CHEMNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-Guided Distant Supervision

Xuan Wang1*, Vivian Hu1*, Xiangchen Song2, Shweta Garg1, Jinfeng Xiao1 and Jiawei Han1

1University of Illinois at Urbana-Champaign, IL, USA
2Carnegie Mellon University, PA, USA

1{xwang174,vivianh2,shwetag2,jxiao13,hanj}@illinois.edu
2xiangchensong@cmu.edu

Abstract

Scientific literature analysis needs fine-grained named entity recognition (NER) to provide a wide range of information for scientific discovery. For example, chemistry research needs to study dozens to hundreds of distinct, fine-grained entity types, making consistent and accurate annotation difficult even for crowds of domain experts. On the other hand, domain-specific ontologies and knowledge bases (KBs) can be easily accessed, constructed, or integrated, which makes distant supervision realistic for fine-grained chemistry NER. In distant supervision, training labels are generated by matching mentions in a document with the concepts in the knowledge bases (KBs). However, this kind of KB-matching suffers from two major challenges: incomplete annotation and noisy annotation. We propose CHEMNER, an ontology-guided, distantly-supervised method for fine-grained chemistry NER to tackle these challenges. It leverages the chemistry type ontology structure to generate distant labels with novel methods of flexible KB-matching and ontology-guided multi-type disambiguation. It significantly improves the distant label generation for the subsequent sequence labeling model training. We also provide an expert-labeled, chemistry NER dataset with 62 fine-grained chemistry types (e.g., chemical compounds and chemical reactions). Experimental results show that CHEMNER is highly effective, outperforming substantially the state-of-the-art NER methods (with .25 absolute F1 score improvement).

1 Introduction

Named entity recognition (NER) is a fundamental step in scientific literature analysis to build AI-driven systems for molecular discovery, synthetic strategy designing, and manufacturing (Xie et al., 2013; Szklarczyk et al., 2015; Huang et al., 2015; Szklarczyk et al., 2017; de Almeida et al., 2019). It aims to locate and classify entity mentions (e.g., “Suzuki-Miyaura cross-coupling reactions”) from unstructured text into pre-defined categories (e.g., “coupling reactions”). In the chemistry domain, previous NER studies are mostly focused on one coarse-grained entity type (i.e., chemicals) (Krallinger et al., 2015; He et al., 2020; Watanabe et al., 2019) and rely on large amounts of manually-annotated data for training deep learning models (Chiu and Nichols, 2016; Ma and Hovy, 2016; Lam-ple et al., 2016; Wang et al., 2019b; Devlin et al., 2019; Liu et al., 2019).

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

Still, challenges exist for correctly recognizing the entity boundaries and accurately typing entities with distant supervision. In distant supervision, training labels are generated by matching the mentions in a document with the concepts in the knowledge bases (KBs). However, this kind of KB-matching suffers from two major challenges: (1) incomplete annotation where a mention in a document can be matched only partially or missed completely due to an incomplete coverage of the KBs (Figure 1a), and (2) noisy annotation where a mention can be erroneously matched due to the potential matching of multiple entity types in the KBs (Figure 1b). Due to the complex name structures (e.g., nested naming structures and long chemical
With these chiral nucleophiles, Suzuki-Miyaura cross-coupling reactions were carried out with various ary1- and hetaryl chlorides in good to excellent yields. Inorganic phosphorus compounds

**Coupling Reactions**

Although it was necessary to employ a stoichiometric quantity of palladium, it is noteworthy that the cross-coupling proceeded in the presence of a wide array of functional groups.

**Functional Groups**

(a) Incomplete Annotation

(b) Noisy Annotation

Figure 1: Two major challenges of distant supervision for fine-grained chemistry NER: (a) incomplete annotation, and (b) noisy annotation. The KB-matching labels are marked in red and the true labels are marked in blue.

formulas) of chemical entities, these challenges lead to severe low-precision and low-recall for fine-grained chemistry NER with distant supervision.

Several studies have attempted to address the incomplete annotation problem in distantly-supervised NER. For example, AutoNER (Shang et al., 2018b) introduces an “unknown” type that can be skipped during training to reduce the effect of false negative labeling with distant supervision. BOND (Liang et al., 2020) leverages the power of pre-trained language models and a self-training approach to iteratively incorporate more training labels and improve the NER performance. However, previous methods assume a high precision and reasonable coverage of KB-matching for distant label generation. For example, the KB-matching on the CoNLL03 dataset (Liang et al., 2020) reported over 80% on precision and over 60% on recall. These methods do not work well with fine-grained chemistry NER that has severe low precision and low recall with KB-matching. Previous studies also largely ignore the noisy annotation problem by simply discarding those multi-labels during the KB-matching process (Liang et al., 2020). However, the noisy labels cannot be simply ignored for the chemistry entities because they consist of a large portion of distant training labels. We observe that more than 60% of the entities have multiple labels during KB-matching in the chemistry domain.

