Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature

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

University of Illinois at Urbana-Champaign, IL, USA

University of Delaware, DE, USA

Outline

- Introduction
- Framework
- Evaluation
- Conclusion

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
  - 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.**

Example

- **Pre-treatment of [ATRA]$_{\text{CHEMICAL}}$ can decrease the overexpression of [cyclin_D1]$_{\text{GENE}}$ and [E2F-1]$_{\text{GENE}}$ induced by [B(a)P]$_{\text{CHEMICAL}}$.**

- 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

Relation Extraction of Existing Studies

- OpenIE
  - Using linguistic features to discover all types of relations
- <Pre-treatment of [ATRA]$_{\text{CHEMICAL}}$, **can decrease**, the overexpression of [cyclin_D1]$_{\text{GENE}}$ and [E2F-1]$_{\text{GENE}}$ induced by [B(a)P]$_{\text{CHEMICAL}}$>
- Pros
  - No pre-defined types
  - The context is kept
- Limitations
  - The extraction structure can be further improved

How Human Structure the Information

- Pre-treatment of [ATRA]$_{\text{CHEMICAL}}$ **can decrease** the overexpression of [cyclin_D1]$_{\text{GENE}}$ and [E2F-1]$_{\text{GENE}}$ induced by [B(a)P]$_{\text{CHEMICAL}}$

- Pre-treatment of [ATRA]$_{\text{CHEMICAL}}$, **can decrease**, the overexpression of ([cyclin_D1]$_{\text{GENE}}$, [E2F-1]$_{\text{GENE}}$), where ([cyclin_D1]$_{\text{GENE}}$, [E2F-1]$_{\text{GENE}}$), **induced by**, [B(a)P]$_{\text{CHEMICAL>}}$

- < [ATRA]$_{\text{CHEMICAL}}$, **decrease**, ([cyclin_D1]$_{\text{GENE}}$, [E2F-1]$_{\text{GENE}}$)
  < ([cyclin_D1]$_{\text{GENE}}$, [E2F-1]$_{\text{GENE}}$), induced by, [B(a)P]$_{\text{CHEMICAL>}}$>>

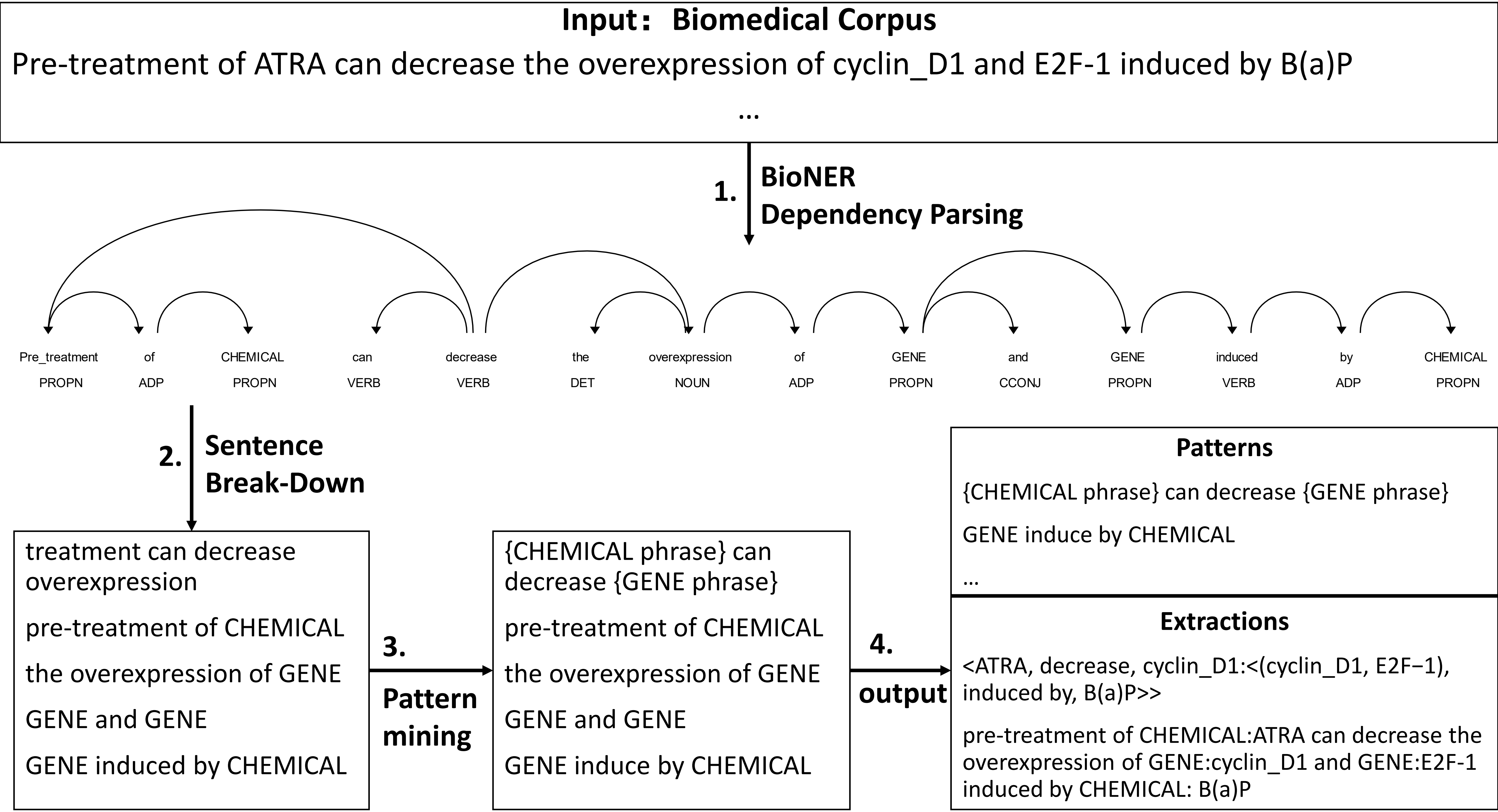
- Hierarchical structure

Outline

- Introduction
- Framework
- Evaluation
- Conclusion

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** …



Input： Biomedical Corpus

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

...

