Chem-FINESE: Validating Fine-Grained Few-shot Entity Extraction through Text Reconstruction

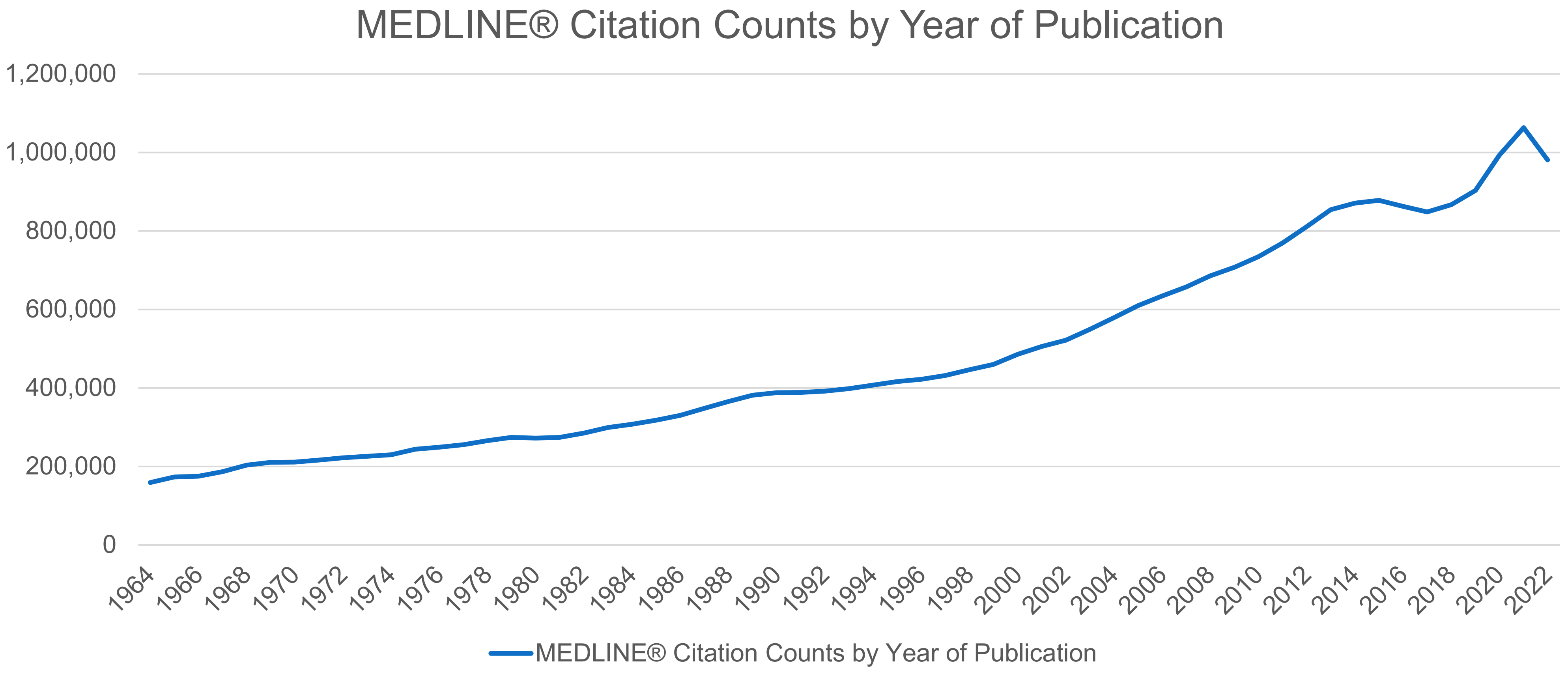
Qingyun Wang, Zixuan Zhang, Hongxiang Li, Xuan Liu, Jiawei Han, Huimin Zhao, Heng Ji

University of Illinois at Urbana-Champaign

BLENDER | Cross-source Information Extraction Lab

Information Overload

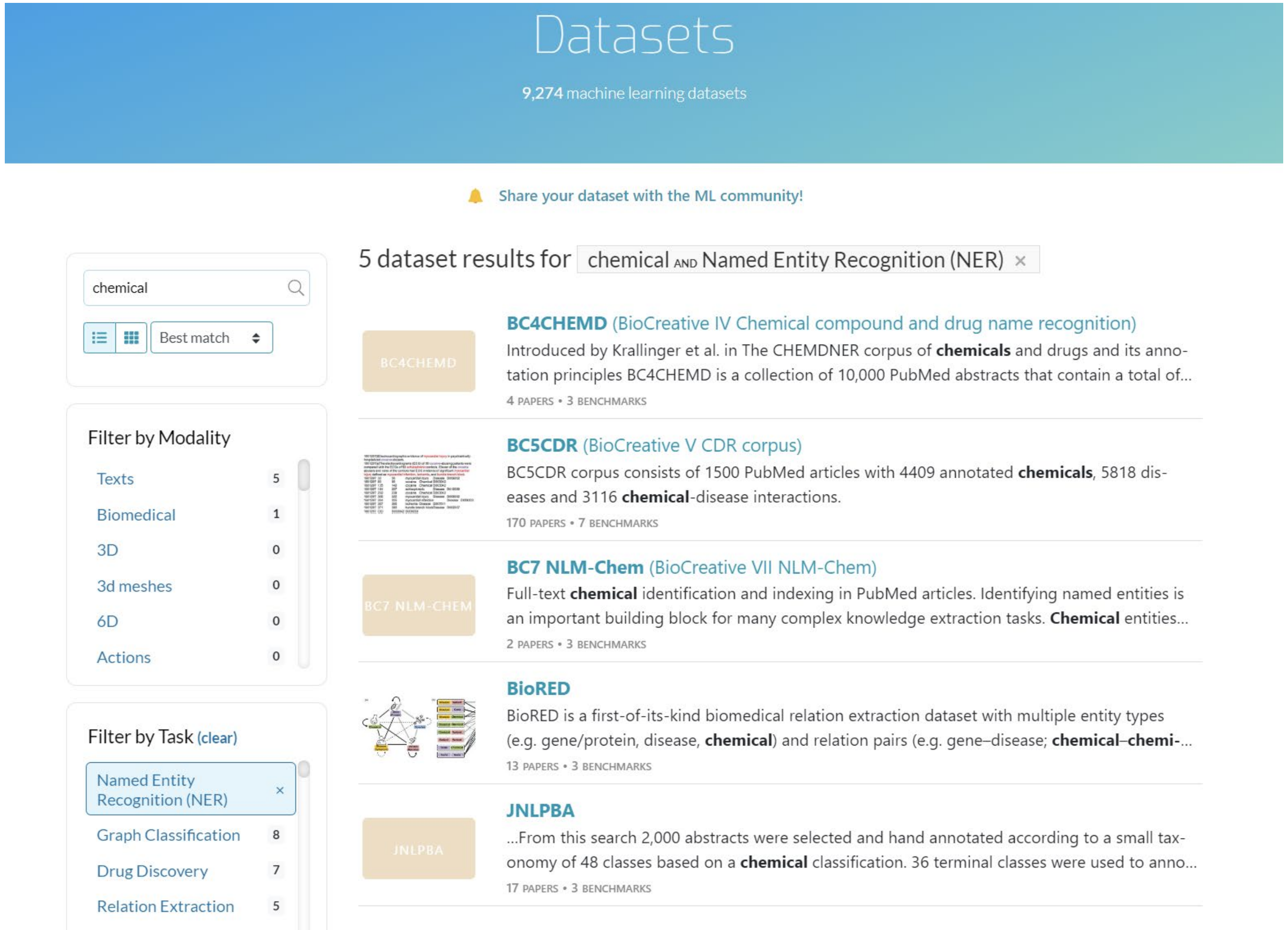
- Millions of scientific papers are published every year
- Human’s reading ability keeps almost the same across years



[1] https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html
[2] Landhuis, E. (2016). Scientific literature: Information overload. Nature, 535(7612), 457-458.