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

2 Related Work

**Distantly-Supervised NER.** Aiming to reduce expensive manual annotation, distant supervision has been used to generate training labels automatically by utilizing the entity information from existing KBs. The major research efforts lie in dealing with the incomplete annotation problem caused by an incomplete coverage of the KBs (Fries et al., 2017; Shang et al., 2018b; Peng et al., 2019; Wang et al., 2019a, 2020a,b; Liang et al., 2020).

AutoNER (Shang et al., 2018b) proposes a “tie-or-break” tagging scheme to leverage distant supervision from entity dictionaries. Compared with the traditional “BIOES” tagging scheme, the “tie-or-break” tagging scheme introduces an “unknown” type that can be skipped during training to reduce the effect of false negative labeling brought by the incomplete KB-matching. However, AutoPhrase often misses low-frequency phrases for the “unknown” entity generation using a phrase mining method AutoPhrase (Shang et al., 2018a). Positive and unlabeled learning (PU-learning) is used in distantly-supervised NER to provide an unbiased and consistent estimator of the objective function (Peng et al., 2019). However, there are two limitations in using PU-learning for distantly-supervised NER. First, PU-learning uses the prior distribution for each entity type, a parameter that is estimated from an existing human-annotated test set that is not always available for new entity types. Second, the performance of PU-learning is highly sensitive to the class-imbalance rate for each entity type.

---

1 [https://github.com/xuanwang91/ChemNER](https://github.com/xuanwang91/ChemNER)
parameter that is heuristically determined. It is difficult to apply PU-learning to distantly-supervised NER tasks on new entity types in new domains due to the above two limitations. BOND (Liang et al., 2020) leverages the power of pre-trained language models (e.g., BERT and RoBERTa) and a self-training approach to iteratively incorporate more training labels and improve the NER performance. However, they do not work well with fine-grained chemistry entities that have a severe low-precision and low-recall problem with KB-matching. They also largely ignore the noisy annotation problem by simply discarding those multi-labels during the KB-matching process.

Other Related Tasks. One similar task to fine-grained NER is entity linking (Francis-Landau et al., 2016; Gupta et al., 2017; Raiman and Raiman, 2018; Le and Titov, 2018) that maps a candidate entity in the text to a concept identifier in the knowledge bases. However, entity linking cannot deal with new entities that do not exist in the background knowledge bases. Another similar task is fine-grained entity typing (FET) (Hoffart et al., 2011; Yosef et al., 2012; Ling and Weld, 2012; Del Corro et al., 2015; Ren et al., 2015; Choi et al., 2018) that has been extensively studied in the general domain. FET aims at classifying an entity mention into a wide range of entity types by disambiguating the pre-identified entity mentions into a set of candidate entity types. It is formulated as a multi-class, multi-label classification problem and does not assume type exclusiveness. The fine-grained NER task targets both entity boundary detection and entity type recognition and assumes each entity to be tagged with only one type in a given context. In this study, we focus on the fine-grained NER task in the chemistry domain.

3 The CHEMNER Framework

We propose CHEMNER, an ontology-guided distantly-supervised NER method for fine-grained chemistry NER (Figure 2). It includes distant label generation (entity span detection, flexible KB-matching, and ontology-guided multi-type disambiguation) and sequence labeling model training.

3.1 Data Preparation

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

Corpus Collection. For this study, we collected a chemistry literature corpus from PubChem\(^2\). This corpus contains 4,608 papers, among which 319 papers have the full-text and all have the title and abstract. There are 71,406 sentences in this corpus.

Type Ontology and Dictionary Collection. We collected a fine-grained chemistry type ontology from Wikipedia categories rooted under the Chemistry category\(^3\). We treat the Wikipedia category pages as types and the titles of the pages associated

\(^2\)https://pubchem.ncbi.nlm.nih.gov/
\(^3\)https://en.wikipedia.org/wiki/Category:Chemistry

Figure 2: The overall framework of CHEMNER. It includes a distant label generation (entity span detection, flexible KB-matching, and ontology-guided multi-type disambiguation) and a sequence labeling model training.
with each category as the entity dictionary for each type. We further remove irrelevant types and merge some fine-grained types to their coarse-grained parent types based on their term frequencies in the corpus. We also expand the entity dictionaries with synonyms collected from the PubChem knowledge base. Finally, we obtained a fine-grained chemistry entity type ontology with 62 types and its associated dictionaries with 10,551 entities. Figure 3 shows a subset of our chemistry type ontology. The complete fine-grained chemistry type ontology with 62 types can be found in Appendix A.1.