treatment can decrease overexpression

pre-treatment of CHEMICAL

the overexpression of GENE

GENE and GENE

GENE induced by CHEMICAL

{CHEMICAL phrase} can decrease {GENE phrase}

pre-treatment of CHEMICAL

the overexpression of GENE

GENE and GENE

GENE induce by CHEMICAL

Patterns

{CHEMICAL phrase} can decrease {GENE phrase}

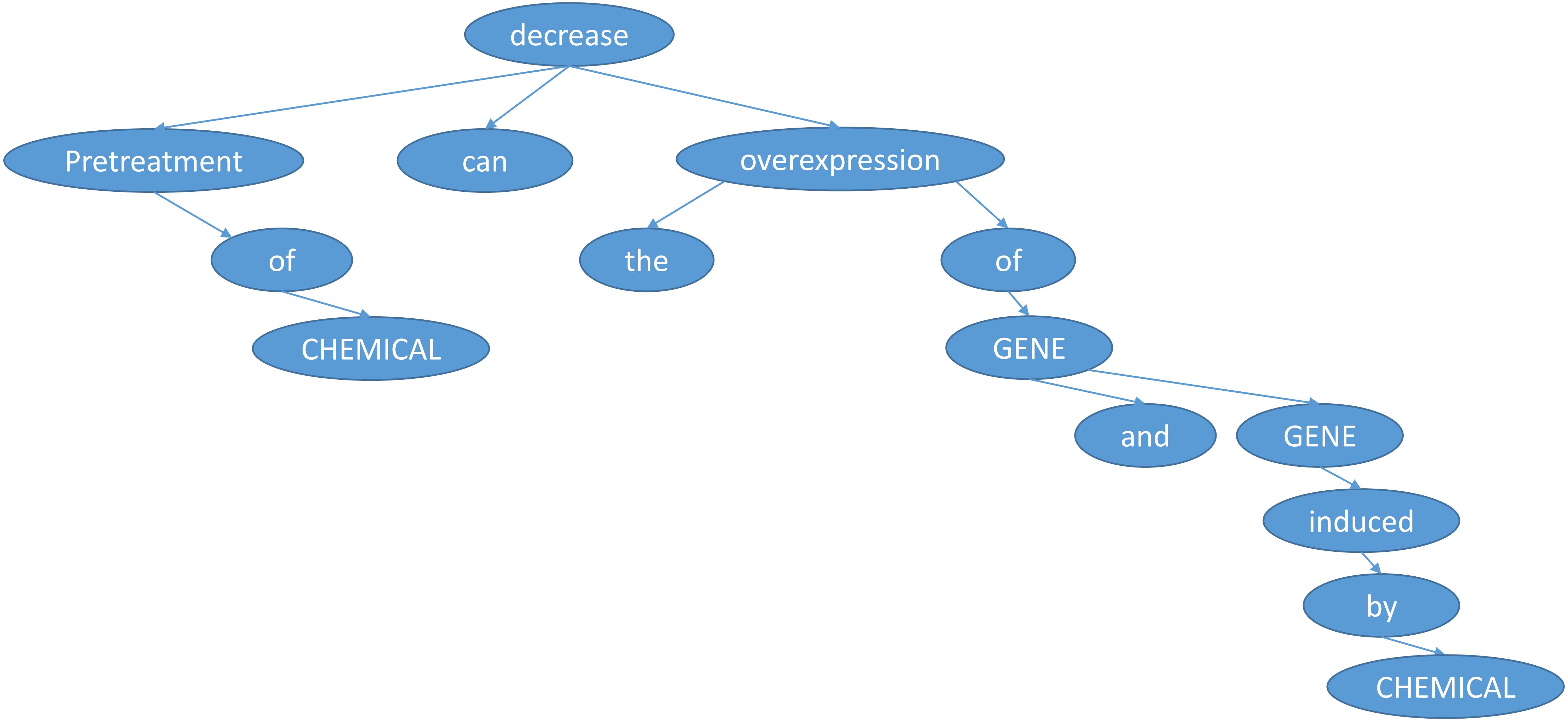
GENE induce by CHEMICAL

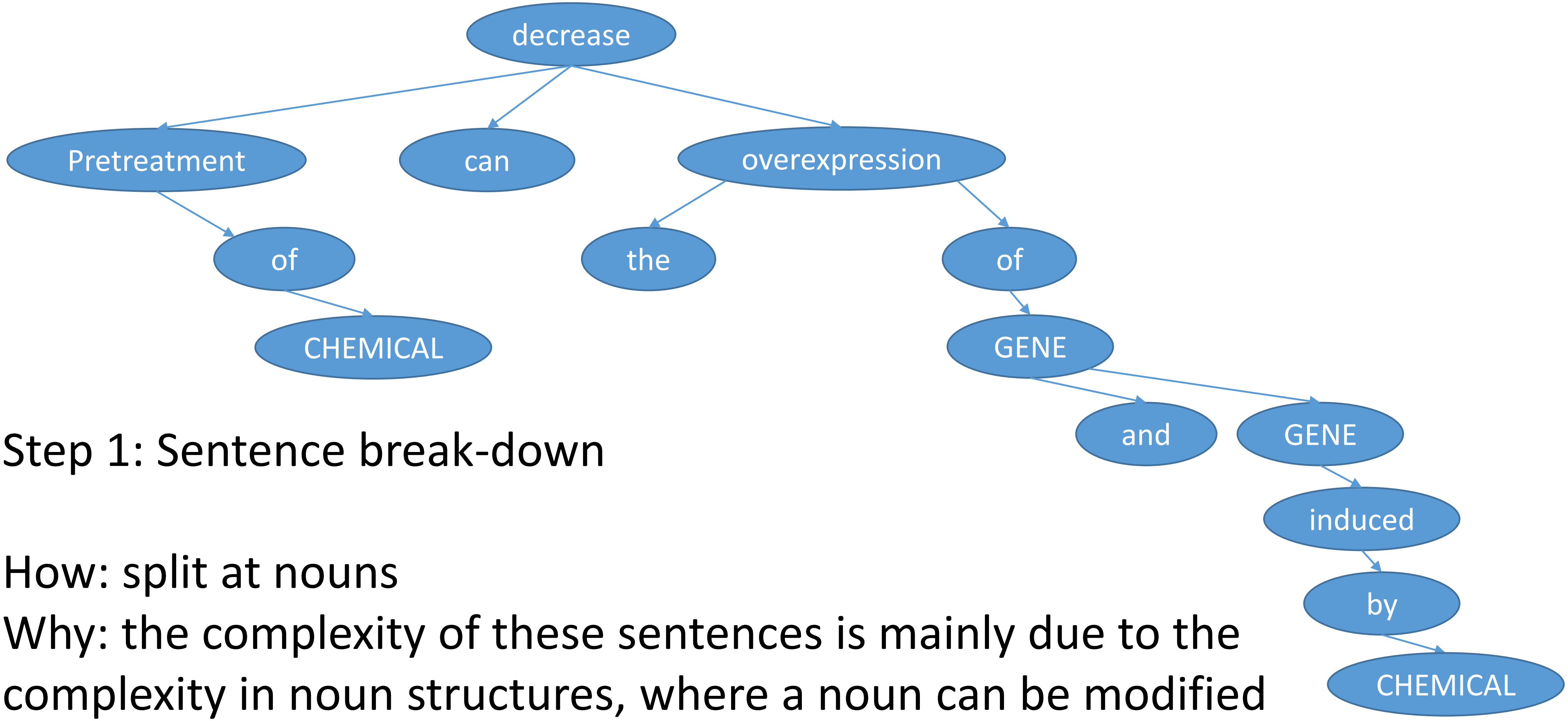
...

Extractions

<ATRA, decrease, cyclin_D1:<(cyclin_D1, E2F−1), induced by, B(a)P>>

pre-treatment of CHEMICAL:ATRA can decrease the overexpression of GENE:cyclin_D1 and GENE:E2F-1 induced by CHEMICAL: B(a)P

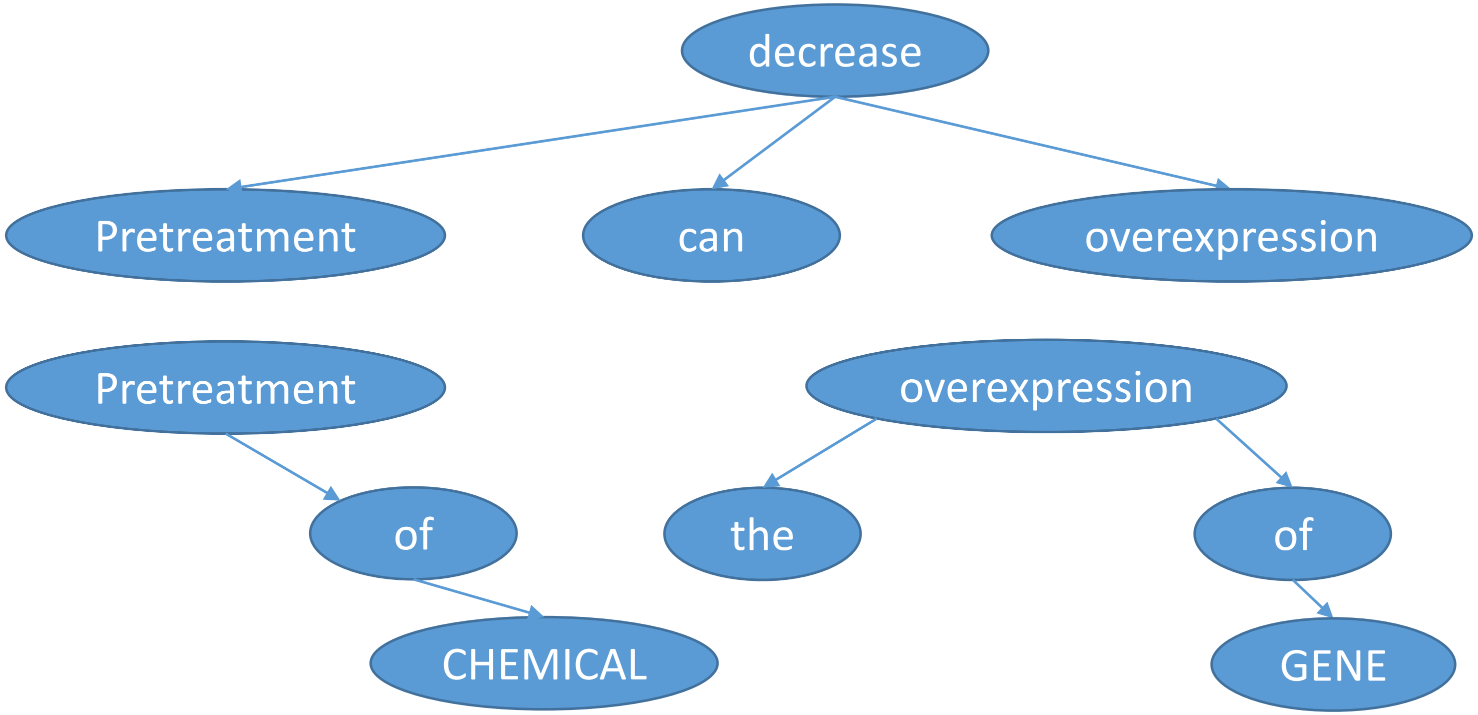

decrease
Pretreatment
can
overexpression
of
the
of
CHEMICAL
GENE
and
GENE
induced
by
CHEMICAL