Challenges in Fine-grained Chemical Entity Extractions

- Few benchmarks are publicly available



Datasets

9,274 machine learning datasets

Share your dataset with the ML community!

chemical

Best match

Filter by Modality

Texts	5
Biomedical	1
3D	0
3d meshes	0
6D	0
Actions	0

Filter by Task (clear)

Named Entity Recognition (NER)	×
Graph Classification	8
Drug Discovery	7
Relation Extraction	5

5 dataset results for chemical AND Named Entity Recognition (NER) ×

BC4CHEMD (BioCreative IV Chemical compound and drug name recognition)
Introduced by Krallinger et al. in The CHEMDNER corpus of **chemicals** and drugs and its annotation principles BC4CHEMD is a collection of 10,000 PubMed abstracts that contain a total of...
4 PAPERS • 3 BENCHMARKS

BC5CDR (BioCreative V CDR corpus)
BC5CDR corpus consists of 1500 PubMed articles with 4409 annotated **chemicals**, 5818 diseases and 3116 **chemical**-disease interactions.
170 PAPERS • 7 BENCHMARKS

BC7 NLM-Chem (BioCreative VII NLM-Chem)
Full-text **chemical** identification and indexing in PubMed articles. Identifying named entities is an important building block for many complex knowledge extraction tasks. **Chemical** entities...
2 PAPERS • 3 BENCHMARKS

BioRED
BioRED is a first-of-its-kind biomedical relation extraction dataset with multiple entity types (e.g. gene/protein, disease, **chemical**) and relation pairs (e.g. gene–disease; **chemical–chemi-**...
13 PAPERS • 3 BENCHMARKS

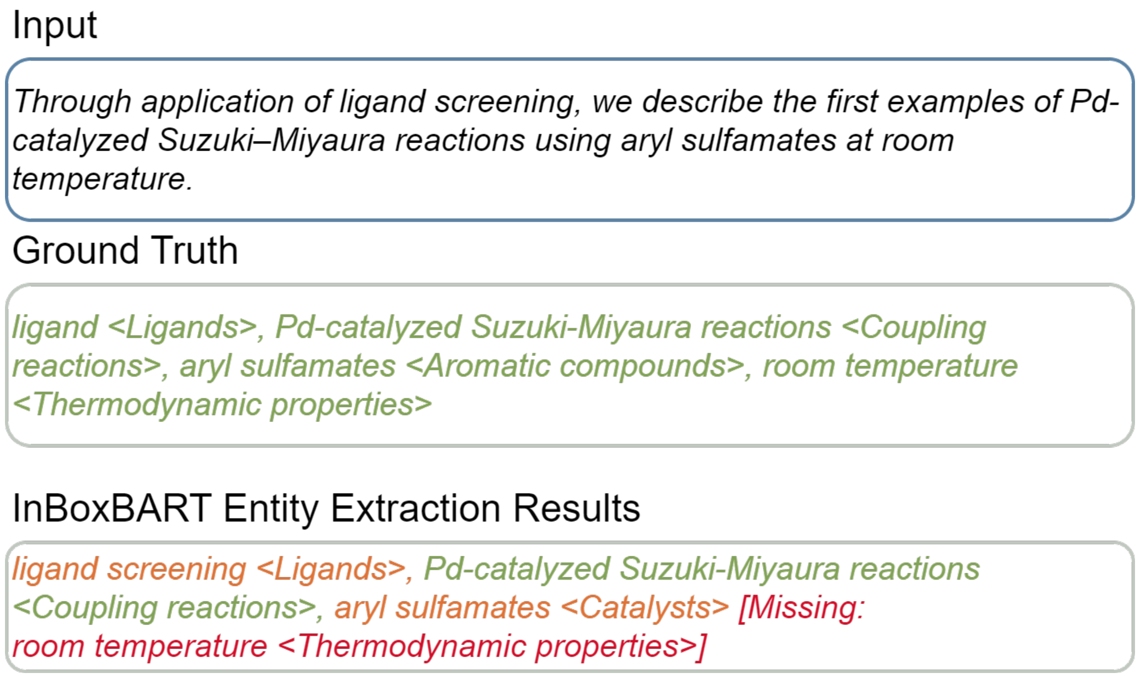
JNLPBA
...From this search 2,000 abstracts were selected and hand annotated according to a small taxonomy of 48 classes based on a **chemical** classification. 36 terminal classes were used to anno...
17 PAPERS • 3 BENCHMARKS

https://paperswithcode.com/datasets?q=chemical&v=lst&o=match&task=named-entity-recognition-ner&page=1

Challenges in Fine-grained Chemical Entity Extractions

✗ Missing mentions

- Scientific documents contain more entities per sentence compared the sentence in general domain (3.1 in ChemNER+ vs 1.5 in CONLL2003)



Input

Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.

Ground Truth

ligand <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, aryl sulfamates <Aromatic compounds>, room temperature <Thermodynamic properties>

