Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature

Qi Li
Univ. of Illinois at Urbana-Champaign
Urbana, IL, USA
E-mail: qili5@illinois.edu

Xuan Wang
Univ. of Illinois at Urbana-Champaign
Urbana, IL, USA
E-mail: xwang174@illinois.edu

Yu Zhang
Univ. of Illinois at Urbana-Champaign
Urbana, IL, USA
E-mail: yuz9@illinois.edu

Fei Ling
Univ. of Illinois at Urbana-Champaign
Urbana, IL, USA
E-mail: fling2@illinois.edu

Cathy H. Wu
Univ. of Delaware
Newark, DE, USA
E-mail: wuc@udel.edu

Jiawei Han
Univ. of Illinois at Urbana-Champaign
Urbana, IL, USA
E-mail: hanj@illinois.edu

Abstract—**Open information extraction is an important task in Biomedical domain. The goal of the OpenIE is to automatically extract structured information from unstructured text with no or little supervision. It aims to extract all the relation tuples from the corpus without requiring pre-specified relation types. The existing tools may extract ill-structured or incomplete information, or fail on the Biomedical literature due to the long and complicated sentences. In this paper, we propose a novel pattern-based information extraction method for the wide-window entities (WW-PIE). WW-PIE utilizes dependency parsing to break down the long sentences first and then utilizes frequent textual patterns to extract the high-quality information. The pattern hierarchical grouping organize and structure the extractions to be straightforward and precise. Consequently, comparing with the existing OpenIE tools, WW-PIE produces structured output that can be directly used for downstream applications. The proposed WW-PIE is also capable in extracting n-ary and nested relation structures, which is less studied in the existing methods. Extensive experiments on real-world biomedical corpus from PubMed abstracts demonstrate the power of WW-PIE at extracting precise and well-structured information.**

I. Introduction

Information extraction (IE) aims to extract structured information from unstructured text. It is an important task with many application. The biomedical domain is especially in huge demand of automatic IE systems, as it is too costly for manual curation to keep up with the rapid growth of the literature. Many BioIE (Biomedical Information Extraction) systems adopt various machine learning models to conduct specific IE tasks. For example, in the recent BioCreative V challenge [1], various machine learning models are applied to conduct a chemical-induced disease (CID) relation extraction task [2]–[4], where a ranked list of $\langle chemical, disease \rangle$ with "induced" relation is expected as output. Such systems are very helpful to enrich the knowledge in bioinformatics databases, such as the Comparative Toxicogenomics Database (CTD) [5], and can greatly assist downstream applications such as identifying potential toxicity.

However, there are several challenges that the aforementioned BioIE system cannot solve. First, the extractions are not precise and do not provide conditions. For example, a certain chemical only induces a disease if it is a long-term high-dose exposure. It is imprecise to simply claim that this chemical induces a disease. Second, these machine learning based models still rely on time and labor consuming human annotation for training data generation. Obtaining sufficient training data can be costly. Third, the systems can only work on pre-specified tasks and cannot be extended to new ones.

To address the above challenges, in this paper, we focus on semi-supervised Bio-OpenIE (Biomedical Open Information Extraction) tasks, which do not pre-specify relation types but aims to extract all the relation tuples from a large biomedical literature corpus with little human effort. Pattern-based methods are widely used for semi-supervised OpenIE in the general domain. These methods discover and organize frequent textual patterns with typed entities to extract information from large corpora [6]–[8]. For example, meta pattern "COUNTRY president PERSON" can be used to extract "president" relation for entities, where COUNTRY can be "USA", "Russia", etc, and PERSON can be "Trump", "Putin", etc. Though the pattern-based methods can often achieve good precision, they may suffer from low recall in biomedical literature due to the length of the sentence and their complicated structure used in the literature. For exmaple, "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 two chemicals, and "cyclin_D1" and "E2F-1" are genes. The existing methods can discover the meta pattern "GENE and GENE" and "CHEMICAL can decrease the CHEMICAL", but not "Pre-treatment of CHEMICAL can decrease the overexpression of GENE and GENE induced by CHEMICAL", because it is too long and infrequent.

Therefore, we propose a novel wide-window pattern-based IE method for biomedical literature, called WW-PIE. There are three challenges that we need to address: (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 pattern mining. After discovering high quality meta patterns, we group patterns hierarchically to better understand and organize the patterns. The extractions will be presented in two formats: the tuple format and the expression format. Using the above sentence as an example, the tuple format of the extraction will be $\langle$ATRA, decrease, cyclin_D1:$\langle$(cyclin_D1, E2F-1), induced by, B(a)P$\rangle\rangle$, and the expression format will be "pre-treatment of CHEMICAL:ATRA can decrease the over-expression of GENE:cyclin_D1 and GENE:E2F-1 induced by CHEMICAL:B(a)P ".

In summary, we make the following contributions in this paper:

- We identify the pitfall and challenge of the existing OpenIE methods in biomedical literature: long distanced entities and the need for structured n-ary and hierarchical relation types.
- We formulate a meta-pattern-based approach that utilizes dependency parsing to resolve the complex sentence structures and utilizes frequency pattern mining to discover high-quality extractions.
- We propose a novel hierarchical pattern grouping to better organize the extractions, keeping both simplicity and the structure of the relationships.

This paper is organized as follows. In the next section, we briefly introduce some related work of Bio-OpenIE. In Section III, we introduce the overall framework of WW-PIE. Then we introduce the detailed methods for meta-pattern extraction in Section IV. Some quantitative and qualitative experiments are introduced in Section V, followed by discussion and conclusions in Sections VI and VII, respectively.