3.2 Flexible KB-Matching

Taking the input corpus, chemistry type ontology and associated entity dictionaries collected from the KBs, we first develop a flexible KB-matching method to resolve the incomplete annotation problem. Chemistry entities usually have complex name structures, such as nested naming structures (e.g., “aryl chloride” where “aryl” is a FUNCTIONAL GROUP, “chloride” is a HALIDE but altogether is an ORGANIC COMPOUND). Long chemical formulas (e.g., “Methyl 3’-(((Trifluoromethyl)sulfonyl)oxy)-[1,1’-biphenyl]-4-carboxylate”), that are quite flexible and cannot be fully covered by the KBs. Simple KB-matching such as the ones in previous distantly-supervised NER methods (Shang et al., 2018b; Liang et al., 2020) cannot match those complex chemistry entities that do not exist in the KBs, which leads to a severe low precision and low recall for labeling the fine-grained chemistry entities.

We propose to first conduct entity span detection with chemistry phrase chunking tools followed by a flexible KB-matching to resolve the incomplete KB-matching problem. We use two phrase chunking tools, ChemDataExtractor (Swain and Cole, 2016) and Genia Tagger (Tsuruoka and Tsujii, 2005), to generate candidate entity spans in the input corpus (e.g., in Figure 2 sentence S2, the phrase chunking tools find “Suzuki-Miyaura cross-coupling reactions” as a candidate entity span.) Based on the detected candidate entity spans, we develop a flexible KB-matching method with TF-IDF-based majority voting to resolve the incomplete annotation problem.

The flexible KB-matching method can match long and complex chemistry entities (e.g., chemical compounds) that do not exist in the KBs. Specifically, we label each candidate entity span by letting each word token in the entity span vote for several entity types that are most likely to involve this word token. For example, in Figure 2 sentence S1, “[Methyl-14C]S-Thd”, which is short for “4’-[methyl-14C]thiodymidine” according to the original document, is an author-defined abbreviation that cannot be covered by the existing KBs. However, since “Methyl-” is a common functional group that is usually the prefix of the organic compounds, this word token in “[Methyl-14C]S-Thd” helps vote for the types “ORGANIC COMPOUNDS” and “ORGANIC POLYMERS”. Another example is sentence S2, where three (“suzuki”, “coupling”, “reaction”) out of the five word tokens in “Suzuki-Miyaura cross-coupling reactions” help vote for the type “COUPLING REACTIONS”.

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

$$TF-IDF(w, t) = TF(w, t) \times IDF(w, t),$$

$$TF(w, t) = \frac{f(w, D_t)}{\sum_{w' \in y} f(w', D_t)},$$

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

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

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

3.3 Ontology-Guided Multi-Type Disambiguation

Based on the output of flexible KB-matching and the chemistry type ontology structure, we develop an ontology-guided multi-type disambiguation method to resolve the noisy annotation problem. An intuition of multi-type disambiguation is that the entities in the same sentence, paragraph or document usually follow a focused topic. For example, if a sentence is talking about organic chemistry, the entities in this sentence are more likely to have types related to organic chemistry. Following this intuition and the chemistry type ontology structure (Section 3.1), we draw two insights for an automated multi-type disambiguation: (1) the entity types in one sentence are usually confined to one big branch on the chemistry type ontology (e.g., organic or inorganic chemistry), and (2) the type of an entity under local context should be close to the types of the surrounding entities in the same sentence on the chemistry type ontology. For example, in Figure 2, sentence S3 contains one entity “palladium” that has two candidate types: “CATALYSTS” that falls under “CHEMICAL REACTIONS” branch, on the type ontology. So the which is also under the larger “CHEMICAL REACTIONS” and “TRANSITION METALS” that falls under “CHEMICAL ELEMENTS”. By looking at its surrounding entities (e.g., “cross-coupling”), we see that the surrounding entity types (e.g., “COUPLING REACTIONS” for “cross-coupling”) fall under the “ORGANIC REACTIONS” branch, which is also under the larger “CHEMICAL REACTIONS” branch, on the type ontology. So the sentence S3 is likely talking about chemical reaction and “palladium” is more suitable to have a type “CATALYSTS” instead of “TRANSITION METALS” based on the local context.

Formally, let \( s = [e_1, e_2, \ldots, e_n] \), where \( s \) denotes a sentence and \( e_i \) ith entity mention in it that has been assigned an initial label set \( T_{e_i} = \{t_{e_{i1}}, \ldots, t_{e_{im}}\} \) with flexible KB-matching. For an entity \( e_i \) with multiple candidate types (\( |T_{e_i}| > 1 \)) to be resolved, we calculate the inverse distance between this candidate type and the distribution of the surrounding types on the type ontology. The disambiguation score for each candidate type \( S_d(t_{e_i}) \) is defined as follows:

\[
S_d(t_{e_i}) = \frac{\sum_{k \in [1..n], k \neq i, |T_{e_k}| = 1 \text{dep}(lca(t_{e_k}, t_{e_i}))} {n * \text{dep}(t_{e_i})},
\]

where \( \text{lca}(-, -) \) denotes the lowest common ancestor of two types on the type ontology and \( \text{dep}(\cdot) \) denotes the depth of the type on the type ontology. \( S_d(t_{e_i}) \in (0, 1) \) and a larger score indicates that the candidate type \( t_{e_i} \) is more likely to be the correct type for the entity \( e_i \) in sentence \( s \).

