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

Xuan Wang*, Yu Zhang*, Qi Li, Cathy H. Wu, Jiawei Han

University of Illinois at Urbana-Champaign, IL, USA

University of Delaware, DE, USA

Outline

- Introduction
- Framework
- Evaluation
- Conclusion

Nested Entity Structures

- PMID 10190572:
  - "… although each of the agents alone caused only slight increase in the [[alanine]$_{CHEMICAL}$ aminotransferase]$_{PROTEIN}$ activity."
  - PubTator recognizes "***alanine***" as a CHEMICAL but misses "***alanine aminotransferase***" as a PROTEIN.

Why are Nested Entities Important?

- The nested entity structure is quite common in biomedical literature.
  - 17% of the entities in the GENIA dataset are embedded with another entity.
- Many downstream tasks require us to detect not just the inner-most entity.
  - E.g., PMID 9256163:
  - "… **Forskolin** (10 microM), …, also increased **renin mRNA release**."
  - NER (PubTator): "CHEMICAL_Forskolin (10 microM), …, also increased GENE_renin mRNA release."
  - Result of Relation Extraction: (Forskolin, increase, renin)
  - Correct Tuple: (Forskolin, increase, renin mRNA release)
  - Incompleteness in NER causes errors in RE.

Previous Studies

- “Flat” BioNER ([1], [2], [3], etc.)
  - Common sequence modeling frameworks cannot detect entities with overlapping tokens.
- Supervised Nested BioNER ([4], [5], [6], etc.)
  - Need massive training data
  - Hard to transfer to new entity types (e.g., the GENIA corpus only contains genes/protein, DNA, RNA, cell lines and cell types. What if we need chemicals and diseases?)

[1] PubTator: a web-based text mining tool for assisting biocuration. NAR 2013

[2] TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics 2016.

[3] Cross-type biomedical named entity recognition with deep multi-task learning. Bioinformatics 2019 (To appear).

[4] Nested named entity recognition. EMNLP 2009.

[5] Labeling gaps between words: Recognizing overlapping mentions with mention separators. EMNLP 2017.

[6] Nested named entity recognition revisited. NAACL 2018.

This Paper

- Nested BioNER **with very weak supervision**
- Idea: Nested structure as a **pattern-level** phenomenon

CHEMICAL aminotransferase = PROTEIN
GENE mRNA release = PROCESS

- **Unsupervised** pattern extraction
- **Few-shot** nested entity recognition for each type

Outline

- Introduction
- Framework
- Evaluation
- Conclusion

Framework Overview

Input Corpus

ID	Sentence
1	TERT encodes the reverse transcriptase subunit of human telomerase.
2	The FGFR-2 receptor is a membrane-spanning tyrosine kinase.
3	Each of the agents alone caused only slight increase in the alanine aminotransferase activity.
…	…

Flat Entity Recognition

ID	Sentence
1	GENE_TERT encodes the reverse transcriptase subunit of SPECIES_human telomerase.
2	The GENE_FGFR-2 receptor is a membrane-spanning CHEMICAL_tyrosine kinase.
3	Each of the agents alone caused only slight increase in the CHEMICAL_alanine aminotransferase activity.
…	…

Meta-Pattern Extraction

Meta-Pattern
SPECIES telomerase
CHEMICAL kinase
CHEMICAL aminotransferase
GENE level
male SPECIES
several DISEASE
…

