Scientific Text Mining and Knowledge Graphs

Chapter 1
Part 1: Phrase Mining

Presenter: Jingbo Shang

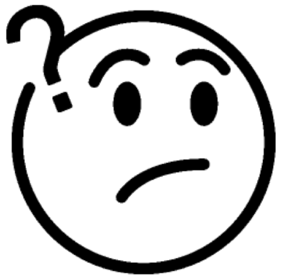
University of California, San Diego

jshang@ucsd.edu

Why Phrase Mining?

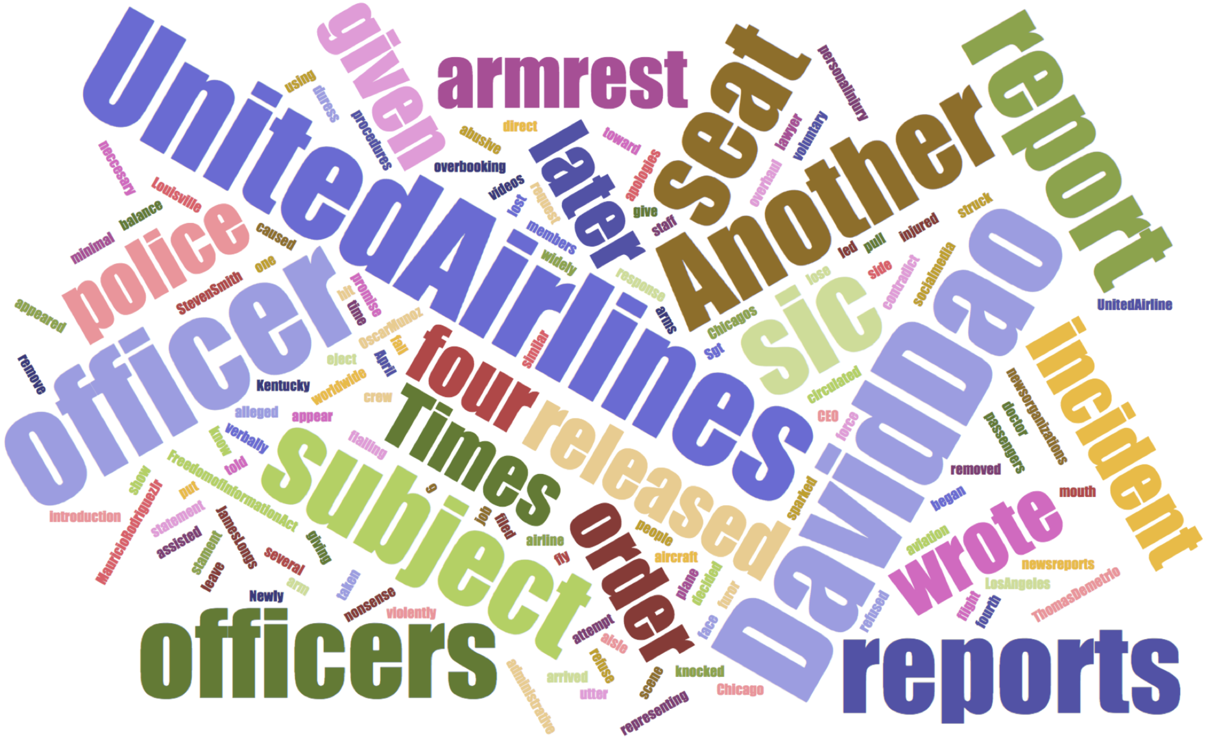
- Analyzes US news articles on April 9, 2017
- **Before Phrase Mining**

- Which “United”?
  - United States?
  - United Parcel Service?
- What’s “Dao”?
  - A person?
  - A place?



Why Phrase Mining?

- Analyzes US news articles on April 9, 2017
- **After Phrase Mining**



- United Airline
- David Dao → A person

United Express Flight 3411 incident

Phrase Mining: A Keystone

- Phrase mining is a **keystone** towards understanding texts
  - Entities, Relational Phrases, **…**

- It can facilitate various **applications** in Natural Language Processing (NLP), Information Retrieval (IR), Text Mining
  - Document Analysis
  - Indexing in Search Engine
  - Key-Phrases for Topic Modeling
  - Summarization
  - Text-based Predictive Analytics
  - **…**

Quality Phrase Mining from Massive **Domain-Specific** Corpora

- Quality phrase mining seeks to extract a **ranked list** of phrases with decreasing quality from a **large collection of documents**
- Examples:

Scientific Papers

Expected Results
…
data mining
machine learning
information retrieval
…
support vector machine
…

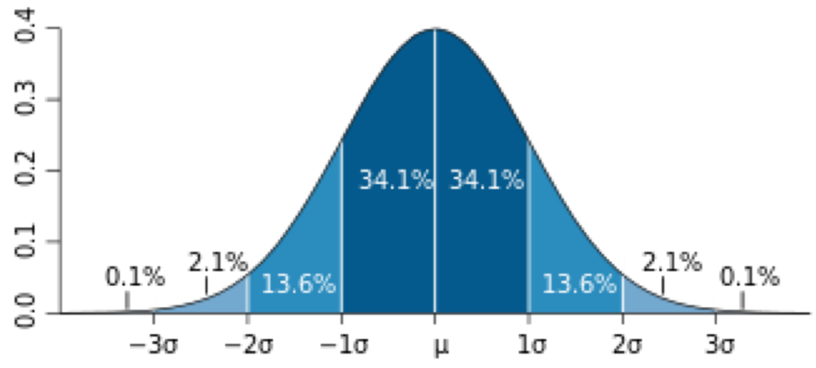
News Articles

Expected Results
…
US President
Anderson Cooper
Barack Obama
…
Obama administration
…

- Existing NLP-based methods (e.g., noun phrase chunking, parsing) rely on ***extensive annotations from domain experts***

TopMine (VLDB'15): Our Unsupervised Pioneer on "Segmentation"

- Statistical signal: Significance score (α)
  - How many standard deviations (σ) away from the mean?
  - $\alpha(\mathrm{A}, \mathrm{B}) \approx \frac{\mathrm{freq}(A \oplus B) - E[\mathrm{freq}(A \oplus B)]}{\sqrt{\mathrm{freq}(A \oplus B)}}$
- A greedy merge algorithm to segment sentences:
  - $\alpha(\mathrm{A}, \mathrm{B}) >$ the given threshold ➔ "A B" is a phrase (a merge)



Segmentation: | Markov Blanket | Feature Selection | for | Support Vector Machine

Support Vector

α > the threshold

Raw Input: | Markov | Blanket | Feature | Selection | for | Support | Vector | Machine

SegPhrase (SIGMOD'15): Quality Estimation using Expert Labels

Input Raw Corpus

Multiple Statistical Signals

1. Significance score
2. $\text{PMI}(A, B) = \log \frac{P(A \oplus B)}{P(A)P(B)}$
3. Chi-square: $\chi^2 = \sum \frac{(O-E)^2}{E}$
4. Inverse Term Frequency (IDF)
5. Stopword ratio
6. Capitalization signals

... ...

0	this paper
0	experiments show
……	
1	support vector machine
1	np hard
……	

Quality Estimation based on Expert-Provided Labels

Estimated Quality Score for Every Phrase

0.999, support vector machine
0.999, objective function
0.999, source code
0.999, upper bound
0.999, optical flow
0.999, hidden markov model
…
0.851, period doubling
0.851, digital game based learning
0.850, bus architecture
0.850, parabolic partial differential equations
…
0.000, this paper
…

SegPhrase (SIGMOD'15): Phrasal Segmentation using Viterbi Algo

Raw Input: Markov | Blanket | Feature | Selection | for | Support | Vector | Machine

Count **Raw** N-gram **Frequency**: "Markov"+=1, "Markov Blanket"+=1, "Markov Blanket Feature"+=1, …

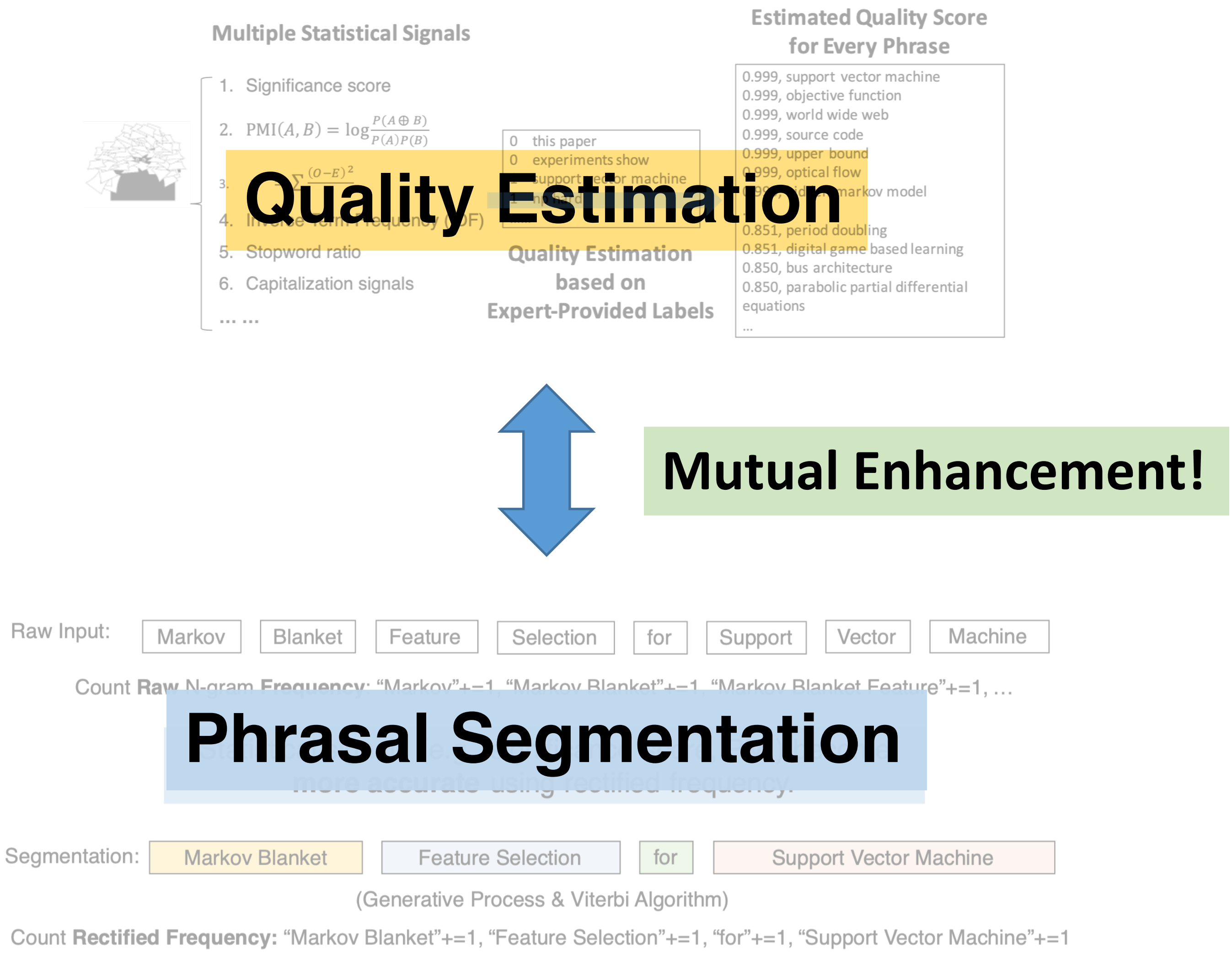
Statistical signals (e.g., significance score, PMI) become **more accurate** using rectified frequency