II. RELATED WORK

A. Open-Domain Information Extraction.

OpenIE aims to extract tuples with any types of relations from a corpus with no or little human supervision. The task has been studied in the NLP area for a decade since the first introduction by Banko et al. [9]. The current OpenIE work mainly follows two lines: clause-based methods and pattern-based methods.

In clause-based methods, linguistic features, such as dependency parsing and POS tagging, are used to discover wide-window relationships. For example, ReVerb [10] and Ollie [11] identify relational phrases via part-of-speech-based regular expressions. ClausIE [12] uses grammar rules to rewrite sentences into standardized formats, such as SVO (Subject-verb-object), SVA (Subject-verb-adverbials), and SVC (Subject-verb-complement), while Stanford OpenIE [13] splits clauses using distant training. Generally speaking, the clause-based methods have high recall but many of the extractions are redundant or uninformative. To handle this issue, Stanford OpenIE, MinIE [14] and ReMine [15] uses different statistical measures to reduce the amount of uninformative or incoherent extractions.

Pattern-based methods can be traced back to Hearst patterns [16], which are still widely used for hyponymy relation extraction. Some pattern-based methods also try to discover meta-patterns (i.e., relational patterns between entity types). Patty [7] and HighLife [17] extract typed lexical patterns along the shortest dependency path, where Patty focuses on binary relation types and HighLife focuses on n-ary relation types. MetaPAD [6] finds candidate meta-patterns (i.e., relational patterns between entity types) through sequential frequent pattern mining and then uses a trained classifier to identify the high quality ones. TruePIE [8] adopts truth discovery ideas to improve the meta-pattern synonymy grouping for specific relation extraction tasks. Generally speaking, when using these meta-patterns to match the corpus, the extractions usually have higher precision, but the meta-patterns can be sparse if the corpus is full of long sentences.

B. Biomedical Information Extraction.

BioIE includes many NLP subtasks on biomedical literature such as NER (named entity recognition), relation extraction, and event extraction [18]–[27]. The task that is the most related to ours is biomedical relation extraction.

The mainstream is to use supervised methods, which rely on annotated corpora to discover certain relation types between entities. For example, Bundchus et al. [28] apply conditional random fields to detect relations between diseases and treatments from PubMed abstracts. Rink et al. [29] adopt support vector machine to classify relations between medical records and treatments. Percha et al. [30] propose a random forest classifier to predict drug-drug interactions. Besides classification approaches, rule-based and pattern-based methods are also studied. For example, Hackenberg et al. [31] extract syntactical patterns from labeled examples and use them to detect interactions between genes and proteins.

OpenIE begins to attract great attention in BioIE domain [32]. Bio-OpenIE aims to extract tuples with any relation types. One example is the seminal work by Kulick et al. [33] that extracts information on drug development and cancer genomics. Nebot et al. [34], [35] utilize a set of Lexico-syntactic patterns (LSP) to first extract binary relationship instances and then infer the semantic types of the extracted relationships. De Silva et al. [36] adapt Ollie to discover inconsistencies in PubMed abstracts through ontology-based information extraction.

III. OVERVIEW

The overall framework of WW-PIE is shown in Figure 1. The inputs are sentences in the biomedical corpus. We firstly conduct the biomedical named entity recognition (BioNER), which recognizes the biomedical entities and provides their entity types, e.g., gene, chemical, and disease. The typed entity mentions are then replaced by their types in the following process. Then each sentence is broken into shorter sentences and segments based on dependency parsing (see Section IV-B). These short segments and sentences are the candidates for sequential frequent pattern mining. Next, we extract the high-quality meta patterns and these patterns are further grouped

