# PENNER: Pattern-enhanced Nested Named Entity Recognition in Biomedical Literature

Xuan Wang[1*], Yu Zhang[1*], Qi Li[1], Cathy H. Wu[2], Jiawei Han[1]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[2]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA
{xwang174, yuz9, qili5, hanj}@illinois.edu, wuc@udel.edu

***Abstract*—Many biomedical entity mentions contain other entity mentions nested inside. Most current named entity recognition (NER) systems deal with only flat entities and ignore such nested entities, which may introduce errors to subsequent tasks such as relation extraction and knowledge base completion. Recently, fully supervised methods are proposed for nested named entity recognition. Despite their success on benchmark datasets, supervised methods rely on human annotation and lead to highly specialized systems that cannot be easily adapted to new entity types. In this study, we propose PENNER, a novel and effective pattern-enhanced nested named entity recognition method that relies on massive corpora plus only very weak supervision. We compare PENNER with a state-of-the-art BioNER system, PubTator, and observe great improvement at recognizing genes, chemicals, diseases and species. PENNER can also accurately extract new types of entities, such as biological process and treatment, that are not annotated by PubTator.**

***Index Terms*—nested named entity recognition, meta-pattern discovery, pattern mining, multi-set expansion**

## I. INTRODUCTION

Biomedical named entity recognition (BioNER) is a task that identifies text spans associated with proper names and classifies them into a set of semantic classes, such as genes, proteins, chemicals and diseases. BioNER is a fundamental step in the biomedical information extraction pipeline. It facilitates many downstream tasks such as relation extraction [2], [18] and knowledge base completion [9], [28], [29], [36].

The common way for BioNER is to formulate the task as a sequence labeling problem. Various approaches have been successfully applied to BioNER, from feature-based [14], [16] to neural network methods [4], [8], [34]. However, most of such methods are unable to handle *nested named entities*. Figure 1 shows an example. A CHEMICAL entity (i.e., "*alanine*") is nested in a PROTEIN entity (i.e., "*alanine aminotransferase*"). The flat entity recognition methods cannot detect both correctly. For example, PubTator [35], a state-of-the-art BioNER system, recognizes "*alanine*" as a CHEMICAL, but misses "*alanine aminotransferase*" as a PROTEIN.

Emphasis should be put on nested named entities for two reasons: (1) The nested entity structure is quite common in biomedical literature. For example, 17% of the entities in the GENIA [30] dataset are embedded within another entity. And

..... although each of the agents alone caused only slight increase in the [[alanine]$_{\text{CHEMICAL}}$ aminotransferase]$_{\text{PROTEIN}}$ activity.

Fig. 1. An example of biomedical named entities with nested naming architecture from PubMed (PMID: 10190572).

(2) many downstream tasks require NER to detect not just the inner-most entity. For example, in Figure 1, we would like to know that a protein activity is being discussed instead of a chemical one. Failing to do so may introduce errors to subsequent relation extraction and knowledge base completion.

Several approaches have been proposed for handling nested named entity recognition [1], [6], [11], [12], [19], [23]. All of these methods are fully supervised, which require human effort for feature engineering or data annotation. Feature-based approaches [1], [6], [19], [23] rely on handcrafted features carefully designed for each entity type. Neural network models [11], [12] 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 biomedical nested named entity recognition, five types of entities (i.e., gene/protein, DNA, RNA, cell type and cell line) are annotated. Despite the success of the supervised models on the GENIA dataset, it remains unknown whether they perform well at detecting nested structures for other important types of entities such as chemicals and diseases. No results are shown in their papers, or even no proper benchmark datasets are found for recognizing nested structures for chemicals and diseases.

In this paper, we propose PENNER (Pattern-enhanced Nested Named Entity Recognition), which relies on massive corpora and unsupervised pattern mining to tackle the problems mentioned above. Our model takes massive corpora as input, with entities pre-tagged by any existing flat named entity recognition tools. 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 seed patterns and perform automatic type set expansion. Note that in this step, we only need **very weak supervision** (two user-specified seed patterns in contrast with a large nested annotated training corpus). The top-ranked meta-patterns for each type are treated as correctly typed meta-patterns and are used to extract the correct outer entities from the input corpus. Compared with previous methods, our method greatly

*The first two authors contributed equally to this work and should be considered as joint first authors.