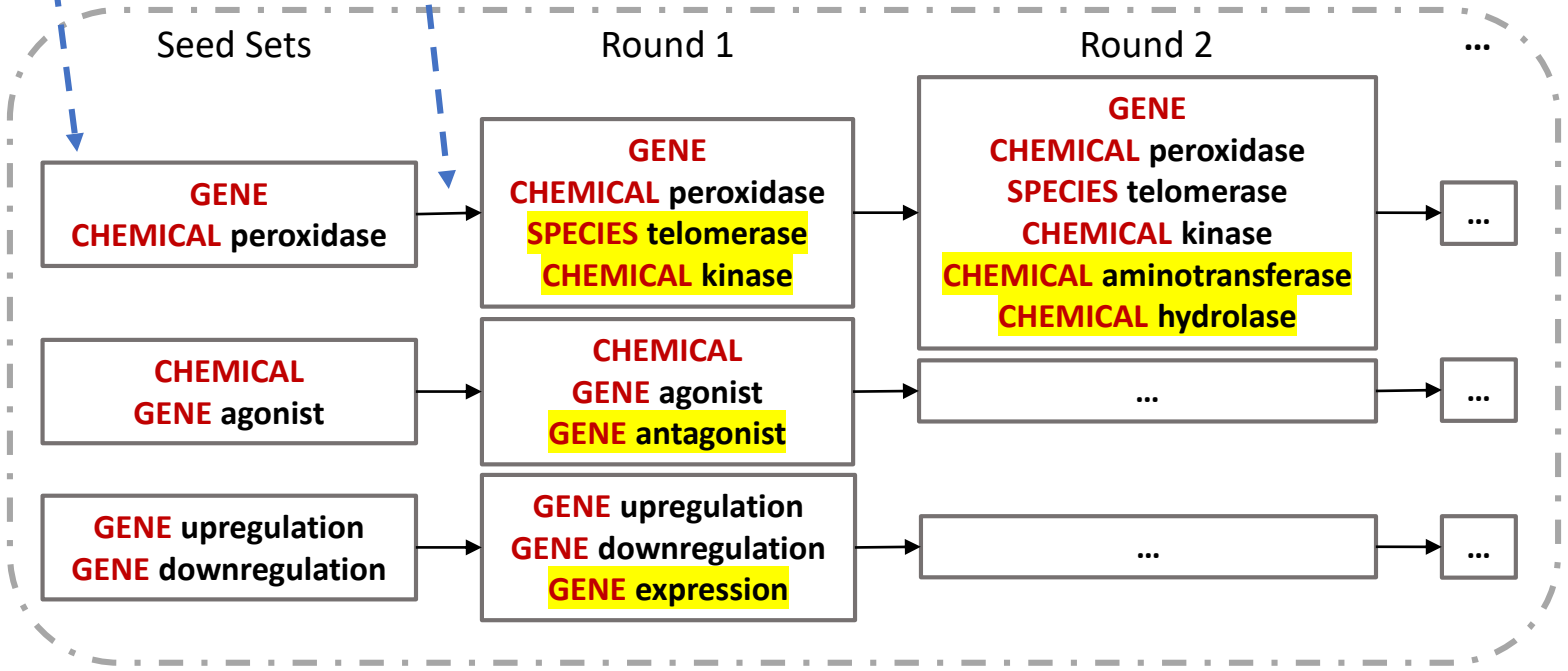
Nested Entity Recognition

ID	Sentence
1	TERT encodes the reverse transcriptase subunit of GENE_human_telomerase.
2	The FGFR-2 receptor is a membrane-spanning GENE_tyrosine_kinase.
3	Each of the agents alone caused only slight increase in the GENE_alanine_aminotransferase activity.
…	…