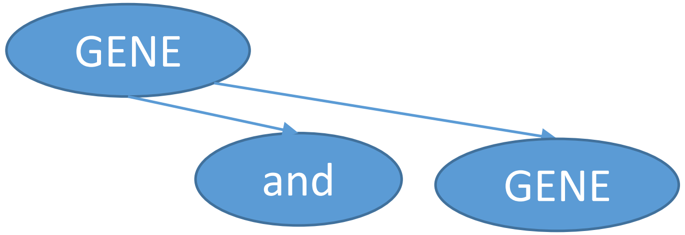


Step 1: Sentence break-down

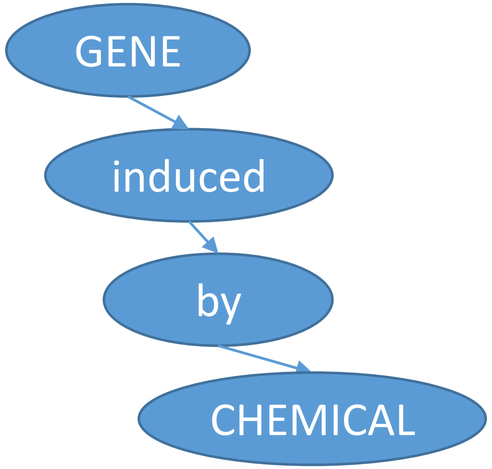
How: split at nouns

Why: the complexity of these sentences is mainly due to the complexity in noun structures, where a noun can be modified by other nouns, adjectives, adjectival clauses, etc.

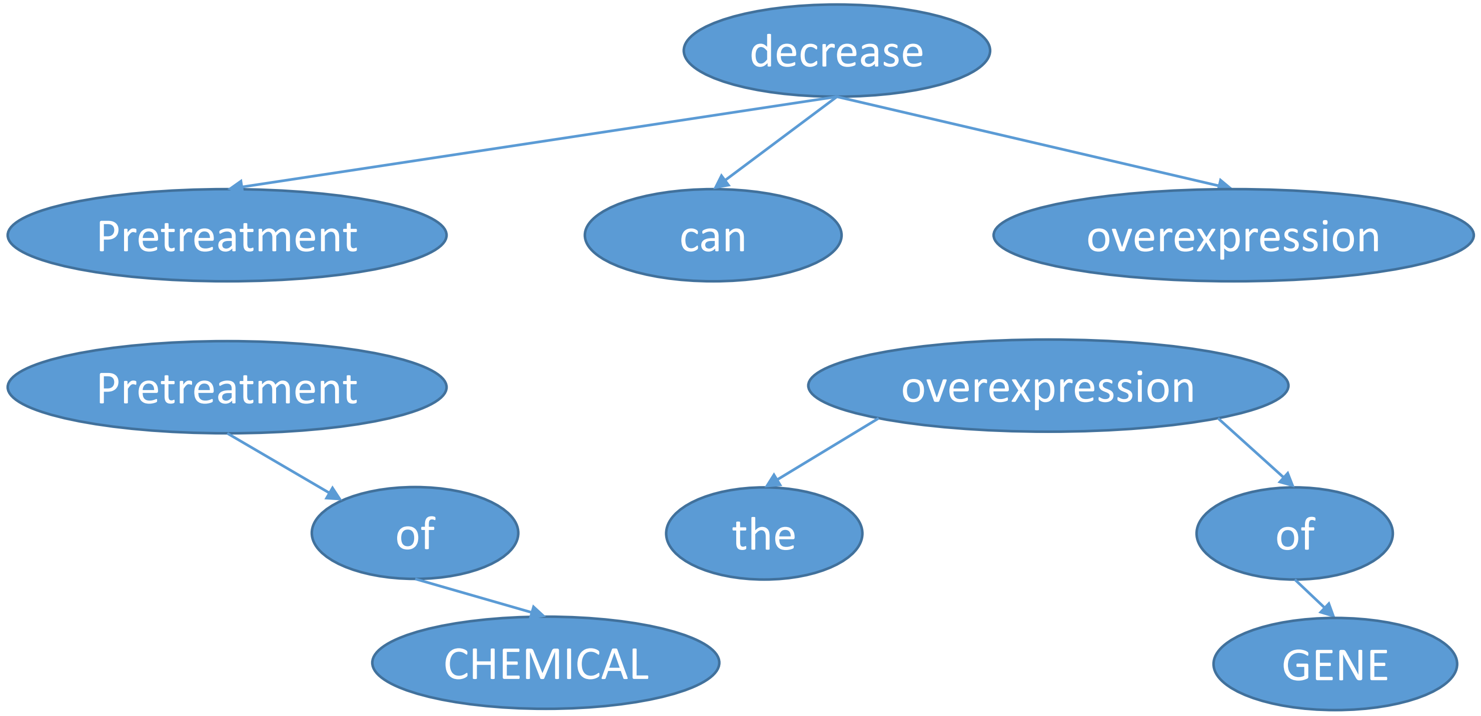


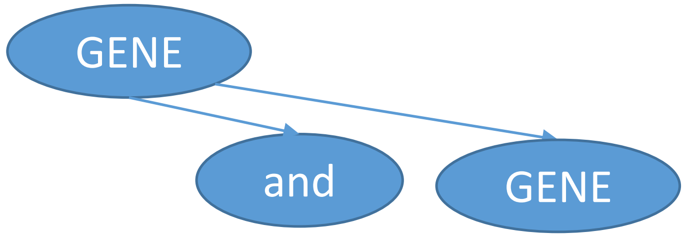


Step 1: Sentence break-down



decrease: treatment can decrease overexpression
Pretreatment: Pretreatment of **CHEMICAL**
overexpression: the overexpression of **GENE**
GENE: **GENE** and **GENE**
GENE: **GENE** induced by **CHEMICAL**





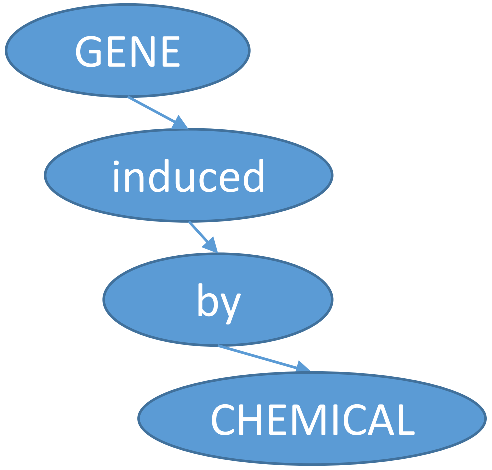
Step 2: Pattern mining on short sentences