Segmentation: Markov Blanket | Feature Selection | for | Support Vector Machine

(Viterbi Algorithm based on Generative Process & Estimated Quality Scores)

Count **Rectified Frequency:** "Markov Blanket"+=1, "Feature Selection"+=1, "for"+=1, "Support Vector Machine"+=1

SegPhrase (SIGMOD'15): Quality Estimation ⇔ Phrasal Segmentation



- Key Ideas of SegPhrase
  - Weak Supervision
    - Unifies multiple stat-signals
  - Text Segmentation
    - Rectifies phrase frequency

SegPhrase (SIGMOD'15): Reliance on Expert-Provided Labels

Multiple Statistical Signals

1. Significanc
2. PMI(A, B)
3. Chi-square
4. Inverse Te
5. Stopword ratio
6. Capitalization signals

… …

Input Raw Corpus

0 this paper

Expert-Provided Labels

Estimated Quality Score for Every Phrase

0.851, digital game based learning
0.850, bus architecture
0.850, parabolic partial differential equations
…
0.000, this paper
…

- 300 *well-designed* binary labels for ~1GB input
- Labeling 300 phrases is an easy task
- But knowing which 300 phrases **from millions of candidate phrases** to be labeled is a problem

AutoPhrase (TKDE'18): Automated Phrase Mining without Expert Labels

Input Raw Corpus

Multiple Statistical Signals

1. Significance score
2. $\mathrm{PMI}(A, B) = \log \frac{P(A \oplus B)}{P(A)P(B)}$
3. Chi-square: $\chi^2 = \sum \frac{(O-E)^2}{E}$
4. Inverse Term Frequency (IDF)
5. Stopword ratio
6. Capitalization signals

… …

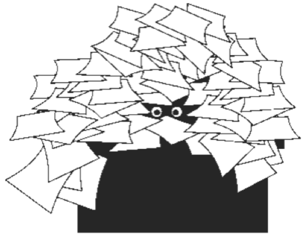
Quality Estimation based on Knowledge Bases

Estimated Quality Score for Every Phrase

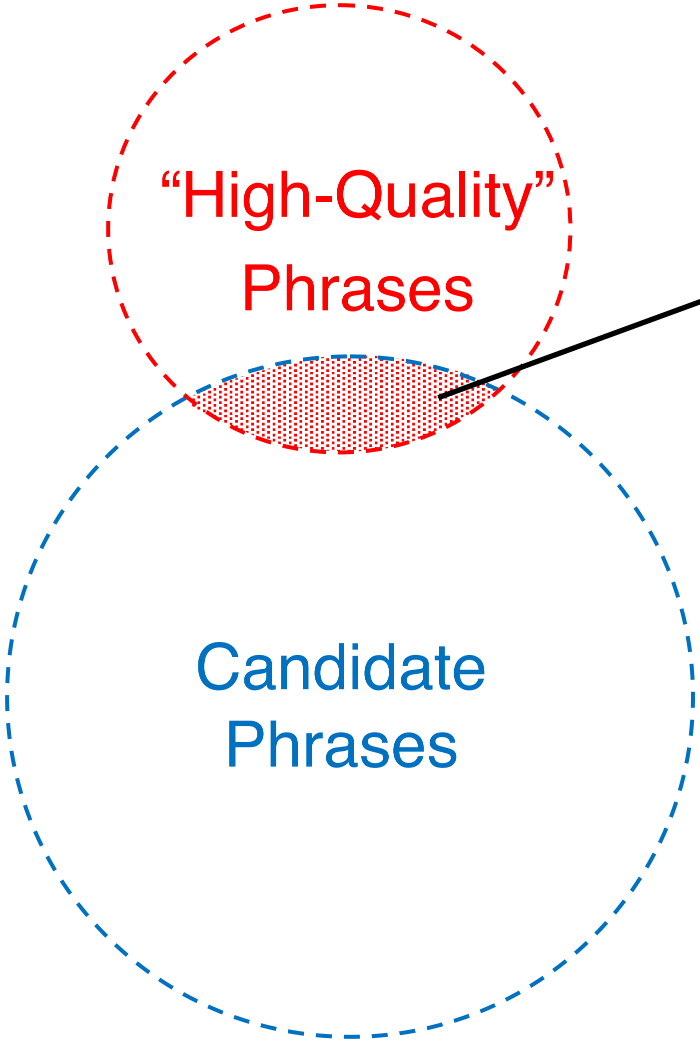
0.999, support vector machine
0.999, objective function
0.999, source code
0.999, upper bound
0.999, optical flow
0.999, hidden markov model
…
0.851, period doubling
0.851, digital game based learning
0.850, bus architecture
0.850, parabolic partial differential equations
…
0.000, this paper
…

AutoPhrase (TKDE'18): Distant Supervision & its Challenges

Knowledge Bases



Input Raw Corpus

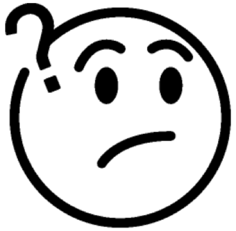


- A ***free***, clean pool of high-quality phrases

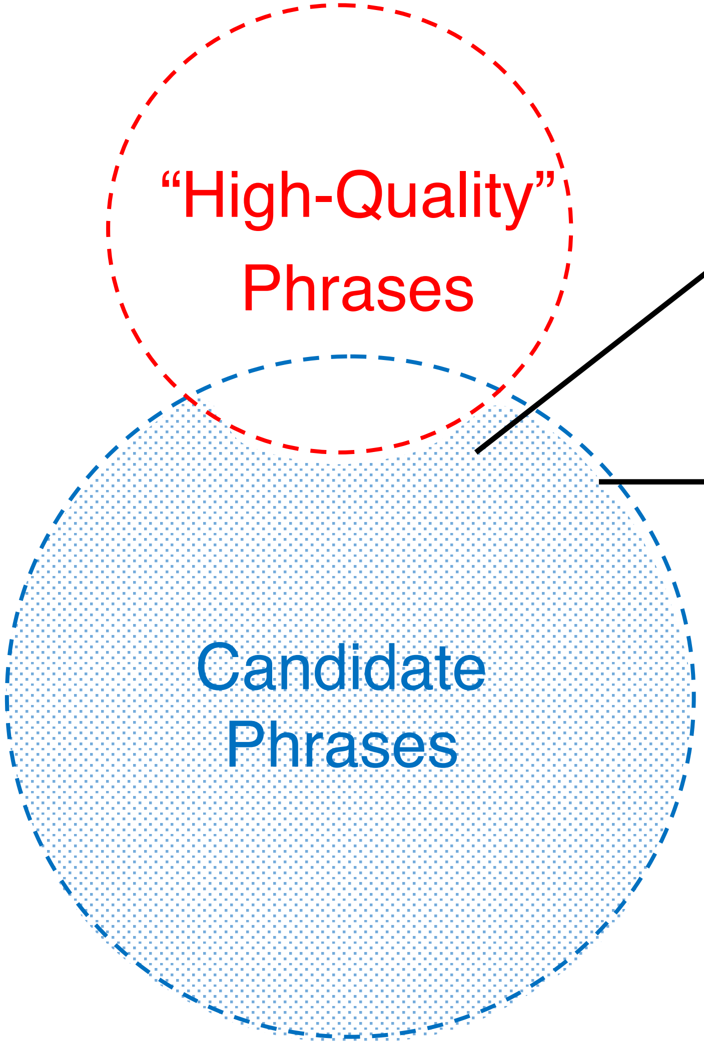
Positive Pool

- Knowledge bases are usually ***incomplete***
  - Much more phrases to be discovered from candidates

There is no mention about "low-quality" phrases in knowledge bases



AutoPhrase (TKDE'18): Negative Sampling from Noisy Negative Pool



~10%

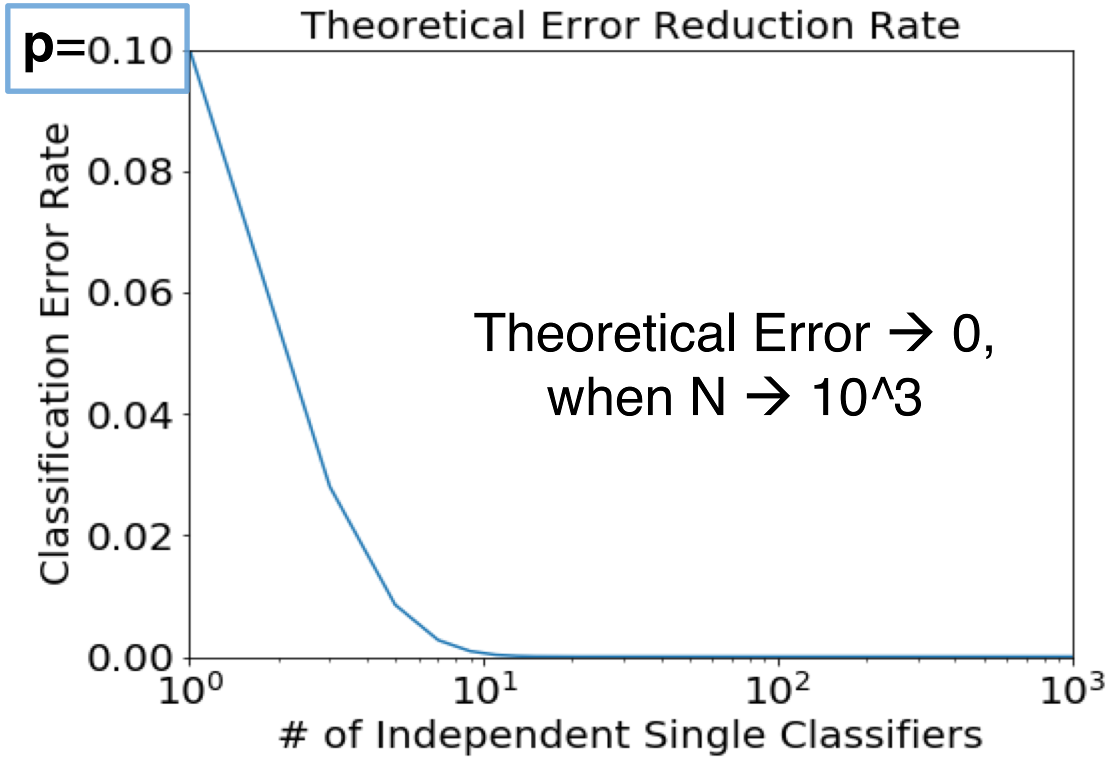
(based on human evaluation on a random sample from candidates)

NOISY Negative Pool

- Negative sampling becomes feasible
- Key Ideas of AutoPhrase
  - Distant Supervision → ***Automatic***
  - Ensemble on massive noisy negative examples

