Text Mining IV: Meta-Pattern Discovery for NER and Others

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

MARCH 12, 2020
Meta-Pattern Discovery for Information Extraction and Text Mining

MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora

TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction

WW-PIE: Pattern Discovery for Wide-Window Open Information Extraction

PENNER: Pattern-Enhanced Nested Named Entity Recognition

Summary
MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora

Meng Jiang, Jingbo Shang, Xiang Ren, Taylor Cassidy, Lance Kaplan, Timothy Hanratty, and Jiawei Han, “MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora”, KDD 2017

Motivation: Given a sentence in a large corpus, “President Blaise Compaoré’s government of Burkina Faso was founded...”, ...

We may find:

- Blaise Compaoré: $PERSON.POLITICIAN
  - age: 65
- Burkina Faso: $COUNTRY
  - president
  - population: 17 million

Attribute Discovery: Two tasks

Task 1: (entity, attribute name, attribute value)
- (Burkina Faso, president, Blaise Compaoré)
- (Burkina Faso, population, 17 million)
- (Blaise Compaoré, age, 65)

Task 2: (entity type, attribute name)
- ($COUNTRY, president)
- ($COUNTRY, population)
- ($PERSON, age)

Instance-level

Type-level
Previous Work on Finding E-A-V and Typed Patterns

- Task 1: Finding E-A-V at the Instance Level
  - Stanford OpenIE [ACL’15], Al²’s Open IE-Ollie [EMNLP’12]
  - Learn syntactic and lexical patterns of expressing relations
  - Input: “President Blaise Compaoré’s government of Burkina Faso was founded...”
  - Output: ⟨President Blaise Compaoré, have, government of Burkina Faso⟩

- Task 2: Finding Typed Patterns
  - Google’s Biperpedia+ARI [VLDB’14, WWW’16], ReNoun [EMNLP’15]:
    - “Barack Obama, President of U.S.,” → “O, A of S,”, “S A O”

Query log: Highly constrained and unavailable

- Input: “...Sunday night, Burkina Faso...” and the “A, E” pattern
- Output: ⟨$COUNTRY, Sunday night⟩
Our Meta-Pattern Methodology

Generate patterns with massive instances in the data

Our Meta-Pattern Methodology

Generate patterns with massive instances in the data

(#1) “President Blaise Compaoré’s government of Burkina Faso was founded ...”
(#2) “President Barack Obama’s government of U.S. claimed that...”
(#3) “U.S. President Barack Obama visited ...”

Meta pattern segmentation

Meta patterns:

president $PERSON POLITICIAN’s government of $LOCATION COUNTRY was founded...

$LOCATION COUNTRY president $PERSON POLITICIAN ...

$COUNTRY, {president}, $POLITICIAN

Adjust types for appropriate granularity

Generate massive triples by matching the meta patterns

Joint extraction

Group synonymous patterns by massive triples

No heavy annotation required
No domain knowledge required
No query log required if we can recognize and type the entities in the same manner...
Pattern Discovery by Phrase Mining and Entity Typing

“President Blaise Compaoré’s government of Burkina Faso was founded ...”

Phrase mining (SegPhrase and AutoPhrase)

“president blaise_compaoré’s government of burkina_faso was founded ...”

Entity recognition and typing with Distant Supervision (ClusType)

“president $PERSON’s government of $LOCATION was founded ...”

Fine-grained typing (PLE by Ren et al. KDD’16)

“president $PERSON.POLITICIAN’s government of $LOCATION.COUNTRY was founded ...”
Meta-Pattern Quality Assessment and Segmentation

A rich set of features:
✓ Frequency
✓ Concordance: “$PERSON’s wife”
✓ Completeness: “$COUNTRY president” vs. “$COUNTRY president $POLITICIAN”
✓ Informativeness: “$PERSON and $PERSON” vs. “$PERSON’s wife, $PERSON”

Regression Q(.): random forest with only 300 labels

Q(.)↑

president $PERSON.POLITICIAN ’s government of $LOCATION.COUNTRY
Grouping Synonymous Patterns

($COUNTRY, president, $POLITICIAN)

$COUNTRY president $POLITICIAN

president $POLITICIAN of $COUNTRY

(United States, Barack Obama)

($PERSON, {age, -year-old}, $DIGIT)

$PERSON, $DIGIT,

$PERSON's age is $DIGIT

$PERSON, a $DIGIT-year-old

(Barack Obama, 55)

(Justin Trudeau, 43)
Adjusting Types in Meta Patterns for Appropriate Granularity

$\text{PERSON, } \text{DIGIT,}$

$\text{PERSON's age is } \text{DIGIT}$

$\text{PERSON, a } \text{DIGIT -year-old}$

$\text{COUNTRY president } \text{POLITICIAN}$

$\text{president } \text{POLITICIAN of } \text{COUNTRY}$

- $\text{$\text{PERSON}}$
- $\text{$\text{ATTACKER}}$
- $\text{$\text{ARTIST}}$
- $\text{$\text{ATHLETE}}$
- $\text{$\text{POLITICIAN}}$
- $\text{$\text{VICTIM}}$

- $\text{$\text{LOCATION}}$
- $\text{$\text{COUNTRY}}$
- $\text{$\text{ETHNICITY}}$
- $\text{$\text{CITY}}$

- $\text{$\text{PERSON}}$
- $\text{$\text{POLITICIAN}}$
- $\text{$\text{ARTIST}}$
### Results: Patterns, Entities and Attribute Values in News Corpus

<table>
<thead>
<tr>
<th>Meta patterns</th>
<th>Entity</th>
<th>Attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{COUNTRY President } $\text{POLITICIAN} $\text{COUNTRY' s president } $\text{POLITICIAN} $\text{President } $\text{POLITICIAN of } $\text{COUNTRY} ... $\text{POLITICIAN' s government of } $\text{COUNTRY}</td>
<td>United States</td>
<td>Barack Obama</td>
</tr>
<tr>
<td></td>
<td>Russia</td>
<td>Vladimir Putin</td>
</tr>
<tr>
<td></td>
<td>France</td>
<td>Francois Hollande</td>
</tr>
<tr>
<td></td>
<td>Burkina Faso</td>
<td>Blaise Compaoré</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Meta patterns</th>
<th>Entity</th>
<th>Attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{COMPANY CEO } $\text{PERSON} $\text{COMPANY chief executive } $\text{PERSON} $\text{PERSON, the } $\text{COMPANY CEO, ...} $\text{COMPANY former CEO } $\text{PERSON} $\text{PERSON, the } $\text{COMPANY former CEO, ...} $\text{COMPANY}</td>
<td>Apple</td>
<td>Tim Cook</td>
</tr>
<tr>
<td></td>
<td>Facebook</td>
<td>Mark Zuckerberg</td>
</tr>
<tr>
<td></td>
<td>Hewlett-Packard</td>
<td>Carly Fiorina</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Infor</td>
<td>Charles Phillips</td>
</tr>
<tr>
<td></td>
<td>Afghan Citadel</td>
<td>Roya Mahboob</td>
</tr>
</tbody>
</table>
Patterns and Entities Found in Medical Science Corpus