If the surrounding types in the sentence still draw ties for the candidate type resolution, we could further enlarge the scope to a few surrounding sentences, the paragraph, the document or the corpus. We introduce a corpus-level global popularity score for each type based on our experimental observations. As shown in Figure 2, we calculate the frequency of each type in our initially labeled corpus with flexible KB-matching. “CATALYSTS” is globally more popular with a frequency of 18,707 compared to “TRANSITION METALS” with a frequency of 9,618. The global popularity score for each candidate type \( S_g(t_{e_i}) \) is defined as follows:

\[
S_g(t_{e_i}) = \frac{f_c(t_{e_i})} {\sum_{v \in T} f_c(v)},
\]

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

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

\[
S(t_{e_i}) = S_d(t_{e_i}) * S_g(t_{e_i}) \in (0, 1).
\]

We choose the type \( t_{e_i} \) for the entity \( e_i \) that has a highest score \( S(t_{e_i}) \) for multi-type disambiguation.

3.4 Sequence Labeling Models

The flexible KB-matching and multi-type disambiguation still rely on the signals from the KBs and ontologies, which cannot cover all the new entities in the corpus. Taken the output from the above two steps as distant supervision, we further train a sequence labeling model to solve the sparsity labeling problem. For example, in Figure 2 sentence 4, “BBA” is a new entity that cannot be labeled by flexible KB-matching since there is no obvious token-level signals. However, there is a
“boronic acid” entity with the type “OXOACIDS” in its surrounding context. The sequence labeling models will be able to capture those context patterns such as “either ... or ...” that usually connect entities with similar types. Thus they are likely to recognize “BBA” with the type “OXOACIDS”.

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

\[ l = \arg \min_{\theta} \sum_{i} \text{loss}(h_{\theta}(x_{i}), y), \]

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

4 Experiments

4.1 Dataset

We provide a chemistry NER dataset covering 62 fine-grained chemistry types such as chemical compounds and chemical reactions. This dataset can be used to benchmark distantly supervised NER methods for the fine-grained chemistry NER task. The input for training includes two parts: (1) a chemistry literature corpus with 69,806 unlabeled sentences, and (2) a chemistry type ontology with 62 fine-grained chemistry types and associated entity dictionaries for each type (Section 3.1). The test set contains 1,600 expert-annotated sentences on the fine-grained chemistry types. We use this test set to compare the performance of different NER methods in our experiments. We report the entity-level micro-precision, micro-recall, and micro-F1 scores\(^4\) of each NER method on the human-annotated test set. More details of the dataset preparation can be found in Appendix A.1.

4.2 Baselines

We compare the performance of CHEMNER with different groups of baseline methods. More details of the parameter settings and runtime analysis of each model can be found in Appendix A.2.

**KB-Matching**: This baseline is a simple string matching as (Peng et al., 2019). It is a greedy search algorithm that walks through a sentence trying to find the longest strings that match the entities in the dictionaries. For the strings matched with multiple types, we simply discard those multi-labels as (Liang et al., 2020).

**KB-Matching (freq)**: This baseline is a simple improvement of KB-Matching. For the strings matched with multiple types, we choose the type that has the highest frequency in the corpus.

**BiLSTM-CRF**: This baseline is the BiLSTM-CRF model (Ma and Hovy, 2016) that takes the results of KB-Matching (freq) as distant supervision.

**AutoNER**: This baseline is the AutoNER model (Shang et al., 2018b) that directly takes the raw corpus and the dictionaries as the input. It has a built-in KB-matching algorithm that maximizes the total number of matched tokens on each sentence to generate distant supervision. For the strings matched with multiple types, it assigns equal probabilities to each candidate type during training.

**RoBERTa**: This baseline is the RoBERTa model (Liu et al., 2019) that takes the results of KB-Matching (freq) as distant supervision.

**ChemBERTa**: This baseline is the ChemBERTa model (Chithrananda et al., 2020) that takes the results of KB-Matching (freq) as distant supervision. The ChemBERTa language model is pre-trained on the SMILE strings of the chemical molecule structures instead of the chemistry corpus. To our knowledge, there is no domain-specific pre-trained language model on the chemistry corpus.