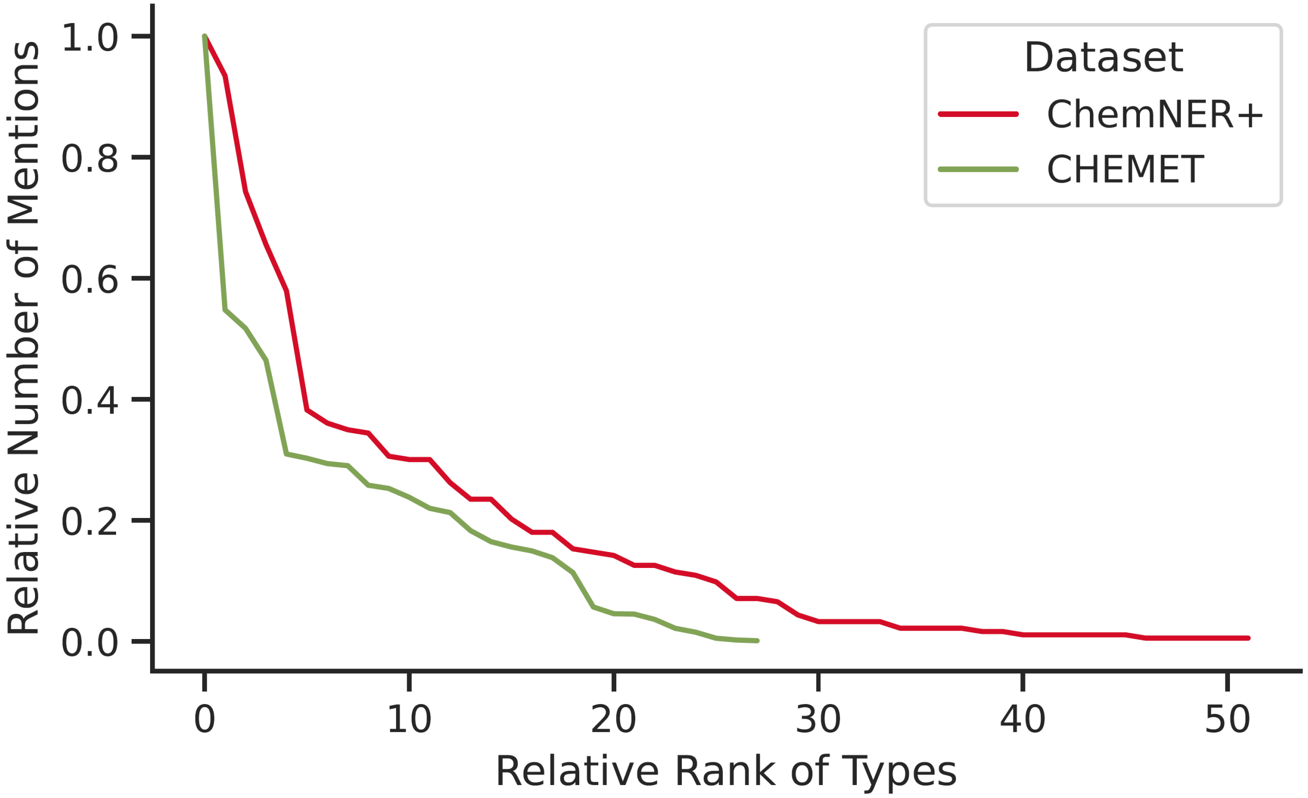
InBoxBART Entity Extraction Results

ligand screening <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, aryl sulfamates <Catalysts> [Missing: room temperature <Thermodynamic properties>]

Challenges in Fine-grained Chemical Entity Extractions

✗ Incorrect long-tail predictions

- Long-tail problems are more prevalent in scientific domain compared to general domain



Goal of Information Extraction

- If the model extracts knowledge precisely, readers should be able to reconstruct the original sentence using the extraction results precisely and faithfully
- Self-validation module to reconstruct the original sentences based on entity extraction results

Input

Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.

Ground Truth

ligand <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, aryl sulfamates <Organic compounds>, room temperature <Thermodynamic properties>

Sentence Reconstructed from Ground Truth

Ligands play a crucial role in Pd-catalyzed Suzuki-Miyaura reactions, which are coupling reactions that enable the synthesis of diverse organic compounds such as aryl sulfamates at room temperature, exploiting their favorable thermodynamic properties.

InBoxBART Entity Extraction Results

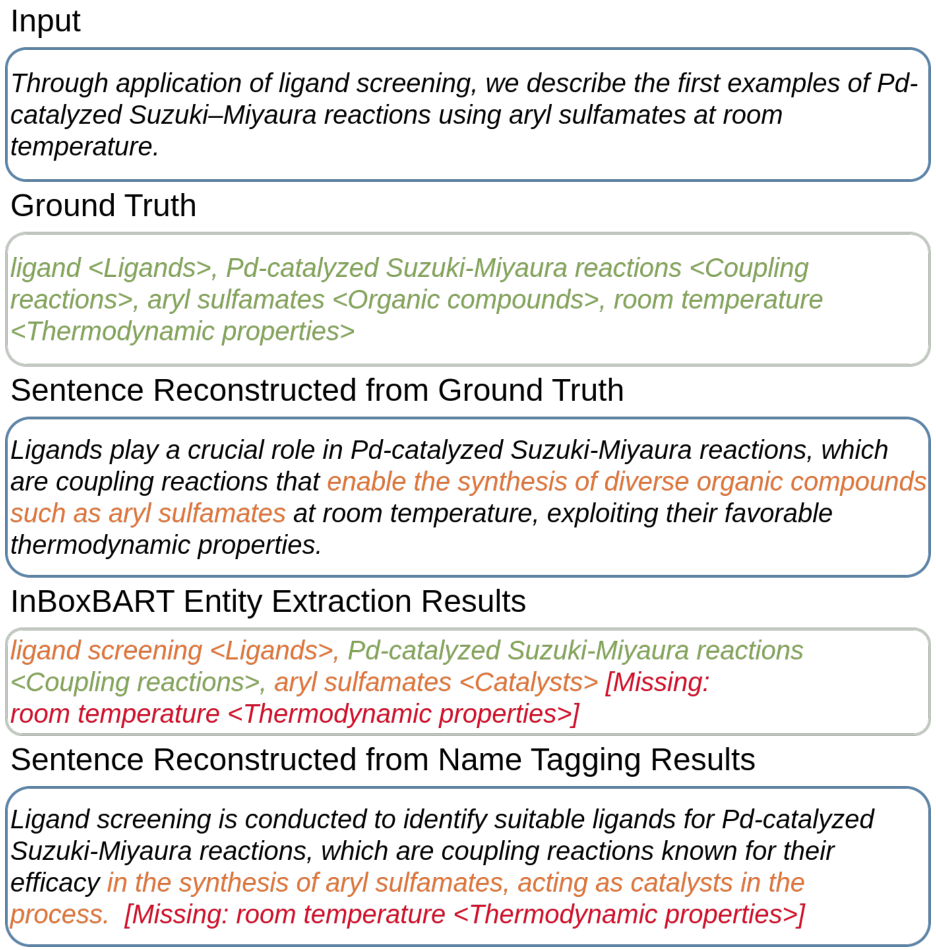
ligand screening <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, aryl sulfamates <Catalysts> [Missing: room temperature <Thermodynamic properties>]

Sentence Reconstructed from Name Tagging Results

Ligand screening is conducted to identify suitable ligands for Pd-catalyzed Suzuki-Miyaura reactions, which are coupling reactions known for their efficacy in the synthesis of aryl sulfamates, acting as catalysts in the process. [Missing: room temperature <Thermodynamic properties>]

Seq2Seq Entity Extraction

- Excessively copy from original sentence
- A new entity decoder contrastive loss to control the mention spans



Input

Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.

Ground Truth

ligand <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, aryl sulfamates <Organic compounds>, room temperature <Thermodynamic properties>

Sentence Reconstructed from Ground Truth

Ligands play a crucial role in Pd-catalyzed Suzuki-Miyaura reactions, which are coupling reactions that enable the synthesis of diverse organic compounds such as aryl sulfamates at room temperature, exploiting their favorable thermodynamic properties.