<table>
<thead>
<tr>
<th>Meta patterns</th>
<th>Entity</th>
<th>Attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{TREATMENT was used to treat }$\text{DISEASE}</td>
<td>zoledronic acid therapy</td>
<td>Paget’s disease of bone</td>
</tr>
<tr>
<td>$\text{DISEASE using the }$\text{TREATMENT}</td>
<td>bisphosphonates</td>
<td>osteoporosis</td>
</tr>
<tr>
<td>$\text{TREATMENT has been used to treat }$\text{DISEASE}</td>
<td>calcitinin</td>
<td>Paget’s disease of bone</td>
</tr>
<tr>
<td>$\text{TREATMENT of patients with }$\text{DISEASE}</td>
<td>calcitinin</td>
<td>osteoporosis</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Meta patterns</th>
<th>Entity</th>
<th>Attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{BACTERIA was resistant to }$\text{ANTIBiotics}$</td>
<td>corynebacterium striatum BM4687</td>
<td>gentamicin</td>
</tr>
<tr>
<td>$\text{BACTERIA are resistant to }$\text{ANTIBiotics}$</td>
<td>corynebacterium striatum BM4687</td>
<td>tobramycin</td>
</tr>
<tr>
<td>$\text{BACTERIA is the most resistant to }$\text{ANTIBiotics}$</td>
<td>methicillin-susceptible S aureus</td>
<td>vancomycin</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>$\text{BACTERIA, particularly those resistant to }$\text{ANTIBiotics}$</td>
<td>multidrug-resistant enterobacteriaceae</td>
<td>gentamicin</td>
</tr>
</tbody>
</table>
Outline

- Meta-Pattern Discovery for Information Extraction and Text Mining
- MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora
- TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction
- WW-PIE: Pattern Discovery for Wide-Window Open Information Extraction
- PENNER: Pattern-Enhanced Nested Named Entity Recognition
- Summary
Pattern-based Information Extraction

- Information extraction from text: Automatically extracting structured information from unstructured and/or semi-structured documents

- Major methods
  - Machine learning methods
    - Use linguistic features and train machine learning models on a labeled corpus
  - Textual pattern methods
    - Based on statistics on a large corpus, such as frequency

- Pattern-based IE methods
  - Extract a collection of <Entity, Attribute, Value>-tuples
    - Example: MetaPAD (Jiang et al, 2017)
      - Formation
      - Grouping
      - Extraction
Error Analysis of Previous Work

- Grouping issues of existing pattern-based IE methods
  - Group patterns by trigger words (e.g., “married”)
  - Include wrong patterns: [$Person married $Person’s daughter]
  - Miss good patterns: [wedding of $Person and $Person]
- Group patterns by agreement on extractions
  - Patterns with little or no overlapped extractions may express the same relation
    - E.g., [president $Person with first lady $Person] vs.
      - [wedding of Prince $Person and Princess $Person]
**TruePIE System Overview**

- **Given a** text corpus, a few seed patterns for a specific extraction task on *attribute*.
- **Find** as many as possible reliable patterns and extract $<entity \ e, \ attribute \ a, \ value \ v>$.

- Q. Li, M. Jiang, X. Zhang, M. Qu, T. Hanratty, J. Gao, and J. Han, “TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction”, KDD’18
- Discover reliable patterns and extract quality EAV-tuples from text data.
An Intuitive Solution: Jointly Consider Pattern Constructing Words and Extractions

- Reliable patterns are semantically similar to the seed patterns
  - Jointly consider pattern constructing words and extractions
    - Ex.: $\text{Person}$, president of $\text{Country}$
      - Constructing words: president, of
      - Extractions: <Russia, Putin>, <China, Xi>, <USA, Trump>, ...
  - Pattern embedding
    - Adapting word embedding technique
      - $v_p = [v_{pw}, v_{pa}]$
    - Reliable patterns are those who are close to the seed patterns

- Issues: Lack of supervision to determine an accurate boundary
  - Solution: Use the pattern embedding as features, build a training set from the seed patterns (positive: highly reliable patterns vs. negative: highly unreliable patterns), and then train a classifier
How to Detect Negative Samples?

- Challenge: Open world assumption
  - Ex.: The seed pattern does not extract <US, president, Putin> nor <Nigeria, president, Buhari>

- Arity-constraint
  - Constraint on degrees of entities and values in an entity-value bipartite graph

- Detect negative tuples (e.g., using arity-constraint)
  - May significantly boost the performance of the information extraction
The arity-constraint is equivalent to setting constraints on the degree of entities $C_e$ and degree of values $C_v$:

- $C_e^a: \text{deg}(e) \leq \text{median}(f_e)$
- $C_v^a: \text{deg}(v) \leq \text{median}(f_v)$

### Hard arity-constraint:
- If the $\text{median}(f) = \beta$-Quantiles($f$), we set it as hard arity-constraint
- For hard arity-constraint, no violation is allowed
  - e.g., #country of a president = 1

### Soft arity-constraint:
- If the $\text{median}(f) < \beta$-Quantiles($f$), we set it as soft arity-constraint
- For soft arity-constraint, some violations are allowed
  - e.g., #president of a country
- If a tuple has a high reliability score, we can add it into the truth tuple set even it may violate the soft arity-constraint
Arity-Constraint based Conflict Finding

 Tuple’s Polarity
- A tuple \( t \) is positive, if \( t \in T \) (i.e., the true tuple set)
- \( t \) is negative, if \( t \notin T \), and adding \( t \) to \( T \) will cause violation of arity-constraints
- \( t \) is unknown, if \( t \notin T \) and \( t \) is not negative

 Pattern reliability score
- Extension of precision
  - Number of positive tuples
  - Number of unknown tuples
  - Total number of tuples
- Positive and negative patterns
  - Positive patterns: \( \rho_p > \theta \)
  - Negative patterns: \( \rho_p < 1 - \theta \)

\[
\rho_p = \frac{|T_p \cap T|}{|T_p|} + \frac{1}{2} \frac{|T_p^u|}{|T_p|}
\]
Tuple Reliability

\[ \tau_t = \sum_{\{p: p \in P\}} \rho_p \times n^p_t \]