AutoPhrase (TKDE'18): Sampling + Ensemble → Robust Classifier

- **Single Classifier** based on randor
  - Error rate **p** ≈ 0.1 based on our evalua
- **Ensemble (N Independent Class**
  - Average their predictions → Denoise
  - The majority of these classifiers make
    - $Error(N) = \sum_{n=\lceil \frac{N}{2} \rceil}^{N} \binom{N}{n} p^n (1-p)^{N-n}$



Phrase Mining: Empirical Evaluation – Precision Recall Curve

- Output: a **ranked list** of phrases with decreasing quality

Scientific Papers

Output Results

…
data mining
machine learning
information retrieval
np hard in a strong sense
…
support vector machine
np hard in a strong
…

Evaluation Process

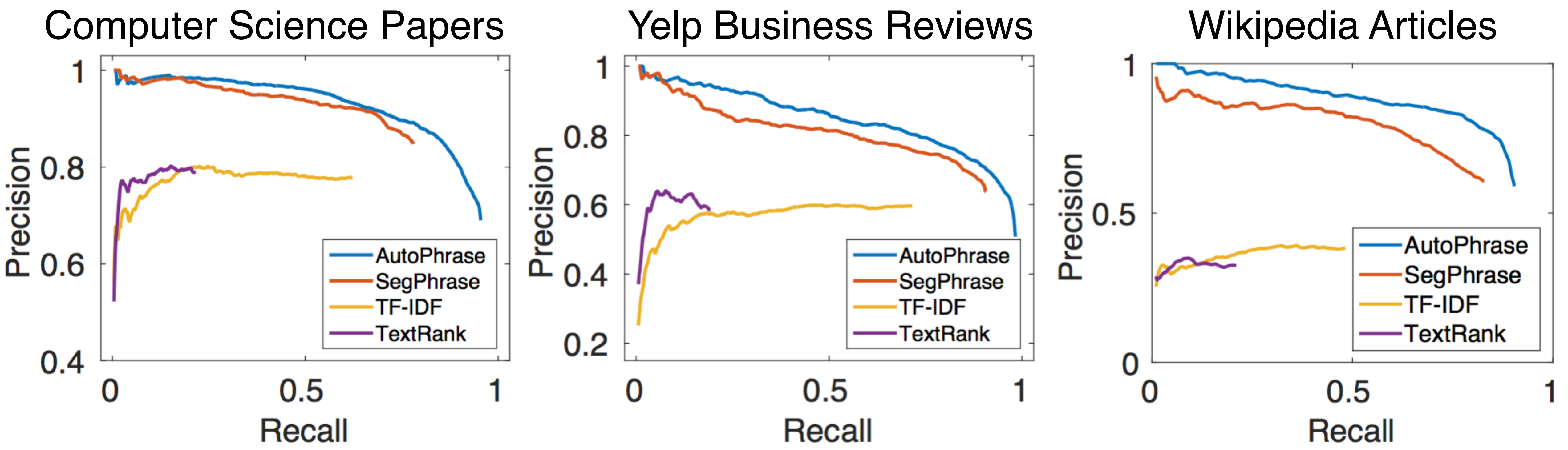
Domain Experts

Evaluation Results

…
1
1
1
1
…
1
0
…

Compute precision & recall at each rank

AutoPhrase (TKDE'18): Cross-Domain Evaluation Results

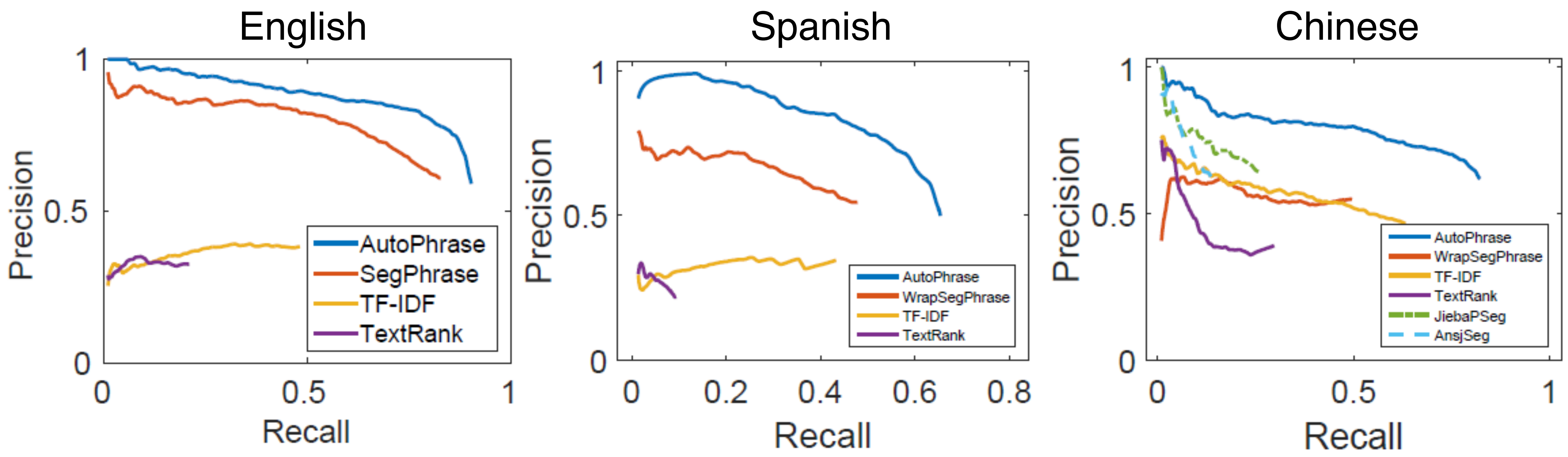


SegPhrase (SIGMOD'15): Outperformed TopMine (VLDB'15) and many other methods

TF-IDF: Stanford NLP Parser (LREC'16) + Ranked by TF-IDF

TextRank (ACL'04): Stanford NLP Parser (LREC'16) + Ranked by TextRank

AutoPhrase (TKDE'18): Cross-Language Evaluation Results



WrapSegPhrase: non-English characters → English letters & SegPhrase

JiebaSeg: Specifically for Chinese; Dictionaries & Hidden Markov Models

AnsjSeg: Specifically for Chinese; Dictionaries & Conditional Random Fields

AutoPhrase (TKDE'18): Results of Chinese Phrases from Wiki Articles

Phrase's Rank	Phrase	Translation (Explanation)
1	江苏_舜_天	(the name of a soccer team)
2	苦_艾_酒	Absinthe
3	白发_魔_女	(the name of a novel/TV-series)

- The size of positive pool is about 29,000
- AutoPhrase finds more than 116,000 quality phrases (quality score > 0.5)
  - Much more!

Phrase's Rank	Phrase	Translation (Explanation)
99,995	恒_天然	Fonterra (a company)
99,996	中国_作家_协会_副_主席	The Vice President of Writers Association of China
99,997	维他命_b	Vitamin B
99,998	舆论_导向	controlled guidance of the media
…	…	…

Phrase Mining: Impact in Various Domains

AutoPhrase to news & analyst reports

...d facts extracted from the review text

Features for "**Catch a Show**" collection

1	broadway shows
2	beacon theater
3	broadway dance center
4	broadway plays
5	david letterman show
6	radio city music hall
7	theatre shows

Features for "**Near The High Line**" collection

1	high line park
2	chelsea market
3	highline walkway
4	elevated park
5	meatpacking district
6	west side
7	old railway

...es text-based analysis from the word level to the phrase level

http://engineering.tripadvisor.com/using-nlp-to-find-interesting-collections-of-hotels/

Scientific Text Mining and Knowledge Graphs

Chapter 1
Part 2: Named Entity Recognition and Neural Language Models

Presenter: Jingbo Shang

University of California, San Diego

jshang@ucsd.edu

What’s Named Entity Recognition?

- Wikipedia:
  - **Named-entity recognition (NER)** is a subtask of **information extraction (IE)** that seeks to **locate** and **classify named entities** in text into **pre-defined categories**.
  - In IE, A **named entity** is a real-world object.
- Example
  - Input
    - Jim bought 300 shares of Acme Corp. in 2006.
  - Output
    - [Jim]$_{\text{Person}}$ bought 300 shares of [Acme Corp.]$_{\text{Organization}}$ in [2006]$_{\text{Time}}$.

Supervised Methods: **Training Data**

- Sequence labeling framework
- Two popular schemes
  - BIO: **B**egin, **I**n, **O**ut
  - BIOES: **B**egin, **I**n, **O**ut, **E**nd, **S**ingleton
  - BIOES is arguably better than BIO (Ratinov and Roth, ACL 09)
- Example:
  - LABELS: [Jim]$_{\text{Person}}$ bought 300 shares of [Acme Corp.]$_{\text{Organization}}$ in [2006]$_{\text{Time}}$.

TOKNES:	Jim	bought	300	shares	of	Acme	Corp.	in	2006	.
BIO:	B-PER	O	O	O	O	B-ORG	I-ORG	O	B-Time	O
BIOES:	S-PER	O	O	O	O	B-ORG	E-ORG	O	S-Time	O

Supervised Methods: **Neural Models**

- Two pioneer models
  - LSTM-CRF (Lample et al., NAACL'16)
  - LSTM-CNN-CRF (Ma and Hovy, ACL'16)

	LSTM-CRF	LSTM-CNN-CRF
Word-Level	Bidirectional LSTMs	Bidirection LSTMs
Character-Level	Bidirectional LSTMs	Convolutional NN

  - The first neural model that outperforms the models based on handcrafted features

“Data-Driven” Philosophy

- Key
  - Enhance NER performance without introducing any additional human annotations

- Questions
  - Can massive raw texts help?
  - Can dictionaries help?
  - Are human annotations always correct?
  - Is Tokenizer always good?
  - …

Questions

- ❑ Can massive raw texts help?
- ❑ Can dictionaries help?