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

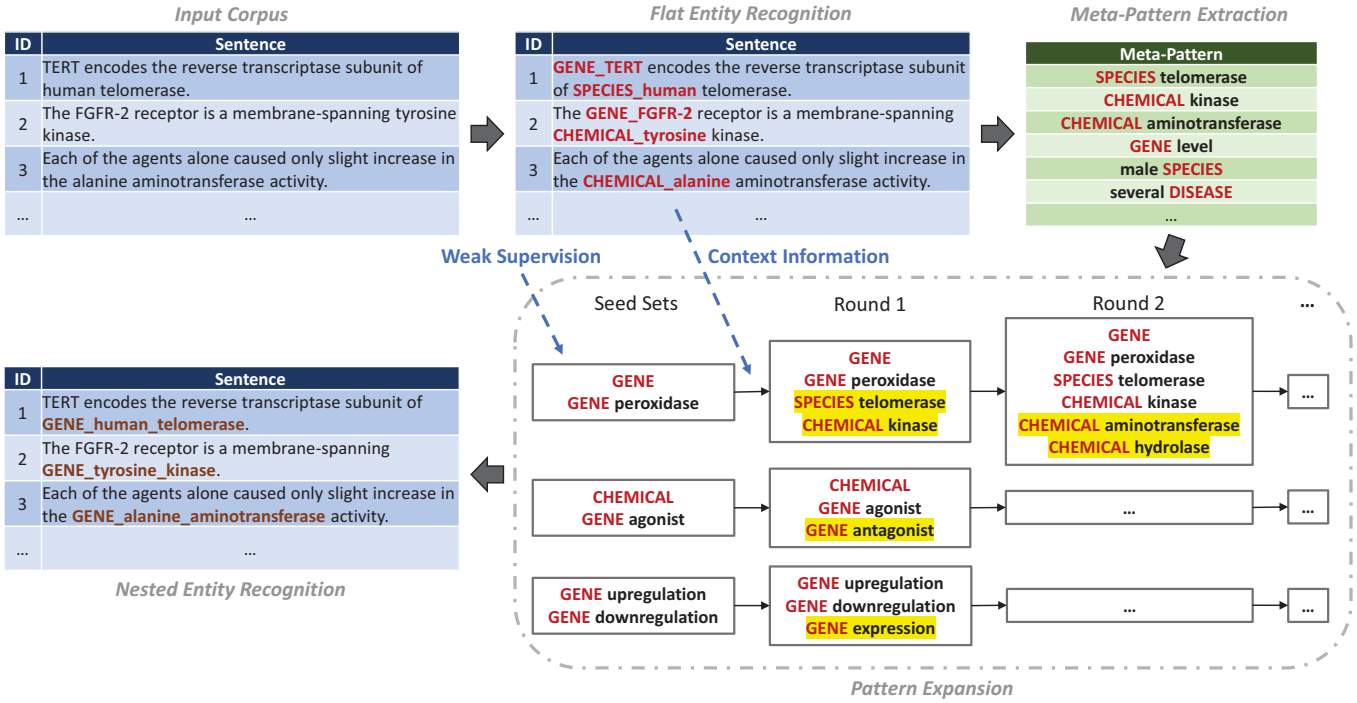


Fig. 2. Framework overview of PENNER.

enhances nested named entity recognition without any human effort for feature engineering or data annotation. Additionally, our novel set expansion approach gives us the advantage to discover new entity types that are not annotated in the pre-tagged input corpus. We compare our method with a state-of-the-art BioNER system, PubTator [35], and observed great improvement in recognizing the outer entities with four types: gene, chemical, disease and species. Our method is also able to accurately extract biological process and treatment entities that are originally not annotated by PubTator in the input corpus.

## II. Related Work

Several methods have been proposed for flat named entity recognition. Early techniques in the supervised domain have been based on hidden markov models (HMMs) [37] or, later, conditional random fields (CRFs) [21]. Recently, recurrent neural networks (RNNs) have been widely applied to several sequence labeling tasks achieving state-of-the-art results. Lample et al. [15] proposed neural models based on long short-term memory networks (LSTMs) and CRFs for flat named entity recognition and achieve state-of-the-art performance.