**Tuple reliability score**

- Edge weight of the entity-value bipartite graph
  - Positive patterns’ reliability score
  - Frequency

- Optimization problem: Find the bipartite graph with the maximal sum of edge weights under the arity-constraints
  - Hard arity-constraint: no violation allowed, \( +\infty \) penalty
  - Soft arity-constraint: violation allowed with a positive penalty
Evaluation Process

$Country president $Person

$Country president $Person

$Country president $Person

$PERSON, president of $LOCATION
$PERSON, the $LOCATION’s president

$PERSON, the elected $LOCATION leader
$LOCATION’s dictator $PERSON
$PERSON ruled $LOCATION
President $PERSON arrived in $LOCATION
$LOCATION: president $PERSON

$PERSON, president of $LOCATION
$PERSON, the $LOCATION’s president

$PERSON, the elected $LOCATION leader
$LOCATION’s dictator $PERSON
$PERSON ruled $LOCATION
President $PERSON arrived in $LOCATION
$LOCATION: president $PERSON

classification
Experimental Evaluation

- Corpus
  - English Gigaword Fourth Edition LDC2009T13
  - 25.7 GB of size including 9.9 million documents and 4.0 billion words
- State-of-the-art pattern-based IE baselines
  - PATTY, MetaPAD
- Performance measure
  - Precision
    - randomly select 10 sets of 50 extracted tuples and label their correctness
  - Coverage
    - Randomly choose 100 corrected tuples from each method and combine them. Check how many are covered by each method
## Performance Comparison and Case Study

<table>
<thead>
<tr>
<th>Task</th>
<th>Patty</th>
<th>MetaPAD</th>
<th>TRUEPIE</th>
<th>Task</th>
<th>Patty</th>
<th>MetaPAD</th>
<th>TRUEPIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Extracted Tuples</td>
<td>2752</td>
<td>4067</td>
<td>2317</td>
<td>#Extracted Tuples</td>
<td>7801</td>
<td>4917</td>
<td>1490</td>
</tr>
<tr>
<td>Average Precision</td>
<td>Leader</td>
<td>0.59 ± 0.05</td>
<td>0.43 ± 0.07</td>
<td>0.87 ± 0.05</td>
<td>President</td>
<td>0.38 ± 0.08</td>
<td>0.30 ± 0.06</td>
</tr>
<tr>
<td>Top 10% Precision</td>
<td>Leader</td>
<td>0.89 ± 0.17</td>
<td>0.66 ± 0.30</td>
<td>0.99 ± 0.03</td>
<td>Director</td>
<td>0.59 ± 0.29</td>
<td>0.42 ± 0.15</td>
</tr>
<tr>
<td>Top K Precision</td>
<td>Leader</td>
<td>0.67 ± 0.12</td>
<td>0.56 ± 0.10</td>
<td>0.99 ± 0.01</td>
<td>Director</td>
<td>0.56 ± 0.27</td>
<td>0.33 ± 0.07</td>
</tr>
<tr>
<td>Coverage Rate</td>
<td>Leader</td>
<td>0.56</td>
<td>0.59</td>
<td>0.61</td>
<td>Director</td>
<td>0.87</td>
<td>0.63</td>
</tr>
<tr>
<td>#Extracted Tuples</td>
<td>Capital</td>
<td>1316</td>
<td>4371</td>
<td>428</td>
<td>Director</td>
<td>10313</td>
<td>14234</td>
</tr>
<tr>
<td>Average Precision</td>
<td>Capital</td>
<td>0.37 ± 0.07</td>
<td>0.27 ± 0.10</td>
<td>0.97 ± 0.02</td>
<td>Director</td>
<td>0.54 ± 0.08</td>
<td>0.56 ± 0.07</td>
</tr>
<tr>
<td>Top 10% Precision</td>
<td>Capital</td>
<td>0.54 ± 0.25</td>
<td>0.47 ± 0.16</td>
<td>1 ± 0</td>
<td>Director</td>
<td>0.63 ± 0.31</td>
<td>0.65 ± 0.20</td>
</tr>
<tr>
<td>Top K Precision</td>
<td>Capital</td>
<td>0.51 ± 0.18</td>
<td>0.47 ± 0.16</td>
<td>0.98 ± 0.02</td>
<td>Director</td>
<td>0.63 ± 0.32</td>
<td>0.67 ± 0.31</td>
</tr>
<tr>
<td>Coverage Rate</td>
<td>Capital</td>
<td>0.67</td>
<td>0.92</td>
<td>0.68</td>
<td>Director</td>
<td>0.52</td>
<td>0.6</td>
</tr>
</tbody>
</table>

### Positive Patterns

- $\$$LOCATION \text{ president } $\$$PERSON$
- $\$$LOCATION \text{ prime minister } $\$$PERSON$
- $\$$LOCATION \text{ military ruler } $\$$PERSON$
- $\$$LOCATION \text{'s chancellor, } $\$$PERSON$
- $\$$LOCATION \text{, the } $\$$LOCATION \text{ administrator}$
- $\$$LOCATION \text{ president sworn in } $\$$LOCATION$
- $\$$LOCATION \text{, } $\$$LOCATION$
- $\$$PERSON \text{’s widower } $\$$PERSON$
- $\$$LOCATION \text{ president } $\$$PERSON \text{ and first lady } $\$$PERSON$
- $\$$PERSON \text{ wedding of prince } $\$$PERSON \text{ and princess } $\$$PERSON$
- $\$$PERSON \text{, the } $\$$LOCATION \text{ senator } $\$$PERSON$

### Negative Patterns

- $\$$LOCATION \text{ leader told } $\$$PERSON$
- $\$$LOCATION \text{ scoring leader } $\$$PERSON$
- $\$$PERSON \text{, son of the } $\$$LOCATION \text{ leader}$
- $\$$LOCATION \text{’s cricket chief, } $\$$PERSON$
- $\$$LOCATION \text{ leader $\$$PERSON \text{ will visit } $\$$LOCATION}$
- $\$$LOCATION \text{ embassy of } $\$$LOCATION \text{ in } $\$$LOCATION$
- $\$$PERSON \text{’s lover } $\$$PERSON$
- $\$$PERSON \text{’s affair with } $\$$PERSON$
- $\$$PERSON \text{’s girlfriend, } $\$$PERSON$
- $\$$PERSON \text{’s brother, } $\$$PERSON$
- $\$$PERSON \text{’s husband } $\$$PERSON$
- $\$$PERSON \text{’s trial in } $\$$YEAR$
- $\$$PERSON \text{ fired him in } $\$$YEAR$
- $\$$PERSON \text{’s husband died in } $\$$YEAR$
- $\$$PERSON \text{ left in } $\$$YEAR$
Error Analysis and Future Work