**BOND**: This baseline is the BOND model (Liang et al., 2020) that takes the results of KB-Matching (freq) as distant supervision. The original distant supervision is our KB-Matching baseline according to the BOND paper. Here we use the improved KB-Matching (freq) baseline to give the BOND baseline an improved performance.

**CHEMNERF**: This is an ablation model of CHEMNER with the flexible KB-Matching only. For the strings matched with multiple types, we simply discard those multi-labels.

**CHEMNERFM**: This is an ablation model of CHEMNER with the flexible KB-Matching and the ontology-guided multi-type resolution.

**CHEMNERBiLSTM-CRF**: This is a variation of CHEMNER that takes the results of CHEMNERFM as distant supervision and trains a BiLSTM-CRF model for the final prediction.

**CHEMNERRoBERTa**: This is a variation of CHEMNER that takes the results of CHEMNERFM as...
Ablation Study.

(32%) and low recall (5%) for labeling the fine-matching and ontology-guided multi-type resolution of our proposed novel methods, flexible KB-recall. These two ablations show the effectiveness of the KB-Matching significantly improves the precision and recall over the best performing baseline methods: KB-Matching and RoBERTa. We also show the prediction results of our ablation models, ChemNERF and ChemNERFM, to demonstrate the contribution of each component and how the ChemNER full model achieves the best performance step by step.

KB-Matching can only match entities that exactly appear in the KB dictionaries, which often leads to incomplete or missing annotations. Based on the results of KB-Matching, RoBERTa learns to give one context-specific label for each entity. For example, in Sentence # 1, KB-Matching failed to give one context-specific label for each entity. As a result, the precision and recall of KB-Matching are significantly lower than those of the baseline methods.

Parameter Study. Table 3 shows the effect of different minimum TF-IDF threshold $\theta$ on the performance of ChemNER. This threshold $\theta$ is used to eliminate common word tokens from voting for the candidate entity types during the flexible KB-Matching. We observe that $\theta = 0.02$ gives the best performance of ChemNER FM. Table 4 shows the effect of different enlarged scopes on the performance of ChemNER FM. This enlarged scope is used to control the performance of ontology-guided multi-type disambiguation. We observe that when the context types in one sentence still draw ties for multi-type disambiguation, it is more effective to directly go to the corpus-level to look at the popularity scores for each type instead of extending the ontology-guided multi-type disambiguation mechanism to the document level.

### Table 1: Overall results (%) on the test set.

| Model          | Prec | Rec  | F1  |
|----------------|------|------|-----|
| KB-Matching    | 32.26| 4.95 | 8.58|
| KB-Matching (freq) | 20.51| 11.88| 15.05|
| BiLSTM-CRF (2016) | 21.88| 10.40| 14.09|
| AutoNER (2018b) | 20.51| 3.96 | 6.64|
| RoBERTa (2019) | 23.55| 17.74| 20.24|
| ChemBERTa (2020) | 17.54| 12.28| 14.45|
| BOND (2020)    | 18.84| 12.87| 15.29|
| ChemNER        | 69.47| 34.34| 45.96|

### Table 2: Results (%) of ChemNER ablations.

| Model          | Prec | Rec  | F1  |
|----------------|------|------|-----|
| ChemNER        | 71.90| 32.83| 45.08|
| ChemNER F      | 74.76| 29.06| 41.85|
| ChemNER FM     | 58.78| 29.06| 38.89|
| ChemNER BiLSTM-CRF | 48.65| 17.82| 26.09|
| ChemNER RoBERTa | 69.47| 34.34| 45.96|
| ChemNER ChemBERTa | 58.78| 29.06| 38.89|
| ChemNER BOND   | 52.21| 26.79| 35.41|

Table 3: Results (%) with different minimum TF-IDF threshold $\theta$ for the flexible KB-Matching.

| Model          | Prec | Rec  | F1  |
|----------------|------|------|-----|
| ChemNERFM      | 74.76| 29.06| 41.73|
| $\theta = 0.005$ | 66.67| 24.15| 35.46|
| $\theta = 0.02$  | 74.76| 29.06| 41.85|
| $\theta = 0.05$  | 71.19| 28.81| 41.43|

Table 4: Results (%) with different enlarged scopes for the ontology-guided multi-type resolution.

### 4.3 Experimental Results

Overall Results. Table 1 shows the overall results on the test set of our fine-grained chemistry NER dataset. ChemNER achieves .25 absolute F1 score improvement over the best performing baseline model RoBERTa. As we have discussed, the KB-Matching method suffers from severe low precision (32%) and low recall (5%) for labeling the fine-grained chemistry entities, which greatly limits the performance of the baseline NER methods that use KB-Matching for distant supervision.

