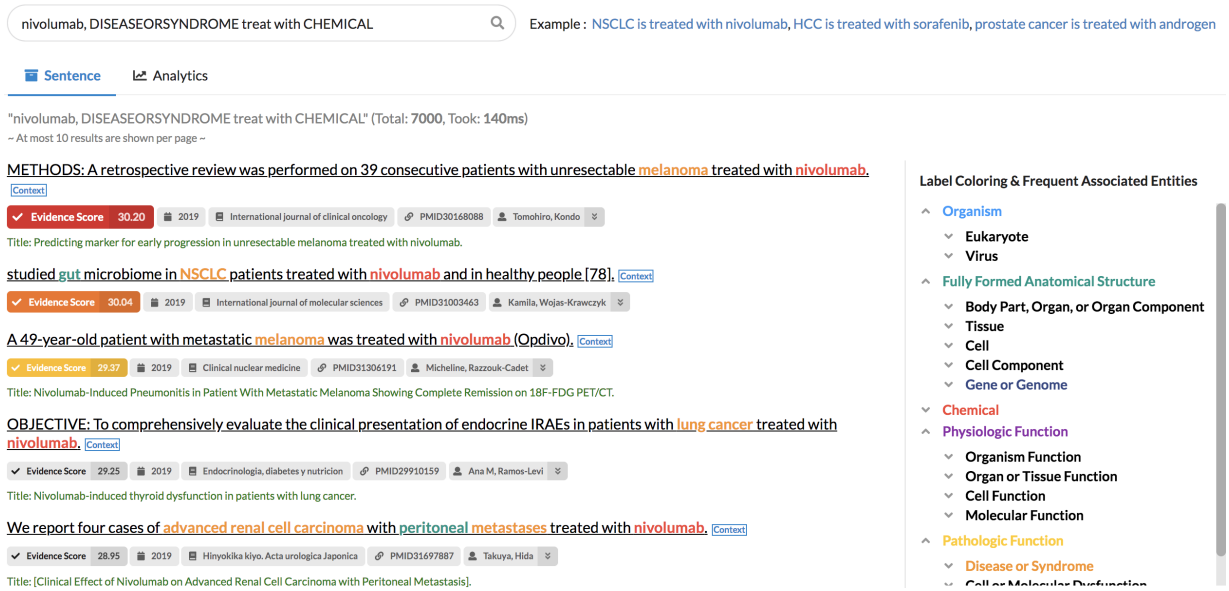
Disease, and Pathological Function) from UMLS as the entity types to be annotated. To tackle the problem of limited coverage of the input dictionary, we first apply a data-driven phrase mining algorithm, AutoPhrase [68], to extract high-quality phrases as additional entity candidates. Then we automatically expand the dictionary with a novel dictionary expansion method [4]. The expanded dictionary is used to label the input corpora with the 17 fine-grained entity types to train a neural model. We apply AutoNER [62], a state-of-the-art distantly supervised NER method that effectively deals with noisy distant supervision. Comparing with PubTator [8], a state-of-the-art BioNER system trained with extensive human annotation on 5 biomedical entity types, EvidenceMiner can automatically annotate 17 fine-grained entity types with high quality without any human effort for training data annotation.

Meta-Pattern-Based Open Relation Extraction Based on the entity annotation results above, meta-patterns can be automatically discovered from the corpora to support textual evidence discovery. Meta-patterns are defined as sub-sequences in an entity-type-replaced corpus with at least one entity type token in it. For example, "PPAR gamma agonist" and "caspase 1 agonist" are two word-sequences in the raw corpus. If we replace all the entities (i.e., "PPAR gamma" and "caspase 1") with their corresponding entity types (i.e., $GENE) in the raw corpus, "PPAR gamma agonist" and "caspase 1 agonist" are represented as one meta-pattern "$GENE agonist" in the entity-type-replaced corpus. Meta-patterns containing at least two entity types (e.g., "$CHEMICAL induce $DISEASE") are relational meta-patterns. Quality relational meta-patterns can serve as informative textual patterns that guide textual evidence discovery. We apply two state-of-the-art meta-pattern discovery methods, CPIE [42] and WW-PIE [122], to extract high-quality meta-patterns from the NER-tagged corpora. Both methods are specifically designed to better deal with the long and complex sentence structures in the biomedical literature. In EvidenceMiner, we combine the meta-pattern extraction results from CPIE and WW-PIE as our informative meta-patterns to guide textual evidence retrieval. We use Elasticsearch[23] to create the index for each sentence for fast online retrieval. In addition to indexing the keywords, we index each sentence with the meta-patterns it matches and the corresponding entities extracted by the meta-patterns in the sentence.

[23] `https://www.elastic.co/`



(a) Query: *melanoma is treated with nivolumab*



(b) Query: *(nivolumab, DISEASEORSYNDROME treat with CHEMICAL)*

Figure 5.2: The search interface with the textual evidence retrieved. The evidence score indicates the confidence of each retrieved sentence being a supporting evidence of the input query.

5.2.2 Textual Evidence Retrieval and Analysis

The textual evidence retrieval and analysis pipeline retrieves textual evidence given a user-input query statement and the indexed corpora. The retrieved evidence sentence can be easily located in the original text. The entity and relation annotation results are also

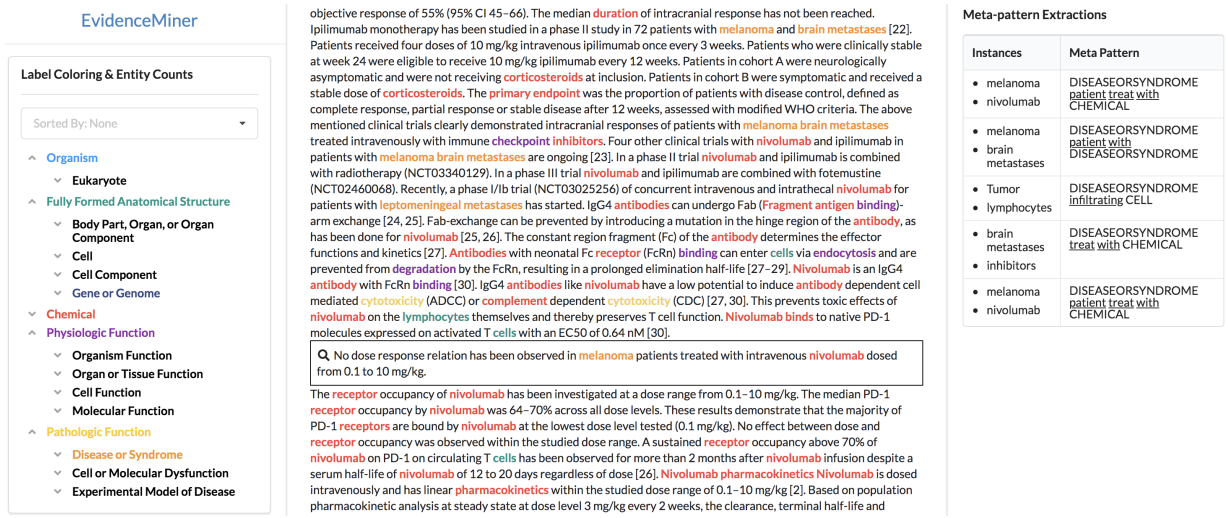


Figure 5.3: The annotation interface with all the entity and relation annotation results.

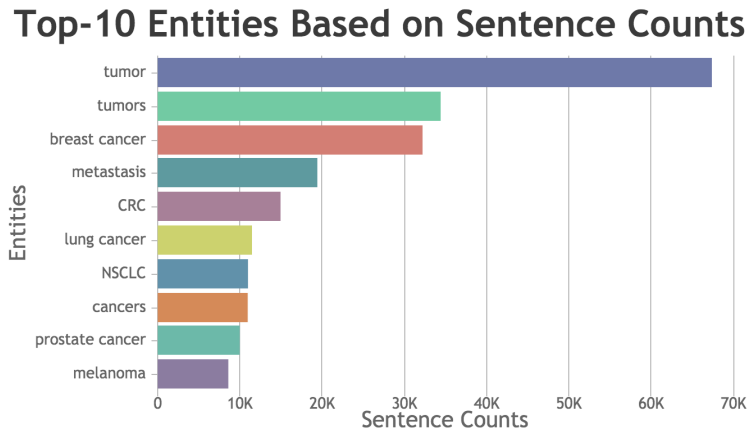
highlighted in the text for better visualization. EvidenceMiner also includes analytic functionalities such as finding the most frequent entities and relations as summarization.

Textual Evidence Sentence Retrieval Given a user-input query statement and the indexed corpora, EvidenceMiner retrieves and ranks the candidate sentences with a combined approach of keyword weighting and meta-pattern weighting. Taking the input query and the background corpus indexed with the extracted entities and relationships, we first retrieve all the candidate evidence sentences that cover the words or entities in the input query. Then we rank the candidate evidence sentences by a confidence score of it being textual evidence for the input query. The confidence score is designed to reflect how well the candidate sentence covers the key entities and expresses the relation between the key entities in the input query. The confidence score is a weighted combination of three scores: a word score, an entity score, and a pattern score. The three scores are designed to satisfy the following three criteria:

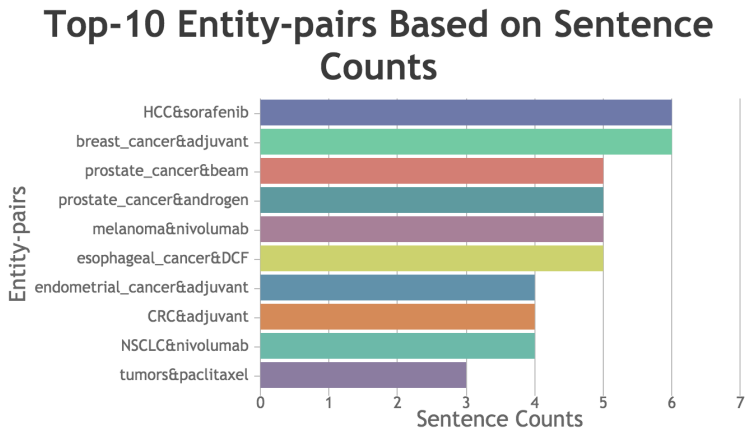
1. Candidate evidence sentences covering the query entities should be ranked higher than those covering only the synonyms to reflect the coverage of the words in the query.

2. Candidate evidence sentences covering the query entities should be ranked higher than those covering only the query words to reflect the coverage of the entities in the query.

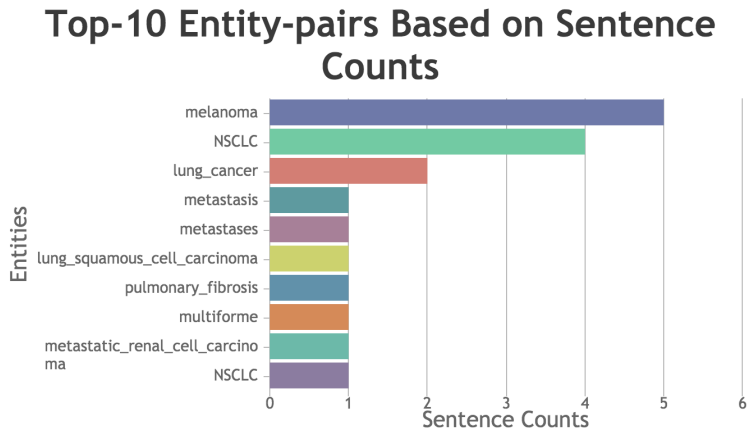
3. Candidate evidence sentences covering more query-matched meta-patterns should be ranked higher to reflect the expression of the relation between the entities in the query.



(a) Query: entity_type=DISEASEORSYNDROME



(b) Query: pattern=DISEASEORSYNDROME treat with CHEMICAL



(c) Query: entity=nivolumab&pattern=DISEASEORSYNDROME treat with CHEMICA

Figure 5.4: The analytic interface with the entity and relation summarization results.

- **Word Score**: We define the word score to satisfy the first criteria: candidate evidence sentences covering the exact query entities will be ranked higher than those covering only the entity synonyms. We use the BM25 [123] score as the word score to measure the relatedness between the query and the candidate evidence sentence. BM25 is a commonly used ranking score for information retrieval. Given a query $q_e = \langle h, r, t \rangle$, where $h \in \mathcal{E}$ is the head entity, $r = \langle w_1, w_2, ...w_{|r|} \rangle$ is the relation, and $t \in \mathcal{E}$ is the tail entity, the BM25 score of a candidate evidence sentence $s \in \mathcal{D}$ is

$$S_w(q_e, s) = \sum_{i=1}^{n} IDF(w_i) \cdot \frac{f(w_i, s) \cdot (k+1)}{f(w_i, s) + k \cdot (1 - b + b \cdot \frac{|s|}{avgsl})}, \tag{5.1}$$

  where $f(w_i, s)$ is the term frequency of w_i in the sentence s, $|s|$ is the length of the sentence s, $avgsl$ is the average length of all the sentences and k and b are two free parameters chosen by the user. $IDF(w_i)$ is the inverse document frequency of w_i,

$$IDF(w_i) = log \frac{N - n(w_i) + 0.5}{n(w_i) + 0.5}, \tag{5.2}$$

  where N is the number of sentences and $n(w_i)$ is the number of sentences containing w_i.