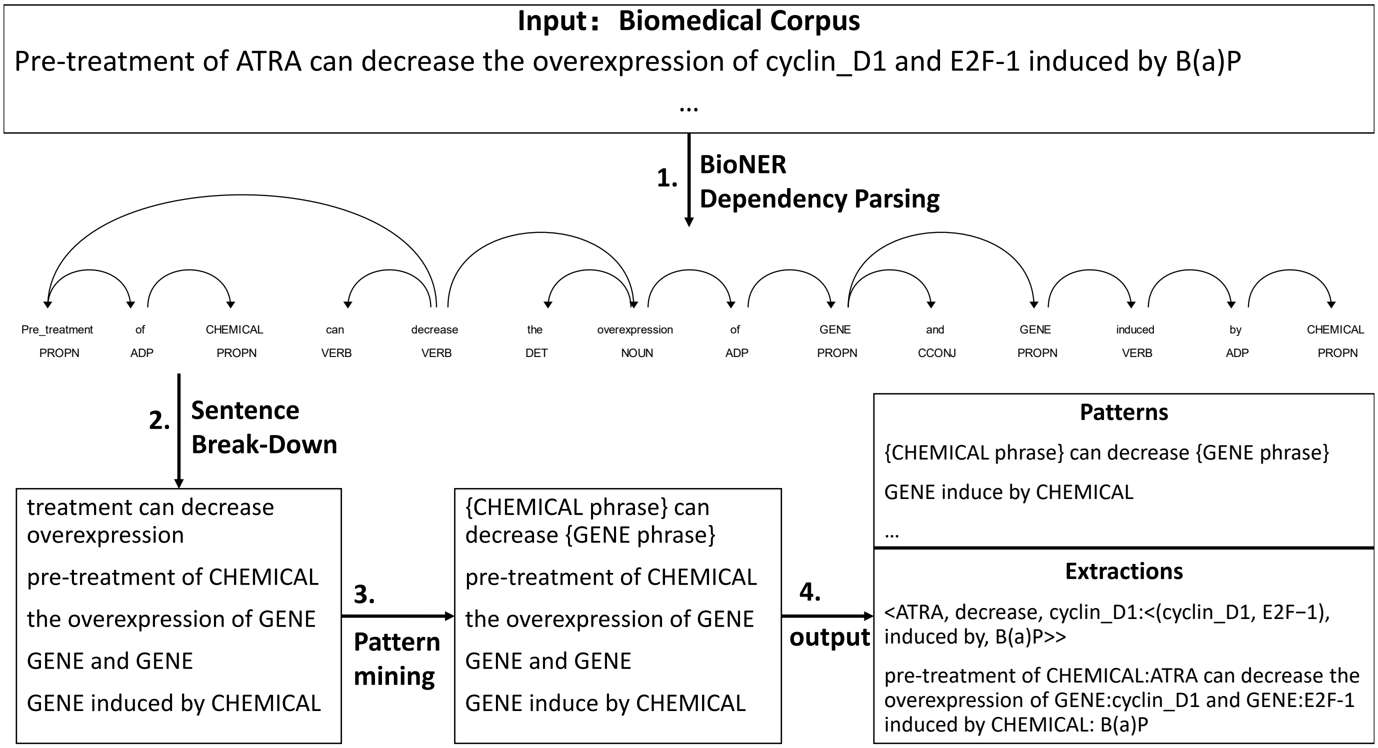

Input：Biomedical Corpus
Pre-treatment of ATRA can decrease the overexpression of cyclin_D1 and E2F-1 induced by B(a)P
...
1.
BioNER
Dependency Parsing
Pre_treatment PROPN
of ADP
CHEMICAL PROPN
can VERB
decrease VERB
the DET
overexpression NOUN
of ADP
GENE PROPN
and CCONJ
GENE PROPN
induced VERB
by ADP
CHEMICAL PROPN
2.
Sentence Break-Down
treatment can decrease overexpression
pre-treatment of CHEMICAL
the overexpression of GENE
GENE and GENE
GENE induced by CHEMICAL
3.
Pattern mining
{CHEMICAL phrase} can decrease {GENE phrase}
pre-treatment of CHEMICAL
the overexpression of GENE
GENE and GENE
GENE induce by CHEMICAL
4.
output
Patterns
{CHEMICAL phrase} can decrease {GENE phrase}
GENE induce by CHEMICAL
...
Extractions
<ATRA, decrease, cyclin_D1:<(cyclin_D1, E2F−1), induced by, B(a)P>>
pre-treatment of CHEMICAL:ATRA can decrease the overexpression of GENE:cyclin_D1 and GENE:E2F-1 induced by CHEMICAL: B(a)P

Fig. 1: The overall framework of WW-PIE: Wide-Window Pattern-based Information Extraction.

hierarchically (see Section IV-C). Finally, the quality meta-patterns are used to extract corresponding relationship instances from the original input sentences, and form both the tuple format and the expression format extractions. The extraction and the quality meta-patterns are the output of our framework.

IV. Meta Pattern Extraction

A. Data Preprocessing

The data preprocessing contains two steps: named entity recognition and sentence dependency parsing. We use Pubtator [37], a state-of-the-art BioNER tool, to first generate a typed corpus. Pubtabtor recognizes five biomedical entity types including gene/protein, chemical, disease, species, and SNP. Pubtator also provides a fully annotated version of the PubMed abstracts[1]. After the BioNER step, we replace the recognized entity mentions by their types, and then use a dependency parser to parse each sentence. The reason of applying BioNER first is that the named entities in the biomedical literature may consist of one or more words and may be out of vocabulary of the dependency parser, and thus causes mistakes in parsing. In this paper, we use spaCy[2], a python NLP library, for dependency parsing. The parsing result has a tree structure, which indicates a set of directed syntactic relations between the words in the sentence (See example in Figure 1).

After the data preprocessing, each sentence s is a sequence of word tokens t, i.e., $s = t_1 t_2 ... t_m$, in which $t_j \in \mathcal{T} \cup \mathcal{W}$ ($\mathcal{T}$ is the set of entity types and $\mathcal{W}$ is the set of non-type words). The dependency parsing outputs a tree structure of tokens $Tree(s)$, and each token's POS (part-of-speech) tag P_{t_j}

B. Sentence Break-Down

As mentioned earlier, sentences in biomedical literature can be long and tokens between entities of interest can be lengthy. Such issues can cause sparsity and incompleteness of the patterns. Therefore, it is important to break down the long and complicated sentences before conducting pattern mining.

In order to break down the sentences, we need to understand the sentence structure and English grammar. In linguistics, words can be categorized as content words and function words. Content words, including nouns, most verbs, and adjectives, are those which have statable lexical meanings, referring to some objects, actions, or characteristic. Function words, on the other hand, are used for grammatical purposes. We observe that in biomedical literature, the complexity of the sentences is mainly due to the complexity in noun structures, where a noun can be modified by other nouns, adjectives, adjectival clauses, etc. Using the sentence in Figure 1 as an example. In this sentence, the object is "the overexpression of Gene and Gene induced by Chemical", which consists of noun phrases, noun conjunctions, and an adjectival clause. Therefore, in this paper, we focus on resolving the complexity of noun structures.

[1] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/

[2] https://spacy.io/

For the dependency parse tree $Tree(s)$, we visit the tree from the root and repeatedly split the tree at the noun nodes, where the nouns are those tokens with POS tags as NN, NNS, NNP, and NNPS. These noun nodes are repeated and kept in the original tree as a leaf and in the subtree as the root. Then keeping the original order of the tokens, each tree can be written as a short sentence. In this way, the original long sentence can be partitioned hierarchically into shorter ones. This sentence break-down step not only enhances the pattern mining step, but also makes the nested relation extraction possible.