Ablation Study. Table 2 shows the results of ablation studies on the test set of our fine-grained chemistry NER dataset. We compared our ChemNER full model with several ablations and variations. Our ablation model ChemNERF significantly improves the precision and recall over KB-matching and ChemNERFM further improves the recall. These two ablations show the effectiveness of our proposed novel methods, flexible KB-matching and ontology-guided multi-type resolution, for fine-grained chemistry NER under distant supervision. The four full model variations further shows that RoBERTa is the best sequence labeling model that takes the output of ChemNERFM as distant supervision.

Parameter Study. Table 3 shows the effect of different minimum TF-IDF threshold $\theta$ on the performance of ChemNERF. This threshold $\theta$ is used to eliminate common word tokens from voting for the candidate entity types during the flexible KB-Matching. We observe that $\theta = 0.02$ gives the best performance of ChemNERF.

Table 4 shows the effect of different enlarged scopes on the performance of ChemNERFM. This enlarged scope is used to control the performance of ontology-guided multi-type disambiguation. We observe that when the context types in one sentence still draw ties for multi-type disambiguation, it is more effective to directly go to the corpus-level to look at the popularity scores for each type instead of extending the ontology-guided multi-type disambiguation mechanism to the document level.

Qualitative Analysis. Table 5 shows some example sentences from our test set. We compare the prediction results of ChemNER with two baseline methods: KB-Matching and RoBERTa. We also show the prediction results of our ablation models, ChemNERF and ChemNERFM, to demonstrate the contribution of each component and how the ChemNER full model achieves the best performance step by step.
The total synthesis of "narciclasine" to two types (i.e., "ALKALOIDS", "FUNCTIONAL GROUPS"). Although "FUNCTIONAL GROUPS" is indeed a more suitable type and can be detected by the sequence labeling model trained on top of CHEMNERFM. For example, in Sentence # 1, CHEMNER finally chooses "ORGANOHALIDES" over "OXOACIDS" instead probably because the sequence labeling model captures the pattern on the co-occurrence of "ORGANOHALIDES" and "OXOACIDS". Interestingly, from the perspective of chemistry, organohalides and organoboron species (a sector of oxoacids) are the exact two couplers of the Suzuki Coupling reaction.

5 Conclusions and Future Work

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

Acknowledgment

This work was supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, US DARPA KAIROS Program No. FA8750-19-2-1004, SocialSim Program No. W911NF-17-C-0099, and INCAS Program No. HR00121C0165, and National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation, DARPA or the U.S. Government.

Ethics/Impact Statement

We provide an expert-labeled, chemistry NER dataset with 62 fine-grained chemistry types on 1,600 sentences. The text corpus is collected from an open-source chemistry database PubChem\(^5\). The entity types are collected from Wikipedia\(^6\). We recruited 5 undergraduate annotators from the Chemistry Department in our university. Each of the annotators is compensated at an hourly salary of $15. Annotators are voluntary participants who were aware of any risks of harm associated with their participation and had given their informed consents. Our project is subjected to the review of and approved by the IRB at our university. This dataset can be used to benchmark the named entity recognition performance on fine-grained chemistry NER, which contains 1,600 carefully annotated sentences. Each sentence is labeled with ground-truth entities with both the entity boundaries and the entity types. We ask three domain experts to annotate each sentence. We provide the annotators with an auto-complete drop-down menu consisting of our entity type vocabulary. Each pair of annotators reach a substantial agreement with a Fleiss’s \(\kappa\) of 0.72. The conflicts among annotators are resolved by another senior domain expert in the final annotated test set. We’ve described many characteristics of the dataset in Section 3.1. More details of the dataset and the steps taken during the data collection and preparation process can be found in Appendix A.1.

References

Adam M Azman. 2012. A chemistry spell-check dictionary for word processors. ACS Publications.

Seyone Chithrananda, Gabe Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. ArXiv preprint, abs/2010.09885.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 4:357–370.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 87–96, Melbourne, Australia. Association for Computational Linguistics.

A Filipa de Almeida, Rui Moreira, and Tiago Rodrigues. 2019. Synthetic organic chemistry driven by artificial intelligence. Nature Reviews Chemistry, 3(10):589–604.

Luciano Del Corro, Abdalghani Abujabal, Rainer Gemulla, and Gerhard Weikum. 2015. FINET: Context-aware fine-grained named entity typing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 868–878, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1256–1261, San Diego, California. Association for Computational Linguistics.

---

\(^5\)https://pubchem.ncbi.nlm.nih.gov/

\(^6\)https://en.wikipedia.org/wiki/Category:Chemistry
Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. ArXiv preprint, abs/1704.06360.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2681–2690, Copenhagen, Denmark. Association for Computational Linguistics.