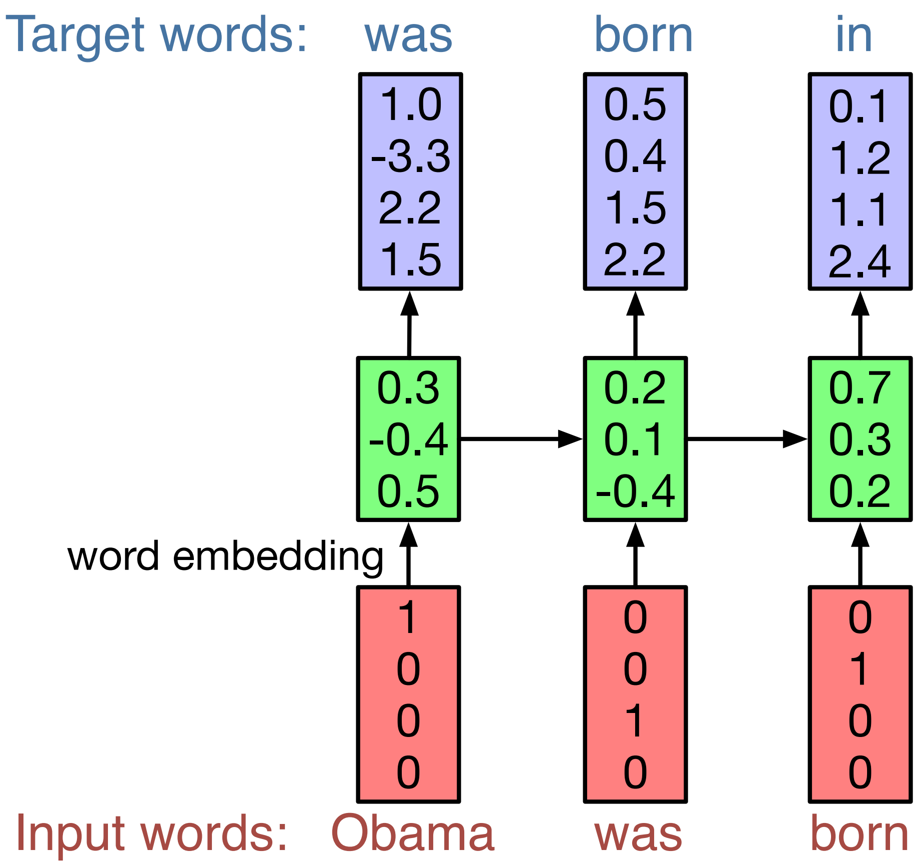
Word Embedding → Language Model (LM)

- Using **Language Model** for better representations:
  - <u>Word-level</u> Language Model:
    - ELMo (Peters et al., NAACL'18, **best paper**)
    - LD-Net (Liu et al., EMNLP'18)
  - <u>Char-level</u> Language Model:
    - LM-LSTM-CRF (Liu et al., AAAI' 18)
    - Flair (Akbik et al., COLING'18)
  - <u>Hybrid</u> Language Model:
    - Cross View Training (Clark et al., EMNLP' 2018)
    - BERT (Devlin et al., NAACL'19, **best paper**)

What's (Neural) **Language Model?**

- Describing the generation of text:
  - predicting the next word based on previous contexts

- Pros:
  - Does not require any human annotations
    - Nearly **unlimited training data**!
  - Resulting models can generate sentences of an unexpectedly high quality



Neural LM: **Example Generations**

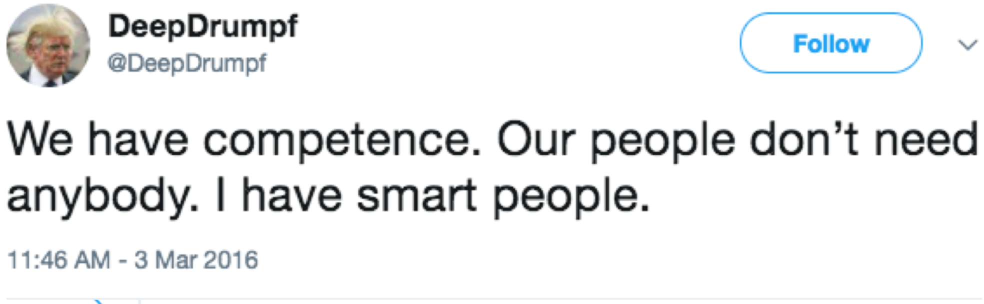
- Char-by-Char Markdown Generations:

```
'''See also''': [[List of ethical consent processing]]

== See also ==
*[[Iender dome of the ED]]
*[[Anti-autism]]

===[[Religion|Religion]]===
*[[French Writings]]
*[[Maria]]
*[[Revelation]]
```

→ **Valid Syntax!**

Neural LM: Example Generations

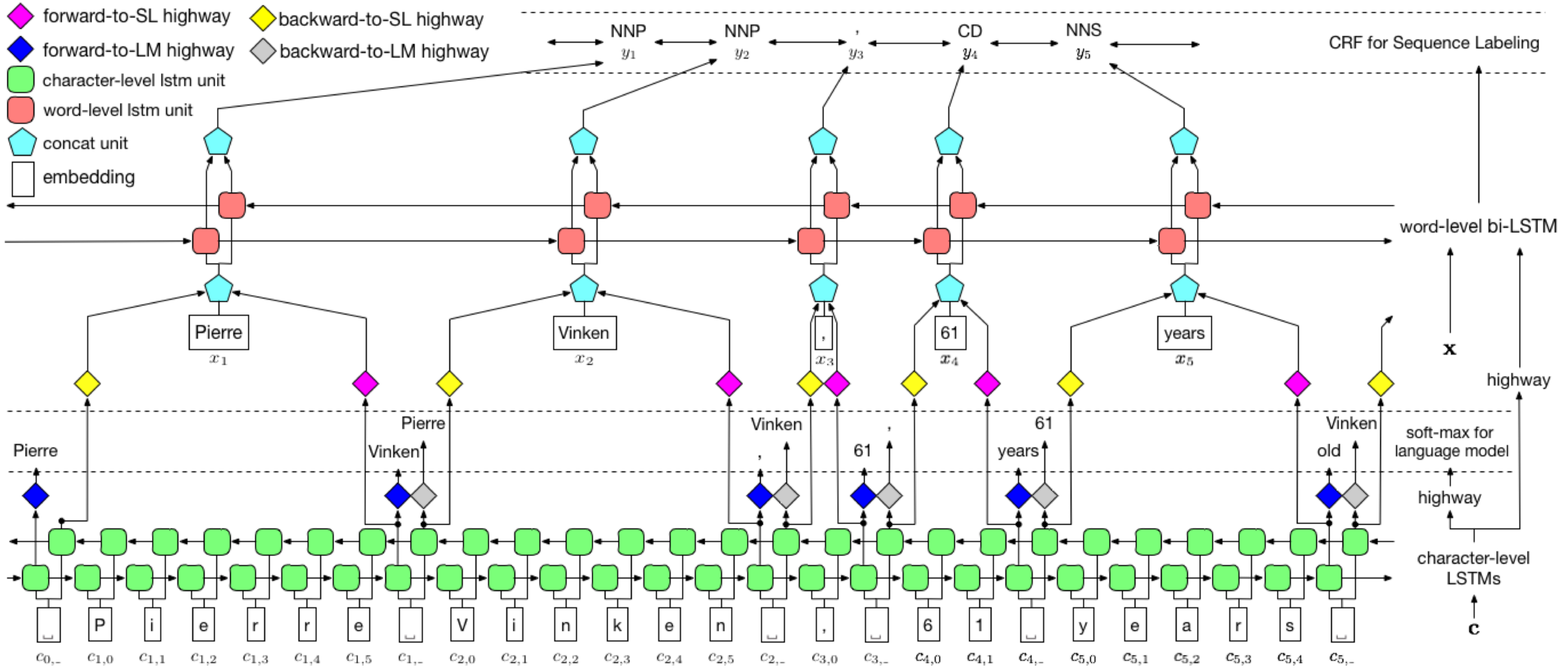
- Deep “Donald Trump”: Mimic President Trump



Fooled many twitter users

LM-LSTM-CRF: Co-Train Neural LM

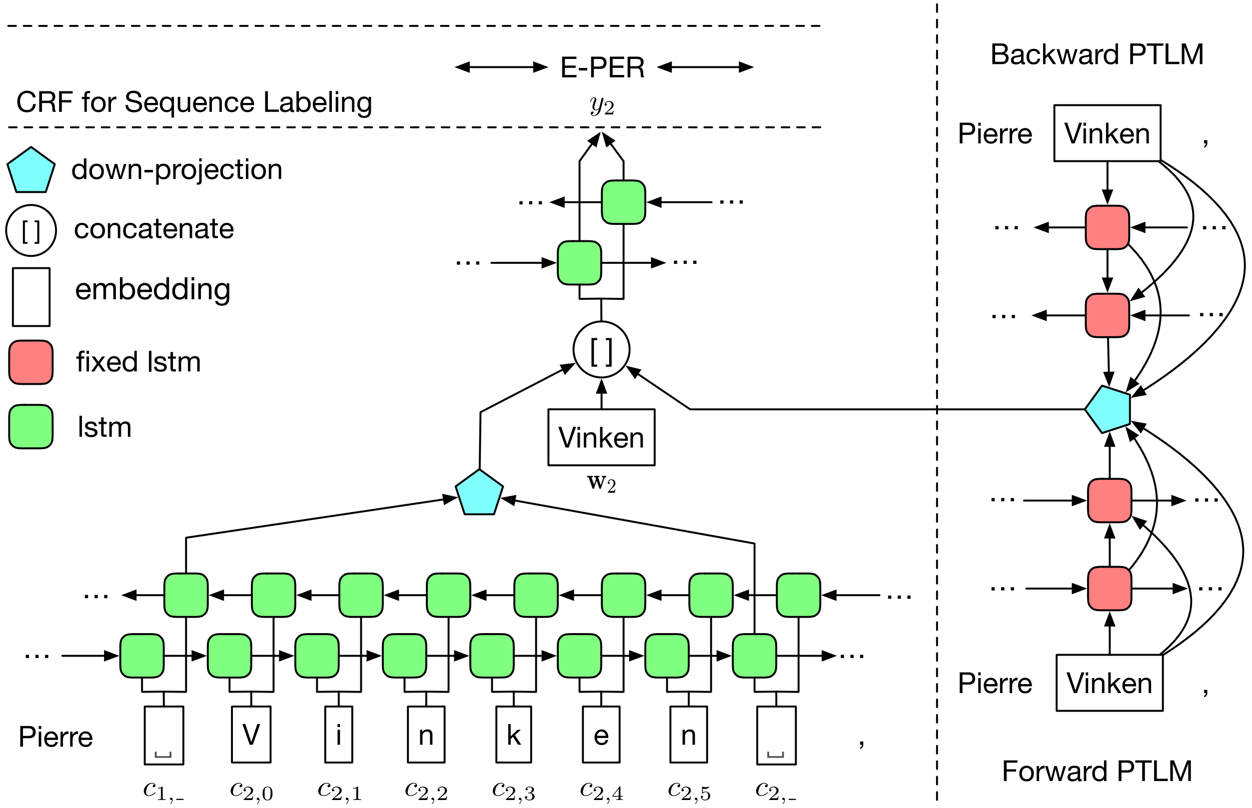
- Propose to use **Character-level language model** as a **Co-Training** objective
- Why character-level?
  - More efficient & More robust to pre-processing



ELMo: Pre-train Word-Level Neural LM

- Add ELMo at the input of RNN. For some tasks (SNLI, SQuAD), including ELMo at the output brings further improvements