- Information sparsity
  - Pattern sparsity: extract little information
  - Entity sparsity: appears infrequent in the corpus
- Information ambiguity
  - Fine-grained typing
    - Ex.: ‘$Country senator $Person’ is semantically different from ‘$State senator $Person’
  - Entity linking or entity normalization
    - ‘John’, ‘John Kennedy’, ‘Kennedy’, ‘J. H. Kennedy’... are they the same person?
    - Are ‘John’ and ‘John’ the same person?
Outline

- Meta-Pattern Discovery for Information Extraction and Text Mining
  - MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora
- TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction
- WW-PIE: Pattern Discovery for Wide-Window Open Information Extraction
- PENNER: Pattern-Enhanced Nested Named Entity Recognition
- Summary
Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature

- Qi Li, Xuan Wang, Yu Zhang, Fei Ling, Cathy H. Wu, Jiawei Han, “Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature”, BIBM’18

- Characteristics of Language in Bio-Literature
  - Long and complicated
  - Some sentences can be as long as a paragraph
  - Need parsing to understand the structure of the sentences
  - “Formal” language (e.g., grammar, terms) in most bio-literature
  - Good news for pattern extraction

- Example: “Pre-treatment of ATRA can decrease the overexpression of cyclin_D1 and E2F-1 induced by B(a)P.”

  NER result: Pre-treatment of [ATRA]CHEMICAL can decrease the overexpression of [cyclin_D1]GENE and [E2F-1]GENE induced by [B(a)P]CHEMICAL.

- Task: Find relationships among the entities
Relation Extraction of Existing Studies

- Supervised methods: Relying on annotated corpora to discover certain relation types between entities
- Distantly supervised methods: Using existing knowledge-bases or databases to annotate corpora: `<ATRA, cyclin_D1, decrease?>` ➔ `<ATRA, cyclin_D1, decrease>`
- Limitations: Pre-defined relation types; relation is pair-wise; the context is ignored
- OpenIE (Open information extraction: Stanford)
  - Extraction of relation tuples, typically binary relations, from plain text
    - Ex.: `(Mark Zuckerberg; founded; Facebook)`
    - Ex.: `<Pre-treatment of [ATRA]_CHEMICAL, can decrease, the overexpression of [cyclin_D1]_GENE and [E2F-1]_GENE induced by [B(a)P]_CHEMICAL>`
- Pros:
  - No pre-defined types: No relation schema need to be specified in advance
  - The context is kept: The relation name is just the text linking two arguments
How Does Human Structure Information?

- Can we do better? Maybe the extraction structure can be further improved
- Let’s examine how human structures the information
- Example:
  - Pre-treatment of $\text{ATRA}_{\text{CHEMICAL}}$ can decrease the overexpression of $\text{cyclin\_D1}_{\text{GENE}}$ and $\text{E2F-1}_{\text{GENE}}$ induced by $\text{B(a)P}_{\text{CHEMICAL}}$
  - Pre-treatment of $\text{ATRA}_{\text{CHEMICAL}}$, can decrease, the overexpression of $([\text{cyclin\_D1}_{\text{GENE}}, \text{E2F-1}_{\text{GENE}}])$, where $([\text{cyclin\_D1}_{\text{GENE}}, \text{E2F-1}_{\text{GENE}}])$, induced by, $\text{B(a)P}_{\text{CHEMICAL}}$
  - $<\text{ATRA}_{\text{CHEMICAL}}, \text{decrease, ([cyclin\_D1}_{\text{GENE}}, \text{E2F-1}_{\text{GENE}})\text{, induced by, [B(a)P}_{\text{CHEMICAL}}>>}$
- People can find and use hierarchical structures
Meta-Pattern Extraction

- What are meta-patterns?
- A mixed sequence of entity types and non-type words in the corpus
  - E.g., pattern: CHEMICAL decrease GENE
    - instance: CHEMICAL = B(a)P, ATRA, ...
    - GENE = cyclin_D1, E2F-1, ...
- New innovation: Hierarchical pattern grouping
  - pattern: \{CHEMICAL\} decrease \{GENE\}
    - sub-patterns: \{CHEMICAL\} = Pretreatment of CHEMICAL, ...
    - \{GENE\} = the overexpression of GENE, GENE induced by CHEMICAL ...
Pre-treatment of ATRA can decrease the overexpression of cyclin_D1 and E2F-1 induced by B(a)P...
Pretreatment can decrease overexpression of CHEMICAL. Pretreatment: Pretreatment of CHEMICAL. Overexpression: the overexpression of GENE. GENE: GENE and GENE. GENE: GENE induced by CHEMICAL.
Step 2: Pattern mining on short sentences
Constraint: words in a pattern should be connected on the tree
Eg. “pretreatment of CHEMICAL” ✓ “and GENE” ×
Constraint: pattern should contain (one entity + one non-stop-word), or more than one entity
Constraint: frequency is high
With pattern mining, we get a new pattern: "\{CHEMICAL\} can decrease \{GENE\}"

With Pattern grouping, we can perform clustering to group synonymous meta patterns
\{CHEMICAL\} can decrease \{GENE\}
\{CHEMICAL\} decrease \{GENE\}
\{GENE\} be decreased by \{CHEMICAL\}
Experiments

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

<table>
<thead>
<tr>
<th>Basic statistics of the subset corpus.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstracts</td>
</tr>
<tr>
<td>Gene</td>
</tr>
<tr>
<td>28007</td>
</tr>
</tbody>
</table>

- Baselines:
  - ClausIE: adopts clause patterns to handle long-distance relationships
  - Stanford OpenIE: learns a clause splitter via distant training data
  - Ollie: utilizes open pattern learning and extracts patterns over dependency path and part-of-speech tags
  - MinIE: refines tuples extracted by ClausIE by identifying and removing parts that are considered overly specific
Performance Comparison with the State-of-the-Art OpenIE Systems

- Randomly sample 96 sentences for human labeling
- one tuple will be judged as correct if it reads smoothly and meets the fact described in the sentence

<table>
<thead>
<tr>
<th>System</th>
<th># Correct extractions</th>
<th># Valid extractions</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClausIE [12]</td>
<td>21</td>
<td>142</td>
<td>0.15</td>
</tr>
<tr>
<td>Stanford [13]</td>
<td><strong>120</strong></td>
<td>277</td>
<td>0.43</td>
</tr>
<tr>
<td>Ollie [11]</td>
<td>43</td>
<td>84</td>
<td>0.51</td>
</tr>
<tr>
<td>MinIE [14]</td>
<td>77</td>
<td>126</td>
<td>0.61</td>
</tr>
<tr>
<td>WW-PIE</td>
<td>110</td>
<td>150</td>
<td><strong>0.73</strong></td>
</tr>
</tbody>
</table>