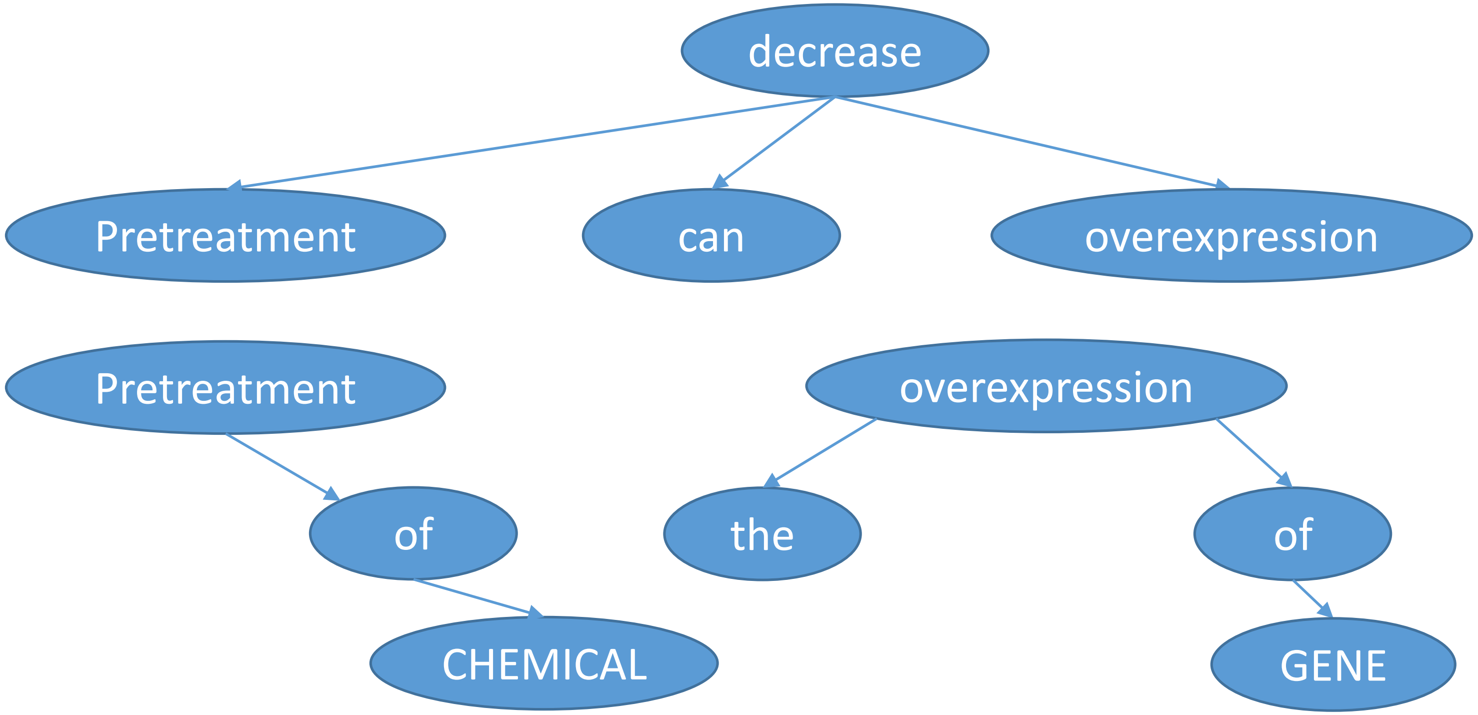
Constraint: words in a pattern should be connected on the tree

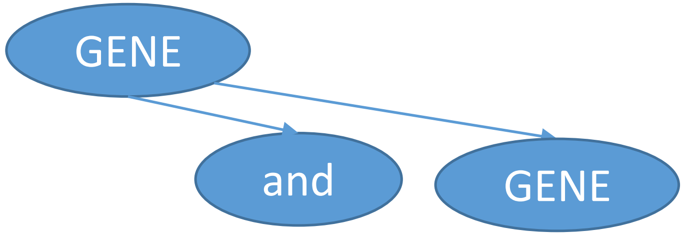
Eg. “pretreatment of CHEMICAL” √ “and GENE” ×

Constraint: pattern should contain (one entity + one non-stop-word), or more than one entity

Constraint: frequency is high

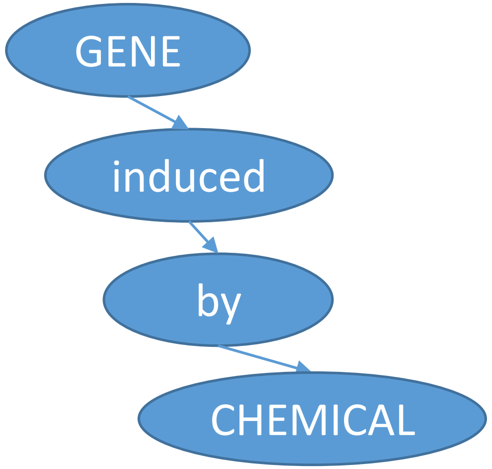


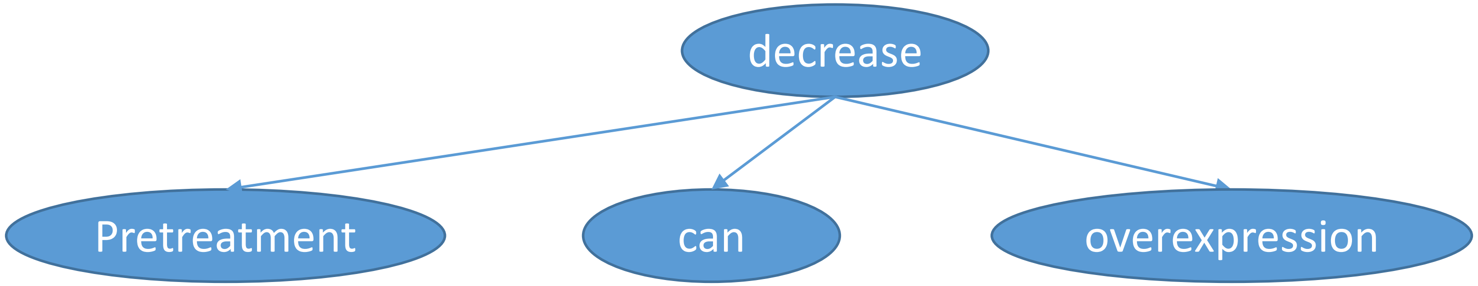


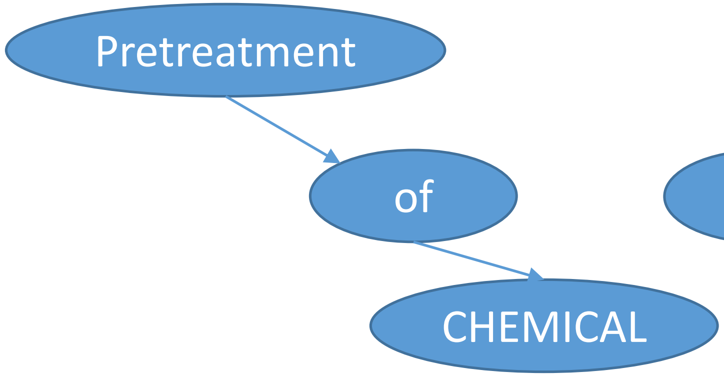


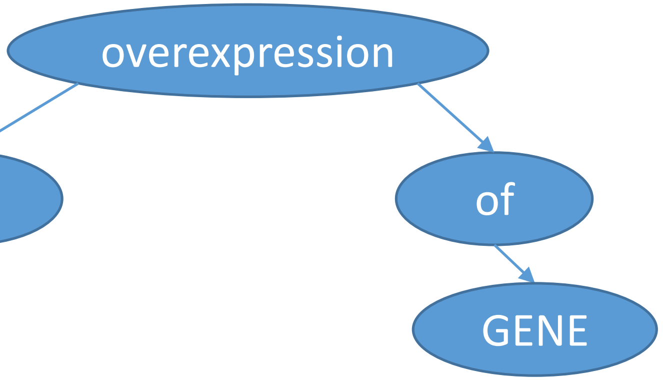
Step 2: Pattern mining on short sentences

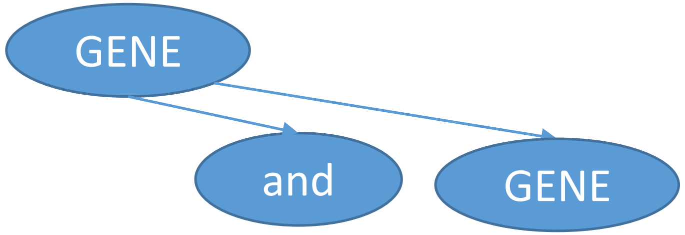
pretreatment of **CHEMICAL**
the overexpression of **GENE**
GENE and **GENE**
GENE induced by **CHEMICAL**