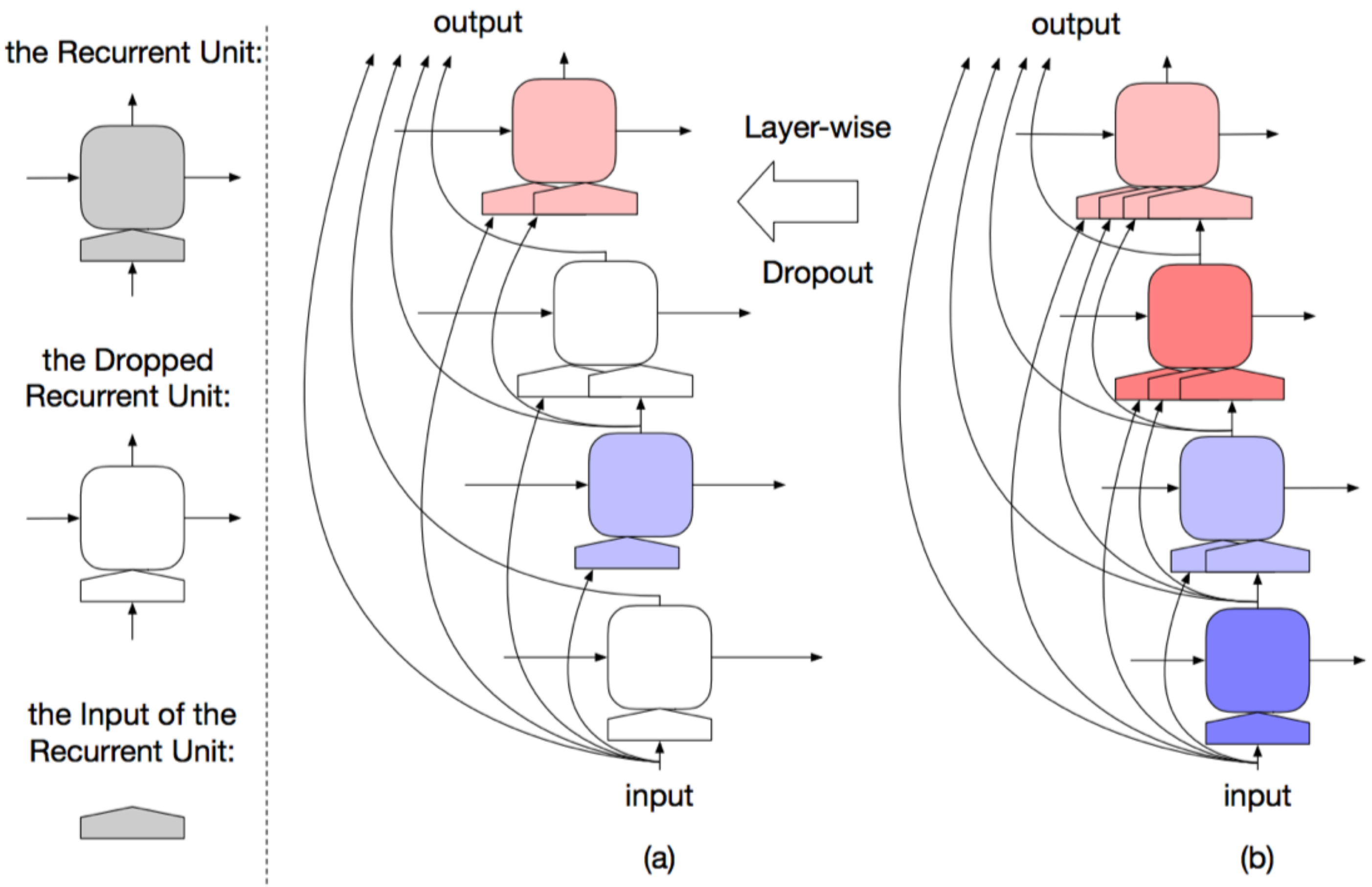
- Key points:
  - **Freeze** the weight of the biLM
  - Regularization are necessary



LD-Net: An efficient version of ELMo

- Make the contextualized represent ***efficient without much loss of effectiveness***

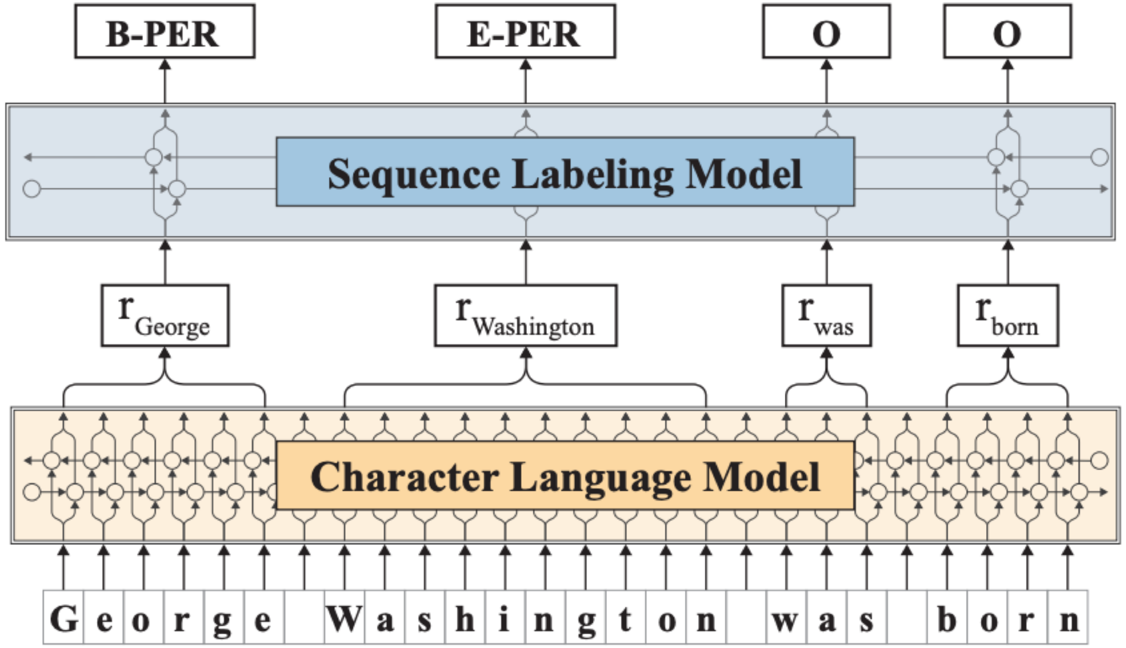
Network (PTLMs Ind.#)	Avg. ppl	#FLOPs ($\cdot 10^6$)	F_1 score (avg±std)
NoLM (/)	/	3	90.78±0.24
O-ELMo (3)	39.70	607†	92.22±0.10
R-ELMo (6)	40.27	215	91.99±0.24
R-ELMo (7)	48.85	135	91.54±0.10
TagLM (5)	47.50	87†	91.62±0.23
LD-Net (8)	45.14	98	91.76±0.18
LD-Net (9)	50.06	98	91.86±0.15
LD-Net (8*)	origin	98	91.95
	pruned	6	91.55±0.06
LD-Net (9*)	origin	98	92.03
	pruned	6	91.84±0.14



Flair: Pre-Train Neural LM at All Levels

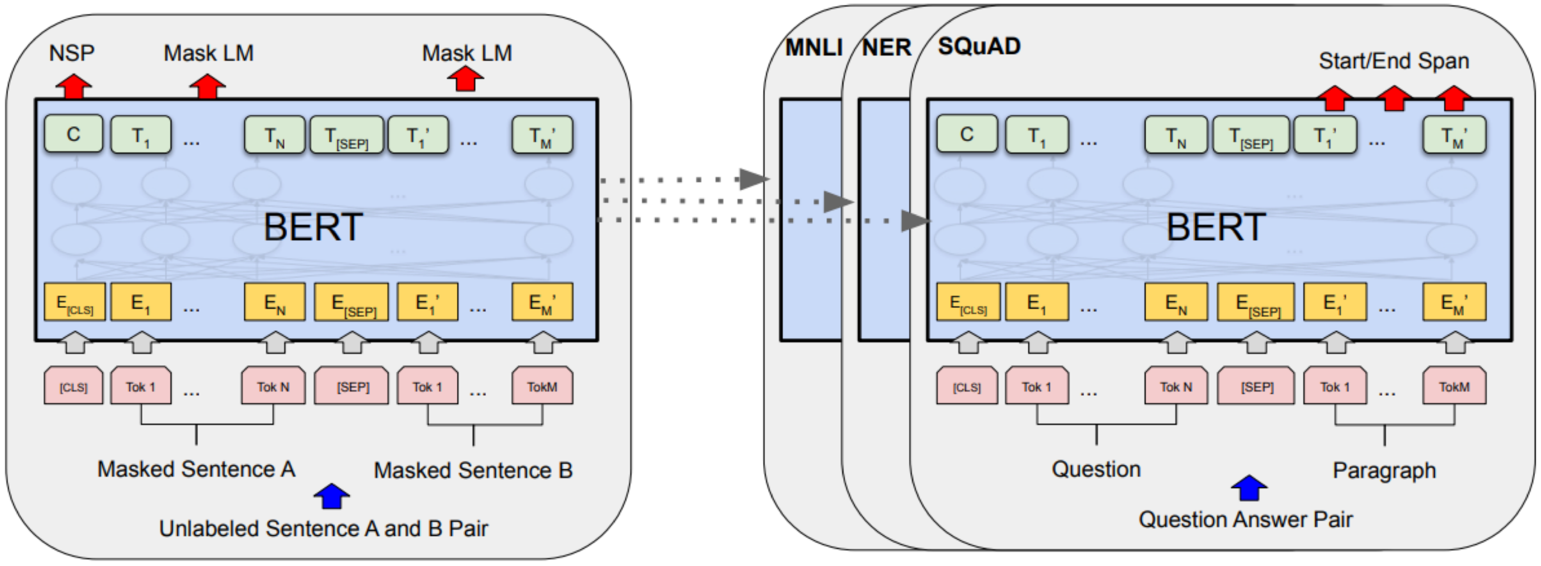
- Even for character-level language model, pre-training is very important.
- The structure is the same with LM-LSTM-CRF, the difference is the pre-training conducted on additional training corpus.

Task	PROPOSED	Previous best
NER English	**93.09**±0.12	*92.22*±0.1 (Peters et al., 2018)
NER German	**88.32**±0.2	*78.76* (Lample et al., 2016)
Chunking	**96.72**±0.05	*96.37*±0.05 (Peters et al., 2017)
PoS tagging	**97.85**±0.01	*97.64* (Choi, 2016)



BERT: Introduce Transformer

- Introduce Transformers, use masked language model + next sentence prediction
- Conduct fine-tuning after pre-training on each task (necessary for sentence-level tasks, NER is a word level task).



Pre-training
Fine-Tuning

System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	**93.1**
Fine-tuning approach		
$\text{BERT}_{\text{LARGE}}$	96.6	92.8
$\text{BERT}_{\text{BASE}}$	96.4	92.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

Table 7: CoNLL-2003 Named Entity Recognition results. Hyperparameters were selected using the Dev set. The reported Dev and Test scores are averaged over 5 random restarts using those hyperparameters.

New State-of-the-arts

- Using **Language Model** for better representations:	F1 on CoNLL03
- Word-level Language Model:	
- ELMo (Peters et al., NAACL'18, **best paper**)	92.2
- LD-Net (Liu et al., EMNLP'18)	92.0, ~5X faster
- Char-level Language Model:	
- LM-LSTM-CRF (Liu et al., AAAI' 18)	91.4
- Flair (Akbik et al., COLING'18)	**93.1**
- Hybrid Language Model:	
- Cross View Training (Clark et al., EMNLP' 2018)	92.6
- BERT (Devlin et al., NAACL'19, **best paper**)	92.4 / 92.8

Questions

- Can massive raw texts help? → Neural language model
- Can dictionaries help?

Distantly Supervised NER

- Input
  - Unlabeled Raw Texts
  - An Entity Dictionary
    - entity type, canonical name, [synonyms_1, synonyms_2, ..., synonyms_k]

- Output
  - A NER model to recognize the entities of the entity types appeared in the given dictionary.
    - Note that the entities to be recognized can be unseen entities.