- Note: we observe that Stanford OpenIE produces over 60 extractions for one sentence, which may be undesirable for some applications
## Pattern and Extraction Examples

<table>
<thead>
<tr>
<th>Meta Pattern</th>
<th>Excerpts in Expression Format</th>
<th>Excerpts in Tuple Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>{CHEMICAL} reduce {DISEASE}</td>
<td>Ranitidine reduce ischemia/reperfusion-induced liver_injury in rats</td>
<td>(Ranitidine, reduce, liver_injury: ischemia/reperfusion, induce, liver_injury, in, rats)</td>
</tr>
<tr>
<td></td>
<td>resveratrol reduce brain_injury</td>
<td>(resveratrol, reduce, brain_injury)</td>
</tr>
<tr>
<td></td>
<td>Resveratrol reduce renal_and_lung_injury cause by sepsis in rats</td>
<td>(Resveratrol, reduce, renal_and_lung_injury: sepsis, cause, renal_and_lung_injury, in, rats)</td>
</tr>
<tr>
<td></td>
<td>Resveratrol reduce TNF-a-induced U373MG human glioma_cell_invasion</td>
<td>(Resveratrol, reduce, glioma_cell_invasion: TNF-a, induce, human glioma_cell_invasion)</td>
</tr>
<tr>
<td></td>
<td>caffeine treatment reduce glioma cell proliferation</td>
<td>(caffeine, reduce, glioma)</td>
</tr>
<tr>
<td>{CHEMICAL} inhibit {GENE}</td>
<td>Progesterone inhibit COX-2 expression</td>
<td>(Progesterone, inhibit, COX-2)</td>
</tr>
<tr>
<td></td>
<td>NAC treatment inhibit phosphorylation of Akt</td>
<td>(NAC, inhibit, Akt)</td>
</tr>
<tr>
<td></td>
<td>ATRA inhibit the expression of CcnB1 and CcnA1</td>
<td>(ATRA, inhibit, (CcnB1, CcnA1))</td>
</tr>
<tr>
<td></td>
<td>Cypermethrin inhibit the interaction between the AR_AFI and SRC-1</td>
<td>(Cypermethrin, inhibit, AR_AFI:(AR_AFI, interaction, SRC-1))</td>
</tr>
<tr>
<td></td>
<td>PGF and H2O2 inhibit SOD1 protein expression and activity</td>
<td>(PGF,H2O2), inhibit, SOD1)</td>
</tr>
<tr>
<td>{GENE} cause {DISEASE}</td>
<td>mutations in the CSB gene cause Cockayne_syndrome</td>
<td>(CSB, cause, Cockayne_syndrome)</td>
</tr>
<tr>
<td></td>
<td>mutations in FOXP2 cause developmental_verbal_dyspraxia (DVD)</td>
<td>(FOXP2, cause, developmental_verbal_dyspraxia: (developmental_verbal_dyspraxia, abbr, DVD))</td>
</tr>
<tr>
<td></td>
<td>mutations in the hENT3 gene cause an autosomal_recessive_disorder in humans</td>
<td>(hENT3, causes, autosomal_recessive_disorder: (autosomal_recessive_disorder, in, humans))</td>
</tr>
<tr>
<td></td>
<td>germline mutations in DIS3L2 cause the Perlman syndrome_of_overgrowth and Wilms_tumor susceptibility</td>
<td>(DIS3L2, cause, Perlman syndrome_of_overgrowth, Wilms_tumor)</td>
</tr>
</tbody>
</table>
More Experimental Results

- **Top 10 Single Entity Patterns**

<table>
<thead>
<tr>
<th>Meta Patterns with Single Entity</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISEASE cell</td>
<td>11210</td>
</tr>
<tr>
<td>effect of CHEMICAL</td>
<td>9507</td>
</tr>
<tr>
<td>GENE expression</td>
<td>6551</td>
</tr>
<tr>
<td>expression of GENE</td>
<td>4940</td>
</tr>
<tr>
<td>CHEMICAL treatment</td>
<td>4896</td>
</tr>
<tr>
<td>GENE gene</td>
<td>4229</td>
</tr>
<tr>
<td>CHEMICAL exposure</td>
<td>3957</td>
</tr>
<tr>
<td>the effect of CHEMICAL</td>
<td>3721</td>
</tr>
<tr>
<td>GENE mRNA</td>
<td>3211</td>
</tr>
<tr>
<td>CHEMICAL level</td>
<td>3076</td>
</tr>
</tbody>
</table>

- **Helpful in named entity recognition tasks**
- **See: PENNER: Pattern-enhanced Nested Named Entity Recognition in Biomedical Literature**

- **Synonymous Pattern Group Examples**

<table>
<thead>
<tr>
<th>Synonymous group</th>
<th>Meta Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEMICAL induced inhibition of GENE</td>
<td>GENE inhibition by CHEMICAL</td>
</tr>
<tr>
<td></td>
<td>CHEMICAL block GENE</td>
</tr>
<tr>
<td></td>
<td>GENE inhibitor, CHEMICAL</td>
</tr>
<tr>
<td></td>
<td>GENE inhibitor CHEMICAL</td>
</tr>
<tr>
<td>CHEMICAL activate GENE</td>
<td>CHEMICAL_activated GENE</td>
</tr>
<tr>
<td></td>
<td>GENE activator CHEMICAL</td>
</tr>
<tr>
<td></td>
<td>GENE agonist CHEMICAL</td>
</tr>
<tr>
<td></td>
<td>GENE agonist, CHEMICAL</td>
</tr>
<tr>
<td></td>
<td>GENE ligand CHEMICAL</td>
</tr>
<tr>
<td></td>
<td>GENE ligand, CHEMICAL</td>
</tr>
<tr>
<td>DISEASE cause by CHEMICAL</td>
<td>CHEMICAL_induced DISEASE</td>
</tr>
<tr>
<td></td>
<td>CHEMICAL can cause DISEASE</td>
</tr>
<tr>
<td></td>
<td>CHEMICAL induce DISEASE</td>
</tr>
<tr>
<td></td>
<td>CHEMICAL cause DISEASE</td>
</tr>
<tr>
<td></td>
<td>DISEASE be induce by CHEMICAL</td>
</tr>
<tr>
<td></td>
<td>DISEASE induce by CHEMICAL</td>
</tr>
<tr>
<td></td>
<td>DISEASE produce by CHEMICAL</td>
</tr>
<tr>
<td>SPECIES treat with CHEMICAL</td>
<td>CHEMICAL administration to SPECIES</td>
</tr>
<tr>
<td></td>
<td>CHEMICAL_treated SPECIES</td>
</tr>
<tr>
<td></td>
<td>CHEMICAL_exposed SPECIES</td>
</tr>
<tr>
<td></td>
<td>CHEMICAL treat SPECIES</td>
</tr>
<tr>
<td></td>
<td>SPECIES be inject with CHEMICAL</td>
</tr>
<tr>
<td></td>
<td>SPECIES be administer with CHEMICAL</td>
</tr>
</tbody>
</table>
Summary: WW-PIE