There are fewer approaches, however, addressed the problem of nested entities. Alex et al. [1] presented several techniques based on CRFs for nested named entity recognition for 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. Their approach could not identify nested entities of the same type. Finkel and Manning [6] 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 [19] 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 [19] 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 [23] introduced the notion of mention separators for nested entity mention detection. In contrast to the hypergraph representation that Lu and Roth [19] 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 named entity recognition are recently proposed as extensions to the state-of-the-art RNN-based models for flat named entity recognition. Katiyar and Cardie [12] proposed to learn a hypergraph representation for nested entities using features extracted from a recurrent neural network. Ju et al. [11] proposed to dynamically stack flat NER layers and recognize outer entities with additional information from their inner entities. The neural network models save efforts for manual feature generation, but they still require a large amount of training data and is not easily

adaptable to new entity types.

## III. Framework

### A. Overview

We lay out our PENNER framework in Figure 2. Given a raw corpus as input, we first use PubTator [35] to recognize and classify biomedical entities. The detected "flat" entities will be replaced by their types. In the second step, we extract quality sequential patterns with entity type tokens, which are also called *meta-patterns* [10]. From our perspective, a quality meta-pattern is assumed to be frequent, informative and complete. In Figure 2, the green box shows some examples of extracted quality meta-patterns. The third step is pattern expansion. For each entity type to be detected, we take two user-specified seeds as weak supervision, and expand the seed set iteratively. At each round, we find the meta-patterns sharing high context similarity with patterns already in the seed set. For example, in Figure 2, the weak supervision is *"GENE"* and *"GENE peroxidase"* for the GENE type. During the first round, *"SPECIES telomerase"* and *"CHEMICAL kinase"* are considered to be similar with those two seeds. Then they will be put into the GENE seed set for future expansion. After getting the meta-patterns for each entity type, we go back to the original corpus to find their concrete instances, which are naturally nested entities.

### B. Meta-pattern Extraction

Given a flat NER result of a corpus, we can replace the spans of tokens with their entity type names. After the replacement, the corpus will be a mixed sequence of entity types and non-type words. A meta-pattern is a sub-sequence of the corpus which contains at least one entity type token. For example, "*human telomerase*" and "*mouse telomerase*" may be two patterns in the original corpus, and they will be represented as one meta-pattern "*SPECIES telomerase*" after the replacement.

Studying meta-patterns is necessary for nested NER from two perspectives: (1) The nested structure is more like a pattern-level phenomenon than an instance-level one. For example, "*SPECIES telomerase*" is a protein entity no matter we are talking about humans or mice. (2) A meta-pattern has the aggregated context information of all of its instances, which helps us learn its semantics in a more accurate way.

Meta-pattern discovery has been studied by [10], [17] and [33]. They propose a set of statistical features (e.g., pattern frequency, IDF score, etc.) to train a classifier that estimates the quality of each candidate meta-pattern. However, they do not rely on deeper semantic analysis of the sentence structure. In biomedical literature, sentences are usually long and with formal language styles. To better utilize this feature, we improve their methods by incorporating dependency parsing [13] into PENNER.

We use SpaCy[1] for dependency parsing. The output parsing tree has a set of directed syntactic relations between the words in a sentence. Figure 3 shows an example. 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`).

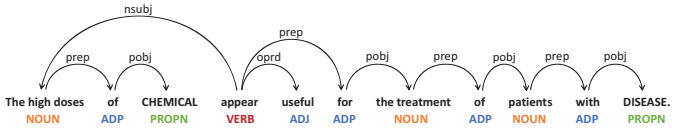


Fig. 3. An example sentence with dependency parsing tree using SpaCy.

According to the parsing tree as well as the original corpus, we propose the following four criteria to select quality meta-patterns:

**(1) Frequency.** A quality meta-pattern should occur frequently. In this paper, we require each meta-pattern candidate to appear for more than 10 times in the corpus.

**(2) 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., *"patients with DISEASE"*). In this paper, since we focus on NER, meta-patterns with two or more entity mentions (e.g., "*CHEMICAL induces DISEASE*") will not be considered, but they will be useful in relation extraction.

**(3) Syntactic Completeness.** For a quality meta-pattern, all of its tokens in the parsing tree should form a connected subgraph. We take the sentence in Figure 3 as an example. "*CHEMICAL appear useful*" is not complete since "*CHEMICAL*" and "*appear*" are separated by other nodes. In contrast, "*patients with DISEASE*" is complete.

**(4) Semantic Completeness.** Since we are doing NER, the extracted pattern should be a complete noun phrase. Recall the previous example. "*of CHEMICAL*" is syntactic complete, but it may not be a complete noun phrase. To alleviate this problem, 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 will serve as the leaf of the current chunk as well as the root of the next chunk. In the example, the sentence will be divided into 4 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.