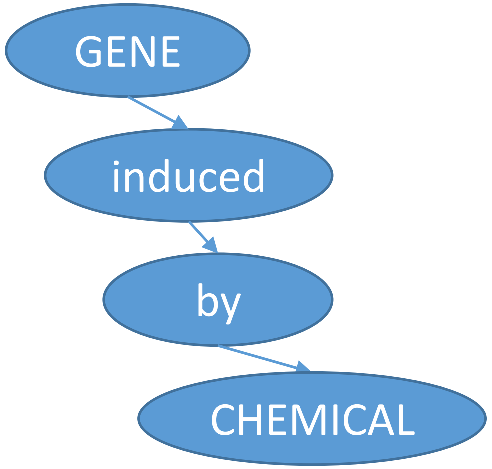
Step 3: Pattern grouping

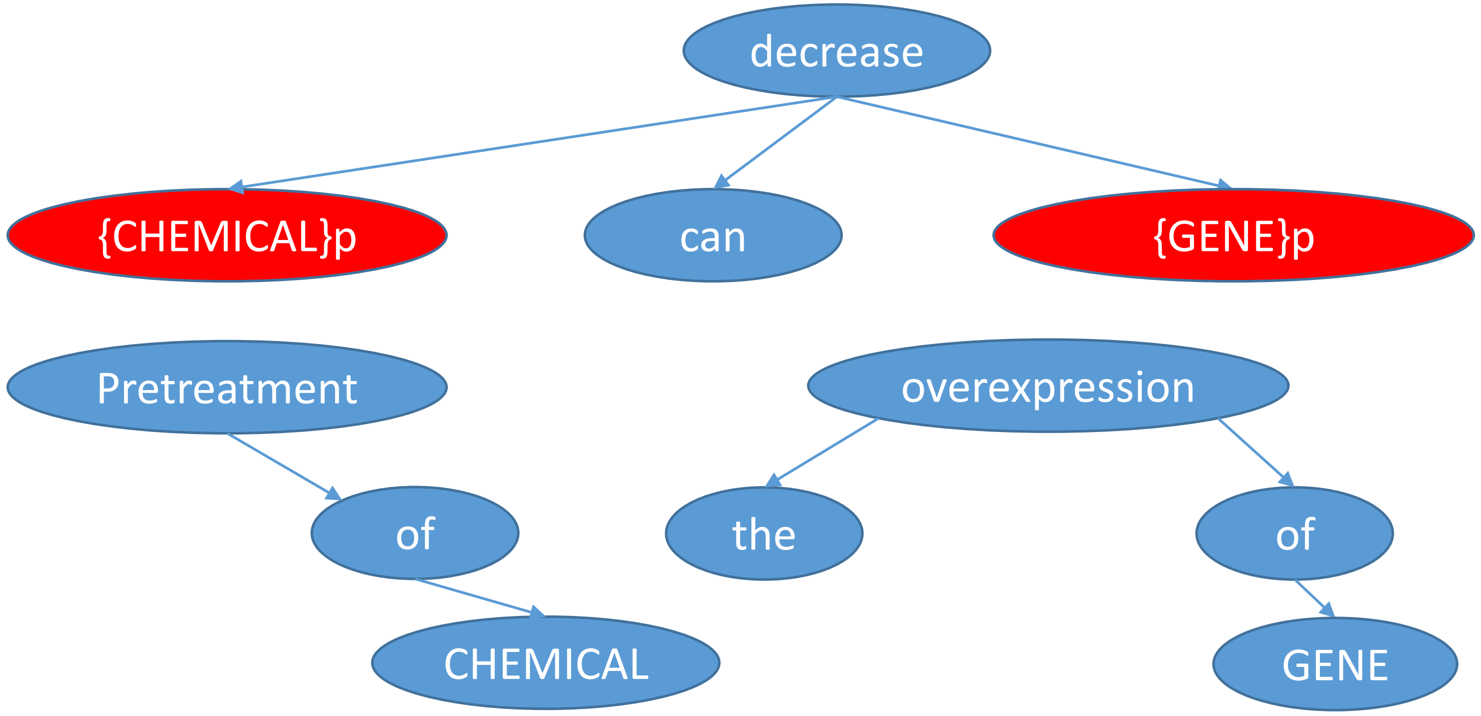
pretreatment of **CHEMICAL** ← **CHEMICAL** phrase

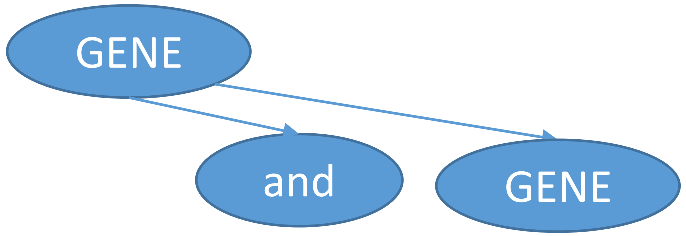
the overexpression of **GENE** ← **GENE** phrase

GENE and **GENE** ← **GENE** phrase

GENE induced by **CHEMICAL** ← **GENE** phrase







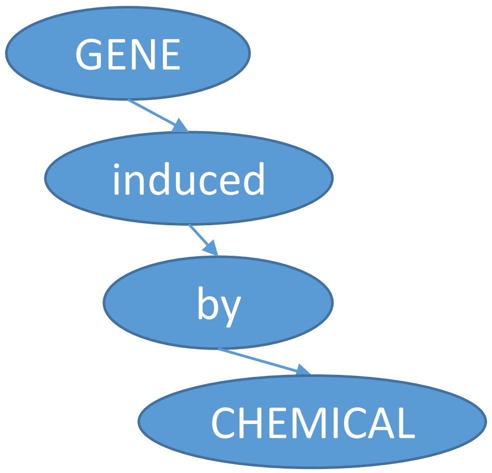
Step 3: Pattern grouping

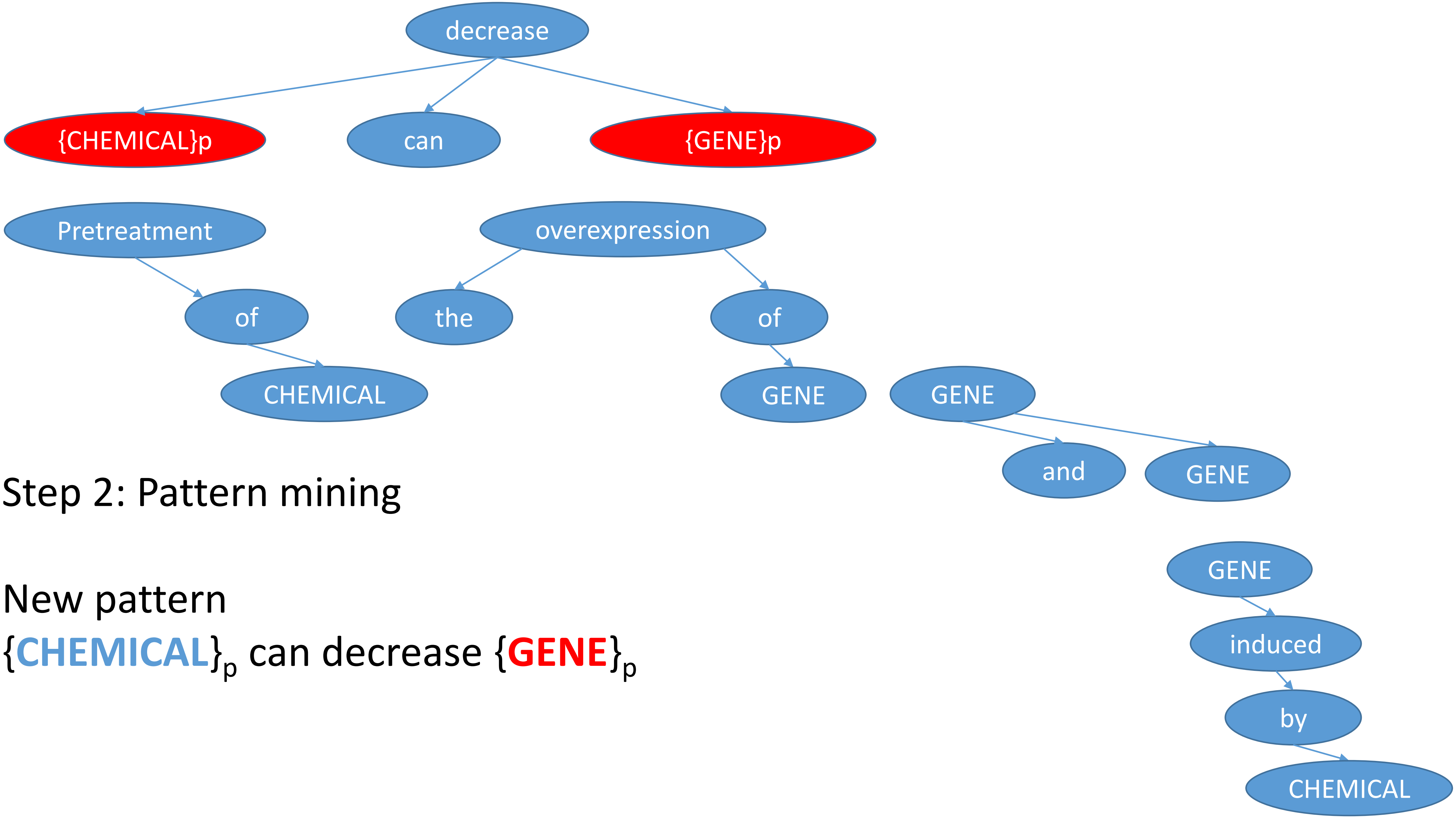
pretreatment of **CHEMICAL** ← **CHEMICAL** phrase