- **Entity Score**: We define the entity score to satisfy the second criteria: candidate evidence sentences covering the query entities will be ranked higher than those covering only the query words. Similarly, we use the BM25 score as the entity score to measure the relatedness of the query and the candidate evidence sentence. Given the query q_e containing the entities $\langle e_1, e_2, ..., e_m \rangle$, the BM25 score of a candidate evidence sentence $s \in \mathcal{D}$ is

$$S_e(q_e, s) = \sum_{i=1}^{m} IDF(e_i) \cdot \frac{f(e_i, s) \cdot (k+1)}{f(e_i, s) + k \cdot (1 - b + b \cdot \frac{|s|}{avgsl})}, \tag{5.3}$$

  where $f(e_i, s)$ is the term frequency of e_i in the sentence s, $IDF(e_i)$ is the inverse document frequency of e_i,

$$IDF(e_i) = log \frac{N - n(e_i) + 0.5}{n(e_i) + 0.5}, \tag{5.4}$$

  where $n(e_i)$ is the number of sentences containing e_i.

- **Pattern Score**: We define the pattern score to satisfy the third criteria: candidate evidence sentences covering more query-matched meta-patterns will be ranked higher. We measure how many times the query meta-pattern can be matched on each candidate evidence sentence. For example, given an input query (e.g., *(resveratrol, inhibit, pancreatic*

cancer)), we first try to convert it into a query meta-pattern (i.e., "CHEMICAL inhibit DISEASE"). Then we measure how many times the query meta-pattern can be matched for each candidate evidence sentence on the query entities (i.e., "resveratrol" and "pancreatic cancer"). Given the query q_e containing the entities $\langle e_1, e_2, ..., e_m \rangle$, the pattern score of a candidate evidence sentence $s \in \mathcal{D}$ is

$$S_p(q_e, s) = \sum_{i=1}^{k} Match(MP_i(q_e), s), \tag{5.5}$$

where $MP_i(q_e)$ is query meta-pattern generated from the query on q_e, $Match(MP_i(q_e), s)$ is an indicator function that measures how much overlap the sentence s has with the query meta-pattern $MP_i(q_e)$ on the query entities, and k indicates how many the times the sentence s matches the query meta-pattern $MP_i(q_e)$.

- **Textual Evidence Score**: The final score of the candidate evidence sentence is a weighted average of the three scores,

$$S(Q, S) = \sigma \cdot S_w + \theta \cdot S_e + \eta \cdot S_p, \tag{5.6}$$

where (σ, θ, η) is the weight vector indicating the importance of each aspect of the information (i.e., word, entity, and pattern). The default weight vector we use is equal weight for the word, entity, and meta-pattern in our experiments. EvidenceMiner is more effective compared with baseline methods, such as LitSense [19], for textual evidence retrieval in biomedical literature.

This ranking mechanism is more effective compared with existing methods (e.g., LitSense) for textual evidence retrieval in biomedical literature. We use Elasticsearch[24] to support a fast evidence retrieval over the indexed background corpora.

In Figure 6.3, we show an example of our search interface. For example, if scientists are interested in finding the textual evidence for "*melanoma is treated with nivolumab*", they can search it in EvidenceMiner and see the top results such as "bicytopenia in primary lung melanoma treated with nivolumab" (Figure 5.2(a)). If they click one of the top results, the retrieved sentence is highlighted in the original article (Figure 5.3) on the annotation interface. Moreover, EvidenceMiner allows more flexible queries, such as a mixture of keywords and relational patterns. For example, if scientists are interested in finding the diseases that can be treated with the chemical "nivolumab", but are not sure which disease

[24] `https://www.elastic.co/`

to search, they may input a query like "*nivolumab, DISEASEORSYNDROME treat with CHEMICAL*". EvidenceMiner automatically finds all the textual evidence indicating a "treatment" relationship with the chemical "nivolumab" (Figure 5.2(b)).

Fine-Grained Entity and Relationship Visualization The annotation interface shows all the annotated entities and relations for better visualization. For example, in Figure 5.3, we color all the annotated entities with different colors for different types. We use five different colors for the five major biomedical entity types and two additional colors for two specific fine-grained types, "Gene or Genome" and "Disease or Syndrome", since those two are the most frequent biomedical entity types. In Figure 5.3, we see that the "melanoma" is colored as a "Disease or Syndrome" and "nivolumab" is colored as a "Chemical". We also list all the meta-pattern instances and meta-patterns that match the sentences in the article. If the user clicks the meta-pattern instances, the corresponding sentences are also highlighted in the article. In Figure 5.3, a meta-pattern "DISEASEORSYNDROME patient treat with CHEMICAL" captures the entity pair "melanoma" and "nivolumab" in the article.

Fine-Grained Entity and Relation Summarization To make our system more user-friendly and interesting, we add analytic functionalities for the most frequent entity and relation summarization. For example, in Figure 5.4, if scientists are interested in finding the most frequent diseases, they can search "entity_type = DISEASEORSYNDROME" in our analytic interface and see the top entities such as *tumor* and *breast cancer*. Similarly, if scientists are interested in finding the most frequent chemical-disease pairs with a treatment relation, they can search "pattern = DISEASEORSYNDROME treat with CHEMICAL" in our analytic interface and see the top entity pairs such as *HCC&sorafenib*. More interestingly, if researchers are interested in finding the most frequent diseases that can be treated by a specific chemical (e.g., nivolumab), they can search "entity = nivolumab & pattern = DISEASEORSYNDROME treat with CHEMICAL" in our analytic interface and see the most frequent diseases, such as *melanoma* and *NSCLC*, that can be treated with nivolumab. With these analytic functionalities, EvidenceMiner can help scientists uncover important research issues, leading to more effective research and more in-depth quantitative analysis.

Table 5.2: Performance comparison of the textual evidence retrieval systems with nDCG@1,5,10.

Method	nDCG@1	nDCG@5	nDCG@10
BM25	0.714	0.720	0.746
LitSense	0.599	0.624	0.658
EvidenceMiner	**0.855**	**0.861**	**0.889**

5.3 EXPERIMENTS

5.3.1 Overall Performance

To demonstrate the effectiveness of EvidenceMiner in textual evidence retrieval, we compare its performance with the traditional BM25 [123] and a recent sentence-level search engine, LitSense [19]. The background corpus is the same PubMed subset for all the compared methods. We first ask domain experts to generate 50 query statements based on the relationships between three biomedical entity types (gene, chemical, and disease) in the Comparative Toxicogenomics Database[25]. Then we ask domain experts to manually label the top-10 retrieved evidence sentences by each method with three grades indicating the confidence of the evidence. We use the average normalized Discounted Cumulative Gain (nDCG) score to evaluate the textual evidence retrieval performance. In Table 6.2, we observe that EvidenceMiner always achieves the best performance compared with other methods. It demonstrates the effectiveness of using meta-patterns to guide textual evidence discovery in biomedical literature.

5.3.2 Future Development

In some cases, a strict query matching may not find sufficiently high-quality answers due to the stringent search requirements or limited available entities that match the search queries. In this case, a smart query processor should automatically kick-in to do an approximate match, such as a graph-based approximate match or an embedding-based semantic match. In other cases, a user may query a set of entities (e.g., genes or diseases) or a timeline. We need to conduct a summary of the major differences among the set of entities or over time by analyzing large text.

[25] `http://ctdbase.org`

5.4 RELATED WORK

Search engines performing sentence-level retrieval have been developed in the biomedical domain. For example, Textpresso [124] highlights the query-related sentences in the retrieved documents. However, the sentence highlighting is only based on query word matching, which does not necessarily find sentences semantically related to the input query. Another example is LitSense [19], which retrieves semantically similar sentences in biomedical literature given a query sentence. It returns best-matching sentences using a combined approach of traditional word matching and neural embedding. However, their neural embeddings are noisy and thus negatively impact the effectiveness in retrieving query-specific evidence sentences. EvidenceMiner is more effective compared with LitSense for textual evidence retrieval in biomedical literature.

Similar tools are also developed for other domains, such as claim mining and argument mining tools on Twitter or news articles. PerspectroScope [119] allows users to query a natural language claim and extract textual evidence in support or against the claim. ClaimPortal [120] is an integrated infrastructure for searching and checking factual claims on Twitter. TARGER [121] is an argument mining framework for tagging arguments in the free input text and keyword-based retrieval of arguments from the argument-tagged corpus. Most of these tools rely on fully supervised methods that require human-annotated training data. It is difficult to directly apply these systems to other domains such as life sciences. Because it is non-trivial to acquire the human-annotated articles and the annotations are usually prone to errors [125].

5.5 SUMMARY

In this chapter, we proposed EvidenceMiner, a web-based system for textual evidence discovery for life sciences. The retrieved evidence sentences can be easily located in the background corpora for better visualization. EvidenceMiner also includes analytic functionalities such as the most frequent entity and relation summarization. We incorporated another corpus on COVID-19 in EvidenceMiner to help boost the scientific discoveries. We plan to further develope EvidenceMiner to be a more intelligent system that can assist in more efficient and in-depth scientific discoveries.

CHAPTER 6: SCIENTIFIC TOPIC CONTRASTING

6.1 INTRODUCTION

Scientific topic contrasting allows scientists to explore multiple topics at the same time to find representative and contrasting knowledge (entities or relationships) for each topic from the scientific literature. A similar scenario, product comparison, has been widely used in e-commerce today. The major difference between them is that product comparison is based on structured databases whereas scientific topic contrasting is based on unstructured text data (e.g., scientific literature).

Scientific topic contrasting is commonly needed in scientific research. For example, in Figure 6.1, the clinical researchers want to develop drugs that precisely treat six main categories of heart diseases: cerebrovascular accident (CVA), ischemic heart disease (IHD), cardiomyopathies (CM), congenital heart disease (CHD), arrhythmias (ARR), and valve disease (VD). To conduct this precision medicine development, researchers need to find the most representative and contrasting proteins for each category as target proteins for treatment. The target proteins for each category of heart diseases (e.g., TSPNA2 for CVA) should be strongly associated with this category (CVA) but weakly associated with other categories (IHD, CM, CHD, ARR, or VD). To find the most representative proteins for each category of heart diseases, researchers often look into biomedical literature for distinctive associations between proteins and heart diseases before they conduct expensive experimental validation. Scientific topic contrasting finds the most representative and contrasting knowledge (e.g., proteins) for multiple comparable topics (e.g., six categories of heart diseases) from the scientific literature.

Scientific topic contrasting, unfortunately, is under-explored in current literature search and analysis systems. Traditional search engines for life sciences (e.g., PubMed) are designed for document retrieval and do not include the scientific topic contrasting function [113]. Large-scale information extraction systems have been constructed to transform massive unstructured text data into structured knowledge [9, 114, 126, 127, 128, 129, 130] in many domains. However, such information extraction systems have not built functions to allow users to query knowledge by comparing across multiple customized topics. Life-iNet [114] provides a function of distinctive entity summarization but the topics that can be compared are restricted to entity types pre-defined by Life-iNet, not allowing users to query customized topics of their interests.

In this chapter, we propose SciContrast that addresses this open problem of scientific

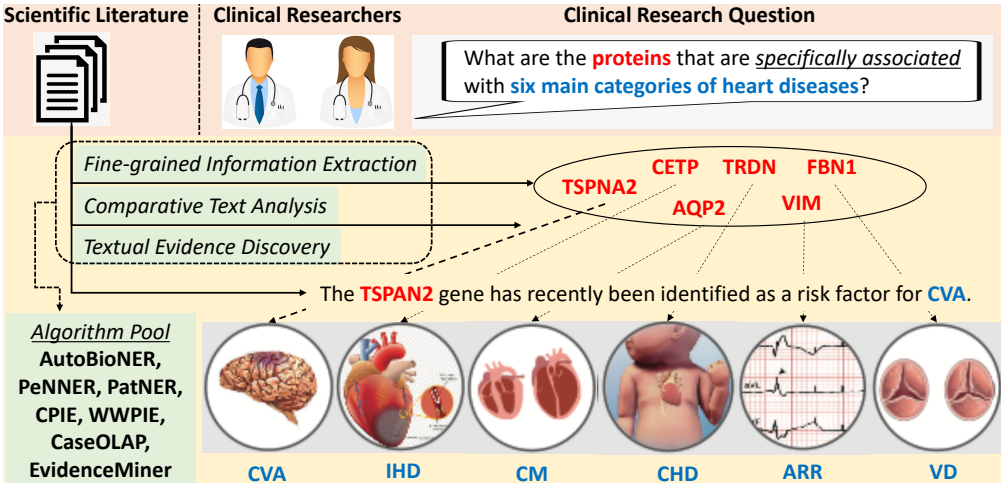


Figure 6.1: An example of scientific topic contrasting from life science literature.

topic contrasting in scientific literature. It allows scientists to explore a set of customized topics of their interests by summarizing the most representative knowledge for each topic as well as providing concrete evidence sentences supporting such knowledge discovery from a scientific corpus. SciContrast is constructed in a completely automated way without any human effort for training data annotation. Taken a background corpus and a domain-specific knowledge base as input, SciContrast relies on the domain-specific knowledge base (e.g., UMLS for biomedicine) to provide distant supervision for fine-grained named entity recognition [2, 4, 5, 7, 80] and meta-pattern-based open relation extraction [42, 122]. Based on the extracted entities and relationships, SciContrast summarizes the most representative knowledge for each topic using comparative text analysis [131]. Concrete evidence sentences [16, 17] are provided to further support knowledge discovery from the scientific literature. SciContrast is resulted from our long-term collaborations with biomedical and clinical researchers to help scientists recognize the major differences among comparative topics and further uncover hidden important issues for effective research [21]. In this study, we demonstrate the power of SciContrast in the biomedical literature. However, SciContrast can be generally applied to any scientific domains requiring a comparative knowledge discovery.