We first use sequential pattern mining [10] to discover all the meta-patterns satisfying our frequency threshold. Then we check them one by one on the informativeness, syntactic completeness and semantic completeness.

### C. Pattern Expansion

Taking the extracted meta-patterns as candidates, we further select patterns associated with the entity types we want to detect. As mentioned above, to get rid of the reliance on entity-type-dependent training corpus, the pattern selection step needs to be solved under **very weak supervision**. Here we adopt the SETEXPAN framework [26]. SETEXPAN takes several user-provided "seed" patterns (e.g., "*GENE*" and "*CHEM-*

[1]https://spacy.io/

*ICAL peroxidase*") and extracts other patterns belonging to the same semantic class (e.g., "*CHEMICAL aminotransferase*", "*CHEMICAL hydrolase*", "*SPECIES telomerase*", etc.). To capture the semantics of each pattern, we utilize **skip-grams** as their features.

Given a target 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 "*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. One advantage of using skip-grams is that it imposes strong positional constraints.

Note that word embedding methods such as word2vec [22] also use skip-gram information. We will show the effectiveness of our method against word2vec in experiments.

SETEXPAN defines the similarity between each pair of pattern $p$ and feature $c$ using the TF-IDF transformation [24]:

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

where $P$ is the set of candidate patterns and $X_{p,c}$ is the raw co-occurrence count between $p$ and $c$. Empirically, [26] shows that such weight scaling outperforms some 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

$$sim(p_1, p_2|F) = \frac{\sum_{c \in F} \min(f_{p_1,c}, f_{p_2,c})}{\sum_{c \in F} \max(f_{p_1,c}, f_{p_2,c})}. \quad (2)$$

Given the seed set $S$, we first score each skip-gram feature $c$ based on its accumulated strength with entities 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

$$score(p|F_i) = \frac{1}{|S|} \sum_{p' \in S} sim(p, p'|F_i). \quad (3)$$

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

$$MRR(p) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{r_{p,i}}. \quad (4)$$

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

In practice, we need to recognize entities of different types **simultaneously**. For instance, we may expand new patterns representing genes and chemicals at the same time, using two seed 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, given in Algorithm 1. Given $Q$ seed sets $S_1, S_2, ..., S_Q$ of different types, MULTISETEXPAN extracts new patterns for each $S_i$ by turns. If a pattern already appears in other seed sets, no matter how large its MRR is, we will not put it in $S_i$.

In our experiments, we find this strategy very useful in avoiding interference among different seed sets.

**Algorithm 1** MULTISETEXPAN($S_1, ..., S_Q, MRR_{thrs}$)

**Input:** $M$ seed sets $S_1, ..., S_Q$ representing $Q$ different entity types.
**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 if**
    **end for**
    **if** nothing added into $S_i$ in this round **then**
      Mark $S_i$ as converged.
    **end if**
  **end for**
**end while**
**Output:** $Q$ expanded sets $S_1, ..., S_Q$

TABLE I
BASIC STATISTICS OF THE SUBSET CORPUS.

| Abstracts | Sentences | Entity Mentions | | | |
|---|---|---|---|---|---|
| | | Gene | Chemical | Disease | Species |
| 28007 | 302736 | 215704 | 314134 | 129931 | 86697 |

## IV. EVALUATION

We aim to answer three questions in the experimental part. First, at the pattern level, does PENNER perform well in meta-pattern expansion? Second, at the instance level, does PENNER perform well in nested named entity recognition? Third, after the pattern enhancement, what is our improvement over PubTator?

### A. Setup

**Dataset.** While PENNER is a general method that can be applied to any set of biomedical literature, we select a subset of PubMed paper abstracts for the evaluation process. The selection is based on the Comparative Toxicogenomics Database (CTD) [5] which 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 I shows some basic statistics of the subset corpus.

**Baselines.** We demonstrate the effectiveness of PENNER by comparing it with two baselines:

TABLE II
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 III
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 |

TABLE IV
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 |

(1) EMBEDDING [22] first 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, other patterns are ranked by the sum of distances away from the seeds, and top-10 patterns will be returned as the result.

(2) SETEXPAN [26] is an ablation of the whole PENNER framework, where the seed sets of every entity types are expanded one by one instead of simultaneously.