Example 1. The sentence in Figure 1 will be rewritten as 5 short sentences:
decrease: treatment can decrease overexpression.
treatment: Pre-treatment of CHEMICAL
overexpression: overexpression of GENE
GENE: GENE and GENE
GENE: GENE induced by CHEMICAL
where the first token of each line indicates the root of the corresponding short sentence.

C. Meta Pattern Mining

After the sentences are partitioned into shorter ones, we then mine frequent sequential text patterns from the set of short sentences. The lemmas of the words are used to avoid the effect of different word forms. The frequent sequential text patterns mp are those sequences of tokens $t_{mp_1}t_{mp_2}...t_{mp_n}$ which appear for more than τ times. The frequency constraint can help identify meaning expressions. However, for a sequential text pattern, the frequency of its any sub-sequence is no less than the frequency of this pattern. Consequently, there exist many noisy and incomplete patterns such as "of CHEMICAL" and "and GENE".

1) Pattern Quality Criteria: In order to reduce the number of the low quality patterns, we further propose three additional criteria to ensure the pattern quality: the patterns should be informative, complete grammarly and complete semantically. To ensure the pattern informativeness, we require that the pattern should either contain one entity mention plus at least one non-stop-word, or contain two or more entity mentions. To ensure the pattern completeness in grammar, we constraint that for a candidate pattern, its tokens should form a connected graph on the dependency parse tree, i.e., $\forall i, j \leqslant n$, there is a path between t_{mp_i} and t_{mp_j} on $Tree(s)$. To ensure the pattern completeness in semantics, we require that the pattern should appear at least once as a short sentence. After the noisy pattern filtering, we get a high quality meta-pattern set $\mathcal{MP}$.

Example 2. Continuing previous example. Pattern "and GENE" is not grammarly complete because these two tokens are siblings and not connected on the dependency parse tree. Pattern "of CHEMICAL" is regarded as incomplete in semantics because there is no short sentence that says "of CHEMICAL". Both patterns are not informative because there is only one entity mention but no other non-stop-word in each pattern.

2) Hierarchical Pattern Grouping: Using the pattern quality criteria, many of the noisy patterns can be removed. However, in some cases, it also breaks the relationship between entities which are in different short sentences. For example, pattern "treatment can decrease overexpression" is considered uninformative due to the lack of entity mentions. However, this pattern is important to extract the relation between the CHEMICAL and the GENE.

The reason that the entity mentions are not expressed in the same short sentence is because grammarly, the entity mentions are modifiers of other nouns. However, semantically speaking, the entity mentions are the emphasis of those noun phrases. For example, "pre-treatment of CHEMICAL", "effect of CHEMICAL", etc, are actions or properties of the CHEMICAL. To tackle this issue, we propose to group patterns hierarchically so that the relations can be extracted and expressed more straightforward. We start from entity phrases: For a noun-rooted short sentence $s = t_1, t_2, ...t_m$, if $s \in \mathcal{MP}$ and has a single entity mention, it is then regarded as the entity phrases. After the grouping of entity phrases, the distance between the entities is further shortened. Then we can re-mine the meta patterns and may discover more high-quality ones.

Example 3. Patterns "overexpression of GENE" and "pre-treatment of CHEMICAL" will be tagged as a GENE phrase and a CHEMICAL phrase respectively. Then "treatment can decrease overexpression" will be rewritten as "{CHEMICAL phrase} can decrease {GENE phrase}". When conducting extraction, the expression format of extraction is then written based on the hierarchy of patterns, entity phrases and expressions according to the parse tree.

For patterns that have more than one entity mention, we want to group them by their semantic meanings. We group them under the following two assumptions: 1) Synonymous meta patterns should have the same entity types and same number of entity mentions; and 2) synonymous meta patterns should share similar extracted tuples and keywords. For the extracted tuples, we adopt the term frequency-inverse document frequency (TF-IDF) vector of its instance sets. Then we perform clustering to group synonymous meta patterns. Since the TF-IDF vector can be sparse, cosine similarity or Jaccard similarity are recommended when constructing the clusters. In the experiment, we adopt the hierarchical clustering approach with cosine distance. The algorithm will keep merging clusters with the minimal cosine distance until the minimal cosine distance is greater than a threshold. Then for each cluster, we use the words with the highest frequency to represent the meaning of the cluster. If two clusters have the same keywords, we further merge them together.

V. EXPERIMENT

To test the effectiveness of WW-PIE, we collect a subset of PubMed paper abstracts as our Biomedical corpus. To ensure that the selected paper contain some biomedical relationships, we used the Comparative Toxicogenomics Database (CTD) [5], a human-curated database containing biological entities

TABLE I: Statistics of the CTD dataset subset used in our experiments.

	# of Entity	# of Relation	# of PMID
Chemical-Gene	Chemical: 7,187 Gene: 667	163,126	26,786
Chemical-Disease	Chemical: 7,187 Disease: 598	84,104	1,619
Disease-Gene	Gene: 669 Disease: 599	834	1,622
Total	**Chemical: 7,187** **Gene: 669** **Disease: 599**	**248,064**	**28,007**

and their relationships, to guide the selection of the PMIDs (PubMed IDs).

We focus on three entity types in the corpus: gene, chemical, and disease. The named entities are recognized and typed by Pubtator [37]. Among these entity types, there are three relation types in CTD: chemical-gene, chemical-disease, and disease-gene relationships. We first randomly select 248,064 relationships from the above three relation types that are associated with experimental evidence in the CTD database. The CTD also provides the PMIDs related to these selected relationships. We then collect all PubMed abstracts that are shown to be associated with the above relationships in CTD as our input corpus. The entity and relationship statistics of our input corpus are shown in Table I.