6.2 THE SCICONTRAST FRAMEWORK

SciContrast consists of three major components: a fine-grained information extraction pipeline, a comparative text analysis pipeline, and a textual evidence discovery pipeline. The details of each component are introduced below.

Table 6.1: Basic statistics of the background corpus. It includes biomedical literature published in 2019 that are specifically related to cancers and heart diseases.

Background Corpora	**Heart Diseases**	**Cancers**
# of PubMed Abstracts	11,766	48,201
# of PMC Full Texts	1,151	7,130
# of Sentences	246,106	1,466,091
# of Entities	400,327	3,315,092
# of Relationships	9,576	29,160

6.2.1 Fine-grained Information Extraction

The fine-grained information extraction pipeline extracts entities with distant supervision from comprehensive biomedical knowledge bases and open relations with meta-pattern discovery methods.

Corpus Collection We have collected a background corpus of biomedical literature from PubMed[26] and PubMed Central[27] (PMC), containing 26 million PubMed papers (titles and abstracts) and 2.2 million PMC full-text papers. For the demonstration purpose, we use a subset corpus containing biomedical literature published in 2019, specifically related to cancers and heart diseases. To collect this subset corpus, we first use a biomedical concept ontology, MeSH[28], to find all the concepts related to cancers ("Neoplasms" in MeSH) and heart diseases ("Cardiovascular Diseases" in MeSH), and then select the papers that contain any of those related concepts for cancers and heart diseases to form the background corpus for SciContrast. Table 6.1 shows the summary statistics of the background corpus for SciContrast.

Distantly Supervised Named Entity Recognition SciContrast utilizes entity information from UMLS[29], a comprehensive biomedical knowledge base for distantly supervised named entity recognition (NER). A major problem of existing distantly supervised NER methods [62] is the limited coverage of the dictionaries, which leads to false-negative labeling errors. We use PatNER [4, 5] that leverages frequent pattern mining to enhance the distantly supervised NER performance. The intuition is that biomedical entities are often named following some principles (e.g., disease entities often contain the words "syndrome" or "disorder") that can be indicative for biomedical entity recognition. PatNER automati-

[26] `https://pubmed.gov/pubmed`
[27] `https://pubmed.gov/pmc`
[28] `https://www.nlm.nih.gov/mesh/`
[29] `https://www.nlm.nih.gov/research/umls/index.html`

cally mines the entity naming principles from the domain-specific dictionaries from UMLS to quantify the entities and candidate phrases. It then labels each candidate entity as a probability distribution over all the entity types to train a fuzzy NER neural model. This fuzzy NER neural model allows us to skip the most uncertain labels during training and significantly reduce the impact of false-negative labeling errors. PatNER achieves comparable NER performance with state-of-the-art supervised BioNER methods, such as BioBERT [20]. However, the supervised BioNER methods cannot be directly applied to recognize new entity types. PatNER requires no human effort for training data annotation and can automatically recognize 17 fine-grained biomedical entity types from UMLS for SciContrast.

Meta-Pattern-Based Open Relation Extraction Following the fine-grained named entity recognition, fine-grained relations need to be further extracted to provide comprehensive knowledge for comparative text analysis. Supervised relation extraction methods cannot cover all possible relation types between the input entity types. To extract all possible relations involving the 17 fine-grained biomedical entity types without human supervision, we use CPIE [42] and WW-PIE [122], two state-of-the-art meta-pattern discovery methods for open relation extraction in the biomedical domain. Meta-pattern-based methods utilize data redundancy to derive informative frequent patterns as relation types for open relation extraction. In a large corpus, such redundancy is abundant. With the entities pre-recognized, entity mentions can be replaced with their entity types, and meta-patterns (textual patterns containing entity type tokens) become apparent (e.g., by replacing "Denosumab" with "DRUG" and "Osteoproposis" with "DISEASE", and so on, "DRUG treat DISEASE" becomes a frequent pattern). By breaking down long sentences into shorter yet meaningful sentences or segments, we can conduct pattern mining, discover high-quality meta-patterns, and group patterns hierarchically to better understand and organize patterns. Then we match the quality meta-patterns back to the corpus for relation instance extraction. The mined meta-patterns (e.g., "DRUG treat DISEASE") can be used as relation types, and their corresponding extractions matched in the text (e.g., (Denosumab, treat, Osteoproposis)) can be used as relation instances.

The fine-grained information extraction pipeline is run offline as a pre-processing step to provide comprehensive knowledge for comparative text analysis. In SciContrast, we have provided the 17 fine-grained biomedical entity types as well as the most frequent 100 meta-pattern relation types for users to select from as the knowledge to be compared across different customized topics.

6.2.2 Comparative Text Analysis

The comparative text analysis pipeline takes the user-input topics and the user-selected knowledge types (entity types and relation types) to perform fast online comparative analysis. It first categorizes the documents in the background corpus by the user-input topics and then summarizes the most representative knowledge for each topic.

Topic-Guided Document Categorization Taken the user-input topics and the documents in the background corpus, we associate each document with its corresponding topics via simple string matching. In particular, a document is associated with a topic if the topic phrase appears at least once in that document. Then we find the most representative entities and relationships within the documents for each topic via comparative text analysis.

Comparative Entity and Relationship Discovery There is no universally accepted standard for measuring the representativeness of entities or relationships for a given topic. Inspired by CaseOLAP [131], we measure the representativeness in terms of two criteria:

- **Popularity**: An entity or relationship is considered popular if it has a large number of occurrences. Representative entities or relationships should appear with some frequency within the documents for that topic for a substantial contribution to the topic semantics.
- **Distinctiveness**: Entities or relationships that appear more discriminatively on one topic in comparison with other topics should have higher discrimination power. Representative entities or relationships should distinguish the target topic from other topics with more salient information.

According to the above two criteria, we define the score for each representative entity or relationship as a combination of two scores: popularity and distinctivity.

The popularity score S_P is defined as follows:

$$S_P(k,t) = \frac{log(c(k,D_t)+1)}{log(\sum_{k' \in K} c(k',D_t))} \in [0,1], \tag{6.1}$$

where $c(k, D_t)$ is the number of occurrences of the entity or relationship k in the documents D_t for the topic t. $K = \{k\}$ is the set of all the entities or relationships given a user-selected knowledge type.

The distinctivity score S_D is defined as follows:

$$S_D(k,t) = \frac{e^{rel(k,t)}}{1+\sum_{t' \in T'} e^{rel(k,t')}} \in (0,1], \tag{6.2}$$

where $T' = T \setminus \{t\}$ is the set of all the user-input topics T excluding the current topic t and $rel(k, t)$ is the relevance score of the entity or relationship k in the documents for the topic t. To better describe the relevance between an entity or a relationship k and a topic t, two normalizations were adopted:

$$rel(k, t) = NormTF(k, t) \times NormDF(k, t). \tag{6.3}$$

$NormTF(k, t)$ is the normalized term frequency and $NormDF(k, t)$ is the normalized document frequency, which are calculated as follows:

$$NormTF(k, t) = \frac{c(k, D_t)(a + 1)}{c(k, D_t) + a(1 - b + bN_t)}, \tag{6.4}$$

where a and b are two weighting constants and N_t is the number of entities or relationships in the documents D_t for the topic t. Specifically for SciContrast, we use $a = 1.2$ and $b = 0.75$.

$$NormDF(k, t) = \frac{log(1 + |\{d \in D_t : k \in d\}|)}{log(1 + max_{k' \in K}(|\{d \in D_t : k' \in d\}|))}, \tag{6.5}$$

where $|\{d \in D_t : k \in d\}|$ is the number of documents for topic t that k occurs and $max_{k' \in K}(|d \in D_t : k' \in d|)$ is the collection of documents in topic t with the largest cardinality.

The combined score S for each entity or relationship is the product of its popularity and distinctivity scores:

$$S = S_P \times S_D \in [0, 1]. \tag{6.6}$$

This ranking function is more effective compared with baseline methods, such as TF-IDF [132] or MCX [133], for representative phrase discovery across multiple comparable document sets [131]. The output of the comparative text analysis pipeline should be ranked lists of the user-selected knowledge (entities or relationships) for the user-input topics.

Figure 6.2 shows an example of the comparative entity and relation discovery results. By default, we show the top-10 representative entities or relationships (ranked by the combined score) for each user-input topic. For example, the top representative "GENE OR GENOME" entity for the topic "breast cancer" is "brca1", one of the most important genes related to breast cancer[30]. Different score components ("distinctivity" and "popularity") can be

[30]About 55%–72% of women who inherit a harmful BRCA1 variant will develop breast cancer by the age of 70–80 [134].

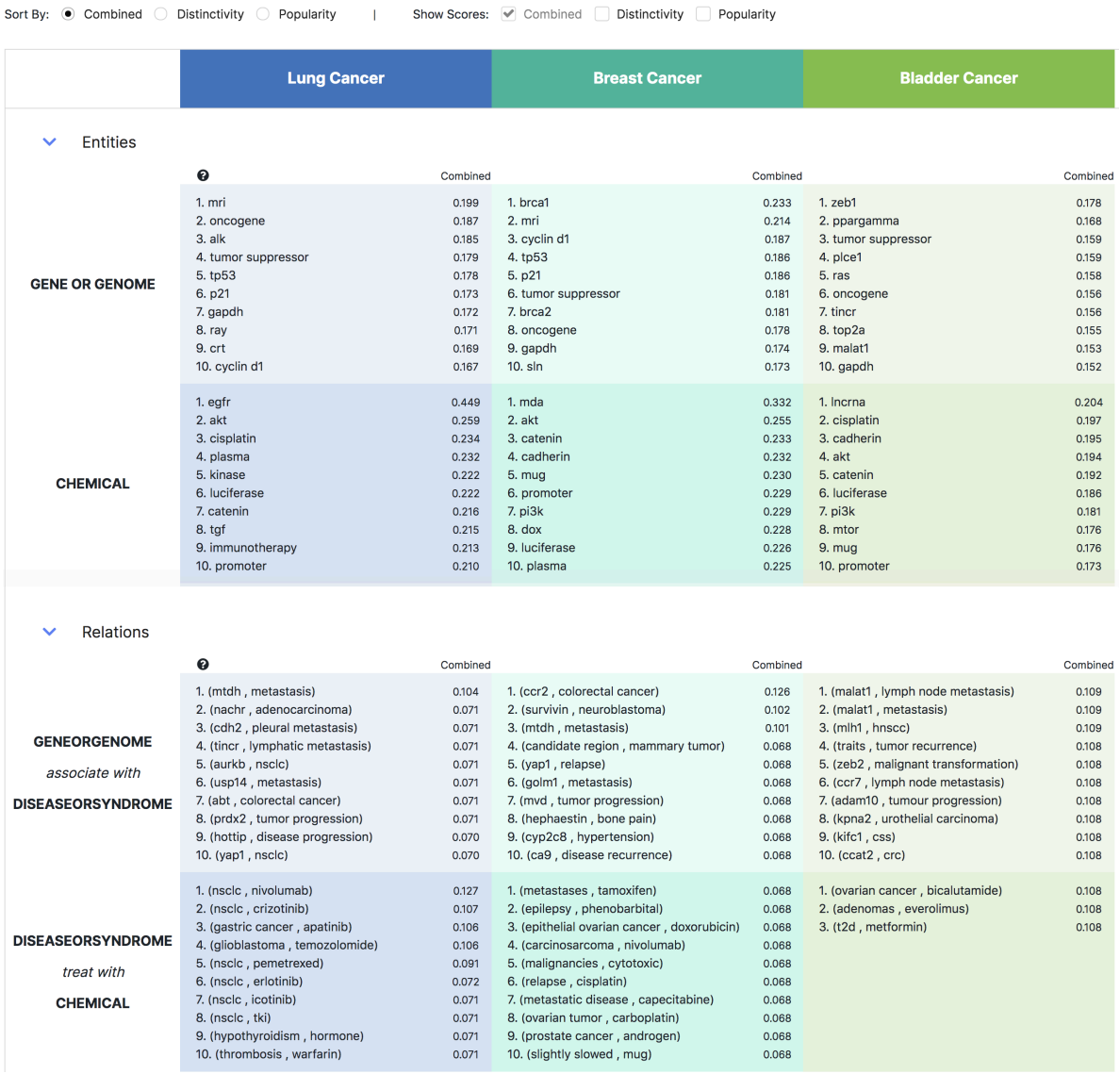


Figure 6.2: Example of the comparative entity and relation discovery results for three user-input topics ("breast cancer", "lung cancer", and "bladder cancer") on two user-selected entity types ("GENE OR GENOME" and "CHEMICAL") and two user-selected relation types ("GENE OR GENOME associate with DISEASE OR SYNDROME" and "DISEASE OR SYNDROME treat with CHEMICAL").

explored by users. A user can select to show "distinctivity" or "popularity", or sort the representative entities or relationships by "distinctivity" or "popularity".