For SETEXPAN and PENNER, we set $M, M_0, N$ to be 200, 120, 10, respectively. If the final expanded list has more than 10 patterns, we only take the first 10 into consideration.

### *B. Extracting New Meta-Patterns*

We first focus on the meta-pattern expansion. As we all know, five major entity types can be recognized by PubTator: gene/protein, chemical, disease, species and SNP. Since SNP is sparse in the corpus, we consider the other four in this paper. For each entity type, we put two patterns in the seed set to see whether the methods can find new nested structures of the same type. The results are shown in Tables II, III and IV.

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 ignores 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 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 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 added into the GENE seed set after the first round. This may cause severe *semantic drift* problem since other disease- and chemical-related patterns will be included in the next few rounds. In contrast, PENNER will never consider *"CHEMICAL"* or *"DISEASE"* as candidates since they already appear in other seed sets.

We now discuss the four sets expanded by PENNER in details.

TABLE V
*NDCG* OF DIFFERENT METHODS ON THE FOUR TYPES.

| | GENE | CHEMICAL | DISEASE | SPECIES |
|---|---|---|---|---|
| EMBEDDING [22] | 0.139 | 0.580 | 0.073 | 0.315 |
| SETEXPAN [26] | 0.602 | 0.312 | **0.754** | 0.417 |
| PENNER | **1.000** | **1.000** | **0.754** | **0.776** |

**(1) GENE.** PENNER detects 10 gene/protein meta-patterns that are all correct. Most of the patterns are enzymes, sharing the same fine-grained type with "*GENE peroxidase*" in the seed set. Note that when entities appear in the same meta-pattern, they are more likely to be similar to each other or belong to the same subtype. For example, the instances of "*CHEMICAL aminotransferase*" include "*alanine aminotransferase*", "*aspartate aminotransferase*", "*tyrosine aminotransferase*", "*ornithine aminotransferase*", etc. All the chemicals here are amino acids.

**(2) CHEMICAL.** PENNER extracts 3 chemical meta-patterns, among which "*GENE antagonist*" is the counterpart of "*GENE agonist*" in the seed set. The instances of "*CHEMICAL chelation*" include "*iron chelation*", "*copper chelation*", "*zinc chelation*", "*EDTA chelation*", etc. We believe that "*CHEMICAL chelation*" as a whole entity is more complete than part of it since metal ions and their chelations are different types of chemicals.

**(3) DISEASE.** PENNER discovers 5 meta-patterns for the disease seeds, among which three are correct and the other two are biological processes. "*hepatic DISEASE*" extracts instances such as "*hepatic fibrosis*", "*hepatic inflammation*", "*hepatic tumor*" and "*hepatic toxicity*". Again, we believe "*hepatic fibrosis*" is a more complete entity than "*fibrosis*" itself. In fact, PubTator recognizes "*liver fibrosis*" or "*liver inflammation*" as a whole (see abstracts with PMIDs 30079841 and 23813842 as two examples). Thus, we also need "*hepatic DISEASE*" for the consistency in recognition.

**(4) SPECIES.** PENNER finds 10 meta-patterns for the species seeds, among which eight are correct and the other two are cell types. From our perspective, including attributes (e.g., *"male"*) for species entities is beneficial to downstream knowledge extraction tasks. For example, one sentence in the abstract of [7] is *"Amphetamine and cocaine decreased susceptibility to myoclonus in young mice and increased susceptibility in mature mice."* If the NER system ignores *"young"* and *"mature"*, the facts extracted from this sentence will be inaccurate and even controversial to each other. Similar sentences can also be observed for male species and female species [27]. In addition, "*Sprague-Dawley rat*" is recognized as a whole by PubTator (see abstracts with PMIDs 24726336 and 25325438 as two examples). Therefore, it is consistent to recognize "*F334 rat*" as well.

### C. Recognizing Nested Entities

Using the extracted meta-patterns for each entity type, we go back to the original corpus to find their concrete instances, which are naturally nested entities.