A. *Results On Relation Extractions*

Baselines. To show the effectiveness of WW-PIE in relation extraction, we compare it with the following state-of-the-art OpenIE approaches:

- **ClausIE** [12] adopts clause patterns to handle long-distance relationships.
- **Stanford OpenIE** [13] learns a clause splitter via distant training data.
- **Ollie** [11] utilizes open pattern learning and extracts patterns over dependency path and part-of-speech tags.
- **MinIE** [14] refines tuples extracted by ClausIE by identifying and removing parts that are considered overly specific.

We also tried pattern-based methods such as METAPAD but they do not work due to pattern sparsity in biomedical corpus.

For the baseline methods, since their extraction are in surface-name form and the results are not structured, we conduct the following post-processing to obtain the structured extractions: (1) use NER tools to recognize the name entities, and (2) keep extractions from the SVO formatted extractions if there is an entity E_s in the subject and an entity E_o in the object. For WW-PIE, we only use patterns with two or more entity mentions to extract the tuples.

Evaluation Metrics. We randomly sample 96 sentences from the 28,007 input PubMed abstracts for performance comparison. The annotator is asked to evaluate without knowing which model produced the results, eliminating potential bias in evaluation. Similar to the settings in previous studies [12], one tuple will be judged as correct if it reads smoothly and meets the fact described in the sentence. For example, both ("DISEASE", "is", "DISEASE") and ("DISEASE", "induced by", "CHEMICAL") are correct. However, ("SPECIES", "with", "DISEASE") will not be counted since they have no relationship.

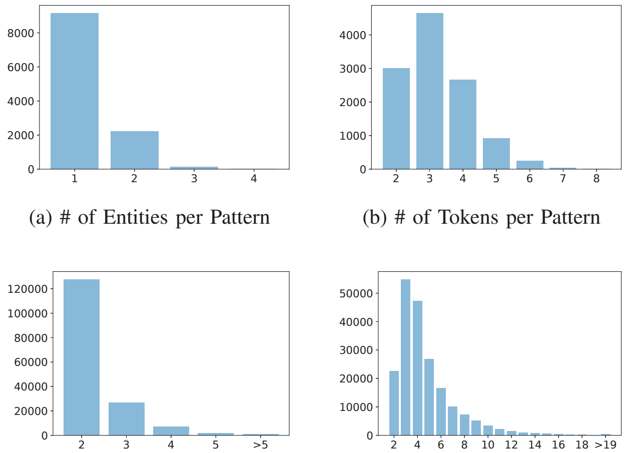
The following measures are adopted: (1) number of correct extractions, (2) number of valid extractions, (3) the precision of the extracted tuples, i.e. # correct extractions divided by # valid extractions. The first measure represents how many correct extractions can be detected from the corpus by each method. Note here, for OpenIE task, it is hard to measure the recall since it is infeasible to know all the "correct" extractions, but the number of correct extraction can reflet the recall in a certain degree. For the second measure, the valid extractions for the baselines refer to the extractions after post-processing, and for WW-PIE, they refer to the extractions after pattern grouping and removing the tuples with conjunction relation types. The third measure represents the precision, the percentage of the correct extractions in the valid extractions. For this measure, the higher the better.

TABLE II: Performance comparison with state-of-the-art OpenIE systems

	# Correct extractions	# Valid extractions	Precision
ClausIE [12]	21	142	0.15
Stanford [13]	**120**	277	0.43
Ollie [11]	43	84	0.51
MinIE [14]	77	126	0.61
WW-PIE	110	150	**0.73**

Results and case studies. Table II summarizes the comparison results on the extracted tuples. It is clear that the proposed WW-PIE method achieves significant improvement in precision. It boosts the precision from 61% of the best baseline (MinIE) to 73%. The number of correct extractions of WW-PIE is also competitive with the baseline methods. Only Standford OpenIE extracts more correct extractions. However, after a careful sentence-by-sentence examination, we find that the information extracted by the baseline OpenIE methods has high redundancy, and the number of correct extractions does not honestly reflect the amount of information detected by these methods.

The major reason is that these methods try to re-write the sentences with various expressions, but different expressions carry almost identical information and do not introduce new information. For example, sentence "YM511 significantly inhibited testosterone-stimulated transcriptional activation of estrogen-responsive element (ERE) in MCF-7 cells transfected transiently with ERE-luciferase reporter plasmid". Stanford OpenIE produces 60 valid extractions, such as ⟨YM511, inhibited, testosterone-stimulated activation⟩ , ⟨YM511, significantly inhibited, testosterone-stimulated activation⟩, and ⟨YM511, significantly inhibited, testosterone-stimulated transcriptional activation in MCF-7 cells⟩. However, these extractions have much information in common. When we evaluate the results, the annotator are asked to give T/F labels for each valid extraction. Therefore, among the 60 valid extractions, 28 are annotated as correct, even though there are only couple of unique meaningful tuples in the original sentence. On the other



(a) # of Entities per Pattern

(b) # of Tokens per Pattern

(c) # of Entities per Expression

(d) # of Tokens per Expression

Fig. 2: Pattern and Expression Distributions.

hand, WW-PIE tries to provide the most complete information in a single tuple, so the redundancy among extractions is very low. We find that almost all correct extractions are unique.