Weak Supervision

Context Information

Pattern Expansion



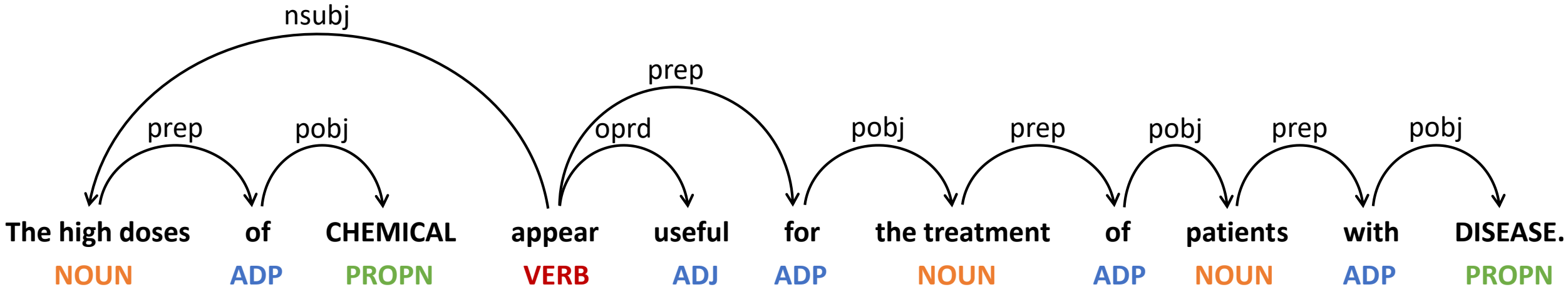
Meta-Pattern Extraction

- What are meta-patterns?
- A mixed sequence of entity types and non-type words in the corpus
  - E.g., pattern: CHEMICAL **aminotransferase**
    instance: CHEMICAL = *alanine, aspartate, tyrosine*…
- Each instance of a meta-pattern has a natural nested structure.
- A meta-pattern has the aggregated context information of all of its instances, which helps us learn its semantics in a more accurate way.

- **Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature. Tomorrow noon, Section 31**

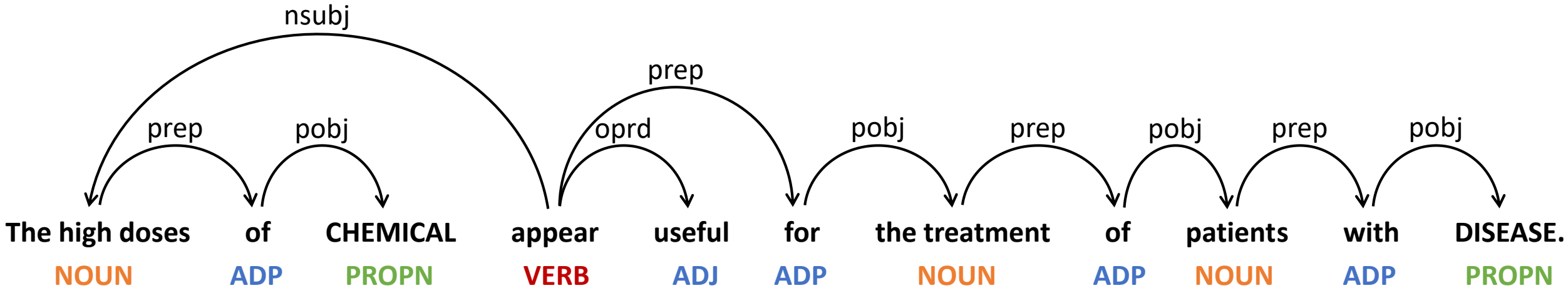
Meta-Pattern Extraction

- How to find quality meta-patterns?
- **Frequency**: Appear more than t times in the corpus
- **Informativeness**: Either a single entity type (e.g., **DISEASE**) or a phrase with one entity type and at least one stopwords (e.g., *patients with* **DISEASE**)
- **Syntactic Completeness**: The tokens form a connected subgraph in the dependency parsing tree. (**CHEMICAL** *appear useful* vs. *patients with* **DISEASE**)