the overexpression of **GENE** ← **GENE** phrase

GENE and **GENE** ← **GENE** phrase

GENE induced by **CHEMICAL** ← **GENE** phrase

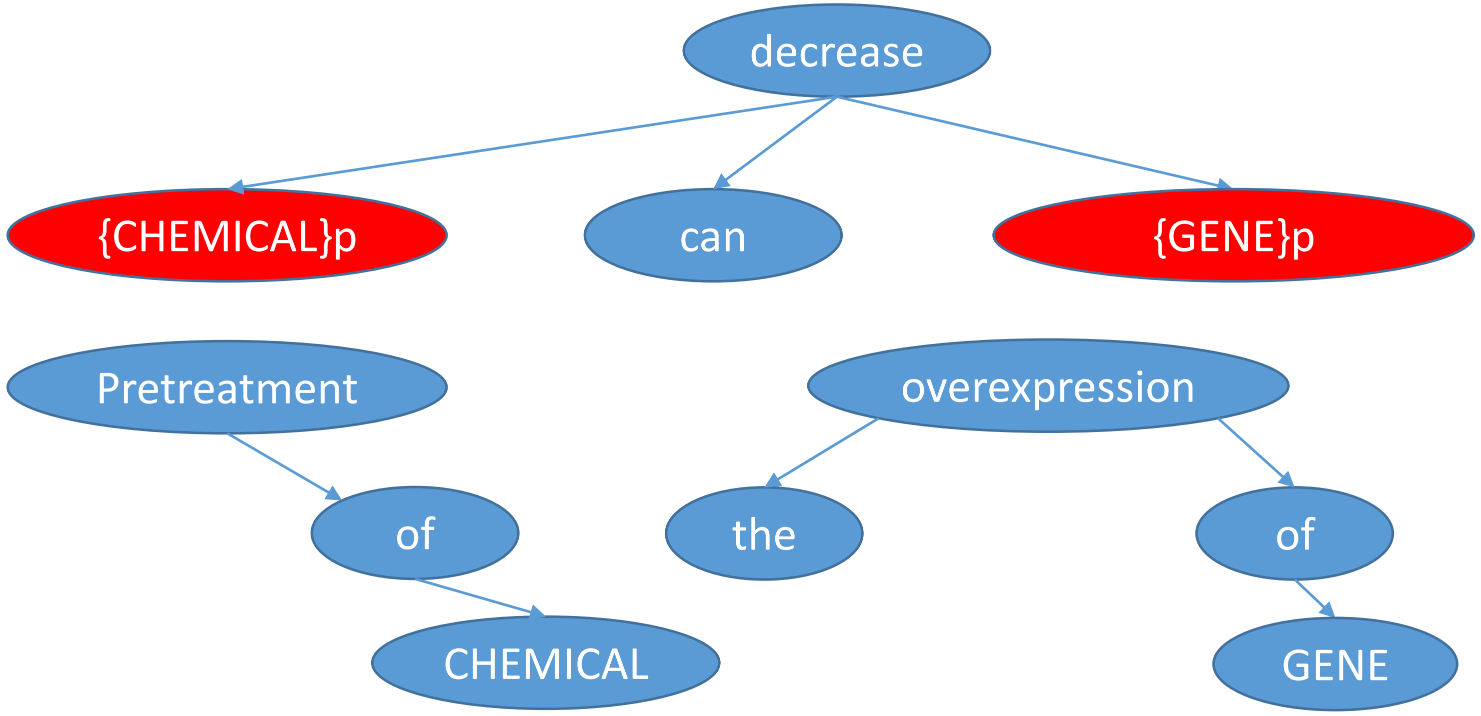


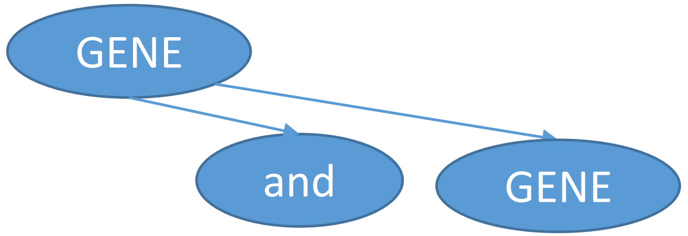


Step 2: Pattern mining

New pattern

{CHEMICAL}$_p$ can decrease {GENE}$_p$

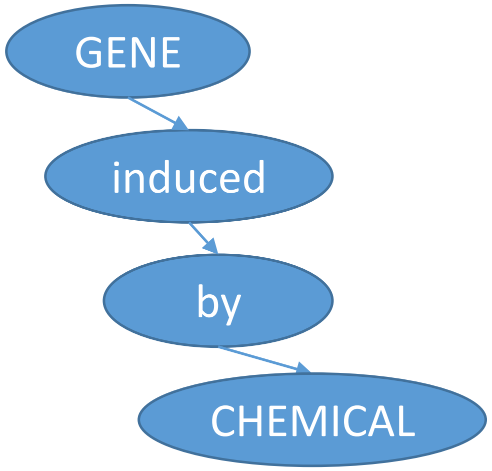




Step 3: Pattern grouping

perform clustering to group synonymous meta patterns

{CHEMICAL}p can decrease {GENE}p
{CHEMICAL}p decrease {GENE}p
{GENE}p be decrease by {CHEMICAL}p



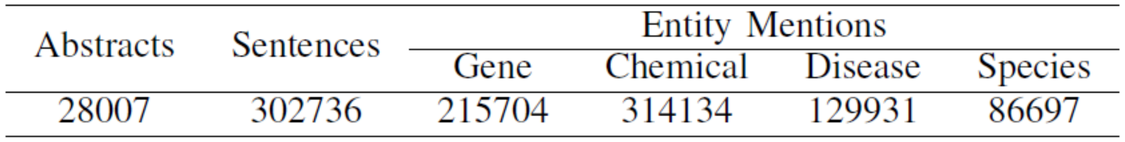
Outline

- Introduction
- Framework
- Evaluation
- Conclusion

Experiments

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

BASIC STATISTICS OF THE SUBSET CORPUS.