Compare with baseline methods, WW-PIE is especially strong in extracting relations expressed in noun phrases. For example, "GENE inhibitor CHEMICAL" is an informative pattern to extract the inhibition relation between the gene mention and the chemical mention. However, due to the lack of verb in this expression, such relation cannot be detected by the baseline methods. The shortcoming of WW-PIE is that some information may be missed due to the infrequent expressions. For example, WW-PIE fail to extract the relation tuples from the sentence in the previous example, as "CHEMICAL-stimulated transcriptional activation" is infrequent even though "CHEMICAL significantly inhibit activation" is frequent. To conquer this shortcoming, we plan to improve WW-PIE with skip-gram pattern extraction.

B. Results On Pattern Extractions

As discussed earlier, it is important that the extractions and the patterns to be simple and straightforward. We examine the meta patterns and the expression formatted extractions by WW-PIE. Figure 2 shows four distributions.

Figure 2a shows the number of patterns with different number of entity mentions. It is clear that most meta patterns from WW-PIE contains a single entity mention, i.e., they are entity phrases. These patterns are not directly used in entity relation extractions, but they can be useful in other tasks (See Table III for more examples). Figure 2b shows the number of patterns with different number of tokens. We can see that most patterns contain less than five tokens, indicating that the patterns are concise. For the two figures at the bottom, we focus on the expression format of extractions, so there are at least two entity mentions in each expression. Figure 2c shows the number of expressions with different number of entity mentions. Though most expressions contain two entities, there are still a good number of extractions contain more entities. Figure 2d shows the number of expressions with different number of tokens. The expressions contain more tokens than the patterns do in general, but most of them still contain less than 5 tokens, further demonstrating the conciseness of the extraction. Comparing with WW-PIE, the average length of the extractions from the four baselines (ClausIE, Standford OpenIE, Ollie, and MinIE) are 13.5, 6.7, 9.4, and 9.1 respectively. We also find that 44% of the extractions are extracted from two or more short sentences, which implies that WW-PIE is effective in getting the complete information.

TABLE III: Top 10 Frequent Meta Patterns with Single Entity.

Meta Patterns with Single Entity	**#**
DISEASE cell	11210
effect of CHEMICAL	9507
GENE expression	6551
expression of GENE	4940
CHEMICAL treatment	4896
GENE gene	4229
CHEMICAL exposure	3957
the effect of CHEMICAL	3721
GENE mrna	3211
CHEMICAL level	3076

TABLE IV: Examples of synonymous meta pattern groups.

Synonymous group	**Meta Patterns**
CHEMICAL_induced inhibition of GENE	GENE inhibition by CHEMICAL
	CHEMICAL block GENE
	GENE inhibitor , CHEMICAL
	GENE inhibitor CHEMICAL
CHEMICAL activate GENE	CHEMICAL_activated GENE
	GENE activator CHEMICAL
	GENE agonist CHEMICAL
	GENE agonist , CHEMICAL
	GENE ligand CHEMICAL
	GENE ligand , CHEMICAL
DISEASE cause by CHEMICAL	CHEMICAL_induced DISEASE
	CHEMICAL can cause DISEASE
	CHEMICAL induce DISEASE
	CHEMICAL cause DISEASE
	DISEASE be induce by CHEMICAL
	DISEASE induce by CHEMICAL
	DISEASE produce by CHEMICAL
SPECIES treat with CHEMICAL	CHEMICAL administration to SPECIES
	CHEMICAL_treated SPECIES
	CHEMICAL_exposed SPECIES
	CHEMICAL treat SPECIES
	SPECIES be inject with CHEMICAL
	SPECIES be administer with CHEMICAL
	...

In Table III, we list the top 10 most frequent informative meta-patterns with single entity and their counts. We can see that these meta patterns are high-quality biomedical phrases. These phrases may be useful for recognizing more entities, entity types, entity functions and entity properties. For example, "GENE expression" and "GENE mrna" may be helpful in named entity recognition tasks: if a proper noun appears frequently in such phrases, there is a high chance it is a gene's name. "CHEMICAL treatment" can be a reliable pattern for recognizing treatments.

In Table IV, we list some examples of the synonymous pattern groups. For example, the first group refers to the

TABLE V: Examples of the extractions from the meta patterns.

Meta Pattern {CHEMICAL} **reduce** {DISEASE}	
Extractions in Expression Format	**Extractions in Tuple Format**
Ranitidine reduce **ischemia/reperfusion**-induced **liver_injury** in **rats**	⟨**Ranitidine**, reduce, **liver_injury**: ⟨**ischemia/reperfusion**, induce, **liver_injury**, in, **rats**⟩ ⟩
resveratrol reduce **brain_injury**	⟨**resveratrol**, reduce, **brain_injury** ⟩
Resveratrol reduce **renal_and_lung_injury** cause by **sepsis** in **rats**	⟨**Resveratrol**, reduce, **renal_and_lung_injury**: ⟨**sepsis**, cause, **renal_and_lung_injury**, in **rats**⟩ ⟩
Resveratrol reduce **TNF-a**-induced U373MG **human glioma_cell_invasion**	⟨**Resveratrol**, reduce, **glioma_cell_invasion**: ⟨**TNF-a**, induce, **human glioma_cell_invasion**⟩ ⟩
caffeine treatment reduce **glioma** cell proliferation	⟨**caffeine**, reduce, **glioma** ⟩
Meta Pattern {CHEMICAL} **inhibit** {GENE}	
Extractions in Expression Format	**Extractions in Tuple Format**
Progesterone inhibit **COX-2** expression	⟨**Progesterone**, inhibit, **COX-2**⟩
NAC treatment inhibit phosphorylation of **Akt**	⟨**NAC**, inhibit, **Akt** ⟩
ATRA inhibit the expression of **Ccnb1** and **Ccna1**	⟨**ATRA**, inhibit, (**Ccnb1**, **Ccna1**)⟩
Cypermethrin inhibit the interaction between the **AR_AF1** and **SRC-1**	⟨**Cypermethrin**, inhibit, **AR_AF1**:⟨**AR_AF1**, interaction, **SRC-1**⟩ ⟩
PGF and **H2O2** inhibit **SOD1** protein expression and activity	⟨(**PGF**,**H2O2**), inhibit, **SOD1**⟩
Meta Pattern {GENE} **cause** {DISEASE}	
Extractions in Expression Format	**Extractions in Tuple Format**
mutations in the **CSB** gene cause **Cockayne_syndrome**	⟨**CSB**, cause, **Cockayne_syndrome**⟩
mutations in **FOXP2** cause **developmental_verbal_dyspraxia** (**DVD**)	⟨**FOXP2**, cause, **developmental_verbal_dyspraxia**: ⟨ ⟨**developmental_verbal_dyspraxia**, abbr, **DVD**⟩ ⟩
mutations in the **hENT3** gene cause an **autosomal_recessive_disorder** in **humans**	⟨**hENT3**, causes, **autosomal_recessive_disorder**: ⟨**autosomal_recessive_disorder**, in, **humans**⟩ ⟩
germline mutations in **DIS3L2** cause the **Perlman_syndrome_of_overgrowth** and **Wilms_tumor** susceptibility	⟨(**DIS3L2**, cause, (**Perlman_syndrome_of_overgrowth**, **Wilms_tumor**) ⟩