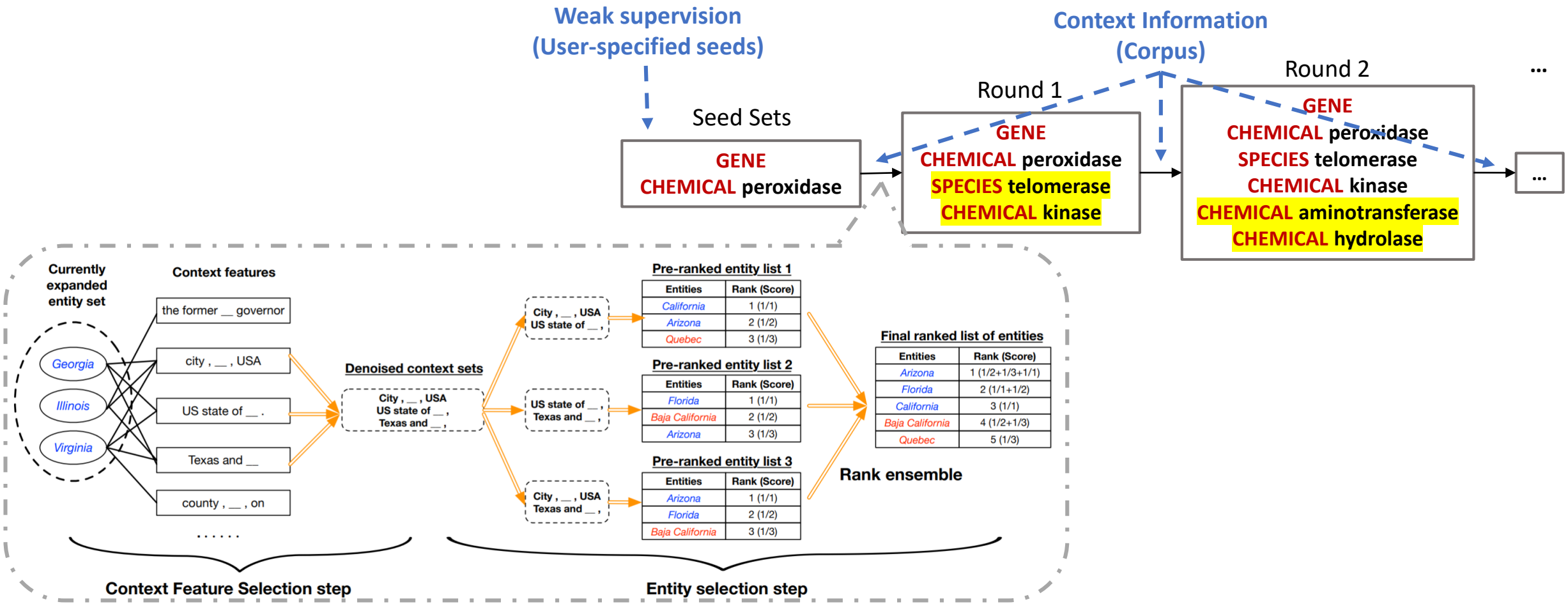
Meta-Pattern Extraction

- How to find quality meta-patterns?
- **Semantic Completeness**: For NER, extracted patterns should be complete noun phrases.
  - Chunking: Iteratively cutting the tree at nouns (i.e., **NOUN** & **PROPN**). Each noun serves as a leaf of the current chunk as well as the root of the next chunk.
  - A semantic complete pattern should be a complete chunk.