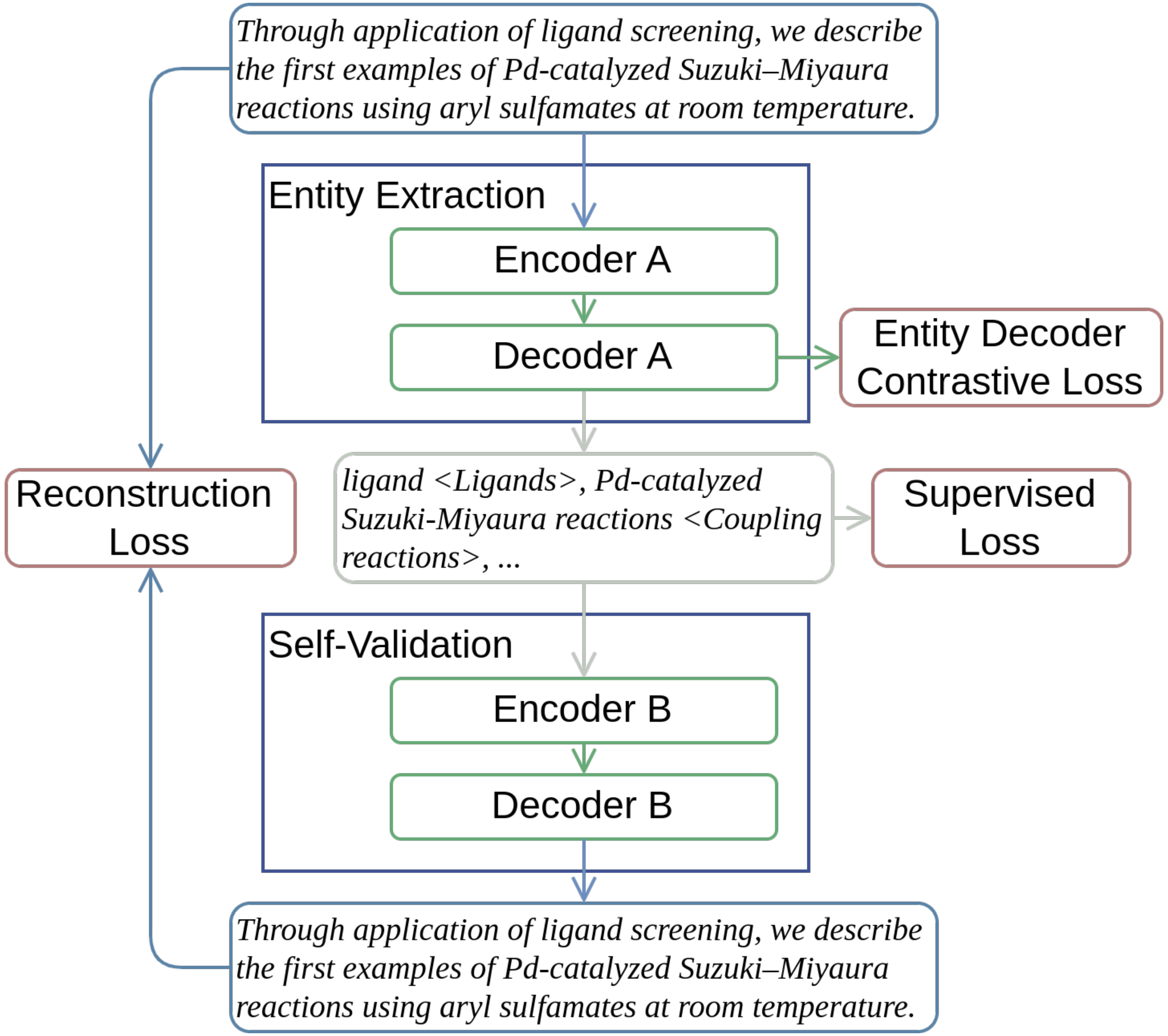
InBoxBART Entity Extraction Results

ligand screening <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, aryl sulfamates <Catalysts> [Missing: room temperature <Thermodynamic properties>]

Sentence Reconstructed from Name Tagging Results

Ligand screening is conducted to identify suitable ligands for Pd-catalyzed Suzuki-Miyaura reactions, which are coupling reactions known for their efficacy in the synthesis of aryl sulfamates, acting as catalysts in the process. [Missing: room temperature <Thermodynamic properties>]

Overview

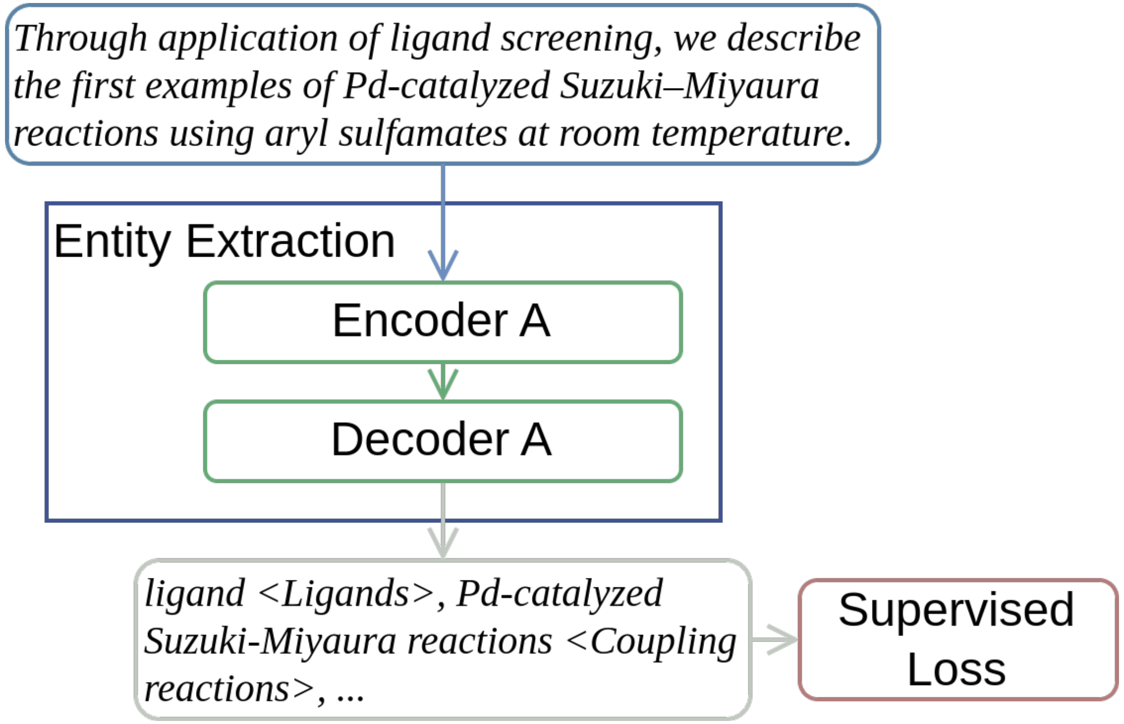


Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.

ligand <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, ...

Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.

Entity Extraction Module



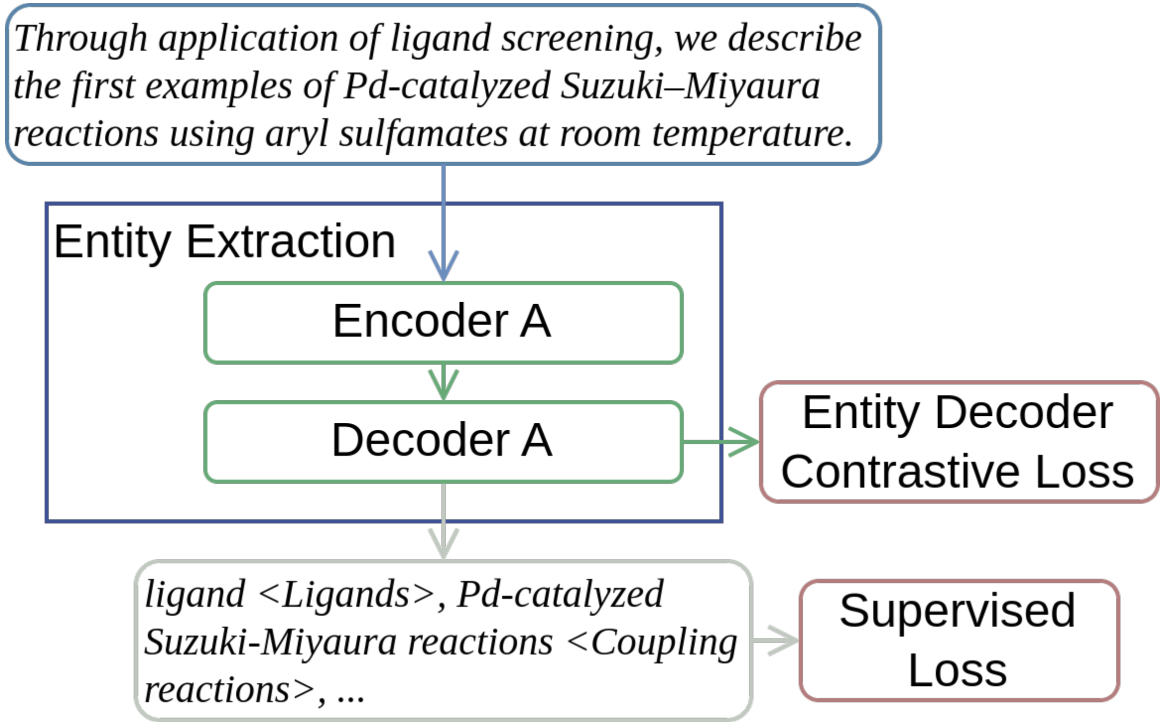
Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.