relation type “GENE inhibitor CHEMICAL”, which includes meta patterns such as “CHEMICAL_induced inhibition of GENE”, “GENE block CHEMICAL”, and “GENE inhibition by CHEMICAL”. We can see that the synonymous patterns may have different key words and may have different entity orders (such as “GENE block CHEMICAL”). However, the meta patterns within each group still have close semantic meaning and each group can be regarded as a specific relation type. By grouping synonymous meta patterns together, we reduce the redundancy of extracted relation types and enrich the relation instances for each type, which makes the output more structured for downstream applications.

Table V shows some examples of the patterns and extractions in both the expression format and the tuple format. The words in bold are the recognized named entities. We can see that for a pattern, the expressions can be very diverse. In existing pattern-based information extraction methods, these expressions will be considered as separated patterns, and consequently, the frequency of each pattern is rather low. These examples also demonstrate three advantages brought by the hierarchical pattern grouping. 1) The pattern grouping merges the diverse expressions into a bigger group, so it can address the challenge of pattern sparsity. 2) It shortens the distance between entity mentions. 3) The hierarchical pattern grouping enables a new structure of extractions, the nested relations, which is not studied in any existing OpenIE methods. As a result, WW-PIE can extract more information and more complete information.

VI. DISCUSSION AND FUTURE WORK

WW-PIE utilizes both dependency parsing and sequential frequent pattern mining in finding typed textual patterns. Both component can be further improved. We test several state-of-the-art parsers which are trained on general domain, but they all experience more difficulty in parsing sentences from the biomedical domain. WW-PIE can tolerant the errors at a certain degree, as the errors are unlikely to appear in a high frequency, but a better dependency parser can enhance the quality of sentence break-down and the relation extraction. The sequential frequent patterns may be carefully extend to frequent skip-gram patterns as discussed in Section V-A. If the infrequent expression “CHEMICAL-stimulated transcriptional activation” can be merged into pattern “CHEMICAL-stimulated {adj} activation”, WW-PIE then can extract the relation tuples from the original sentence, and improve recall overall.

In the experiments, we mainly examine the patterns with multiple entity mentions and focus on relation extraction tasks. However, the patterns can be useful in a wide range of applications. As mentioned in Section V-B, the patterns with single entity can be useful in named entity recognition, entity function discovery, entity property discovery, etc. For example, recent papers [38], [39] point out the importance of expressions such “GENE phosphorylation”, “GENE overexpression”, and “CHEMICAL sensitivity” in extracting genomic anomalies and post-translational modifications (PTMs).

Through the error analysis, we find that both the baseline methods and WW-PIE have trouble for some negation structures. For example, “there is no evidence that ...”, and "apoptosis of DISEASE cell". Such negation structures may change the relation type found in the OpenIE methods. We leave this open challenge to our future work.

VII. CONCLUSION

We propose a novel method WW-PIE for the Bio-OpenIE tasks. It can extract all variety of the relation tuples from large biomedical literature corpora with little human effort. WW-PIE first resolves the long and complicated sentence structures by breaking down the sentences using dependency parse tree into shorter ones. Then WW-PIE discovers frequent textual meta patterns and group them hierarchically to extract n-ary hierarchical tuples with entity type information. Our method achieves the highest precision in comparison with state-of-the-art OpenIE baselines and keeps simplicity and hierarchical structures of the extracted information. These various experiments demonstrate the effectiveness of WW-PIE in handling real-world biomedical literature with complicated sentence structures and rich information.

REFERENCES

[1] C.-H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, and Z. Lu, "Overview of the biocreative v chemical disease relation (cdr) task," in *Proceedings of the fifth BioCreative challenge evaluation workshop*. Sevilla Spain, 2015, pp. 154–166.

[2] Y. Peng, C.-H. Wei, and Z. Lu, "Improving chemical disease relation extraction with rich features and weakly labeled data," *Journal of cheminformatics*, vol. 8, no. 1, p. 53, 2016.

[3] 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 *BIBM'16*. IEEE, 2016, pp. 994–1001.

[4] J. Gu, F. Sun, L. Qian, and G. Zhou, "Chemical-induced disease relation extraction via convolutional neural network," *Database*, vol. 2017, 2017.