TABLE VI
NUMBER OF INSTANCES EXTRACTED BY DIFFERENT METHODS ON THE FOUR ENTITY TYPES.

| | GENE | CHEMICAL | DISEASE | SPECIES |
|---|---|---|---|---|
| EMBEDDING [22] | 79 | 139 | 61 | 45 |
| SETEXPAN [26] | 1734 | **458** | **184** | 2211 |
| PENNER | **5254** | **458** | **184** | **3212** |

TABLE VII
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 | - |

Tables V and VI show the performance of PENNER in nested NER. Since the extracted patterns form a ranking list, we use *normalized discounted cumulative gain* (NDCG) [25] to evaluate the rank-aware precision. Besides precision, we also show the numbers 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 V and VI, we can see that PENNER consistently outperforms the baselines both in precision and recall. For GENE and CHEMICAL, all the instances extracted by PENNER are correct. Besides, PENNER finds 5254 nested GENE entities in 28,007 abstracts (i.e., on average, one nested GENE entity in every 5.33 abstracts), which reflects that in biomedical literature, the nested structure is not an exception, but the norm.

### D. Finding New Types of Entities

We now demonstrate an important advantage of PENNER against fully-supervised methods: finding new types of entities. As we all know, if one entity type is not even mentioned in the training set, it would be extremely difficult for supervised methods to detect entities of this type. In biomedical literature, biomedical processes [3] and treatment entities [31] arouse great concern. 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 [20] and protein modification (PTM) [32]. However, these two types are not originally annotated by PubTator. Under this setting, PENNER shows its power. Similar to the stories above, we just use two seeds for each entity type.

Table VII shows the pattern expansion results on the two new types.

**(1) Biological Process.** Taking *"GENE upregulation"* and *"GENE downregulation"* as seeds, PENNER extracts 10 other

TABLE VIII
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**]$_{\text{CHEMICAL}}$]$_{\text{TREATMENT}}$ ... |

patterns from the text, among which 8 are correct. Similar to the seeds, most of the extracted biological process patterns are describing the activities of genes.

**(2) Treatment.** Taking *"CHEMICAL injection"* and *"CHEMICAL inhalation"* as seeds, PENNER expands the set to a large group of treatment patterns. PENNER makes a mistake here since *"CHEMICAL exposure"* is actually a symptom. In fact, if a chemical appears in an instance of *"CHEMICAL treatment"*, it is more likely to be a drug. (E.g., top-5 frequent instances of *"CHEMICAL treatment"* include "*resveratrol treatment*", "*simvastatin treatment*", "*quercetin treatment*", "*estrogen treatment*" and "*5-FU treatment*".) In contrast, if a chemical appears in an instance of *"CHEMICAL exposure"*, it is usually a toxic entity.

### E. Case Study

In order to further show our improvements over PubTator, we compare the annotation results of several sentences by PENNER and PubTator in Table VIII. In the first two sentences, PubTator can only do flat NER, while PENNER successfully detects CHEMICAL-GENE and GENE-CHEMICAL nested structures. From the second sentence, we can also see that recognizing more complete entities will benefit downstream applications such as relation extraction. For example, it is *"Erk1/2 antagonist"* instead of *"Erk1/2"* inhibiting *"MCP-1"*. Failing to detect *"antagonist"* will lead us to an almost opposite conclusion. In the third and fourth sentences, PENNER finds new types of entities using the biological process and treatment meta-patterns we just obtained. Again, the GENE-PROCESS nested structure helps us realize that it is *"STAT1 phosphorylation"* instead of *"STAT1"* being up-regulated.

While the annotation results have been improved by PENNER, there is still room for future progress. For example, meta-patterns can be utilized in a more general way. PENNER mainly uses meta-patterns with only one entity type token to deal with nested structures. However, meta-patterns with two or more type tokens may also be useful. We still take the sentences in Table VIII 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 structure.

## V. CONCLUSIONS AND FUTURE WORK

This paper presents a framework to find nested entity structures in biomedical literature. Taking a corpus pre-tagged by any existing flat NER tool, we can extract quality meta-patterns in an unsupervised way, and find meta-patterns associated with each entity type under very weak supervision. Experiments show that our PENNER outperforms the baselines by a large margin in finding nested patterns and entities. Besides, it can also be used to find new types of entities with just two user-specified seeds. Moreover, we show that the final annotation results are largely improved over PubTator.

There are several future directions in light of these results. First, as mentioned in Section IV-E, meta-patterns can be utilized in a more general way. Second, it is widely believed that various types of biomedical entities have their formation rules. The unsupervised meta-pattern discovery can actually help us find some entity naming principles. Third, it is interesting to study whether nested entity structures can help meta-pattern

discovery in return. If so, we can run the pattern extraction module and the pattern expansion module iteratively, letting them mutually enhance each other. Besides, we would like to see how the detected complex entity types (e.g., gene variants/mutations, protein phosphorylation) can benefit real use cases.