- WW-PIE
- Extract a variety of the relation tuples from large biomedical literature corpora
- Resolve the long and complicated sentence structures by breaking down the sentences
- Group meta-patterns hierarchically to extract n-ary hierarchical tuples

Discussion and Future Work
- Pattern grouping can be enhanced
- Negation structures
  - For example, “there is no evidence that ...”
- Dependency parser may introduce noise
Outline

❑ Meta-Pattern Discovery for Information Extraction and Text Mining
  ❑ MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora
  ❑ TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction
  ❑ WW-PIE: Pattern Discovery for Wide-Window Open Information Extraction
  ❑ PENNER: Pattern-Enhanced Nested Named Entity Recognition
❑ Summary
PENNER: Pattern-Enhanced Nested Named Entity Recognition in Biomedical Literature

- Xuan Wang*, Yu Zhang*, Qi Li, Cathy H. Wu, Jiawei Han, “PENNER: Pattern-enhanced Nested Named Entity Recognition in Biomedical Literature”, BIBM’18

What is a nested entity structure?

- Example: PID: 10190572:
  - “… although each of the agents alone caused only slight increase in the \([\text{alanine}]_{\text{CHEMICAL}}\ \text{aminotransferase}]_{\text{PROTEIN}}\ activity.”

- PubTator recognizes “alanine” as a CHEMICAL but misses “alanine aminotransferase” as a PROTEIN

Nested entities are very important!

- 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
Previous Studies

- “Flat” BioNER (e.g., [1], [2], [3])
- Common sequence modeling frameworks cannot detect entities with overlapping tokens
- Supervised Nested BioNER (e.g., [4], [5], [6])
- 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?)

[5] Labeling gaps between words: Recognizing overlapping mentions with mention separators. EMNLP 2017
[6] Nested named entity recognition revisited. NAACL 2018
PENNER: Key Ideas of Using Meta-Pattern

- Nested BioNER with very weak supervision
- Idea: Nested structure as a pattern-level phenomenon
  - CHEMICAL aminotransferase = PROTEIN
  - GENE mRNA release = PROCESS

- Framework
  - Taking a corpus pre-tagged by any flat NER tool as input
  - Unsupervised meta-pattern extraction
  - Few-shot nested entity recognition for each type

- Evaluation
  - Outperforming baselines in both meta-pattern extraction and nested NER
  - Detecting new entity types with few seeds
  - Improving annotation results over PubTator
TERT encodes the reverse transcriptase subunit of human telomerase.
The FGFR-2 receptor is a membrane-spanning tyrosine kinase.
Each of the agents alone caused only slight increase in the alanine aminotransferase activity.

**Flat Entity Recognition**

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TERT encodes the reverse transcriptase subunit of human telomerase.</td>
</tr>
<tr>
<td>2</td>
<td>The FGFR-2 receptor is a membrane-spanning tyrosine kinase.</td>
</tr>
<tr>
<td>3</td>
<td>Each of the agents alone caused only slight increase in the alanine aminotransferase activity.</td>
</tr>
</tbody>
</table>

**Meta-Pattern Extraction**

<table>
<thead>
<tr>
<th>Meta-Pattern</th>
<th>SPECIES telomerase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CHEMICAL aminotransferase</td>
</tr>
<tr>
<td></td>
<td>GENE level</td>
</tr>
<tr>
<td></td>
<td>male SPECIES</td>
</tr>
<tr>
<td></td>
<td>several DISEASE</td>
</tr>
</tbody>
</table>

**Nested Entity Recognition**

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TERT encodes the reverse transcriptase subunit of human telomerase.</td>
</tr>
<tr>
<td>2</td>
<td>The FGFR-2 receptor is a membrane-spanning tyrosine kinase.</td>
</tr>
<tr>
<td>3</td>
<td>Each of the agents alone caused only slight increase in the alanine aminotransferase activity.</td>
</tr>
</tbody>
</table>

**Weak Supervision**

**Context Information**

**Seed Sets**

**Round 1**

**Round 2**
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
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)

The high doses of CHEMICAL appear useful for the treatment of 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

The high doses of CHEMICAL appear useful for the treatment of patients with DISEASE.

```
NOUN  ADP  PROPN  VERB  ADJ  ADP  NOUN  ADP  NOUN  ADP  PROPN
prep  prep  prep  prep  prep  prep  prep  prep  prep  prep
nsubj pobj  pobj  pobj  pobj  pobj  pobj  pobj  pobj  pobj
```

50
Weakly-supervised Pattern Expansion

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

![Diagram of weakly-supervised pattern expansion](image)

**Seed Sets**
- Round 1:
  - GENE: peroxidase
  - CHEMICAL: peroxidase
  - SPECIES: telomerase
  - CHEMICAL: kinase
  - CHEMICAL: aminotransferase
  - CHEMICAL: hydrolase

**Context Information (Corpus)**
- Round 2:
  - GENE: telomerase
  - SPECIES: telomerase
  - CHEMICAL: kinase
  - CHEMICAL: aminotransferase
  - CHEMICAL: hydrolase

**Weak supervision (User-specified seeds)**
- **SetExpan** algorithm (Shen et al., ECML-PKDD 2017)
- Skip-gram + Rank Ensemble

**Pre-ranked entity list 1**
- California: 1 (1/1)
- Arizona: 2 (1/2)
- Quebec: 3 (1/3)

**Pre-ranked entity list 2**
- Florida: 1 (1/1)
- Baja California: 2 (1/2)
- Arizona: 3 (1/3)

**Pre-ranked entity list 3**
- Arizona: 1 (1/1)
- Florida: 2 (1/2)
- Baja California: 3 (1/3)

**Final ranked list of entities**
- Arizona: 1 (1/1)
- Florida: 2 (1/2)
- Baja California: 3 (1/3)
- Quebec: 4 (1/4)
- California: 5 (1/5)

**Context Feature Selection step**
- Denoised context sets
  - City, USA, Texas and...
  - Country, USA, on

**Entity selection step**
- 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
Experiments

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

<table>
<thead>
<tr>
<th>Abstracts</th>
<th>Sentences</th>
<th>Entity Mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>28007</td>
<td>302736</td>
<td>Gene 215704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chemical 314134</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease 129931</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Species 86697</td>
</tr>
</tbody>
</table>