Jiayuan He, Dat Quoc Nguyen, Saber A Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoan Fang, Hiyori Yoshikawa, et al. 2020. Overview of chemu 2020: named entity recognition and event extraction of chemical reactions from patents. In International Conference of the Cross-Language Evaluation Forum for European Languages, pages 237–254. Springer.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordini, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, and Jiawei Han. 2015. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, pages 995–1004. ACM.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018a. Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering, 30(10):1825–1837.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018b. Learning named entity tagger using domain-specific dictionary. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.
Matthew C Swain and Jacqueline M Cole. 2016. Chem-dataextractor: a toolkit for automated extraction of chemical information from the scientific literature. Journal of chemical information and modeling, 56(10):1894–1904.

Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezda T Doncheva, Alexander Roth, and Peer Bork. 2017. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research, 45(D1):D362–D368.

Damian Szklarczyk, Alberto Santos, Christian von Mering, Lars Juhl Jensen, Peer Bork, and Michael Kuhn. 2015. Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic acids research, 44(D1):D380–D384.

Yoshimasa Tsuruoka and Jun’ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 467–474, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Xuan Wang, Yingjun Guan, Yu Zhang, Qi Li, and Jiawei Han. 2020a. Pattern-enhanced named entity recognition with distant supervision. In 2020 IEEE International Conference on Big Data (Big Data), pages 818–827. IEEE.

Xuan Wang, Xiangchen Song, Bangzheng Li, Kang Zhou, Qi Li, and Jiawei Han. 2020b. Fine-grained named entity recognition with distant supervision in covid-19 literature. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 491–494. IEEE.

Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019a. Distantly supervised biomedical named entity recognition with dictionary expansion. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 496–503. IEEE.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019b. Cross-type biomedical named entity recognition with deep multi-task learning. Bioinformatics, 35(10):1745–1752.

Taiki Watanabe, Akihiro Tamura, Takashi Ninomiya, Takuya Makino, and Tomoya Iwakura. 2019. Multi-task learning for chemical named entity recognition with chemical compound paraphrasing. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6244–6249, Hong Kong, China. Association for Computational Linguistics.
A Appendix

A.1 Dataset Preparation

We have released all of our data and code for future studies, including the chemistry literature corpus, fine-grained entity type ontology and associated dictionaries collected from Wikipedia-Chemistry, manually-annotated test set for NER performance evaluation, and the code of CHEMNER.

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

Dictionary Collection. We collected a fine-grained chemistry entity type ontology from Wikipedia by treating category pages as types and the titles of the pages associated with each category as the entities for each type. We first conducted depth-first search (DFS) starting from the Chemistry category and found that the search did not stop when one million categories had been visited, and it often happened that a category relevant to Chemistry has irrelevant children. Therefore, we decide to use a technical term list to filter out irrelevant categories. We collected a spell-checker dictionary (Azman, 2012) with over 104,000 technical chemistry terms, and dropped a category from the search if less than 20% of 1-grams in its name and the names of all its direct children were covered by the dictionary. The threshold of 20% was selected empirically. After this step, we obtained a fine-grained chemistry entity type ontology with 3,775 types and 101,415 entities. We future tailor the entity type ontology and their associated entities by removing some irrelevant types and merge some fine-grained types to their coarse-grained parent types based on their frequencies in our chemistry literature corpus. We also expand the entity dictionaries with synonyms collected from the PubChem knowledge base. Finally, we obtained a fine-grained chemistry entity type ontology with 62 types and 10,551 entities. Figure 4 shows our complete chemistry entity type ontology.

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

A.2 Parameter Settings

Runtime with Parameters. We compared all sequence model we adopted during experiments. Our models are trained on a single NVIDIA Titan Xp (12GB) GPU. The details about the average runtime and the number of parameters are given in Table 6. All training hyperparameters follow their original implementation.

| Model     | Ave. runtime | # parameters |
|-----------|--------------|--------------|
| BiLSTM-CRF | 6h           | 2M           |
| AutoNER   | 20h          | 8M           |
| RoBERTa   | 4h           | 110M         |
| ChemBERTa | 3h           | 110M         |
| BOND      | 8h           | 110M         |

Table 6: Runtime and Number of Parameters
ers Python Interface to train the RoBERTa model on the Suzuki Coupling data using the *roberta-base* model with 10 epochs and a batch size of 32. The other hyperparameters are set to default values.