6.2.3 Textual Evidence Discovery

The textual evidence discovery pipeline retrieves textual evidence to support the representative entity or relationship discovery from the scientific literature. It also includes visu-

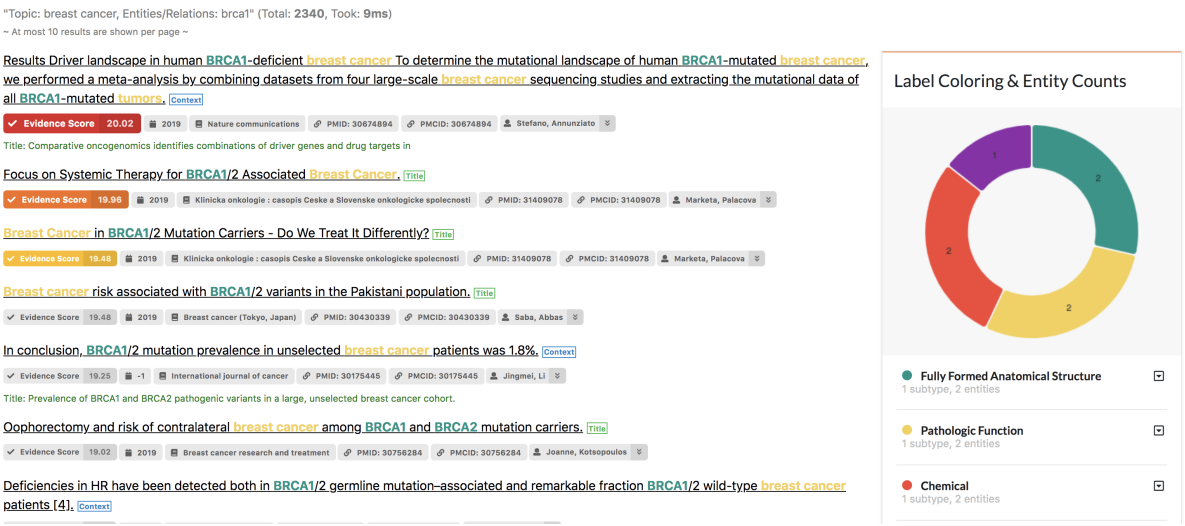


Figure 6.3: Example of the evidence sentence retrieval results for the topic "breast cancer" and its top-1 representative GENE "brca1". The fine-grained entities are highlighted in different colors according to their entity types.

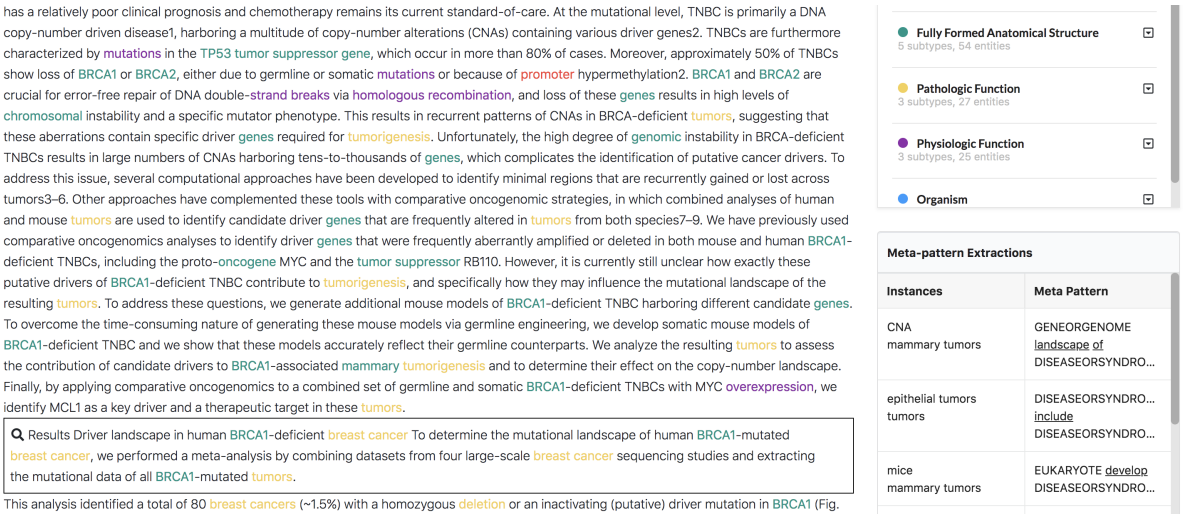


Figure 6.4: Example visualization of a retrieved textual evidence and the fine-grained entity and relation extraction results in the original document. The fine-grained biomedical entities are highlighted in different colors according to their entity types. The meta-pattern relations with their extracted relation instances are also shown on the right.

alization of the retrieved textual evidence and the fine-grained entity and relation extraction results in the original documents.

Textual Evidence Sentence Retrieval Given the top-ranked entities or relationships and their corresponding topics, SciContrast retrieves and ranks the evidence sentences with EvidenceMiner [16, 17] to support this representative entity or relationship discov-

ery from the background corpus. EvidenceMiner is more effective compared with baseline methods, such as LitSense [19], for textual evidence retrieval in biomedical literature. The sentences containing the selected entity or relationship are ranked higher if they are more related to the selected topic.

Figure 6.3 shows an example of the evidence sentence retrieval results. For example, if the scientists are interested in finding the textual evidence for the topic "breast cancer" and its top-1 representative GENE "brca1", SciContrast will return the evidence sentences such as "Breast cancer risk associated with BRCA1/2 variants in the Pakistani population" as the supporting evidence. The evidence score indicates the confidence of each retrieved sentence being supporting evidence for this representative knowledge discovery.

Fine-Grained Entity and Relationship Visualization If the user clicks one of the retrieved evidence sentences, the selected evidence sentence will be highlighted in the original document. Figure 6.4 shows an example of retrieved textual evidence and the fine-grained entity and relation extraction results in the original document. The fine-grained entities are highlighted in different colors according to their entity types. The meta-pattern relations with their extracted relation instances are also shown and can be linked to the sentences where each relation instance is extracted in this document.

6.3 EXPERIMENTS

6.3.1 Overall Performance

To demonstrate the effectiveness of SciContrast in unsupervised representative knowledge discovery, we compare its performance with traditional information retrieval methods, TF-IDF [132], BM25 [123], and QLM-Dirichlet [135], and a recent pre-trained language model in the biomedical domain, BioBERT [20]. We first collect the top 50 proteins for the six main categories of heart diseases (CVA, IHD, CM, CHD, ARR, and VD) generated by all the baseline methods. We then ask the clinical researchers at the UC Davis Medical Center to manually label the proteins with five grades indicating the relevance of the protein to the heart diease category. We use the average normalized Discounted Cumulative Gain (nDCG) score to evaluate the unsupervised representative protein discovery performance. In Table 6.2, we observe that SciContrast always achieves the best performance compared with other baseline methods. Some of our discovered proteins are under experimental evaluation by clinical researchers at the UC Davis Medical Center, looking for novel therapeutic targets that do not respond to the conventional drug treatment used in clinic for heart failure.

Table 6.2: Performance comparison of baseline methods for unsupervised representative protein discovery in six main categories of heart diseases with nDCG@10,50.

Method	nDCG@10	nDCG@50
TF-IDF	0.5439	0.8183
BM25	0.5501	0.8209
QLM-Dirichlet	0.5547	0.8226
BioBERT	0.6054	0.8415
SciContrast	**0.6819**	**0.8721**

Table 6.3: Popularity comparison of baseline methods for unsupervised representative protein discovery in six main categories of heart diseases with nDCG@10,50.

Method	nDCG@10	nDCG@50
TF-IDF	0.5671	0.8371
BM25	0.5789	0.8420
QLM-Dirichlet	0.5724	0.8383
BioBERT	0.6642	0.8705
SciContrast	**0.6704**	**0.8748**

6.3.2 Popularity and Distinctivity

In addition to the general relevance showed in Table 6.2, we also ask the clinical researchers in the UC Davis Heart Failure Program at the UC Davis Medical Center to manually label the proteins with five grades indicating the **popularity** and **distinctivity** of the protein to the heart diease category, respectively. In Table 6.3 and 6.4, we observe that SciContrast always achieves the best performance in both popularity and distinctivity compared with other baseline methods.

6.3.3 Case Study

We also show some case studies of the top-10 proteins discovered by different baseline methods in arrhymias (ARR) in Table 6.5. The good proteins (marked in blue) are the ones that have already been identified with clear clinical meanings for the disease ARR. The bad proteins (marked in red) are the ones identified by the clinical researchers that do not have any relationship with the disease ARR. In Table 6.5, we observe that SciContrast discovers many good proteins and a few bad proteins in the top results. TF-IDF, BM25, and QLM-Dirichlet discover more bad proteins than good proteins in the top results. BioBERT does not discover any good proteins nor bad proteins in the top results. Moreover, we observe that the top results of SciContrast and BioBERT are complementary to each other. One future improvement is to ensemble with or incorporate the BioBERT embedding into the

Table 6.4: Distinctivity comparison of baseline methods for unsupervised representative protein discovery in six main categories of heart diseases with nDCG@10,50.

Method	nDCG@10	nDCG@50
TF-IDF	0.4778	0.7587
BM25	0.4689	0.7551
QLM-Dirichlet	0.4854	0.7613
BioBERT	0.5417	0.7851
SciContrast	**0.6345**	**0.8452**

Table 6.5: Case study of the top-10 proteins discovered by different baseline algorithms in arrhymias (ARR). The clinically relevant proteins are marked in blue and the irrelevant proteins are marked in red based on clinical researcher evaluation.

SciContrast	TF-IDF	BM25	QLM-Dirichlet	BioBERT
methionine synthase	methionine synthase	methionine synthase	beta-2-glycoprotein 1	troponin t cardiac muscle
ryanodine receptor 2	beta-2-glycoprotein 1	guanine nucleotide-binding protein g	amyloid beta a4 protein	troponin i cardiac muscle
potassium voltage-gated channel subfamily h member	guanine nucleotide-binding protein g	cytochrome p450 2c9	cytochrome p450 2c9	natriuretic peptides a
inward rectifier potassium channel 2	cytochrome p450 2c9	amyloid beta a4 protein	guanine nucleotide-binding protein g	endothelin-1 receptor
beta-2-glycoprotein 1	amyloid beta a4 protein	collagen alpha-1	methionine synthase	ryanodine receptor 2
amyloid beta a4 protein	collagen alpha-1	beta-2-glycoprotein 1	collagen alpha-1	endothelial lipase
gap junction alpha-1 protein	natriuretic peptides b	natriuretic peptides b	natriuretic peptides b	natriuretic peptides b
collagen alpha-1	mineralocorticoid receptor	mineralocorticoid receptor	mineralocorticoid receptor	platelet-activating factor acetylhydrolase
guanine nucleotide-binding protein g	ryanodine receptor 2	ryanodine receptor 2	ryanodine receptor 2	endothelin-1
cytochrome p450 2c9	natriuretic peptides a	natriuretic peptides a	angiotensin-converting enzyme	beta-1 adrenergic receptor

comparative text analysis framework of SciContrast for better performance.

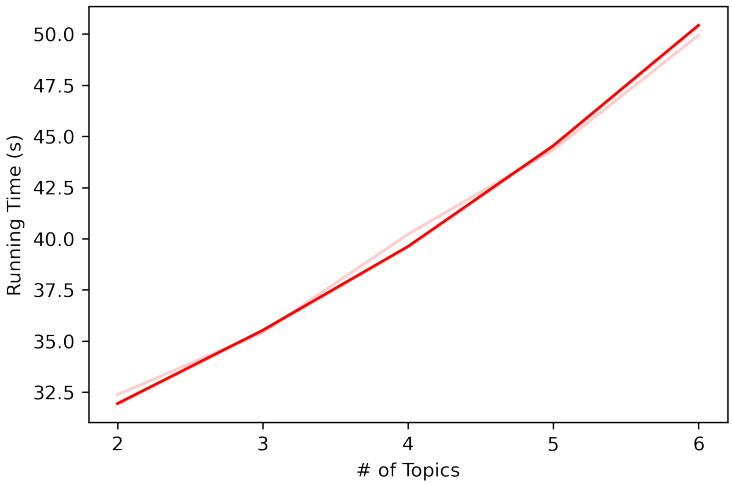


Figure 6.5: Running Time vs. the number of topics

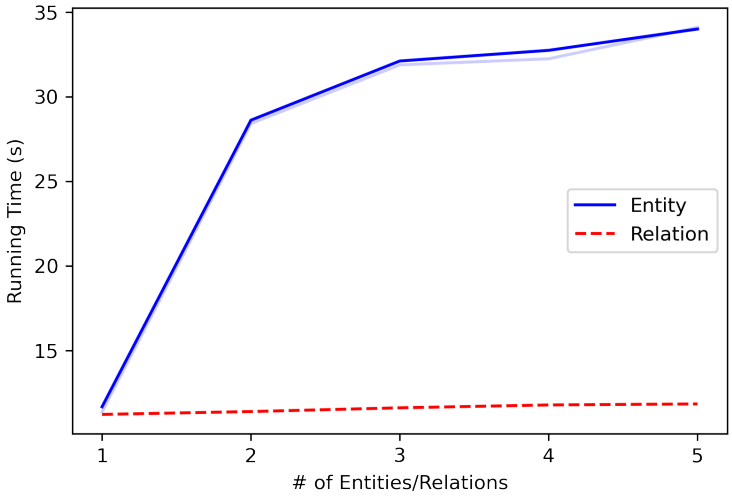


Figure 6.6: Running Time vs. the number of entities/relations

6.3.4 Runtime Analysis

The SciContrast system is run on the Intel Xeon CPU E5-2630 with no GPU involved. There are three types of inputs that users can change for SciContrast (i.e., topics, entities, relations). To avoid system timeout, users can contrast at most six topics, five entities, and five relations. A detailed runtime analysis is shown below.