ligand <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, ...

- Use the state-of-the-art coarse-grained chemical entity extractor InBoXBART as backbone to extract entities from source document

Parmar, M., Mishra, S., Purohit, M., Luo, M., Murad, M. H., & Baral, C. (2022). In-boxbart: Get instructions into biomedical multi-task learning. NAACL 2022 Findings

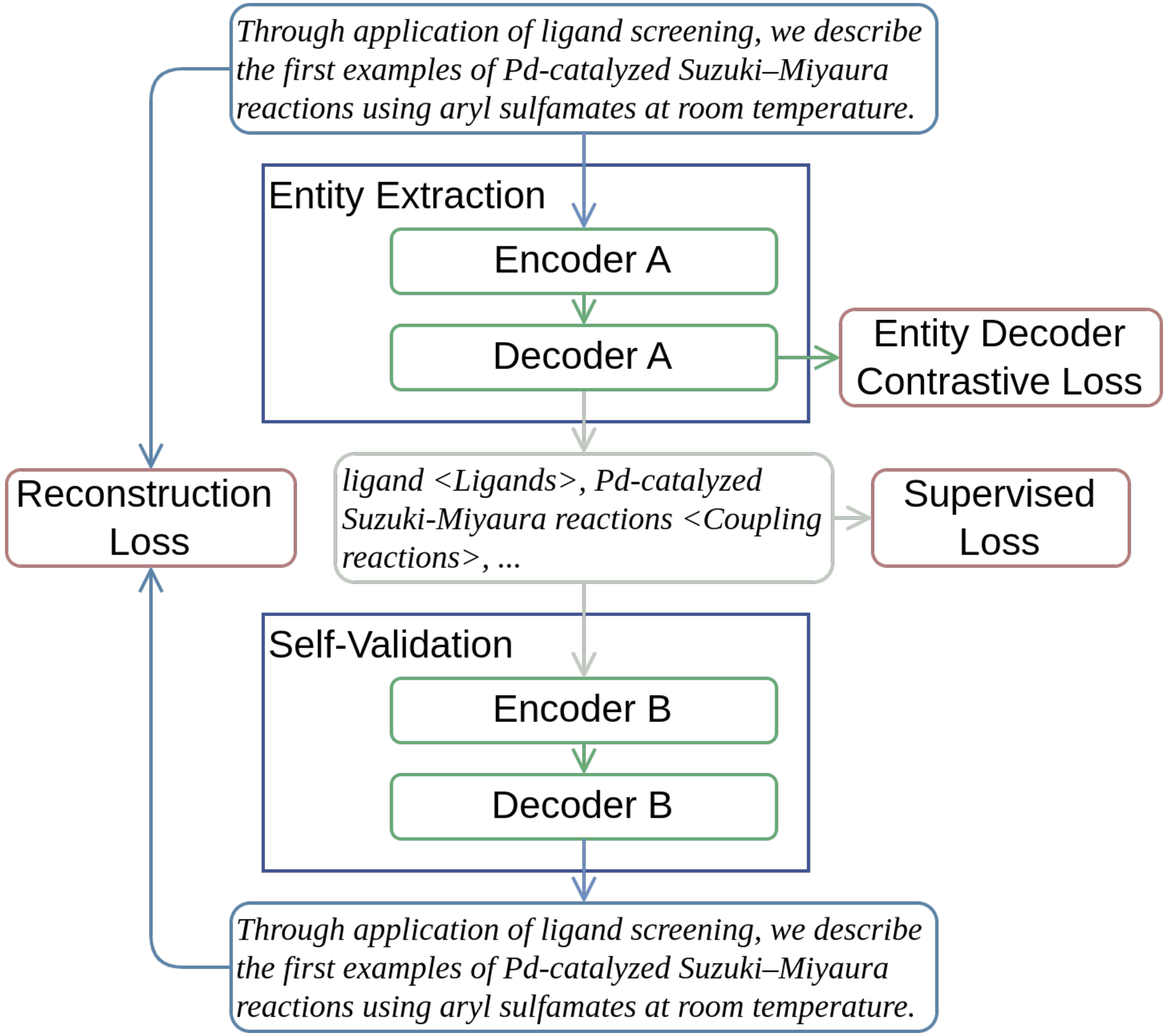
Contrastive Entity Decoding


Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.
Entity Extraction
Encoder A
Decoder A
Entity Decoder Contrastive Loss
ligand <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, ...
Supervised Loss

- Construct decoder negatives by combining mentions with surrounding unrelated contexts

Parmar, M., Mishra, S., Purohit, M., Luo, M., Murad, M. H., & Baral, C. (2022). In-boxbart: Get instructions into biomedical multi-task learning. NAACL 2022 Findings

Self-validation Module


Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.
Entity Extraction
Encoder A
Decoder A
Entity Decoder Contrastive Loss
Reconstruction Loss
ligand <Ligands>, Pd-catalyzed Suzuki-Miyaura reactions <Coupling reactions>, ...
Supervised Loss
Self-Validation
Encoder B
Decoder B
Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.