- 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**

<table>
<thead>
<tr>
<th>Seed</th>
<th>{GENE, GENE peroxidase}</th>
<th>{CHEMICAL, GENE agonist}</th>
<th>{DISEASE, cellular DISEASE}</th>
<th>{SPECIES, female SPECIES}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unassigned: GENE</td>
<td>CHEMICAL receptor modulator (serm)</td>
<td>DISEASE vera</td>
<td>fischer SPECIES</td>
</tr>
<tr>
<td>2</td>
<td>CHEMICAL phosphatase</td>
<td>antagonist of CHEMICAL</td>
<td>potential for DISEASE</td>
<td>SPECIES and adult</td>
</tr>
<tr>
<td>3</td>
<td>(\text{CHEMICAL })</td>
<td>offspring of SPECIES</td>
<td>GENE transcription</td>
<td>exposure to CHEMICAL or</td>
</tr>
<tr>
<td>4</td>
<td>SPECIES cardiomycocyte</td>
<td>CHEMICAL oxidase</td>
<td>SPECIES and adult</td>
<td>SPECIES in vivo</td>
</tr>
<tr>
<td>5</td>
<td>potential against DISEASE</td>
<td>DISEASE chemopreventive agent</td>
<td>growth and DISEASE</td>
<td>CHEMICAL protect</td>
</tr>
<tr>
<td>6</td>
<td>GENE inductor</td>
<td>GENE receptor activity</td>
<td>a common DISEASE</td>
<td>CHEMICAL interfere</td>
</tr>
<tr>
<td>7</td>
<td>effect and mechanism of CHEMICAL</td>
<td>antagonist (CHEMICAL)</td>
<td>rare DISEASE</td>
<td>a cohort of SPECIES</td>
</tr>
<tr>
<td>8</td>
<td>inductor of GENE</td>
<td>CHEMICAL blocker</td>
<td>detection of DISEASE</td>
<td>SPECIES albino</td>
</tr>
<tr>
<td>9</td>
<td>(GENE) antagonist</td>
<td>CHEMICAL substituent</td>
<td>DISEASE as well as</td>
<td>CHEMICAL exposure ,</td>
</tr>
<tr>
<td>10</td>
<td>GENE level and</td>
<td>CHEMICAL vapor</td>
<td>progression and DISEASE</td>
<td>the detrimental effect of CHEMICAL</td>
</tr>
</tbody>
</table>

**SeedExpan**

<table>
<thead>
<tr>
<th>Seed</th>
<th>{GENE, GENE peroxidase}</th>
<th>{CHEMICAL, GENE agonist}</th>
<th>{DISEASE, cellular DISEASE}</th>
<th>{SPECIES, female SPECIES}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SPECIES telomerase</td>
<td>DISEASE chemopreventive agent</td>
<td>hepatic DISEASE</td>
<td>male SPECIES</td>
</tr>
<tr>
<td>2</td>
<td>CHEMICAL</td>
<td>DISEASE</td>
<td>dermal DISEASE</td>
<td>CHEMICAL</td>
</tr>
<tr>
<td>3</td>
<td>DISEASE</td>
<td>SPECIES</td>
<td>clinical DISEASE</td>
<td>DISEASE cell</td>
</tr>
<tr>
<td>4</td>
<td>CHEMICAL acetyltransferase</td>
<td>CHEMICAL chelation</td>
<td>GENE phosphorylation</td>
<td>GENE</td>
</tr>
<tr>
<td>5</td>
<td>CHEMICAL aminotransferase</td>
<td>GENE antagonist</td>
<td>-</td>
<td>SPECIES cell</td>
</tr>
<tr>
<td>6</td>
<td>SPECIES</td>
<td>DISEASE cell</td>
<td>-</td>
<td>pregnant SPECIES</td>
</tr>
<tr>
<td>7</td>
<td>CHEMICAL hydroxase</td>
<td>underlying mechanism of CHEMICAL</td>
<td>-</td>
<td>adult SPECIES</td>
</tr>
<tr>
<td>8</td>
<td>GENE kinase</td>
<td>CHEMICAL exclusion</td>
<td>-</td>
<td>CHEMICAL channel</td>
</tr>
<tr>
<td>9</td>
<td>CHEMICAL kinase</td>
<td>DISEASE cell line</td>
<td>-</td>
<td>DISEASE cell</td>
</tr>
<tr>
<td>10</td>
<td>CHEMICAL influx</td>
<td>10 m CHEMICAL</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**PENNER**
Entity-level Task: Nested NER

• “Precision”: NDCG of the ranking list of expanded entities

\[
\text{DCG}_p = \sum_{i=1}^{p} \frac{2^{\text{rel}_i} - 1}{\log_2 (i + 1)} \quad \text{IDCG}_p = \sum_{i=1}^{|\text{REL}|} \frac{2^{\text{rel}_i} - 1}{\log_2 (i + 1)} \quad \text{nDCG}_p = \frac{\text{DCG}_p}{\text{IDCG}_p}
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>GENE</th>
<th>CHEMICAL</th>
<th>DISEASE</th>
<th>SPECIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMBEDDING [22]</td>
<td>0.139</td>
<td>0.580</td>
<td>0.073</td>
<td>0.315</td>
</tr>
<tr>
<td>SETEXSPAN [26]</td>
<td>0.602</td>
<td>0.312</td>
<td>0.754</td>
<td>0.417</td>
</tr>
<tr>
<td>PENNER</td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>0.754</strong></td>
<td><strong>0.776</strong></td>
</tr>
</tbody>
</table>

• “Recall”: Number of correct instances

<table>
<thead>
<tr>
<th>Method</th>
<th>GENE</th>
<th>CHEMICAL</th>
<th>DISEASE</th>
<th>SPECIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMBEDDING [22]</td>
<td>79</td>
<td>139</td>
<td>61</td>
<td>45</td>
</tr>
<tr>
<td>SETEXSPAN [26]</td>
<td>1734</td>
<td>458</td>
<td>184</td>
<td>2211</td>
</tr>
<tr>
<td>PENNER</td>
<td>5254</td>
<td>458</td>
<td>184</td>
<td>3212</td>
</tr>
</tbody>
</table>

• **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!