- **Fine-grained Information Extraction**: The fine-grained information extraction pipeline is run offline. The extracted entities and relationships are stored and indexed to support a fast online topic contrasting and evidence sentence retrieval.
- **Query Processing**: Taken the user-input topics and selected entity/relation types as

input, the query processing step takes several milliseconds since no query pre-processing is involved.

- **Scientific Topic Contrasting**: The details of the average runtime performance of topic contrasting are shown in Figures 6.5 and 6.6. The running time in Figure 6.5 is the result of holding both the number of entities and the number of relations constantly at four. The average runtime linearly increases as the number of topics increases. The running time in Figure 6.6 is the result of holding both the number of topics constantly at two and the other entity/relation variable at zero. The average runtime has no obvious relationship with the number of entities and relations but is rather related to the number of occurrences of the entity or relation in the documents. Since the occurrences of relations are usually far less than the occurrences of entities in the documents, the increase in average runtime by increasing the number of relations is not as obvious as that of entities.

- **Evidence Sentence Retrieval**: We use Elasticsearch[31] to create the index for each sentence for fast online retrieval. In addition to indexing the keywords, we index each sentence with the meta-patterns it matches and the corresponding entities extracted by the meta-patterns in the sentence. The average runtime of evidence sentence retrieval is about tens of milliseconds for any given query.

6.3.5 Clinical Use Case

We have been collaborating with clinical researchers on using SciContrast for distinctive gene set discovery to identify cardiovascular proteins that are specifically associated with six main categories of heart diseases [21]. Some of our discovered representative proteins are under experimental evaluation by clinical researchers looking for novel therapeutic targets in patients and means to repurpose drugs already used in clinic. Our method may unveil new molecular drug targets in heart diseases that do not respond to the conventional drug treatment used in clinic. We are continuing our collaboration with the UC Davis Medical School to further extend SciContrast for broader clinical applications such as identifying the distinctive genes for different risk factors (e.g., diabetes) that may lead to different sub-categories of heart diseases as well as cancers.

[31] `https://www.elastic.co/`

6.4 RELATED WORK

Large-scale information extraction systems have been constructed to transform massive unstructured text data into structured knowledge [9, 114, 126, 127, 128, 129, 130]. In the biomedical domain, Life-iNet [114] constructs structured networks of factual knowledge from large amounts of biomedical literature. COVID-KG [9] constructs a comprehensive knowledge graph by extracting fine-grained multimedia knowledge elements (entities, relations, and events) from COVID-19 literature. The extracted knowledge (e.g., entities and relationships) helps downstream tasks such as knowledge exploration, question answering, drug target prediction, and drug repurposing report generation. However, existing literature search and analysis systems built on top of these literature information extraction results are designed for users to query knowledge without comparing across different topics or conditions.

Little effort has been made for scientific topic contrasting from the scientific literature. Life-iNet [114] has a function of distinctive entity summarization. However, the categories that can be compared with are confined to some pre-defined categories by the system. It does not allow users to query any customized categories of their interests. Also, Life-iNet only finds distinctive entities without other kinds of knowledge such as distinctive relationships. For the supporting literature evidence, Life-iNet only shows the relevant papers for each distinctive entity without finding concrete evidence sentences supporting the distinctive association between each entity and its corresponding category.

6.5 SUMMARY

In this chapter, we proposed SciContrast, a web-based system for scientific topic contrasting from life science literature. SciContrast summarizes and contrasts the most representative knowledge for each user-input topic as well as providing concrete evidence sentences supporting this representative knowledge discovery from the scientific literature. We have been collaborating with clinical researchers on using the system for real-world clinical studies. SciContrast have high potential to advance both literature analysis techniques and scientific discovery applications.

CHAPTER 7: APPLICATIONS AND CONCLUSIONS

7.1 SCIENTIFIC TEXT MINING: SUMMARY

This thesis focuses on developing effective and scalable text mining *algorithms* and *systems* to enable and accelerate scientific discovery. 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. To address this challenge, we have developed methods that extract information from text with minimal human supervision. With the advanced text mining methods developed, we future study how to enable and accelerate real-world knowledge discovery. We have been collaborating with experts in various science domains (e.g., biomedicine, chemistry, and health) to achieve this goal. Overall, this thesis has made contributions to the following aspects.

Contributions 7.1 We propose **three weak supervision sources** for scientific information extraction with minimal human supervision.

- **Pattern-Enhanced Weak Supervision**: Scientific literature analysis needs dozens to hundreds of distinct, fine-grained entity types, making consistent and accurate annotation difficult even for crowds of domain experts. However, domain-specific ontologies and knowledge bases (KBs) can be easily accessed, constructed, or integrated, making distant supervision realistic for fine-grained scientific NER tasks. For distant supervision, training labels are automatically generated by matching the mentions in text with the concepts in the KBs. A major challenge of distant supervision is the limited coverage of the dictionaries from the KBs, leading to false-negative errors during the distant training label creation. To tackle the challenge *incomplete dictionaries* for distant label generation, we study the problem of biomedical named entity recognition with various weak supervision signals (e.g., distant supervision from knowledge bases and weak supervision from seed textual patterns).
- **Ontology-Guided Distant Supervision**: In addition to the aforementioned incomplete dictionary problem, the distant supervision faces another great challenge of *noisy annotation* where a mention can be erroneously matched due to the potential matching of multiple entity types in the KBs. Previous distantly supervised NER studies largely ignore the noisy annotation problem by simply discarding those multi-labels during the

KB-matching process. However, the noisy labels cannot be simply ignored for the chemistry entities because they consist of a large portion of distant training labels. To tackle the *noisy annotation* challenge for distant label generation, we study the problem of fine-grained chemistry named entity recognition with distant supervision from domain-specific knowledge bases and ontologies for multi-type disambiguation.

- **Cross-Modal Supervision Between Text and Graph**: Scientific knowledge usually resides in *multiple modalities*. For example, chemical compounds can be described with both text descriptions and molecule graphs. It is challenging to learn a scientific entity representation with multi-modal information. On the other hand, we see this multi-modal representation as an opportunity since the information in one modality may benefit the tasks in other modalities. To investigate and better utilize the multi-modal representation of the scientific knowledge, we study the problem of reactant entity classification with supervision from molecule graph matching.

Contributions 7.2 We study **different scientific text mining tasks** including the biomedical entity recognition (Chapter 2), fine-grained chemistry named entity recognition (Chapter 3), chemical reactant entity classification (Chapter 4), scientific textual evidence retrieval (Chapter 5), and scientific topic contrasting (Chapter 6). In particular, we investigate solutions with minimal human supervision using weak supervision from knowledge bases, ontologies, and other data modalities (e.g., graphs).

Contributions 7.3 We have proposed **models** and **algorithms** to solve the above tasks.

- We proposed PeNNER (Chapter 2) to solve the nested biomedical named entity recognition problem. PeNNER relies on massive corpora and unsupervised pattern mining for nested named entity boundary correction.

- We proposed ChemNER (Chapter 3) to solve the fine-grained chemistry named entity recognition problem. ChemNER leverages the chemistry type ontology structure to provide a global topic constraint for context-aware multi-type disambiguation.

- We proposed ReactClass (Chapter 4) to solve the chemical reactant entity classification problem. ReactClass is designed to take two special characteristics, multi-modal representation and knowledge-aware subword correlation, of the chemical molecules into consideration.

- We proposed EvidenceMiner (Chapter 5) to solve the scientific textual evidence retrieval problem. EvidenceMiner incorporates the fine-grained named entity and open relation information to discover textual evidence.

- We proposed SciContrast (Chapter 6) to solve the scientific topic contrasting problem. SciContrast summarizes and contrasts the most representative knowledge for each user-input topic as well as provides concrete evidence sentences supporting this representative knowledge discovery from the scientific literature.

7.2 APPLICATIONS

7.2.1 Open Information Extraction with Meta-Pattern Discovery

Inspired by our pattern-enhanced, weakly supervised BioNER methods, we further propose pattern-guided open information extraction (OpenIE) methods for biomedical literature. OpenIE 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 [39, 41] utilize data redundancy to derive informative frequent patterns and use the derived patterns as relation types for open relation extraction. Compared with existing 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. For example, in Figure 7.1, "Pre-treatment of ATRA can decrease the overexpression of cyclin_D1 and E2F-1 induced by B(a)P", where "ATRA" and "B(a)P" are chemicals and "cyclin_D1" and "E2F-1" are genes. Existing meta-pattern discovery methods discover frequent meta-patterns such as "GENE and GENE" and "CHEMICAL decrease CHEMICAL", but not long and infrequent meta-patterns such as "CHEMICAL decrease GENE and GENE induced by CHEMICAL". To tackle the above challenge, we propose several meta-pattern-guided OpenIE methods (CPIE [42] and WW-PIE [122]) that extract meta-patterns spanning long and complex sentences in biomedical literature.

We propose WW-PIE, a novel wide-window pattern-based OpenIE method for biomedical literature. WW-PIE addresses three challenges: (1) the long sentences with long-distanced entity mentions, (2) the hierarchical or n-ary relations among long-distanced entities mentioned in one sentence, and (3) the completeness of extractions. The key idea is to first break down the long sentences into shorter yet meaningful sentences or segments and then conduct

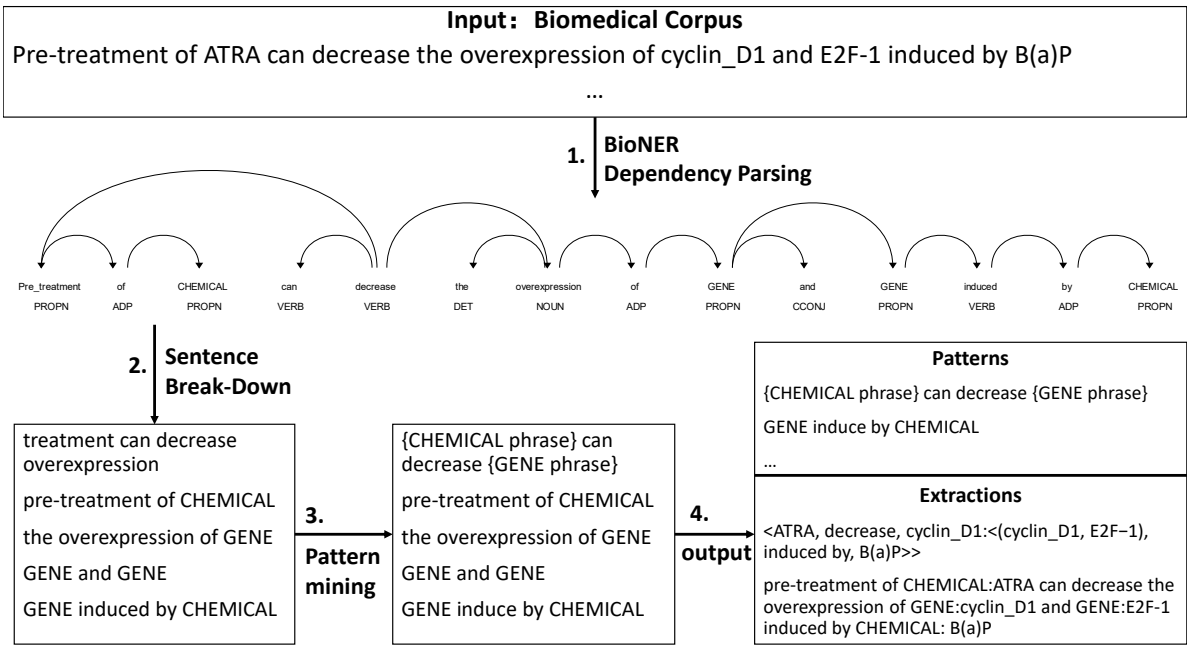


Figure 7.1: The overall framework of WW-PIE: Wide-Window Pattern-based Open Information Extraction. This figure is adapted from [122].

pattern mining. We utilize dependency parsing to resolve complex sentence structures and utilize frequency pattern mining to discover quality meta-patterns. After discovering quality meta-patterns, we propose a novel hierarchical pattern grouping to better organize the extractions, keeping both the simplicity and the structure of the relationships. The output will include two parts: the tuples as relation instances and the meta-patterns as relation types. For example, in Figure 7.1, the tuple extraction is ⟨ATRA, decrease, cyclin_D1:⟨(cyclin_D1, E2F-1), induced by, B(a)P⟩⟩, and the meta-pattern extraction is "pre-treatment of CHEMICAL:ATRA can decrease the overexpression of GENE:cyclin_D1 and GENE:E2F-1 induced by CHEMICAL:B(a)P ". Experiments on real-world biomedical corpus demonstrate the power of WW-PIE at extracting precise and well-structured information. Our meta-pattern-guided OpenIE methods are highly effective in extracting rich information from large-scale biomedical literature. They have been used for downstream applications such as textual evidence discovery in life sciences [16, 17].