[5] 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.

[6] M. Jiang, J. Shang, T. Cassidy, X. Ren, L. M. Kaplan, T. P. Hanratty, and J. Han, "Metapad: Meta pattern discovery from massive text corpora," in *KDD'17*. ACM, 2017, pp. 877–886.

[7] N. Nakashole, G. Weikum, and F. Suchanek, "Patty: a taxonomy of relational patterns with semantic types," in *EMNLP'12*. ACL, 2012, pp. 1135–1145.

[8] Q. Li, M. Jiang, X. Zhang, M. Qu, and J. Han, "Truepie: Discovering reliable patterns in pattern-based information extraction," in *KDD'18*. ACM, 2018.

[9] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web." in *IJCAI'07*. AAAI, 2007, pp. 2670–2676.

[10] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *EMNLP'11*. ACL, 2011, pp. 1535–1545.

[11] M. Schmitz, R. Bart, S. Soderland, O. Etzioni *et al.*, "Open language learning for information extraction," in *EMNLP'12*. ACL, 2012, pp. 523–534.

[12] L. Del Corro and R. Gemulla, "Clausie: clause-based open information extraction," in *WWW'13*. ACM, 2013, pp. 355–366.

[13] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *ACL'15*. ACL, 2015, pp. 344–354.

[14] K. Gashteovski, R. Gemulla, and L. Del Corro, "Minie: minimizing facts in open information extraction," in *EMNLP'17*. ACL, 2017, pp. 2630–2640.

[15] Q. Zhu, X. Ren, J. Shang, Y. Zhang, F. F. Xu, and J. Han, "Open information extraction with global structure constraints," in *WWW'18*. ACM, 2018, pp. 57–58.

[16] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *ACL'92*. ACL, 1992, pp. 539–545.

[17] P. Ernst, A. Siu, and G. Weikum, "Highlife: Higher-arity fact harvesting," in *WWW'18*, 2018, pp. 1013–1022.

[18] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.

[19] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "Frontiers of biomedical text mining: current progress," *Briefings in bioinformatics*, vol. 8, no. 5, pp. 358–375, 2007.

[20] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle *et al.*, "Extracting information from textual documents in the electronic health record: a review of recent research," *Yearb Med Inform*, vol. 35, no. 8, pp. 128–144, 2008.

[21] M. Krallinger, A. Valencia, and L. Hirschman, "Linking genes to literature: text mining, information extraction, and retrieval applications for biology," *Genome biology*, vol. 9, no. 2, p. S8, 2008.

[22] W. W. Chapman and K. B. Cohen, "Guest editorial: Current issues in biomedical text mining and natural language processing," *Journal of biomedical informatics*, vol. 42, no. 5, pp. 757–759, 2009.

[23] M. S. Simpson and D. Demner-Fushman, "Biomedical text mining: A survey of recent progress," in *Mining text data*, C. C. Aggarwal and C. Zhai, Eds. Springer, 2012, pp. 465–517.

[24] U. Hahn, K. B. Cohen, Y. Garten, and N. H. Shah, "Mining the pharmacogenomics literature–a survey of the state of the art," *Briefings in bioinformatics*, vol. 13, no. 4, pp. 460–494, 2012.

[25] D. Zhou, D. Zhong, and Y. He, "Biomedical relation extraction: from binary to complex," *Computational and mathematical methods in medicine*, vol. 2014, pp. 1–18, 2014.

[26] K. B. Cohen and D. Demner-Fushman, *Biomedical natural language processing*. John Benjamins Publishing Company, 2014, vol. 11.

[27] S. Ananiadou, P. Thompson, R. Nawaz, J. McNaught, and D. B. Kell, "Event-based text mining for biology and functional genomics," *Briefings in functional genomics*, vol. 14, no. 3, pp. 213–230, 2014.

[28] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H.-P. Kriegel, "Extraction of semantic biomedical relations from text using conditional random fields," *BMC bioinformatics*, vol. 9, no. 1, p. 207, 2008.

[29] B. Rink, S. Harabagiu, and K. Roberts, "Automatic extraction of relations between medical concepts in clinical texts," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 594–600, 2011.

[30] B. Percha, Y. Garten, and R. B. Altman, "Discovery and explanation of drug-drug interactions via text mining," in *Biocomputing 2012*. World Scientific, 2012, pp. 410–421.

[31] J. Hakenberg, C. Plake, U. Leser, H. Kirsch, and D. Rebholz-Schuhmann, "Lll05 challenge: Genic interaction extraction-identification of language patterns based on alignment and finite state automata." Citeseer.

[32] F. Liu, J. Chen, A. Jagannatha, and H. Yu, "Learning for biomedical information extraction: Methodological review of recent advances," *arXiv preprint arXiv:1606.07993*, 2016.

[33] S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein, L. Ungar, S. Winters, and P. White, "Integrated annotation for biomedical information extraction," in *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, 2004.

[34] V. Nebot and R. Berlanga, "Semantics-aware open information extraction in the biomedical domain," in *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*. ACM, 2011, pp. 84–91.

[35] ——, "Exploiting semantic annotations for open information extraction: an experience in the biomedical domain," *Knowledge and information Systems*, vol. 38, no. 2, pp. 365–389, 2014.

[36] N. de Silva, D. Dou, and J. Huang, "Discovering inconsistencies in pubmed abstracts through ontology-based information extraction," in *ACM-BCB'17*. ACM, 2017, pp. 362–371.

[37] 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.

[38] 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.

[39] 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," in *Protein Bioinformatics*, 2017, pp. 213–232.