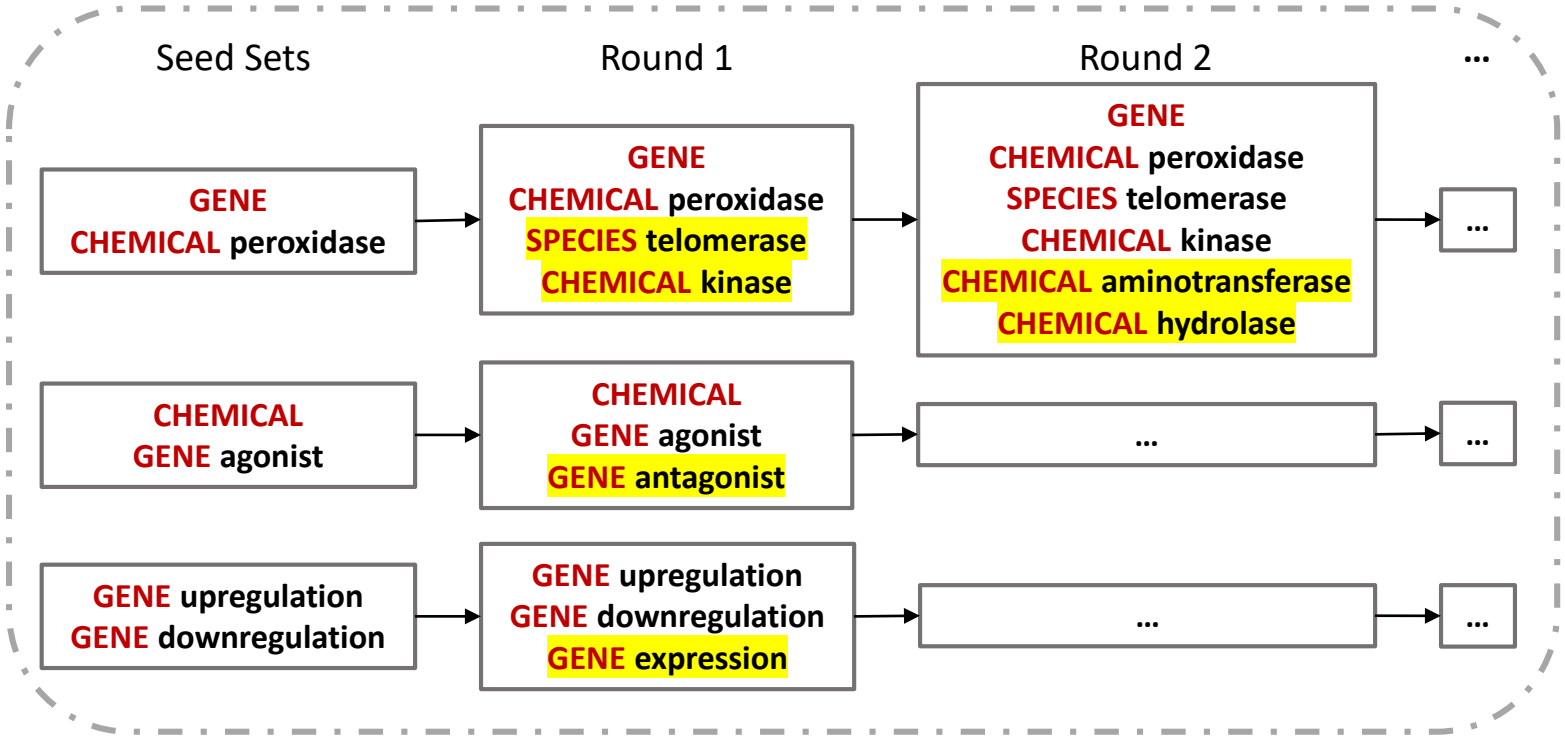
Weakly-supervised Pattern Expansion

- Finding new patterns with few user-specified seeds
- Method: **SetExpan** (Shen et al., ECML-PKDD 2017): Skip-gram + Rank Ensemble



Expanding Multiple Sets Simultaneously

- SetExpan essentially combines frequency and context similarity.
- Unlike entities, some meta-patterns may be extremely frequent (e.g., "CHEMICAL")
- Utilizing the mutual exclusiveness of seed sets.



Outline

- Introduction
- Framework
- Evaluation
- Conclusion

Experiments

- Dataset: A subset of PubMed abstracts, selected using tuples in CTD

BASIC STATISTICS OF THE SUBSET CORPUS.

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

- Baselines:
  - **Embedding**: Using Word2Vec to learn embeddings of meta-patterns, and then searching nearest neighbors for seed patterns
  - **SetExpan**: Expanding different types one by one. No mutual exclusiveness.

Pattern-Level Task: Meta-Pattern Extraction

Embedding

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

SetExpan

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

PENNER

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

Entity-level Task: Nested NER

- "Precision": NDCG of the ranking list of expanded entities

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad \mathrm{IDCG_p} = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad \mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

	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**

- "Recall": Number of correct instances

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

- **Embedding** does not consider frequency. Infrequent patterns may have inaccurate embeddings.
- **SetExpan** does not exploit mutual exclusiveness. Extremely frequent patterns may cause semantic drift during expansion.

Detecting New Entity Types

- Detecting Biological Process and Treatment entities using only two seeds!

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	-

- Fine-grained flat NER may further improve the performance.
  - E.g., pattern1: CHEMICAL **treatment** (Treatment)
    instance: CHEMICAL = *resveratrol, simvastatin, quercetin, …* (drug)
    pattern2: CHEMICAL **exposure** (**symptom rather than treatment**)
    instance: CHEMICAL = *lead, mercury, hydrofluoric acid, …* (toxic)

Case Study

- Nested Structure + New Entity Types

	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**]$_{\text{CHEMICAL}}$]$_{\text{TREATMENT}}$ ...

Outline

- Introduction
- Framework
- Evaluation
- Conclusion

Conclusion

- Framework
  - Taking a corpus pre-tagged by any flat NER tool as input
  - Unsupervised meta-pattern extraction
  - Few-shot pattern expansion
- Evaluation
  - Outperforming baselines in both meta-pattern extraction and nested NER
  - Detecting new entity types with few seeds
  - Improving annotation results over PubTator
- Future Work
  - To use meta-patterns to find biomedical entity naming principles
  - To use nested structures to help meta-pattern discovery in return

Thank you! Questions?