7.2.2 Knowledge Graph Construction and Drug Repurposing Report Generation

Our fine-grained named entity recognition methods (e.g, CORD-NER [6]) and textual evidence discovery methods (e.g., EvidenceMiner [16]) have been used in and inspired a follow-up work COVID-KG [9]: COVID-19 knowledge graph construction and drug repur-

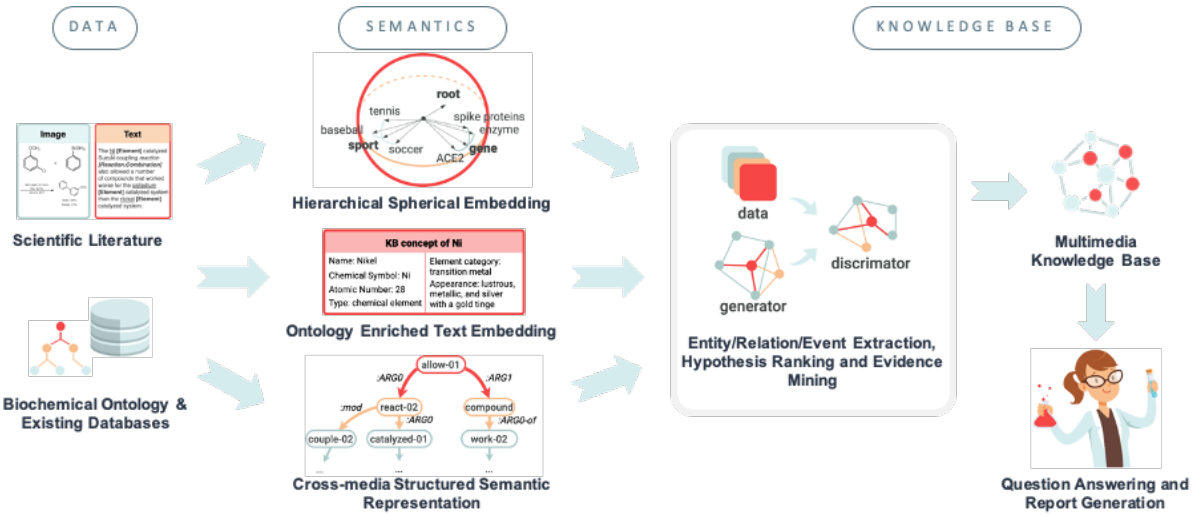


Figure 7.2: COVID-KG overview: from data to semantics to knowledge. This figure is adapted from [9].

posing report generation. COVID-KG has been awarded the **Best Demo Paper Award in 2021 from NAACL**.

Practical progress in combating COVID-19 relies heavily on effective search, discovery, assessment, and extension of scientific research results. However, clinicians and scientists are facing two unique barriers in digesting these research papers. The first challenge is quantity. Such a bottleneck in knowledge access is exacerbated during a pandemic when increased investment in relevant research leads to even faster growth of literature than usual. The resulting knowledge bottleneck contributes to significant delays in the development of vaccines and drugs for COVID-19. More intelligent knowledge discovery technologies need to be developed to enable researchers to more quickly and accurately access and digest relevant knowledge from the literature. The second challenge is quality. Many research results about coronavirus from different research labs and sources are redundant, complementary, or even conflicting with each other, while some false information has been promoted in both formal publication venues as well as social media platforms such as Twitter. As a result, some of the public policy responses to the virus, and public perception of it, have been based on misleading, and at times erroneous claims. The relative isolation of these knowledge resources makes it hard, if not impossible, for researchers to connect the dots that exist in separate resources to gain new insights.

Let us consider drug repurposing as a case study[32]. Besides the long process of clinical

[32]This is a pre-clinical phase of biomedical research to discover new uses of existing, approved drugs that have already been tested in humans and so detailed information is available on their pharmacology, formulation, and potential toxicity.

trials and biomedical experiments, another major cause of the lengthy discovery phase is the complexity of the problem involved and the difficulty in drug discovery in general. The current clinical trials for drug repurposing rely mainly on reported symptoms in considering drugs that can treat diseases with similar symptoms. However, there are too many drug candidates and too much misinformation published in multiple sources. The clinicians and scientists thus urgently need assistance in obtaining a reliable ranked list of drugs with detailed evidence, and also in gaining new insights into the underlying molecular cellular mechanisms on COVID-19 and the pre-existing conditions that may affect the mortality and severity of this disease.

To tackle the above two challenges, we propose a new framework, COVID-KG, to accelerate scientific discovery and build a bridge between the research scientists making use of our framework and clinicians who will ultimately conduct the tests, as illustrated in Figure 7.2. COVID-KG starts by reading existing papers to build multimedia knowledge graphs (KGs), in which nodes are entities/concepts and edges represent relations and events involving these entities, as extracted from both text and images. Given the KGs enriched with path ranking and evidence mining, COVID-KG answers natural language questions effectively. With drug repurposing as a case study, we focus on eleven typical questions that human experts pose and integrate our techniques to generate a comprehensive report for each candidate drug. Preliminary assessments by expert clinicians and medical school students indicate that reports generated by our framework are both informative and sound.

7.2.3 Searching and Mining Literature for Chemical Reactions

Our fine-grained chemistry named entity recognition method (ChemNER [7]) and reactant entity classification method (ReactClass [15]) have been used in and inspired a follow-up work ReactionTracker: Searching and Mining Literature for Chemical Reactions. We expect this ReactionTracker system will significantly benefit the query-based tracking of scientific publications and the downstream scientific experiment design process.

Tracking the latest research on certain chemical reactions is a key challenge in mining and searching chemistry literature for chemical synthesis applications. For example, chemists may have a specific information need of finding papers about Suzuki coupling reactions involving two reactant groups: primary alkyl boronates and primary alkyl halides. They will formulate a query such as "Suzuki coupling between primary alkyl boronates and primary alkyl halides" and put it into a literature search engine. However, current literature search engines (e.g., PubMed and Reaxys) will have a low recall (i.e. missing some papers) on this kind of reaction queries. Because they do not know, for example, "n-octyl-9-BBN" is a

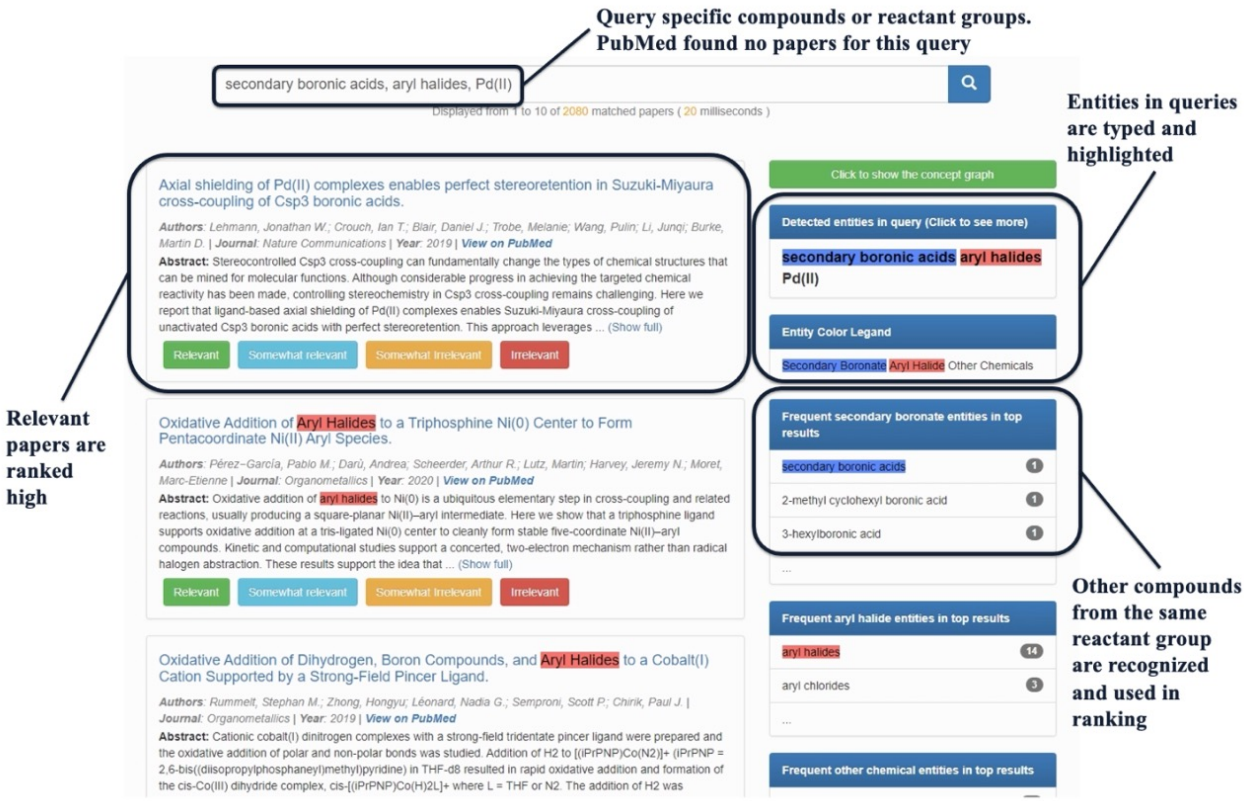


Figure 7.3: The demo interface of ReactionTracker. Papers are ranked based on their relevance to the query reactant groups.

"primary alkyl boronate" and "1-bromododecane" is a "primary alkyl halide".

To tackle the above challenge, we propose a new chemical literature search engine, ReactionTracker, that uses ChemNER and ReactClass for a smart query expansion to enhance the chemical reaction literature tracking performance. A demo interface of ReactionTracker is shown in Figure 7.3. If we input this query of reactant groups into PubMed, the most widely used scientific literature search engine, we will get no paper returned. Because PubMed search is based on keyword matching, while most scientific papers discuss concrete chemical compounds instead of their corresponding reactant groups. In ReactionTracker, we got 2,000 papers returned with the top results highly relevant to our input query. Because ReactionTracker first uses ChemNER to identify all the fine-grained chemistry entities in the background corpus and then uses ReactClass to automatically map those chemical entities to their corresponding reactant groups. As a result, we know what chemical compounds are relevant to the input query and it greatly improved the recall of the ReactionTracker system. In addition to the retrieved papers, ReactionTracker further highlights the query entities and other chemical compounds in the same reactant groups as the query entities for

better visualization. ReactionTracker is one key achievement of the National Science Foundation (NSF)-funded Molecular Maker Lab Institute at the University of Illinois at Urbana-Champaign. It has further been used in AI-driven systems for automatic chemical/material synthesis plan generation and optimization to support intelligent molecule discovery.

7.3 CONCLUSIONS

My research tackles a series of technical challenges for extracting a wide range of fine-grained information from unstructured text for scientific discovery. Our 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, UC Davis, UCSD, USC, Purdue, and Iowa State University). Our algorithms and systems can be generally used for any science domain where a knowledge discovery from massive text data is needed. Finally, our work has been used in the following settings:

- **Used in real world**:
  - **Clinical Domain**: Our text mining methods have been used to find proteins that are specifically associated with six main categories of heart diseases. Our top-ranked proteins match the knowledge of the clinical researchers very well. Some of our discovered proteins are currently under experimental validation by clinical researchers at the UC Davis Medical School. This collaboration has a high potential to unveil novel therapeutic targets in patients and repurpose drugs already used in the clinic.
  - **Chemistry Domain**: Our text mining methods have been used to support an intelligent molecule discovery process in organic chemistry. We have been collaborating with the researchers in the Chemistry Department at UIUC, finding the most representative catalysts and reaction conditions by comparing different organic reaction types. This collaboration leads to AI-driven systems for automatic chemical/material synthesis plan generation and optimization.
- **Taught in classes and conference tutorials**: Our methods on pattern-enhanced weakly-supervised NER (PeNNER), ontology-guided distantly-supervised NER (ChemNER), and cross-modal supervision between text and graph (ReactClass) are being taught in graduate courses, e.g., University of Illinois at Urbana-Champaign (CS 512), and are

introduced as major parts of the conference tutorial in top data mining and database conferences such as SIGKDD, WWW, and IEEE-BigData.