- Self-validation module takes in the entity extraction results and generates a reconstructed input sentence
- Use Gumbel-Softmax (GS) estimator to avoid the non-differentiable issue in explicit decoding

Parmar, M., Mishra, S., Purohit, M., Luo, M., Murad, M. H., & Baral, C. (2022). In-boxbart: Get instructions into biomedical multi-task learning. NAACL 2022 Findings

Benchmark Dataset

- ChemNER+
  - Based on available sentences from ChemNER (Wang et al., 2021)
  - Annotated by two Chemistry Ph.D. students
  - Covering 59 fine-grained chemistry types with 742 sentences
- CHEMET (Sun et al., 2021)
  - Consisting of 30 fine-grained organic chemical types

Dataset	Split	# Pair	Avg. Token	Avg. Entity
ChemNER+	Train	542	32.9	3.10
	Valid	100	39.9	4.57
	Test	100	39.4	4.61
CHEMET	Train	6,561	37.8	1.57
	Valid	520	31.6	2.15
	Test	663	36.6	1.95

Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021a. ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision. EMNLP 2021

C. Sun, W. Li, J. Xiao, N. Parulian, C. Zhai, and H. Ji. 2021. Fine-grained chemical entity typing with multimodal knowledge representation. BIBM 2021

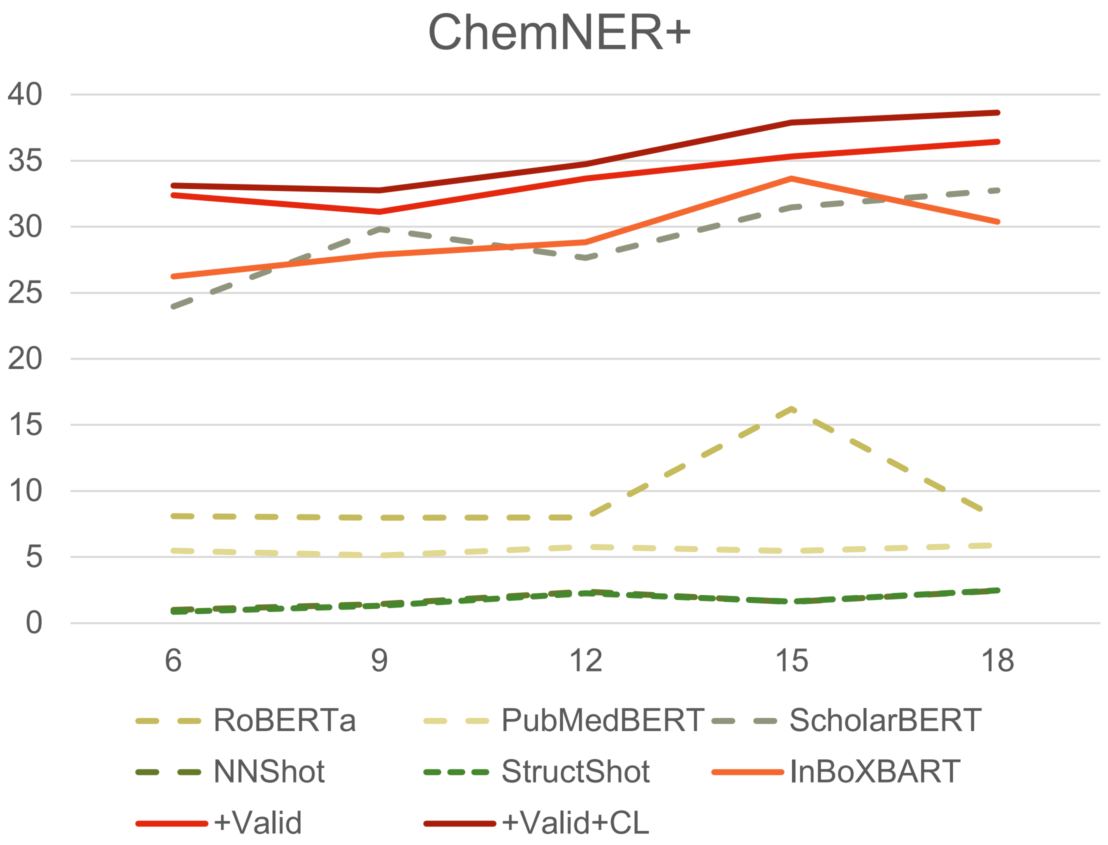
Few-shot Setup

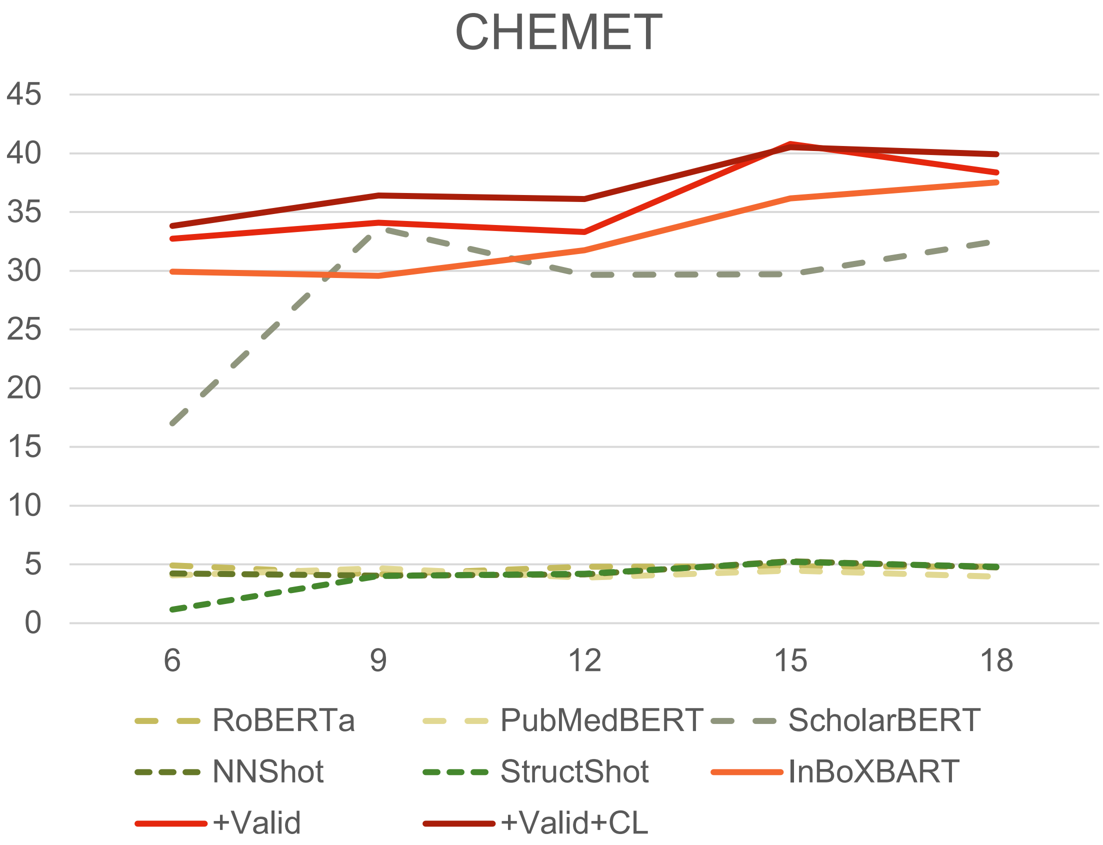
- Set the number of maximum entity mentions *k* for the most frequent entity type
- Randomly sample other types and ensure that the distribution remains the same
- Choose the values *6, 9, 12, 15, 18* as the potential maximum entity mentions for *k*