## ACKNOWLEDGMENT

The research was sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreement No. W911NF-17-C0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HDTRA11810026, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). Any opinions, findings, and conclusions or recommendations expressed in this document are those of the author(s) and should not be interpreted as the views of any U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. We also thank anonymous reviewers for valuable and insightful feedback.

## REFERENCES

[1] 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*, pages 65–72. ACL, 2007.

[2] M. Cokol, I. Iossifov, C. Weinreb, and A. Rzhetsky. Emergent behavior of growing knowledge about molecular interactions. *Nature biotechnology*, 23(10):1243–1247, 2005.

[3] G. O. Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261, 2004.

[4] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368, 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*, 45(D1):D972–D978, 2016.

[6] J. R. Finkel and C. D. Manning. Nested named entity recognition. In *EMNLP'09*, pages 141–150. ACL, 2009.

[7] C. A. Greer and H. P. Alpern. Maturational changes related to dopamine in the effects of d-amphetamine, cocaine, nicotine, and strychnine on seizure susceptibility. *Psychopharmacology*, 64(3):255–260, 1979.

[8] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

[9] 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 *BIBM'15*, pages 975–982. IEEE, 2015.

[10] 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*, pages 877–886. ACM, 2017.

[11] M. Ju, M. Miwa, and S. Ananiadou. A neural layered model for nested named entity recognition. In *NAACL-HLT'18*, pages 1446–1459. ACL, 2018.

[12] A. Katiyar and C. Cardie. Nested named entity recognition revisited. In *NAACL-HLT'18*, pages 861–871. ACL, 2018.

[13] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *ACL'03*, pages 423–430. ACL, 2003.

[14] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, and A. Valencia. Chemdner: The drugs and chemical names extraction challenge. *Journal of cheminformatics*, 7(1):S1, 2015.

[15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *NAACL-HLT'16*, pages 260–270. ACL, 2016.

[16] R. Leaman and Z. Lu. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32(18):2839–2846, 2016.

[17] 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 *KDD'18*, pages 1675–1684. ACM, 2018.

[18] 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*, pages 994–1001. IEEE, 2016.

[19] W. Lu and D. Roth. Joint mention extraction and classification with mention hypergraphs. In *EMNLP'15*, pages 857–867. ACL, 2015.

[20] 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*, 12(12):e0189663, 2017.

[21] R. McDonald and F. Pereira. Identifying gene and protein mentions in text using conditional random fields. *BMC bioinformatics*, 6(1):S6, 2005.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS'13*, pages 3111–3119. MIT Press, 2013.

[23] A. O. Muis and W. Lu. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In *EMNLP'17*, pages 2608–2618. ACL, 2017.

[24] X. Rong, Z. Chen, Q. Mei, and E. Adar. Egoset: Exploiting word ego-networks and user-generated ontology for multifaceted set expansion. In *WSDM'16*, pages 645–654. ACM, 2016.

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

[26] 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 *ECML-PKDD'17*, pages 288–304. Springer, 2017.

[27] R. E. Sorge, J. C. Mapplebeck, S. Rosen, S. Beggs, S. Taves, J. K. Alexander, L. J. Martin, J.-S. Austin, S. G. Sotocinal, D. Chen, et al. Different immune cells mediate mechanical pain hypersensitivity in male and female mice. *Nature neuroscience*, 18(8):1081, 2015.

[28] 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*, 45(D1):D362–D368, 2017.

[29] 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*, 44(D1):D380–D384, 2015.

[30] 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 *Panhellenic Conference on Informatics*, pages 382–392. Springer, 2005.

[31] 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*, 18(5):552–556, 2011.

[32] 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*, pages 213–232. Springer, 2017.

[33] X. Wang, Y. Zhang, Q. Li, Y. Chen, and J. Han. Open information extraction with meta-pattern discovery in biomedical literature. In *ACM-BCB'18*, pages 291–300. ACM, 2018.

[34] 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*, page bty869, 2018.

[35] C.-H. Wei, H.-Y. Kao, and Z. Lu. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522, 2013.

[36] B. Xie, Q. Ding, H. Han, and D. Wu. mircancer: a microrna–cancer association database constructed by text mining on literature. *Bioinformatics*, 29(5):638–644, 2013.

[37] G. Zhou and J. Su. Named entity recognition using an hmm-based chunk tagger. In *ACL'02*, pages 473–480. ACL, 2002.