<table>
<thead>
<tr>
<th>Seed</th>
<th>{GENE upregulation, GENE downregulation}</th>
<th>{CHEMICAL injection, CHEMICAL Inhalation}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GENE expression</td>
<td>CHEMICAL treatment</td>
</tr>
<tr>
<td>2</td>
<td>GENE phosphorylation</td>
<td>CHEMICAL administration</td>
</tr>
<tr>
<td>3</td>
<td>the development of DISEASE</td>
<td>CHEMICAL exposure</td>
</tr>
<tr>
<td>4</td>
<td>GENE induction</td>
<td>treatment with CHEMICAL</td>
</tr>
<tr>
<td>5</td>
<td>CHEMICAL action</td>
<td>CHEMICAL exposure to CHEMICAL</td>
</tr>
<tr>
<td>6</td>
<td>identification of GENE</td>
<td>administration of CHEMICAL</td>
</tr>
<tr>
<td>7</td>
<td>GENE suppression</td>
<td>pretreatment with CHEMICAL</td>
</tr>
<tr>
<td>8</td>
<td>DISEASE reduction</td>
<td>CHEMICAL pretreatment</td>
</tr>
<tr>
<td>9</td>
<td>CHEMICAL production</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>GENE activity</td>
<td>-</td>
</tr>
</tbody>
</table>

• 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)
Comparison with PubTator

• 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]CHEMICAL dismutase (SOD) and aminotransferases like [alanine]CHEMICAL aminotransferase (Ala-AT) and [aspartate]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]CHEMICAL dismutase]GENE (SOD) and aminotransferases like [[alanine]CHEMICAL aminotransferase]GENE (Ala-AT) and [[aspartate]CHEMICAL aminotransferase]GENE in different age groups ... |

PMID: 10919993

| PubTator | Mitogen-activated protein (MAP) kinase [Erk1/2]GENE antagonist mainly inhibited the release of [MCP-1]GENE, whereas MAP kinase [p38]GENE antagonist mainly suppressed the release of [IL-8]GENE and [RANTES]GENE. |
| PENNER | Mitogen-activated protein (MAP) kinase [[Erk1/2]GENE antagonist]CHEMICAL mainly inhibited the release of [MCP-1]GENE, whereas MAP kinase [[p38]GENE antagonist]CHEMICAL mainly suppressed the release of [IL-8]GENE and [RANTES]GENE. |

PMID: 21266192


PMID: 10498651

| PubTator | [COL1A2]GENE expression was decreased by [vitamin E]CHEMICAL treatment or transfection with [manganese superoxide]CHEMICAL dismutase, and was further increased after treatment with [L-buthionine sulfoximine]CHEMICAL ... |
| PENNER | [[COL1A2]GENE expression]PROCESS was decreased by [[vitamin E]CHEMICAL treatment]TREATMENT or transfection with [[manganese superoxide]CHEMICAL dismutase]GENE, and was further increased after [treatment with [L-buthionine sulfoximine]CHEMICAL]TREATMENT ... |
BioNER Demo

Entity type hierarchy: 5 coarse types, 17 fine-grained types

- **Organism**
  - Archaeon
  - Bacterium
  - Eukaryote
  - Virus

- **Fully Formed Anatomical Structure**
  - Body Part, Organ, or Organ Component
  - Tissue
  - Cell
  - Cell Component
  - Gene or Genome

- **Chemical**
- **Physiologic Function**
  - Organism Function
  - Organ or Tissue Function
  - Cell Function
  - Molecular Function

- **Pathologic Function**
  - Disease or Syndrome
  - Cell or Molecular Dysfunction
  - Experimental Model of Disease
Telmisartan: just an antihypertensive agent? A literature review.

PMID: 22077832

Authors: Maurizio, Destro; Francesca, Cagnoni; Giuseppina Patrizia, Dognini; Vincenzo, Galimberti; Carlo, Tietti; Chiara, Cavallieri; Emilio, Galli;

Journal: Expert opinion on pharmacotherapy

Publish Year: 2011

Abstract:

The modulation of the renin angiotensin aldosterone system (RAAS) is an important pathway in managing high blood pressure, and its overexpression plays a key role in target end-organ damage. Telmisartan is an angiotensin II receptor blocker (ARB) with unique pharmacologic properties, including the longest half-life among all ARBs; this leads to a significant and 24-h sustained reduction of blood pressure. Telmisartan has well-known antihypertensive properties, but there is also strong clinical evidence that it reduces left ventricular hypertrophy, arterial stiffness and the recurrence of atrial fibrillation, and confers renoprotection. This paper reviews telmisartan’s pharmacological properties in terms of efficacy for hypertension control and, importantly, focuses on its new therapeutic indications and their clinical implications. ONTARGET (ongoing telmisartan alone and in combination with ramipril global endpoint trial) demonstrated that telmisartan confers cardiovascular protective effects similar to those of ramipril, but with a better tolerability. Moreover, recent investigations focused on the capability of telmisartan to modulate the peroxisome proliferator-activated receptor-gamma (PPAR-γ), an established target in the treatment of insulin resistance, diabetes and metabolic syndrome, whose activation is also correlated to anti-inflammatory and, finally, anti-atherosclerotic properties. Telmisartan shows peculiar features that go beyond blood pressure control. It presents promising and unique protective properties against target end-organ damage, potentially able to open a scenario of new therapeutic approaches to cardiovascular disease.
Outline

- Meta-Pattern Discovery for Information Extraction and Text Mining
  - MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora
- TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction
- WW-PIE: Pattern Discovery for Wide-Window Open Information Extraction
- PENNER: Pattern-Enhanced Nested Named Entity Recognition
- Summary
Meta-Pattern Discovery for Information Extraction and Text Mining

MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora

TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction

WW-PIE: Pattern Discovery for Wide-Window Open Information Extraction

PENNER: Pattern-Enhanced Nested Named Entity Recognition
References

- Alon Halevy, Natalya Noy, Sunita Sarawagi, Steven Euijong Whang, and Xiao Yu, “Discovering structure in the universe of attribute names”, WWW’16

- Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance M. Kaplan, Timothy P. Hanratty, and Jiawei Han, “MetaPAD: Meta Pattern Discovery from Massive Text Corpora”, KDD’17, Aug. 2017

- Qi Li, Meng Jiang, Xikun Zhang, Meng Qu, Timothy Hanratty, Jing Gao and Jiawei Han, "TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction", KDD’18

- Qi Li, Xuan Wang, Yu Zhang, Fei Ling, Cathy Wu, and Jiawei Han, "Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature", Proc. 2018 Int. Conf. on Bioinformatics and Biomedicine (BIBM’18), Madrid, Spain, Dec. 2018

- Xuan Wang, Yu Zhang, Qi Li, Cathy Wu, and Jiawei Han, "PENNER: Pattern-enhanced Nested Named Entity Recognition in Biomedical Literature", Proc. 2018 Int. Conf. on Bioinformatics and Biomedicine (BIBM’18), Madrid, Spain, Dec. 2018