Baselines

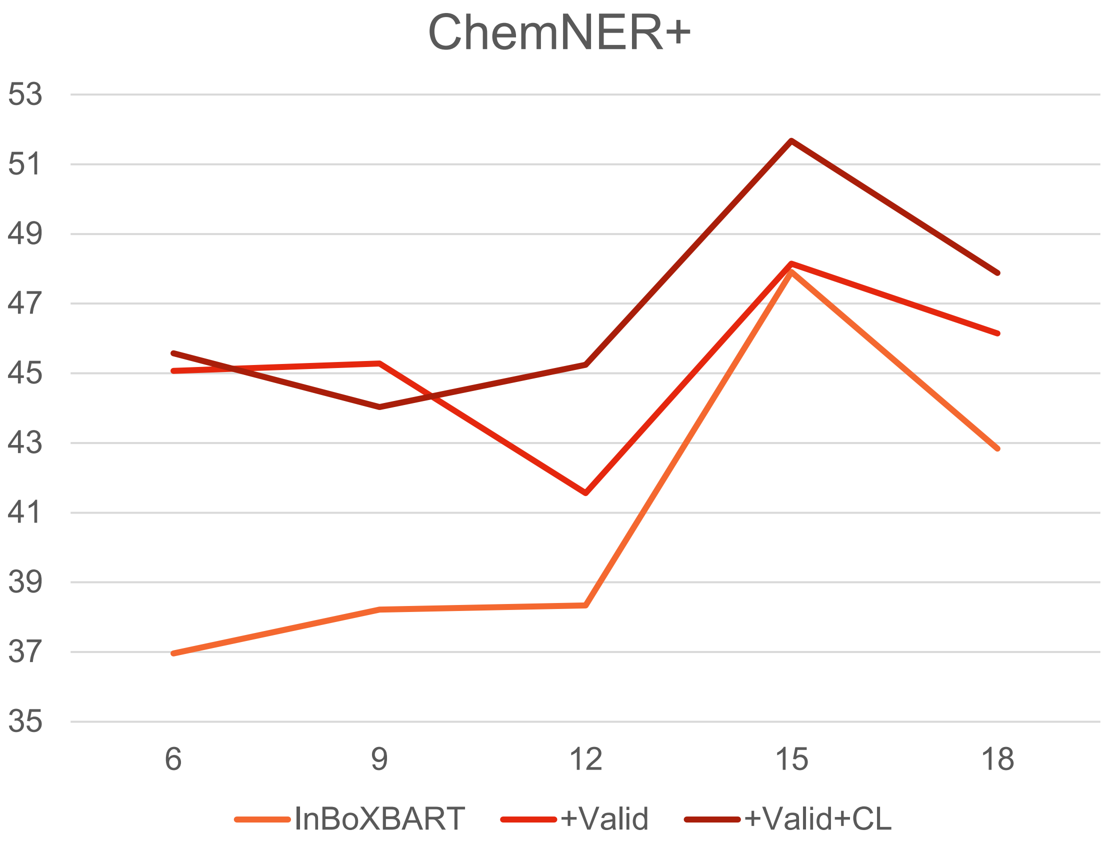
- State-of-the-art pretrained encoder-based models
- Few-shot baselines
- Ablation baselines
  - *Valid* is models with self-validation module
  - *CL* is models with contrastive entity decoding