Distantly Supervised NER **Methods**

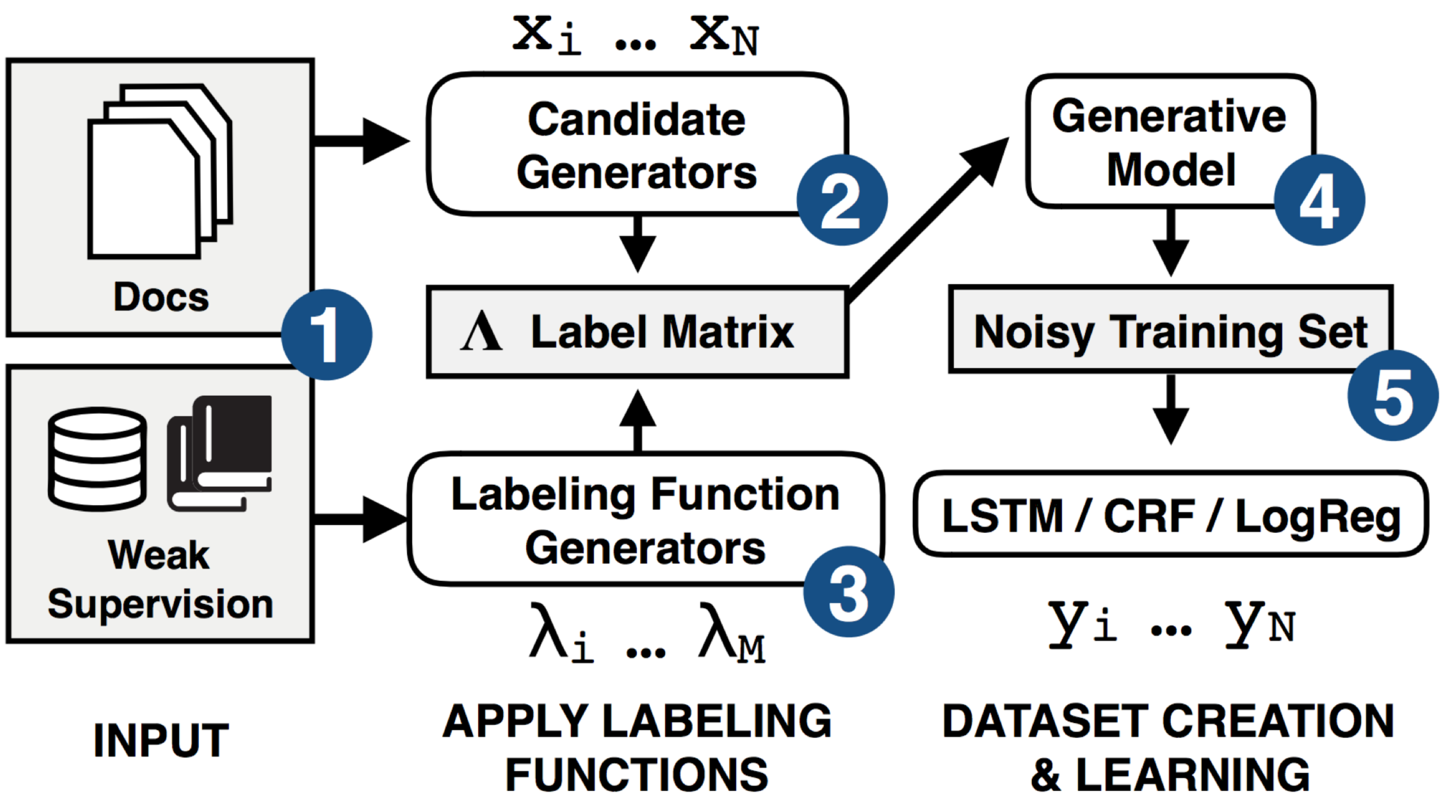
- String-match / rule-based distant supervision generation

- AutoEntity, SwellShark, ClusType, …
  - Leave the entity span detection to experts
  - POS Tag Rule-based (e.g., regular expressions)

- AutoNER
  - A novel “Tie-or-Break” labeling scheme + tailored neural model

SwellShark: Distantly Supervised Typing

- Data Programming for Typing
- Entity Span Detection: Regular expressions based on part-of-speech (POS) tags
  - **Requires expert efforts**
    - Candidate Generators



AutoNER: Dual Dictionaries

- A core dictionary
  - Leads to high-precision but low-recall matches
- A “full” dictionary
  - Leads to high-recall but low-precision matches
  - Introduce out-of-dictionary high-quality phrases as new entities
    - Their types are “unknown”
    - It could be any IOBES + any type

AutoNER: **Fuzzy-LSTM-CRF** Baseline

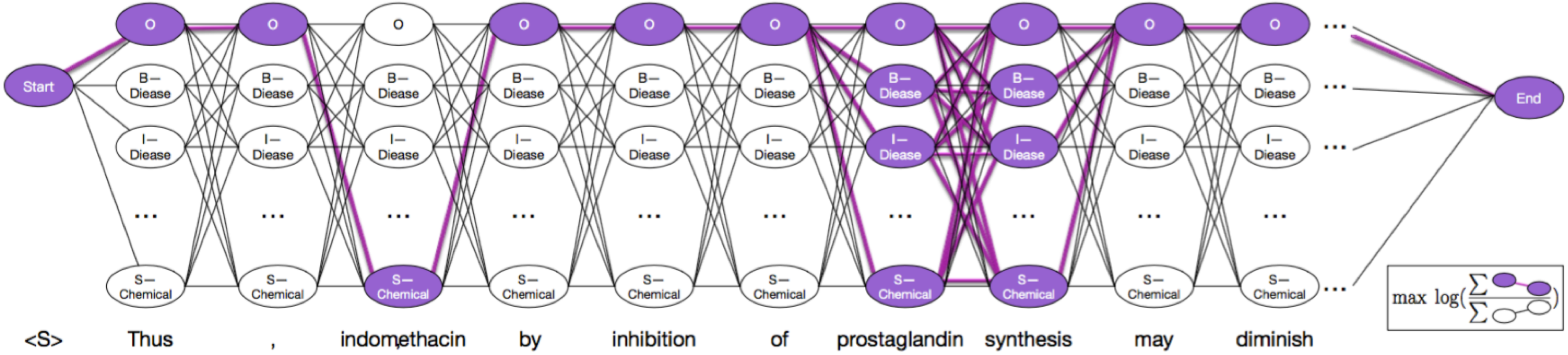


Figure 1: The illustration of the Fuzzy CRF layer with modified `IOBES` tagging scheme. The named entity types are {`Chemical`, `Disease`}. "indomethacin" is a matched `Chemical` entity and "prostaglandin synthesis" is an unknown-typed high-quality phrase. Paths from `Start` to `End` marked as purple form all possible label sequences given the distant supervision.

AutoNER: "Tie or Break"

- Instead of labeling each token, we choose to tag the connection between two adjacent tokens.

- For every two adjacent tokens, the connection between them is labeled as
  - (1) **Tie**, when the two tokens are matched to the same entity
  - (2) **Unknown**, if at least one of the tokens belongs to an unknown-typed high-quality phrase;
  - (3) **Break**, otherwise.

AutoNER: **Tailored Neural Model**

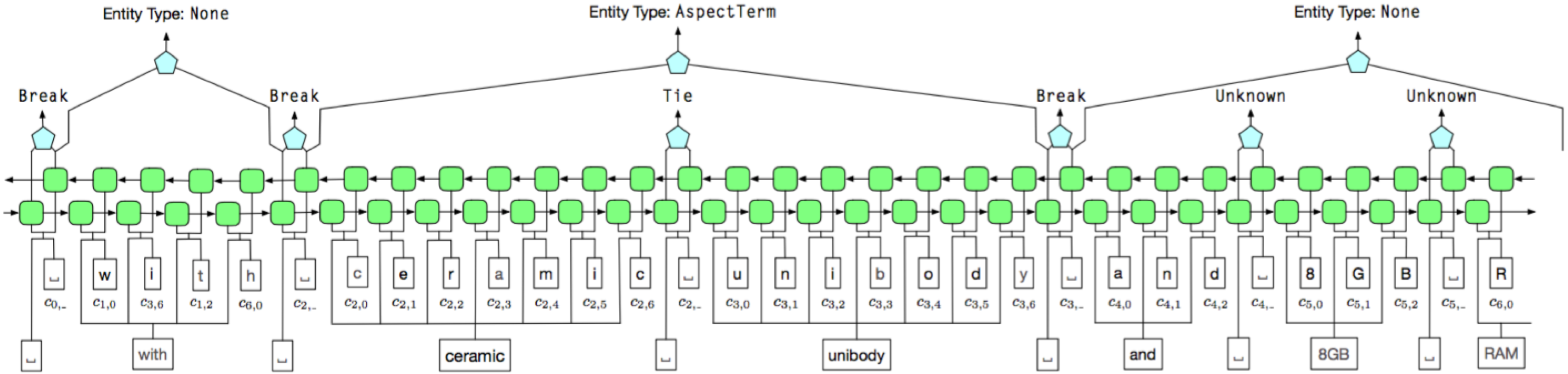


Figure 2: The illustration of AutoNER with `Tie or Break` tagging scheme. The named entity type is {`AspectTerm`}. "ceramic unibody" is a matched `AspectTerm` entity and "8GB RAM" is an unknown-typed high-quality phrase. `Unknown` labels will be skipped during the model training.

Comparison – Biomedical Domain

Table 2: [Biomedical Domain] NER Performance Comparison. The supervised benchmarks on the BC5CDR and NCBI-Disease datasets are LM-LSTM-CRF and LSTM-CRF respectively (Wang et al., 2018). SwellShark has no annotated data, but for entity span extraction, it requires pre-trained POS taggers and extra human efforts of designing POS tag-based regular expressions and/or hand-tuning for special cases.

Method	Human Effort other than Dictionary	BC5CDR Pre	BC5CDR Rec	BC5CDR F1	NCBI-Disease Pre	NCBI-Disease Rec	NCBI-Disease F1
Supervised Benchmark	Gold Annotations	88.84	85.16	**86.96**	86.11	85.49	**85.80**
SwellShark	Regex Design + Special Case Tuning	86.11	82.39	84.21	81.6	80.1	**80.8**
	Regex Design	84.98	83.49	**84.23**	64.7	69.7	67.1
Dictionary Match	None	93.93	58.35	71.98	90.59	56.15	69.32
Fuzzy-LSTM-CRF		88.27	76.75	82.11	79.85	67.71	73.28
AutoNER		88.96	81.00	**84.8**	79.42	71.98	**75.52**

Summary & Q&A

- Using neural language model, massive raw texts can help!
- High-quality dictionaries can help!

Scientific Text Mining and Knowledge Graphs

Chapter 1
Part 3: Relation Extraction and Attribute Discovery

Presenter: Jingbo Shang

University of California, San Diego

jshang@ucsd.edu

Meta-Pattern Mining for Information Extraction

- Meta-Pattern
  - $PERSON (age $DIGITS)
    - $PERSON has an attribute "age"
    - $DIGITS is the value of this attribute
  - $PERSON meets $PERSON
    - A "meet" relation between two people
- Applications
  - Attribute Discovery: both **attribute name** and **attribute value**
  - Relation Extraction: **relation** between entities
  - Named Entity Recognition: entity boundaries & types
    - Using discovered meta-patterns to match new raw texts

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

- Task 1: Finding E-A-V at the Instance Level

Ignore entity-typing information!