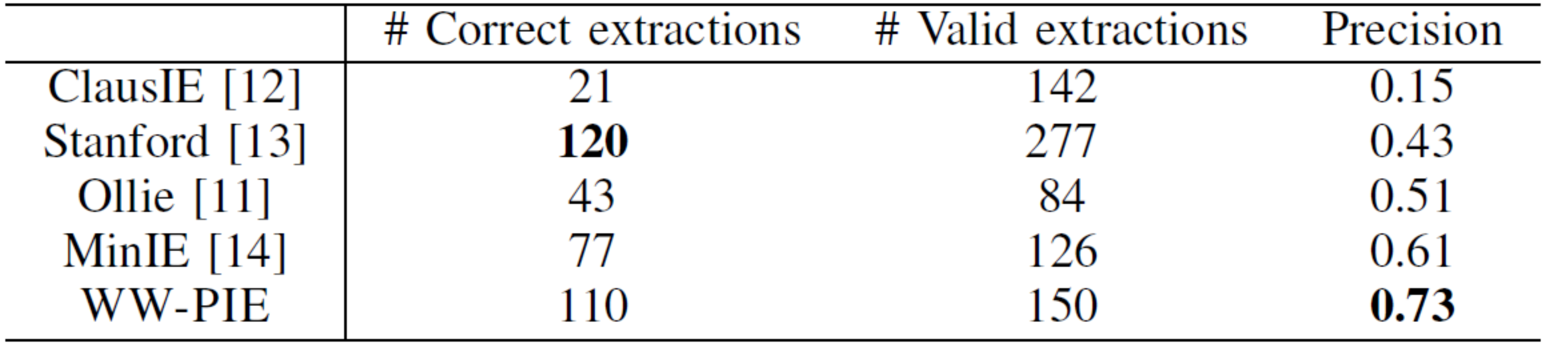


Abstracts	Sentences	Entity Mentions			
		Gene	Chemical	Disease	Species
28007	302736	215704	314134	129931	86697

- Baselines:
  - 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 state-of-the-art OpenIE systems

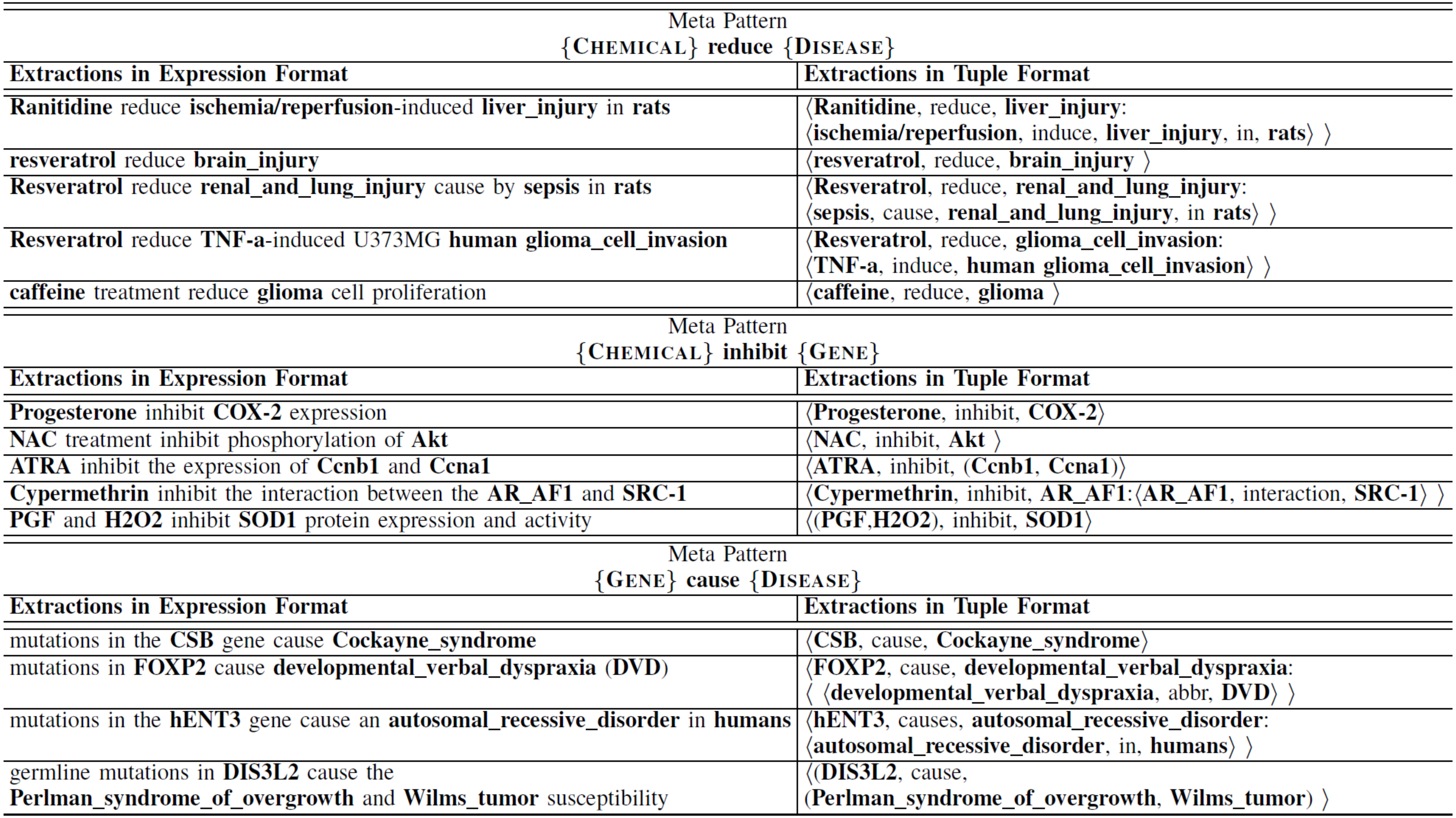
- Randomly sample 96 sentences for human labeling
- one tuple will be judged as correct if it reads smoothly and meets the fact described in the sentence



	# Correct extractions	# Valid extractions	Precision
ClausIE [12]	21	142	0.15
Stanford [13]	**120**	277	0.43
Ollie [11]	43	84	0.51
MinIE [14]	77	126	0.61
WW-PIE	110	150	**0.73**

- Note: we observe that Stanford OpenIE produces over 60 extractions for one sentence, which may be undesired for some applications.

Pattern and Extraction Examples



Meta Pattern
{CHEMICAL} **reduce** {DISEASE}

Extractions in Expression Format	**Extractions in Tuple Format**
Ranitidine reduce **ischemia/reperfusion**-induced **liver_injury** in **rats**	⟨**Ranitidine**, reduce, **liver_injury**: ⟨**ischemia/reperfusion**, induce, **liver_injury**, in, **rats**⟩ ⟩
resveratrol reduce **brain_injury**	⟨**resveratrol**, reduce, **brain_injury** ⟩
Resveratrol reduce **renal_and_lung_injury** cause by **sepsis** in **rats**	⟨**Resveratrol**, reduce, **renal_and_lung_injury**: ⟨**sepsis**, cause, **renal_and_lung_injury**, in **rats**⟩ ⟩
Resveratrol reduce **TNF-a**-induced U373MG **human glioma_cell_invasion**	⟨**Resveratrol**, reduce, **glioma_cell_invasion**: ⟨**TNF-a**, induce, **human glioma_cell_invasion**⟩ ⟩
caffeine treatment reduce **glioma** cell proliferation	⟨**caffeine**, reduce, **glioma** ⟩

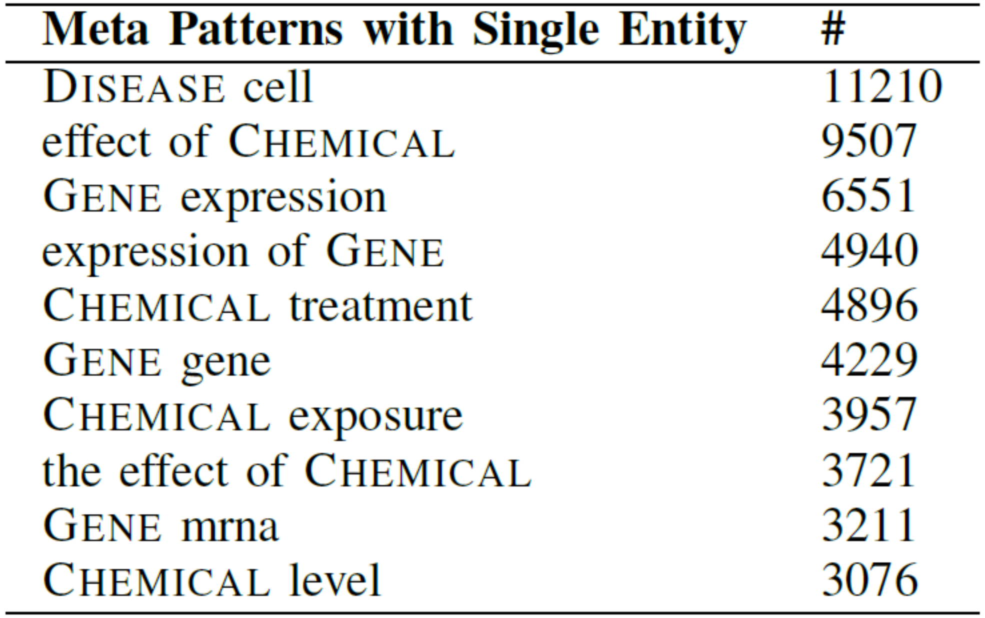
Meta Pattern
{CHEMICAL} **inhibit** {GENE}