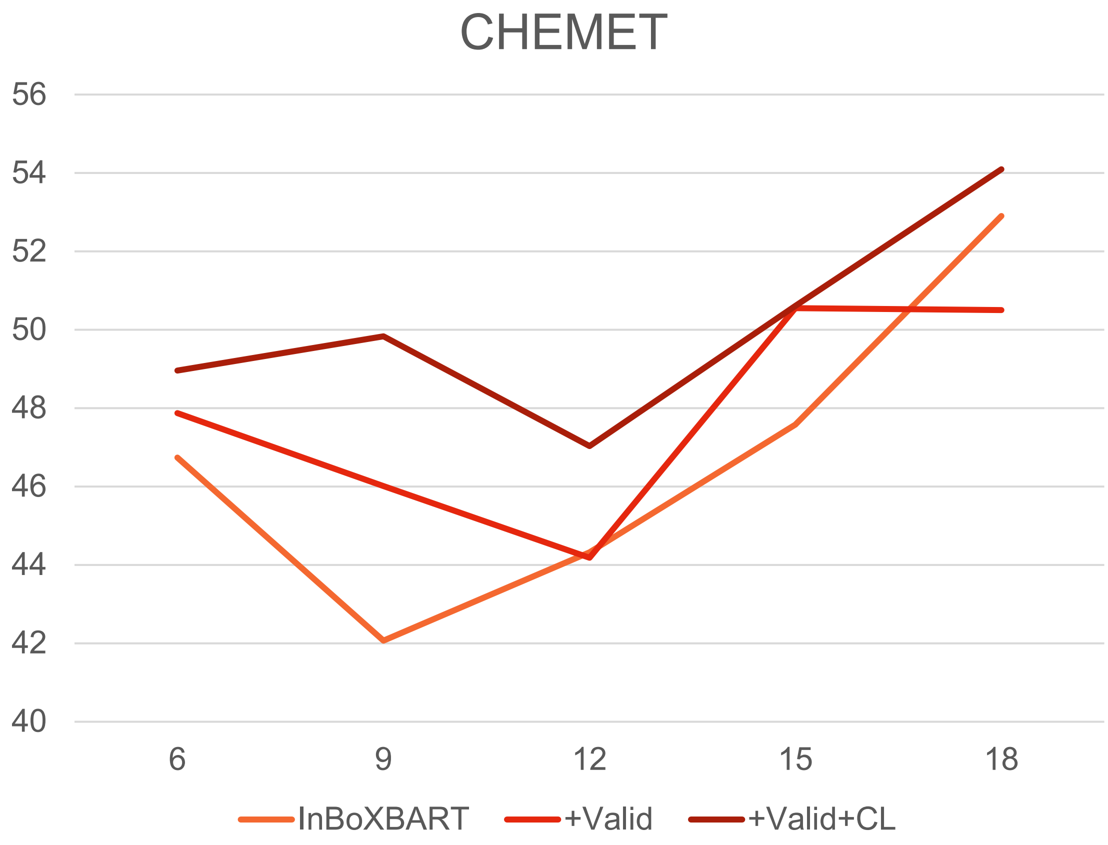
k-shot Micro-F1 Results



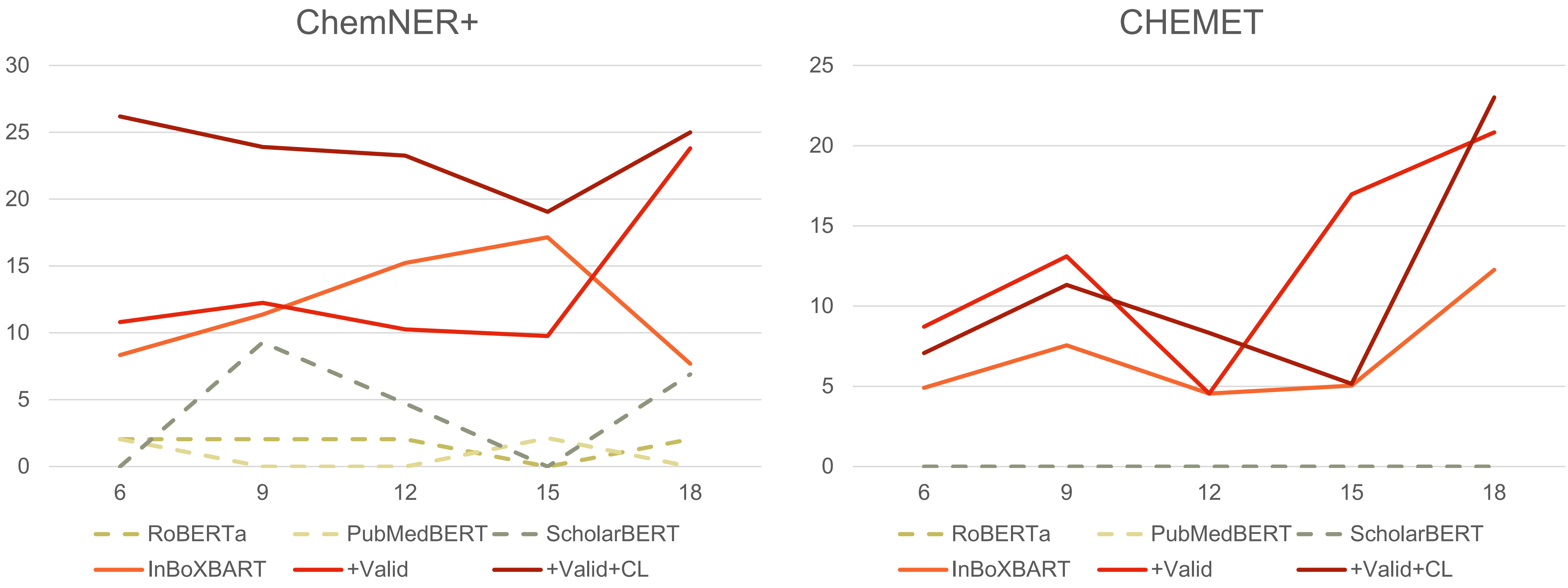


k-shot Mention Micro-F1 Results





k-shot micro-F1 for Long-tail Entity Results



Qualitative Analysis

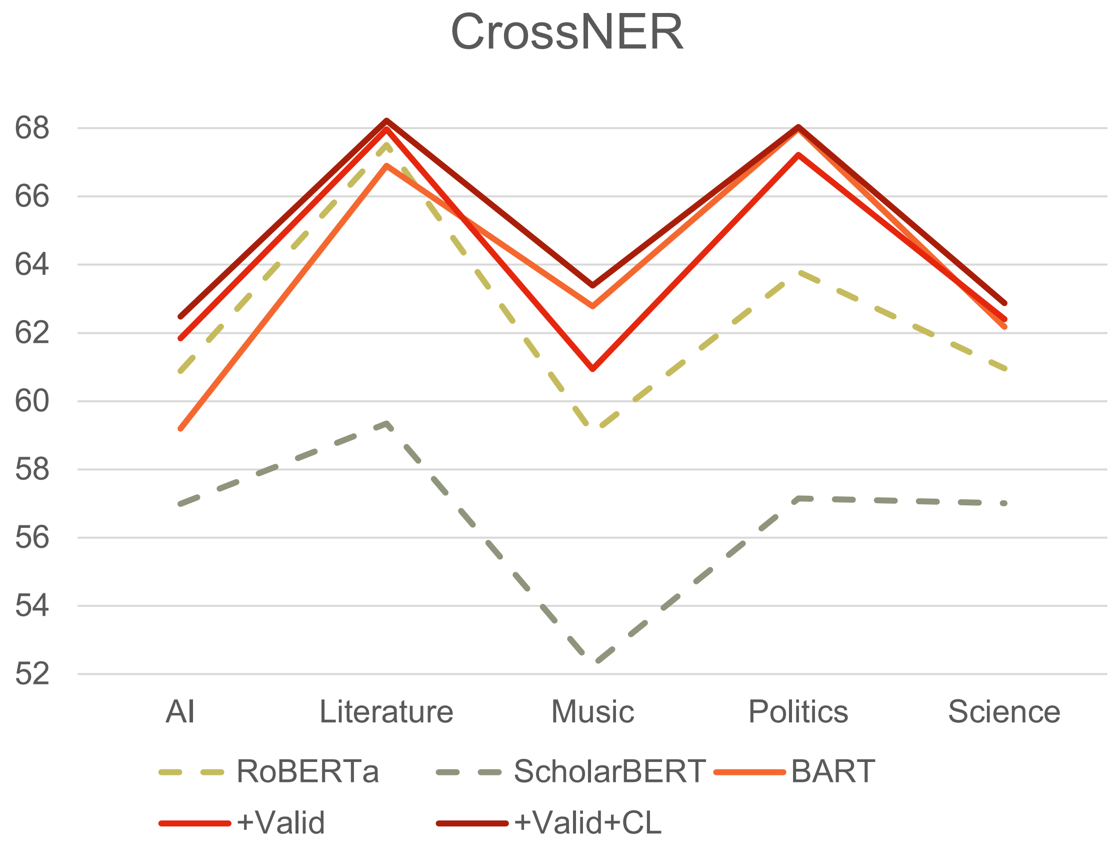
InBoxBART	Several *cyclophanes*, *polycycles*, … have been synthesized by employing a combination of *Suzuki cross-coupling and metathesis* $_{\text{Coupling reactions}}$.
+ Valid	Several *cyclophanes* $_{\text{Heterocyclic Compounds}}$, *polycycles*, … have been synthesized by employing a combination of *Suzuki cross-coupling and metathesis* $_{\text{Organic reactions}}$.
+ Valid + CL	Several *cyclophanes* $_{\text{Heterocyclic Compounds}}$, *polycycles* $_{\text{Biomolecules}}$, … have been synthesized by employing a combination of *Suzuki cross-coupling* $_{\text{Coupling reactions}}$ and *metathesis* $_{\text{Chemical properties}}$.
Ground Truth	Several *cyclophanes* $_{\text{Aromatic Compounds}}$, *polycycles* $_{\text{Organic polymers}}$, … have been synthesized by employing a combination of *Suzuki cross-coupling* $_{\text{Coupling reactions}}$ and *metathesis* $_{\text{Substitution reactions}}$.

Qualitative Analysis

InBoxBART	…with the advantages of *asymmetric catalysis* (step and atom economy) in a rare example of an *enantioselective cross coupling of a racemic electrophile bearing an oxygen leaving group* Catalysis … the identification of a *highly enantioselective process*.
+ Valid	…with the advantages of asymmetric catalysis (step and atom economy) in a rare example of an *enantioselective cross coupling of a racemic electrophile bearing an oxygen leaving group* Organometallic compounds… the identification of a *highly enantioselective process*.
+ Valid + CL	…with the advantages of *asymmetric catalysis* Catalysis (step and atom economy) in a rare example of an *enantioselective cross coupling* of a *racemic electrophile* bearing an *oxygen leaving group* Functional groups … the identification of a highly *enantioselective process* Chemical properties.
Ground Truth	…with the advantages of *asymmetric catalysis* Catalysis (step and atom economy) in a rare example of an *enantioselective cross coupling* Coupling reactions of a *racemic electrophile* Organic compounds bearing an *oxygen leaving group* Functional groups… the identification of a *highly enantioselective process* Catalysis.

Case Study

Domain	Train	Valid	Test	# Type	Avg. Token	Avg. Entity
AI	100	350	430	14	31.5	4.42
Lit.	99	400	416	12	37.6	5.39
Music	100	380	465	13	41.4	7.05
Politics	200	541	654	9	43.5	6.46
Science	200	450	543	17	35.8	5.62



Conclusion

- Propose ***two few-shot chemical fine-grained entity extraction datasets***, based on human-annotated ChemNER+ and CHEMET.
- Propose a new framework to address the **mention coverage and long-tailed entity type problems** in chemical fine-grained entity extraction tasks *through **a novel self-validation module*** and ***a new entity extractor decoder contrastive objective***

Code and Data are public at: github.com/EagleW/Chem-FINESE

Thank you!

Code and Data are public at: github.com/EagleW/Chem-FINESE

BLENDER | Cross-source Information Extraction Lab