- **Awards**: This thesis work has been awarded YEE fellowship from 2020 to 2021 from the University of Illinois at Urbana-Champaign. It has also impacted an application on COVID-19 knowledge graph construction that has been awarded the Best Demo Paper Award in 2021 from NAACL.

CHAPTER 8: VISION AND FUTURE DIRECTIONS

There still remain grand challenges for scientific text mining, 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. In the future, I plan to develop knowledge-enhanced, condition-aware, and multi-modal text mining approaches for scientific discovery to tackle the above challenges. These future directions will involve collaborations 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.

8.1 KNOWLEDGE-ENHANCED SCIENTIFIC INFORMATION COMPREHENSION

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. I propose developing approaches combining deep learning and symbolic patterns to better understand the scientific text. Recently, deep learning-based approaches have led to state-of-art performance on various NLP and text mining tasks. 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, although symbolic pattern matching-based approaches are less accurate on standard test splits than deep learning-based approaches, 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.

Specifically, we can integrate domain-specific knowledge graphs with the representation learning in text in two ways: (1) *using domain-specific knowledge bases to guide the language model pre-training*, and (2) *using the pre-trained language models to enhance the knowledge graph (KG) reasoning*. For the first direction, we can formulate the structured knowledge (e.g., molecule graphs, reaction equations, and numerical properties) into textual sequences and incorporate this additional information into the domain-specific language model pre-training. We expect this additional knowledge will greatly enrich the knowledge of the domain-specific language models and benefit downstream tasks such as information extraction. For the second direction, we can harness the power of pre-trained language models

(PLMs) to add facts, definitions, and attribute information for open knowledge graph reasoning. The mission of open knowledge graph reasoning is to draw new findings from known facts. Existing works that augment such reasoning require either (1) factual triples to directly enrich the current KG or (2) manually crafting prompts to probe knowledge from a PLM, indicating limited performance and expensive expert knowledge respectively. Additionally, most of them only support single-hop reasoning, whereas multi-hop reasoning has a broader range of uses. We can automatically generate decent prompts and support information to fine-tune PLM for the task of multi-hop KG reasoning.

8.2 MULTI-MODAL SCIENTIFIC INFORMATION EXTRACTION

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 can also exist in multi-omics formats in some domains (e.g., genomics and proteomics data in the biology domain). A multi-modal information extraction will significantly benefit 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, bioinformatics, and natural sciences.

Specifically, we can develop cross-media semantic representation learning and information extraction approaches to support complex real-world applications such as multi-modal knowledge base curation and completion. One way to do the multi-modal knowledge integration is *multi-media representation learning*. We can develop methods for mapping knowledge elements from difference spaces of various modalities to the same continuous vector space with much lower dimensionality. Another way to do the multi-modal knowledge integration is *intermediate graph-based knowledge fusion*. We can first convert both text and images into intermediate graph structures (e.g., knowledge graphs for text and scene graphs for images) and then integrate these separate graphs into one comprehensive graph. This fused graph not only integrates knowledge from different data modalities but also facilitates downstream tasks such as knowledge graph reasoning and graph-based text generation. Compared with directly fusing multi-modal information with a multi-media representation learning, we expect this intermediate graph-based knowledge fusion could possibly provide more explainability for downstream task predictions.

8.3 MULTI-DIMENSIONAL SCIENTIFIC INFORMATION ANALYSIS

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 for a disease with a certain dosage or for certain patient groups (e.g., age, gender, or comorbidity with other diseases). I propose to consider different conditions as different dimensions, organizing massive text into multi-dimensional text cube structure, and synthesizing knowledge in the multi-dimensional space.

Specifically, we can develope *multi-dimensional information extraction approaches* to extract entities, relationships, and knowledge graphs by considering user-specified conditions or dimensions. This multi-dimensional information extraction and knowledge organization facilitates complex real-world applications, such as distinctive summarization of entities, relationships, and networks under each dimension and knowledge discovery through cross-dimensional comparison and inference. For example, we can develop a multi-dimensional-cube-based document organization method that benefits downstream tasks such as comparative document summarization and analysis. Massive documents can be organized into multi-dimensional text cubes to facilitate downstream tasks such as search and summarization. For example, the COVID-19 literature can be organized in a three-dimensional text cube of "Virus Type", "Study of Virus", and "Age Group". Each dimension may contain several categories for comparison (e.g., the "Virus Type" dimension may contain categories "COVID-19", "SARS", "MERS", and "Ebola"). The task of multi-dimensional cube-based document search aims to automatically retrieve the most relevant documents (sentences or paragraphs) for each cell in the text cube. We can use category-indicative concept discovery for an explainable cube-based document search.

REFERENCES

[1] 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, vol. 571, no. 7763, pp. 95–98, 2019.

[2] X. Wang, Y. Zhang, Q. Li, C. H. Wu, and J. Han, "Penner: Pattern-enhanced nested named entity recognition in biomedical literature," in Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2018, pp. 540–547.

[3] 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, vol. 35, no. 10, pp. 1745–1752, 2019.

[4] X. Wang, Y. Zhang, Q. Li, X. Ren, J. Shang, and J. Han, "Distantly supervised biomedical named entity recognition with dictionary expansion," in Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2019, pp. 496–503.

[5] X. Wang, Y. Guan, Y. Zhang, Q. Li, and J. Han, "Pattern-enhanced named entity recognition with distant supervision," in Proceedings of the 2020 IEEE International Conference on Big Data. IEEE, 2020, pp. 818–827.

[6] 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 Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2020, pp. 491–494.

[7] 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 Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 5227–5240.

[8] C.-H. Wei, H.-Y. Kao, and Z. Lu, "Pubtator: a web-based text mining tool for assisting biocuration," Nucleic Acids Research, vol. 41, no. W1, pp. W518–W522, 2013.

[9] Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, R. H. Zhang, W. Liu et al., "Covid-19 literature knowledge graph construction and drug repurposing report generation," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, 2021, pp. 66–77.

[10] X. Wang, Y. Zhang, Q. Li, and J. Han, "Taming unstructured big data: automated information extraction from massive text."

[11] X. Wang, H. Wang, H. Ji, and J. Han, "Modern natural language processing techniques for scientific web mining: tasks, data, and tools," in Proceedings of the ACM Web Conference 2022, 2022.

[12] X. Wang, H. Wang, H. Ji, and J. Han, "New frontiers of scientific text mining: tasks, data, and tools," in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 4832–4833.

[13] 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.

[14] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," arXiv preprint arXiv:2010.09885, 2020.

[15] X. Wang, V. Hu, M. Jiang, Y. Zhang, J. Xiao, D. C. Loving, H. Ji, M. Burke, and J. Han, "Reactclass: Cross-modal supervision for subword-guided reactant entity classification," in Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2022.

[16] X. Wang, Y. Guan, W. Liu, A. Chauhan, E. Jiang, Q. Li, D. Liem, D. Sigdel, J. Caufield, P. Ping et al., "Evidenceminer: Textual evidence discovery for life sciences," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 56–62.

[17] X. Wang, Y. Zhang, A. Chauhan, Q. Li, and J. Han, "Textual evidence mining via spherical heterogeneous information network embedding," in Proceedings of the 2020 IEEE International Conference on Big Data. IEEE, 2020, pp. 828–837.

[18] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney et al., "Cord-19: The covid-19 open research dataset," in Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, 2020.

[19] A. Allot, Q. Chen, S. Kim, R. Vera Alvarez, D. C. Comeau, W. J. Wilbur, and Z. Lu, "Litsense: Making sense of biomedical literature at sentence level," Nucleic Acids Research, vol. 47, no. W1, pp. W594–W599, 2019.

[20] 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, vol. 36, no. 4, pp. 1234–1240, 2020.

[21] D. A. Liem, S. Murali, D. Sigdel, Y. Shi, X. Wang, J. Shen, H. Choi, J. H. Caufield, W. Wang, P. Ping et al., "Phrase mining of textual data to analyze extracellular matrix protein patterns across cardiovascular disease," American Journal of Physiology-Heart and Circulatory Physiology, vol. 315, no. 4, pp. H910–H924, 2018.

[22] M. Cokol, I. Iossifov, C. Weinreb, and A. Rzhetsky, "Emergent behavior of growing knowledge about molecular interactions," Nature Biotechnology, vol. 23, no. 10, pp. 1243–1247, 2005.

[23] Z. Li, Z. Yang, H. Lin, J. Wang, Y. Gui, Y. Zhang, and L. Wang, "Cidextractor: A chemical-induced disease relation extraction system for biomedical literature," in Proceeding of the 2016 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2016, pp. 994–1001.

[24] B. Xie, Q. Ding, H. Han, and D. Wu, "mircancer: a microrna–cancer association database constructed by text mining on literature," Bioinformatics, vol. 29, no. 5, pp. 638–644, 2013.

[25] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, "Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data," Nucleic Acids Research, vol. 44, no. D1, pp. D380–D384, 2015.

[26] J. Huang, F. Gutierrez, D. Dou, J. A. Blake, K. Eilbeck, D. A. Natale, B. Smith, Y. Lin, X. Wang, Z. Liu et al., "A semantic approach for knowledge capture of microrna-target gene interactions," in Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2015, pp. 975–982.

[27] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, and P. Bork, "The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible," Nucleic Acids Research, vol. 45, no. D1, pp. D362–D368, 2017.

[28] R. Leaman and Z. Lu, "Taggerone: joint named entity recognition and normalization with semi-markov models," Bioinformatics, vol. 32, no. 18, pp. 2839–2846, 2016.

[29] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, and A. Valencia, "Chemdner: The drugs and chemical names extraction challenge," Journal of Cheminformatics, vol. 7, no. 1, p. S1, 2015.

[30] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," Bioinformatics, vol. 33, no. 14, pp. i37–i48, 2017.

[31] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, "A neural network multi-task learning approach to biomedical named entity recognition," BMC Bioinformatics, vol. 18, no. 1, p. 368, 2017.

[32] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, "Developing a robust part-of-speech tagger for biomedical text," in Proceedings of the Panhellenic Conference on Informatics. Springer, 2005, pp. 382–392.

[33] B. Alex, B. Haddow, and C. Grover, "Recognising nested named entities in biomedical text," in Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. ACL, 2007, pp. 65–72.

[34] J. R. Finkel and C. D. Manning, "Nested named entity recognition," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, pp. 141–150.

[35] W. Lu and D. Roth, "Joint mention extraction and classification with mention hypergraphs," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 857–867.

[36] A. O. Muis and W. Lu, "Labeling gaps between words: recognizing overlapping mentions with mention separators," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2608–2618.

[37] A. Katiyar and C. Cardie, "Nested named entity recognition revisited," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 861–871.

[38] M. Ju, M. Miwa, and S. Ananiadou, "A neural layered model for nested named entity recognition," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1446–1459.

[39] 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 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 877–886.

[40] J. Pei, J. Han, and W. Wang, "Constraint-based sequential pattern mining: the pattern-growth methods," Journal of Intelligent Information Systems, vol. 28, no. 2, pp. 133–160, 2007.

[41] Q. Li, M. Jiang, X. Zhang, M. Qu, T. P. Hanratty, J. Gao, and J. Han, "Truepie: Discovering reliable patterns in pattern-based information extraction," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018, pp. 1675–1684.

[42] X. Wang, Y. Zhang, Q. Li, Y. Chen, and J. Han, "Open information extraction with meta-pattern discovery in biomedical literature," in Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2018, pp. 291–300.

[43] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 2003, pp. 423–430.

[44] J. Shen, Z. Wu, D. Lei, J. Shang, X. Ren, and J. Han, "Setexpan: Corpus-based set expansion via context feature selection and rank ensemble," in Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2017, pp. 288–304.

[45] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the Advances in Neural Information Processing Systems. MIT Press, 2013, pp. 3111–3119.

[46] X. Rong, Z. Chen, Q. Mei, and E. Adar, "Egoset: Exploiting word ego-networks and user-generated ontology for multifaceted set expansion," in Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, 2016, pp. 645–654.

[47] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics database: update 2017," Nucleic Acids Research, vol. 45, no. D1, pp. D972–D978, 2016.

[48] H. Schütze, C. D. Manning, and P. Raghavan, Introduction to information retrieval. Cambridge University Press, 2008.

[49] G. O. Consortium, "The gene ontology (go) database and informatics resource," Nucleic Acids Research, vol. 32, no. suppl_1, pp. D258–D261, 2004.

[50] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/va challenge on concepts, assertions, and relations in clinical text," Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 552–556, 2011.

[51] A. A. Mahmood, S. Rao, P. McGarvey, C. Wu, S. Madhavan, and K. Vijay-Shanker, "egard: Extracting associations between genomic anomalies and drug responses from text," PloS One, vol. 12, no. 12, p. e0189663, 2017.

[52] Q. Wang, K. E. Ross, H. Huang, J. Ren, G. Li, K. Vijay-Shanker, C. H. Wu, and C. N. Arighi, "Analysis of protein phosphorylation and its functional impact on protein–protein interactions via text mining of the scientific literature," Protein Bioinformatics, pp. 213–232, 2017.

[53] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 473–480.

[54] R. McDonald and F. Pereira, "Identifying gene and protein mentions in text using conditional random fields," BMC Bioinformatics, vol. 6, no. 1, p. S6, 2005.