Extractions in Expression Format	**Extractions in Tuple Format**
Progesterone inhibit **COX-2** expression	⟨**Progesterone**, inhibit, **COX-2**⟩
NAC treatment inhibit phosphorylation of **Akt**	⟨**NAC**, inhibit, **Akt** ⟩
ATRA inhibit the expression of **Ccnb1** and **Ccna1**	⟨**ATRA**, inhibit, (**Ccnb1**, **Ccna1**)⟩
Cypermethrin inhibit the interaction between the **AR_AF1** and **SRC-1**	⟨**Cypermethrin**, inhibit, **AR_AF1**:⟨**AR_AF1**, interaction, **SRC-1**⟩ ⟩
PGF and **H2O2** inhibit **SOD1** protein expression and activity	⟨(**PGF**,**H2O2**), inhibit, **SOD1**⟩

Meta Pattern
{GENE} **cause** {DISEASE}

Extractions in Expression Format	**Extractions in Tuple Format**
mutations in the **CSB** gene cause **Cockayne_syndrome**	⟨**CSB**, cause, **Cockayne_syndrome**⟩
mutations in **FOXP2** cause **developmental_verbal_dyspraxia** (**DVD**)	⟨**FOXP2**, cause, **developmental_verbal_dyspraxia**: ⟨ ⟨**developmental_verbal_dyspraxia**, abbr, **DVD**⟩ ⟩
mutations in the **hENT3** gene cause an **autosomal_recessive_disorder** in **humans**	⟨**hENT3**, causes, **autosomal_recessive_disorder**: ⟨**autosomal_recessive_disorder**, in, **humans**⟩ ⟩
germline mutations in **DIS3L2** cause the **Perlman_syndrome_of_overgrowth** and **Wilms_tumor** susceptibility	⟨(**DIS3L2**, cause, (**Perlman_syndrome_of_overgrowth**, **Wilms_tumor**) ⟩

Top 10 Single Entity Patterns

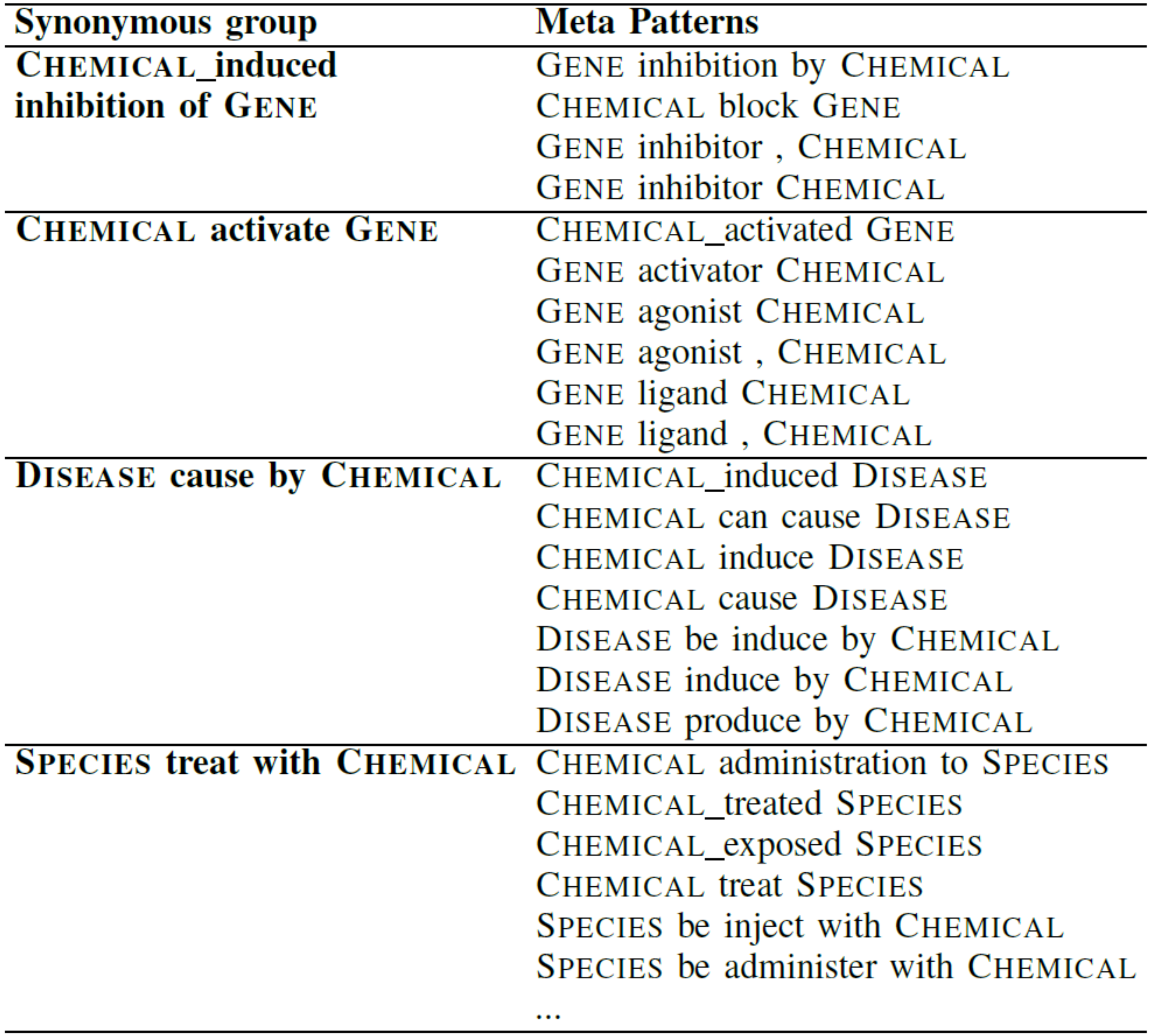


Meta Patterns with Single Entity	#
DISEASE cell	11210
effect of CHEMICAL	9507
GENE expression	6551
expression of GENE	4940
CHEMICAL treatment	4896
GENE gene	4229
CHEMICAL exposure	3957
the effect of CHEMICAL	3721
GENE mrna	3211
CHEMICAL level	3076

- Can be helpful in named entity recognition tasks

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

Synonymous Pattern Group Examples



Synonymous group	**Meta Patterns**
CHEMICAL_induced inhibition of GENE	GENE inhibition by CHEMICAL CHEMICAL block GENE GENE inhibitor , CHEMICAL GENE inhibitor CHEMICAL
CHEMICAL activate GENE	CHEMICAL_activated GENE GENE activator CHEMICAL GENE agonist CHEMICAL GENE agonist , CHEMICAL GENE ligand CHEMICAL GENE ligand , CHEMICAL
DISEASE cause by CHEMICAL	CHEMICAL_induced DISEASE CHEMICAL can cause DISEASE CHEMICAL induce DISEASE CHEMICAL cause DISEASE DISEASE be induce by CHEMICAL DISEASE induce by CHEMICAL DISEASE produce by CHEMICAL
SPECIES treat with CHEMICAL	CHEMICAL administration to SPECIES CHEMICAL_treated SPECIES CHEMICAL_exposed SPECIES CHEMICAL treat SPECIES SPECIES be inject with CHEMICAL SPECIES be administer with CHEMICAL ...

Outline

- Introduction
- Framework
- Evaluation
- Conclusion

Conclusion and Future Work

- WW-PIE
  - can extract all variety of the relation tuples from large biomedical literature corpora
  - resolves the long and complicated sentence structures by breaking down the sentences
  - groups 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

Thank you! Questions?