  - Stanford OpenIE [ACL'15], AI[2]'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]:

"president of united states"	➡	"A of E", "E 's A", "E A", "A, E"
"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⟩ ☹

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

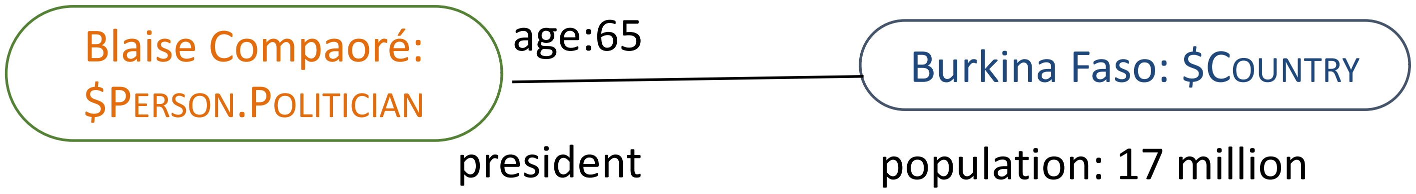
age:65

president

Burkina Faso: $COUNTRY

population: 17 million

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



- 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⟩

Instance-level

Task 2: ⟨entity type, attribute name⟩

⟨$COUNTRY, president⟩

⟨$COUNTRY, population⟩

⟨$PERSON, age⟩

Type-level

Our Meta-Pattern Methodology

(#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 …"

Generate patterns with massive instances in the data

Meta pattern segmentation

No heavy annotation required
No domain knowledge required
No query log required
if we can recognize and type the entities in the same manner…

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

⟨Burkina Faso, {president}, Blaise Compaoré⟩
⟨U.S., {president}, Barack Obama⟩

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:

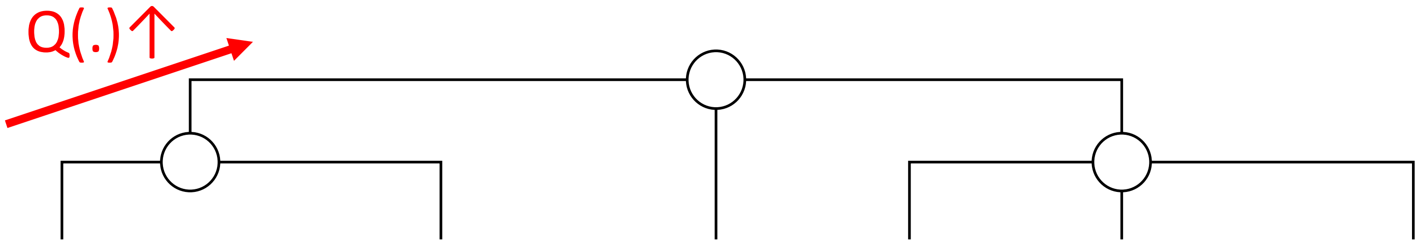
✓Frequency

✓Concordance: "$PERSON's wife"

✓Completeness: "$COUNTRY president" vs. "$COUNTRY president $POLITICIAN"

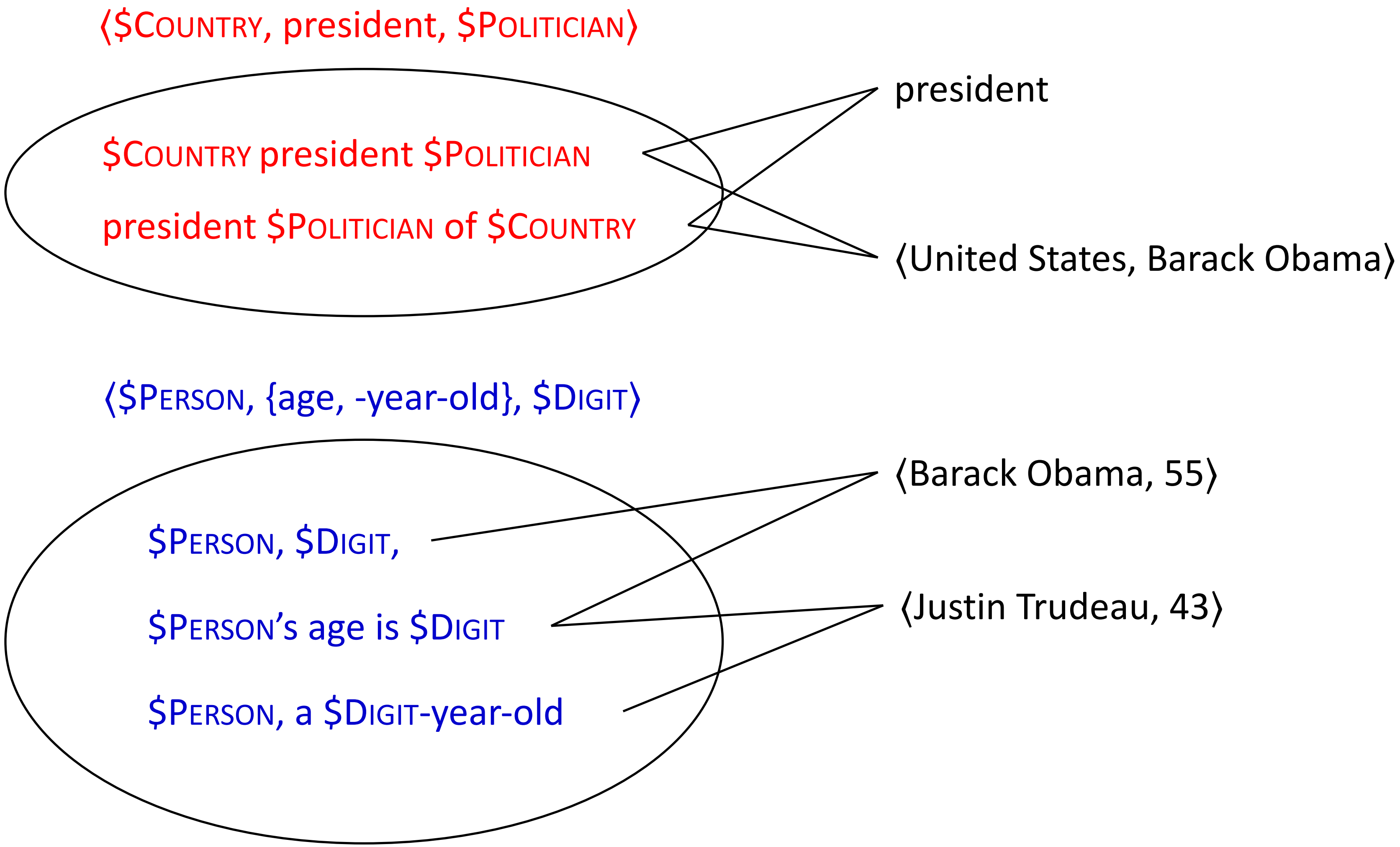
✓Informativeness: "$PERSON and $PERSON" vs. "$PERSON 's wife, $PERSON"

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



president $PERSON.POLITICIAN 's government of $LOCATION.COUNTRY

Grouping Synonymous Patterns



Adjusting Types in Meta Patterns for Appropriate Granularity

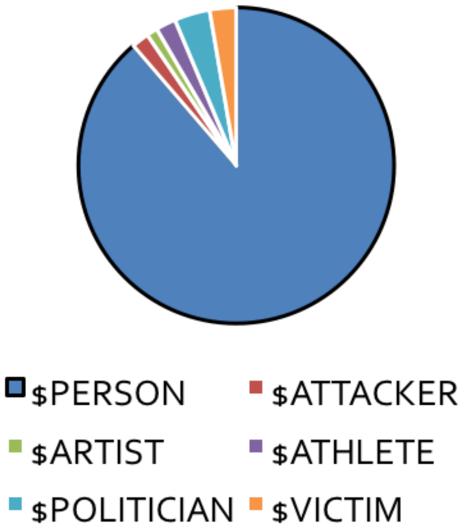
$Person, $Digit,

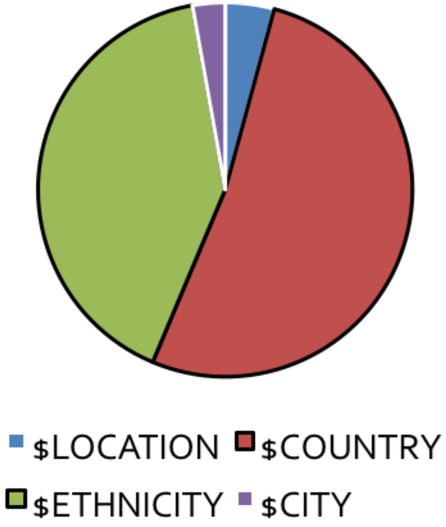
$Person's age is $Digit

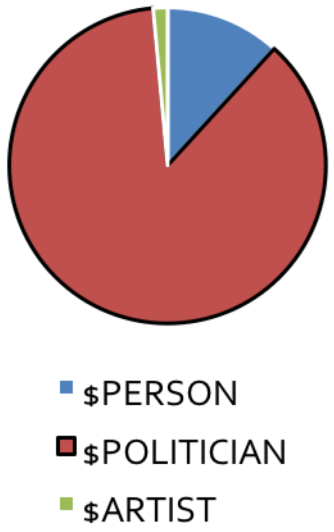
$Person, a $Digit -year-old

$Country president $Politician

president $Politician of $Country







Results: Patterns, Entities and Attribute Values in News Corpus

Meta patterns	Entity	Attribute value
\$COUNTRY President \$POLITICIAN \$COUNTRY's president \$POLITICIAN President \$POLITICIAN of \$COUNTRY … \$POLITICIAN's government of \$COUNTRY	United States	Barack Obama
	Russia	Vladimir Putin
	France	Francois Hollande
	…	…
	Burkina Faso	Blaise Compaoré

Meta patterns	Entity	Attribute value
\$COMPANY CEO \$PERSON \$COMPANY chief executive \$PERSON \$PERSON, the \$COMPANY CEO, … \$COMPANY former CEO \$PERSON \$PERSON, the \$COMPANY former CEO,	Apple	Tim Cook
	Facebook	Mark Zuckerberg
	Hewlett-Packard	Carly Fiorina
	…	…
	Infor	Charles Phillips
	Afghan Citadel	Roya Mahboob

Patterns and Entities Found in Medical Science Corpus

Meta patterns	Entity	Attribute value
$TREATMENT was used to treat $DISEASE $DISEASE using the $TREATMENT $TREATMENT has been used to treat $DISEASE $TREATMENT of patients with $DISEASE …	zoledronic acid therapy	Paget's disease of bone
	bisphosphonates	osteoporosis
	calcitonin	Paget's disease of bone
	calcitonin	osteoporosis
	…	…

Meta patterns	Entity	Attribute value
$BACTERIA was resistant to $ANTIBIOTICS $BACTERIA are resistant to $ANTIBIOTICS $BACTERIA is the most resistant to $ANTIBIOTICS … $BACTERIA, particularly those resistant to $ANTIBIOTICS	corynebacterium striatum BM4687	gentamicin
	corynebacterium striatum BM4687	tobramycin
	methicillin-susceptible S aureus	vancomycin
	multidrug-resistant enterobacteriaceae	gentamicin
	…	…