[55] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 260–270.

[56] A. F. de Almeida, R. Moreira, and T. Rodrigues, "Synthetic organic chemistry driven by artificial intelligence," Nature Reviews Chemistry, vol. 3, no. 10, pp. 589–604, 2019.

[57] J. He, D. Q. Nguyen, S. A. Akhondi, C. Druckenbrodt, C. Thorne, R. Hoessel, Z. Afzal, Z. Zhai, B. Fang, H. Yoshikawa et al., "Overview of chemu 2020: named entity recognition and event extraction of chemical reactions from patents," in International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, 2020, pp. 237–254.

[58] T. Watanabe, A. Tamura, T. Ninomiya, T. Makino, and T. Iwakura, "Multi-task learning for chemical named entity recognition with chemical compound paraphrasing," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6244–6249.

[59] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional list-cans," Transactions of the Association for Computational Linguistics, vol. 4, pp. 357–370, 2016.

[60] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1064–1074.

[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[62] J. Shang, L. Liu, X. Gu, X. Ren, T. Ren, and J. Han, "Learning named entity tagger using domain-specific dictionary," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2054–2064.

[63] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, and C. Zhang, "BOND: bert-assisted open-domain named entity recognition with distant supervision," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2020, pp. 1054–1064.

[64] M. C. Swain and J. M. Cole, "Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature," Journal of Chemical Information and Modeling, vol. 56, no. 10, pp. 1894–1904, 2016.

[65] Y. Tsuruoka and J. Tsujii, "Bidirectional inference with the easiest-first strategy for tagging sequence data," in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 467–474.

[66] A. M. Azman, "A chemistry spell-check dictionary for word processors," in Journal of Chemical Education.

[67] M. Peng, X. Xing, Q. Zhang, J. Fu, and X. Huang, "Distantly supervised named entity recognition using positive-unlabeled learning," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2409–2419.

[68] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han, "Automated phrase mining from massive text corpora," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 10, pp. 1825–1837, 2018.

[69] M. Francis-Landau, G. Durrett, and D. Klein, "Capturing semantic similarity for entity linking with convolutional neural networks," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1256–1261.

[70] N. Gupta, S. Singh, and D. Roth, "Entity linking via joint encoding of types, descriptions, and context," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2681–2690.

[71] J. Raiman and O. Raiman, "Deeptype: Multilingual entity linking by neural type system evolution," in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, the 30th innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, 2018, pp. 5406–5413.

[72] P. Le and I. Titov, "Improving entity linking by modeling latent relations between mentions," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 1595–1604.

[73] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 782–792.

[74] M. A. Yosef, S. Bauer, J. Hoffart, M. Spaniol, and G. Weikum, "HYENA: Hierarchical type classification for entity names," in Proceedings of COLING 2012: Posters, 2012, pp. 1361–1370.

[75] X. Ling and D. S. Weld, "Fine-grained entity recognition," in Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

[76] L. Del Corro, A. Abujabal, R. Gemulla, and G. Weikum, "FINET: Context-aware fine-grained named entity typing," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 868–878.

[77] 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 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 995–1004.

[78] E. Choi, O. Levy, Y. Choi, and L. Zettlemoyer, "Ultra-fine entity typing," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 87–96.

[79] P. Lison, J. Barnes, A. Hubin, and S. Touileb, "Named entity recognition without labelled data: a weak supervision approach," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[80] 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 Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 10 367–10 378.

[81] J. Li, H. Ding, J. Shang, J. McAuley, and Z. Feng, "Weakly supervised named entity tagging with learnable logical rules," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 4568–4581.

[82] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2545–2568.

[83] J. Parker and S. Yu, "Named entity recognition through deep representation learning and weak supervision," in Proceedings of the Findings of the Association for Computational Linguistics, 2021, pp. 3828–3839.

[84] E. K. Mallory, M. de Rochemonteix, A. Ratner, A. Acharya, C. Re, R. A. Bright, and R. B. Altman, "Extracting chemical reactions from text using snorkel," BMC Bioinformatics, vol. 21, pp. 1–15, 2020.

[85] M. Krallinger, O. Rabal, A. Lourenco, J. Oyarzabal, and A. Valencia, "Information retrieval and text mining technologies for chemistry," Chemical Reviews, vol. 117, no. 12, pp. 7673–7761, 2017.

[86] Y. Zhou, B. Zhou, S. Jiang, and F. J. King, "Chemical-text hybrid search engines," Journal of Chemical Information and Modeling, vol. 50, no. 1, pp. 47–54, 2010.

[87] C. Edwards, C. Zhai, and H. Ji, "Text2mol: Cross-modal molecule retrieval with natural language queries," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 595–607.

[88] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, "Prediction of organic reaction outcomes using machine learning," ACS Central Science, vol. 3, no. 5, pp. 434–443, 2017.

[89] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle, "Predicting reaction performance in c–n cross-coupling using machine learning," Science, vol. 360, no. 6385, pp. 186–190, 2018.

[90] D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken, and P. Baldi, "Deep learning for chemical reaction prediction," Molecular Systems Design & Engineering, vol. 3, no. 3, pp. 442–452, 2018.

[91] H. Wang, W. Li, X. Jin, K. Cho, H. Ji, J. Han, and M. Burke, "Chemical-reaction-aware molecule representation learning," in Proceedings of the International Conference on Learning Representations, 2022.

[92] J. Goodman, "Computer software review: Reaxys," 2009.

[93] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, "PubChem in 2021: new data content and improved web interfaces," Nucleic Acids Research, vol. 49, no. D1, pp. D1388–D1395, 2020.

[94] D. M. Lowe, P. T. Corbett, P. Murray-Rust, and R. C. Glen, "Chemical name to structure: Opsin, an open source solution," 2011.

[95] J. Guo, A. S. Ibanez-Lopez, H. Gao, V. Quach, C. W. Coley, K. F. Jensen, and R. Barzilay, "Automated chemical reaction extraction from scientific literature," Journal of Chemical Information and Modeling, 2021.

[96] L. Gong, D. He, Z. Li, T. Qin, L. Wang, and T. Liu, "Efficient training of bert by progressively stacking," in Proceedings of the International Conference on Machine Learning. PMLR, 2019, pp. 2337–2346.

[97] X. Gu, Z. Wang, Z. Bi, Y. Meng, L. Liu, J. Han, and J. Shang, "Ucphrase: Unsupervised context-aware quality phrase tagging," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021, pp. 478–486.

[98] S. Rensi and R. B. Altman, "Flexible analog search with kernel pca embedded molecule vectors," Computational and Structural Biotechnology Journal, vol. 15, pp. 320–327, 2017.

[99] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, and F. Glorius, "A structure-based platform for predicting chemical reactivity," Chem, vol. 6, no. 6, pp. 1379–1390, 2020.

[100] S. Jaeger, S. Fulle, and S. Turk, "Mol2vec: unsupervised machine learning approach with chemical intuition," Journal of Chemical Information and Modeling, vol. 58, no. 1, pp. 27–35, 2018.

[101] B. Fabian, T. Edlich, H. Gaspar, M. Segler, J. Meyers, M. Fiscato, and M. Ahmed, "Molecular representation learning with language models and domain-relevant auxiliary tasks," arXiv preprint arXiv:2011.13230, 2020.

[102] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in Proceedings of the European Conference on Computer Vision. Springer, 2016, pp. 285–301.

[103] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2020, pp. 6829–6833.

[104] S.-W. Chung, H.-G. Kang, and J. S. Chung, "Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision," Proceedings of the Interspeech 2020, pp. 3486–3490, 2020.

[105] N. Sankaran, D. D. Mohan, S. Setlur, V. Govindaraju, and D. Fedorishin, "Representation learning through cross-modality supervision," in Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2019, pp. 1–8.

[106] A. Nagrani, C. Sun, D. Ross, R. Sukthankar, C. Schmid, and A. Zisserman, "Speech2action: Cross-modal supervision for action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10 317–10 326.

[107] C. Sheng, M. Pietikäinen, Q. Tian, and L. Liu, "Cross-modal self-supervised learning for lip reading: when contrastive learning meets adversarial training," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2456–2464.

[108] M. Zambelli, Y. Aytar, F. Visin, Y. Zhou, and R. Hadsell, "Learning rich touch representations through cross-modal self-supervision," in Proceedings of the Conference on Robot Learning. PMLR, 2021, pp. 1415–1425.

[109] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu, "Rodnet: Radar object detection using cross-modal supervision," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 504–513.

[110] M. Li, A. Zareian, Q. Zeng, S. Whitehead, D. Lu, H. Ji, and S.-F. Chang, "Cross-media structured common space for multimedia event extraction," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2557–2568.

[111] T. Linzen, G. Chrupała, Y. Belinkov, and D. Hupkes, "Proceedings of the 2019 acl workshop blackboxnlp: Analyzing and interpreting neural networks for nlp," in Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2019.

[112] T. Kim, J. Choi, D. Edmiston, and S.-g. Lee, "Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction," in Proceedings of the International Conference on Learning Representations, 2019.

[113] Z. Lu, "Pubmed and beyond: a survey of web tools for searching biomedical literature," Database, vol. 2011, 2011.

[114] 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," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2017, pp. 55–60.

[115] J. Shen, J. Xiao, X. He, J. Shang, S. Sinha, and J. Han, "Entity set search of scientific literature: An unsupervised ranking approach," in Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 565–574.

[116] M. Lippi and P. Torroni, "Margot: A web server for argumentation mining," Expert Systems with Applications, vol. 65, pp. 292–303, 2016.

[117] H. Wachsmuth, M. Potthast, K. Al Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, and B. Stein, "Building an argument search engine for the web," in Proceedings of the 4th Workshop on Argument Mining, 2017, pp. 49–59.

[118] C. Stab, J. Daxenberger, C. Stahlhut, T. Miller, B. Schiller, C. Tauchmann, S. Eger, and I. Gurevych, "Argumentext: Searching for arguments in heterogeneous sources," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2018, pp. 21–25.

[119] S. Chen, D. Khashabi, C. Callison-Burch, and D. Roth, "Perspectroscope: A window to the world of diverse perspectives," p. 129–134, 2019.

[120] S. Majithia, F. Arslan, S. Lubal, D. Jimenez, P. Arora, J. Caraballo, and C. Li, "Claimportal: Integrated monitoring, searching, checking, and analytics of factual claims on twitter," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019, pp. 153–158.

[121] A. Chernodub, O. Oliynyk, P. Heidenreich, A. Bondarenko, M. Hagen, C. Biemann, and A. Panchenko, "Targer: Neural argument mining at your fingertips," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019, pp. 195–200.

[122] Q. Li, X. Wang, Y. Zhang, Q. Li, F. Ling, C. Wu H, and J. Han, "Pattern discovery for wide-window open information extraction in biomedical literature," in Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine, 2018, pp. 420–427.

[123] S. Robertson, H. Zaragoza et al., "The probabilistic relevance framework: Bm25 and beyond," Foundations and Trends® in Information Retrieval.

[124] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," PLoS Biology, vol. 2, no. 11, p. e309, 2004.

[125] R. Levy, S. Gretz, B. Sznajder, S. Hummel, R. Aharonov, and N. Slonim, "Unsupervised corpus–wide claim detection," in Proceedings of the 4th Workshop on Argument Mining, 2017, pp. 79–84.

[126] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia-a crystallization point for the web of data," Journal of Web Semantics, vol. 7, no. 3, pp. 154–165, 2009.

[127] R. Navigli and S. P. Ponzetto, "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," Artificial Intelligence, vol. 193, pp. 217–250, 2012.

[128] T. Rebele, F. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum, "Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames," in Proceedings of the International Semantic Web Conference, 2016, pp. 177–185.

[129] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 243–246.

[130] M. Li, A. Zareian, Y. Lin, X. Pan, S. Whitehead, B. Chen, B. Wu, H. Ji, S.-F. Chang, C. Voss et al., "Gaia: A fine-grained multimedia knowledge extraction system," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 77–86.

[131] F. Tao, H. Zhuang, C. W. Yu, Q. Wang, T. Cassidy, L. M. Kaplan, C. R. Voss, and J. Han, "Multi-dimensional, phrase-based summarization in text cubes," Data Engineering, vol. 39, no. 3, pp. 74–84, 2016.

[132] A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing & Management, vol. 39, no. 1, pp. 45–65, 2003.

[133] A. Simitsis, A. Baid, Y. Sismanis, and B. Reinwald, "Multidimensional content exploration," Proceedings of the VLDB Endowment, vol. 1, no. 1, pp. 660–671, 2008.

[134] K. B. Kuchenbaecker, J. L. Hopper, D. R. Barnes, K.-A. Phillips, T. M. Mooij, M.-J. Roos-Blom, S. Jervis, F. E. Van Leeuwen, R. L. Milne, N. Andrieu et al., "Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers," JAMA, vol. 317, no. 23, pp. 2402–2416, 2017.

[135] C. Zhai, "Statistical language models for information retrieval," Synthesis Lectures on Human Language Technologies, vol. 1, no. 1, pp. 1–141, 2008.