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

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

\(^{12}\text{https://github.com/cliang1453/BOND}\)
| Chemistry | Chemical_elements Sets_of_chemical_elements | Halogens |
|------------|--------------------------------------------|----------|
| Chemistry | Chemical_elements Sets_of_chemical_elements | Transition_metals |
| Chemistry | Chemical_elements Sets_of_chemical_elements | Noble_gases |
| Chemistry | Chemical_properties | Thermodynamic_properties |
| Chemistry | Chemical_reactions Catalysis | Catalysts |
| Chemistry | Chemical_reactions Catalysis | Enzymes |
| Chemistry | Chemical_reactions Catalysis | Catalysts |
| Chemistry | Chemical_reactions Catalysis | Homogeneous_catalysis |
| Chemistry | Chemical_reactions Catalysis | Hydrogenation_catalysts |
| Chemistry | Chemical_reactions Chemical_reaction_engineering | Chemical_kinetics |
| Chemistry | Chemical_reactions Organic_reactions | Carbon-carbon_bond_forming_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Coupling_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Functional_modification_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Addition_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Functional_modification_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Elimination_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Functional_modification_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Substitution_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Joining_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Polymerization_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Name_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Organic_redox_reactions |
| Chemistry | Chemical_reactions Organic_reactions | Ring_forming_reactions |
| Chemistry | Chemical_reactions Inorganic_reactions | Ligands |
| Chemistry | Inorganic_chemistry Coordination_chemistry | Coordination_compounds |
| Chemistry | Inorganic_chemistry Coordination_chemistry | Non-coordinating_anions |
| Chemistry | Inorganic_chemistry Inorganic_compounds | Chlorides |
| Chemistry | Inorganic_chemistry Inorganic_compounds | Inorganic_carbon_compounds |
| Chemistry | Inorganic_chemistry Inorganic_compounds | Inorganic_nitrogen_compounds |
| Chemistry | Inorganic_chemistry Inorganic_compounds | Inorganic_phosphorus_compounds |
| Chemistry | Inorganic_chemistry Inorganic_compounds | Inorganic_silicon_compounds |
| Chemistry | Inorganic_chemistry Inorganic_compounds | Metal_halides |
| Chemistry | Inorganic_chemistry Inorganic_compounds | Mineral_acids |
| Chemistry | Inorganic_chemistry Inorganic_compounds | Ossacids |
| Chemistry | Inorganic_chemistry Inorganic_reactions | Organometallic_chemistry |
| Chemistry | Inorganic_chemistry Organometallic_compounds | Organometallic_compounds |
| Chemistry | Inorganic_chemistry Organometallic_chemistry | Cyclometalated_complexes |
| Chemistry | Inorganic_chemistry Organometallic_chemistry | Sandwich_compounds |
| Chemistry | Organic_chemistry Functional_groups | Alkaloids |
| Chemistry | Organic_chemistry Organic_compounds | Aromatic_compounds |
| Chemistry | Organic_chemistry Organic_compounds | Biomolecules |
| Chemistry | Organic_chemistry Organic_compounds | Heterocycle_compounds |
| Chemistry | Organic_chemistry Organic_compounds | Macrocycles |
| Chemistry | Organic_chemistry Organic_compounds | Organic_acids |
| Chemistry | Organic_chemistry Organic_compounds | Organic_polymers |
| Chemistry | Organic_chemistry Organic_compounds | Organohalides |
| Chemistry | Organic_chemistry Organic_compounds | Organonitrogen_compounds |
| Chemistry | Organic_chemistry Organic_compounds | Organophosphorus_compounds |
| Chemistry | Organic_chemistry Organic_compounds | Organosulfur_compounds |
| Chemistry | Organic_chemistry Organic_compounds | Reactive_intermediates |
| Chemistry | Organic_chemistry Organic_compounds | Carbones |
| Chemistry | Organic_chemistry Organic_compounds | Reactive_intermediates |
| Chemistry | Organic_chemistry Organic_compounds | Free_radicals |
| Chemistry | Organic_chemistry Organic_compounds | Spiro_compounds |
| Chemistry | Organic_chemistry Organic_reactions | Carbon-carbon_bond_forming_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Coupling_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Functional_modification_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Addition_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Elimination_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Functional_modification_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Substitution_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Joining_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Polymerization_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Name_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Organic_redox_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Ring_forming_reactions |
| Chemistry | Organic_chemistry Organic_reactions | Organometallic_chemistry |
| Chemistry | Organic_chemistry Organometallic_compounds | Organometallic_compounds |
| Chemistry | Organic_chemistry Organometallic_chemistry | Cyclometalated_complexes |
| Chemistry | Organic_chemistry Organometallic_chemistry | Sandwich_compounds |
| Chemistry | Physical_chemistry Reactive_intermediates | Reactive_intermediates |
| Chemistry | Organic_chemistry Physical_chemistry Reactive_intermediates | Carbones |
| Chemistry | Organic_chemistry Physical_chemistry Reactive_intermediates | Free_radicals |
| Chemistry | Organic_chemistry Reactive_reagents_for_organic_chemistry | Alkylation_reagents |
| Chemistry | Organic_chemistry Reagents_for_organic_chemistry Protecting_groups |
| Chemistry | Organic_chemistry Stereochemistry | Stereochemistry |
| Chemistry | Organic_chemistry Stereochemistry | Isomerism |
| Chemistry | Organic_chemistry Stereochimstry | Substituents |

Figure 4: The complete fine-grained chemistry entity type hierarchy for CHEMNER.