Further Enhancements of MetaPAD

- TruePIE
  - 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

- WW-PIE
  - 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
  - Leverage parsing structures to mine ***long-distance*** patterns

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 **[[alanine]CHEMICAL aminotransferase]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

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

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.
…	…

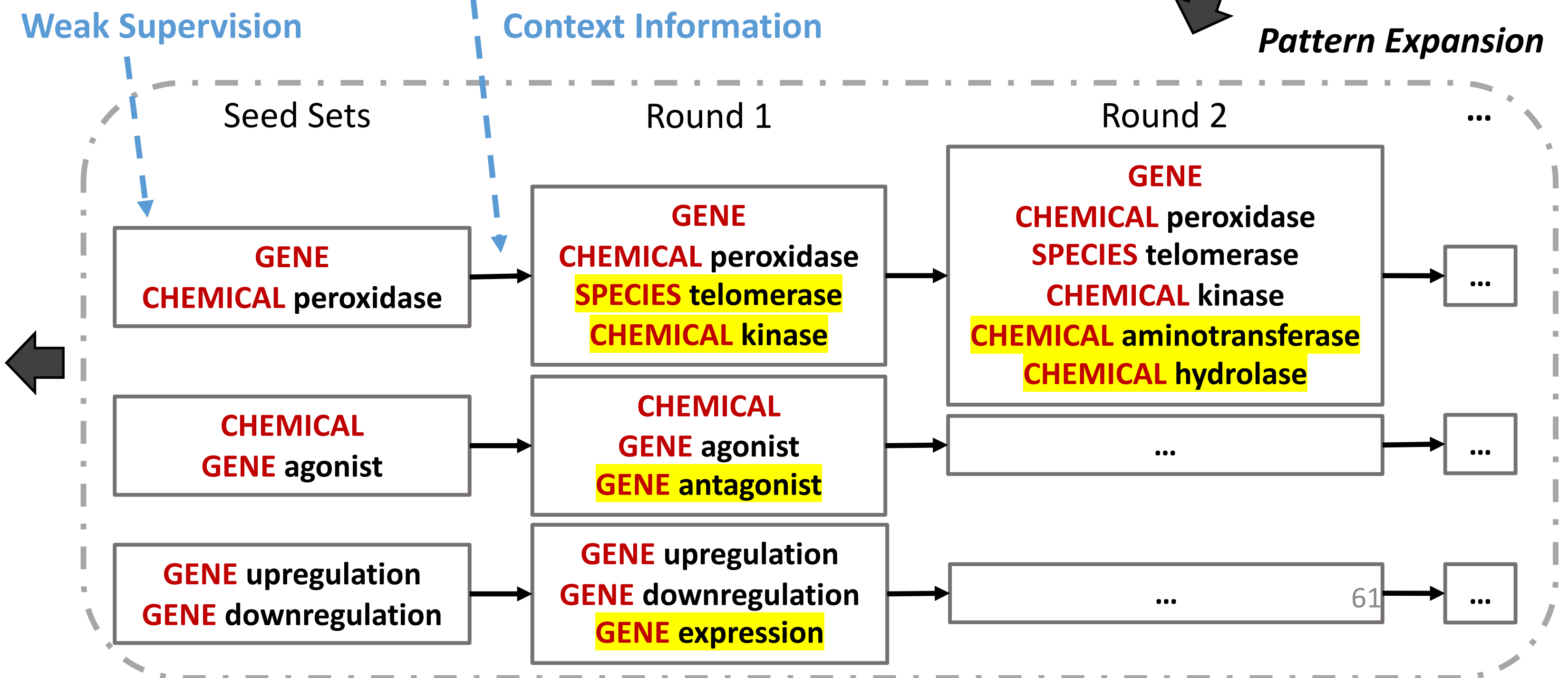
Flat Entity Recognition

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
…

Pattern Expansion

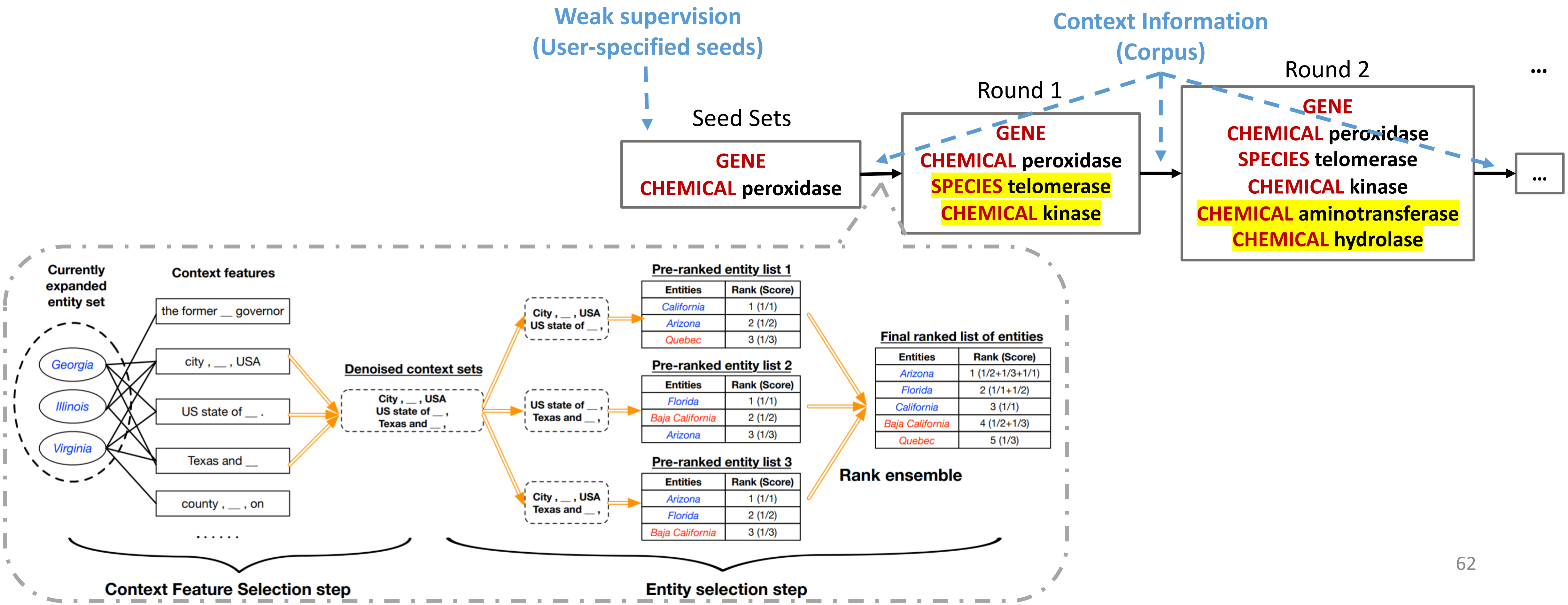


Nested Entity Recognition

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.
…	…

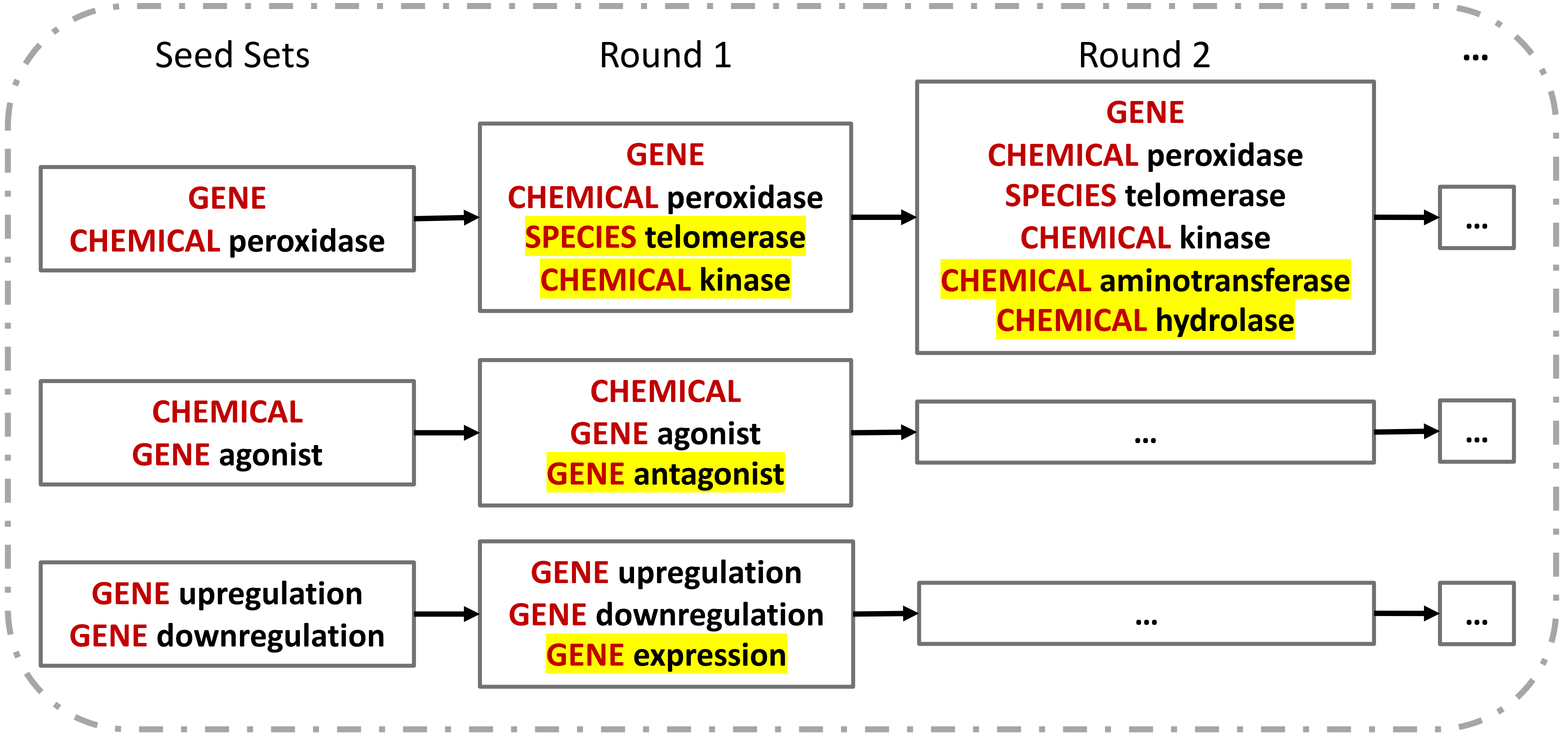
Weakly-supervised Pattern Expansion

- 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



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

Entity-level Task: Nested NER

- 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)

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	
PubTator	... it suppressed [STAT3]GENE and [**STAT5**]GENE phosphorylation in HS-578T cells, whereas it up-regulated [**STAT1**]GENE phosphorylation and down-regulated [**STAT5**]GENE phosphorylation in MCF-7 cells.
PENNER	... it suppressed [STAT3]GENE and [[**STAT5**]GENE **phosphorylation**]PROCESS in HS-578T cells, whereas it up-regulated [[**STAT1**]GENE **phosphorylation**]PROCESS and down-regulated [[**STAT5**]GENE **phosphorylation**]PROCESS in MCF-7 cells.

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 ...