Taming Unstructured Big Data: Automated Information Extraction from Massive Text

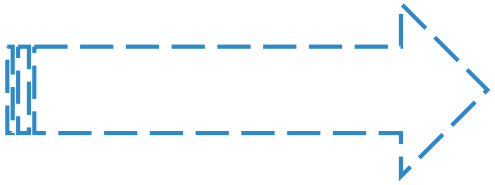
Xuan Wang, Yu Zhang, Qi Li, Jiawei Han

Computer Science, University of Illinois at Urbana-Champaign

December 11, 2019

Mining Knowledge from Big Data

Big Data



Structured Knowledge
& Insights

Massive Unstructured Text Data

News

Social Media

Business & Finance

Scientific Papers

Medical Records

.....

Massive Unstructured Text Data

News

Social Media

Business & Finance

Scientific Papers

Medical Records

.....

Goal: Texts → Knowledge & Insights

- This tutorial focuses on data mining methods that turn **unstructured text data** into **structured knowledge and insights**.

Unstructured Text Data

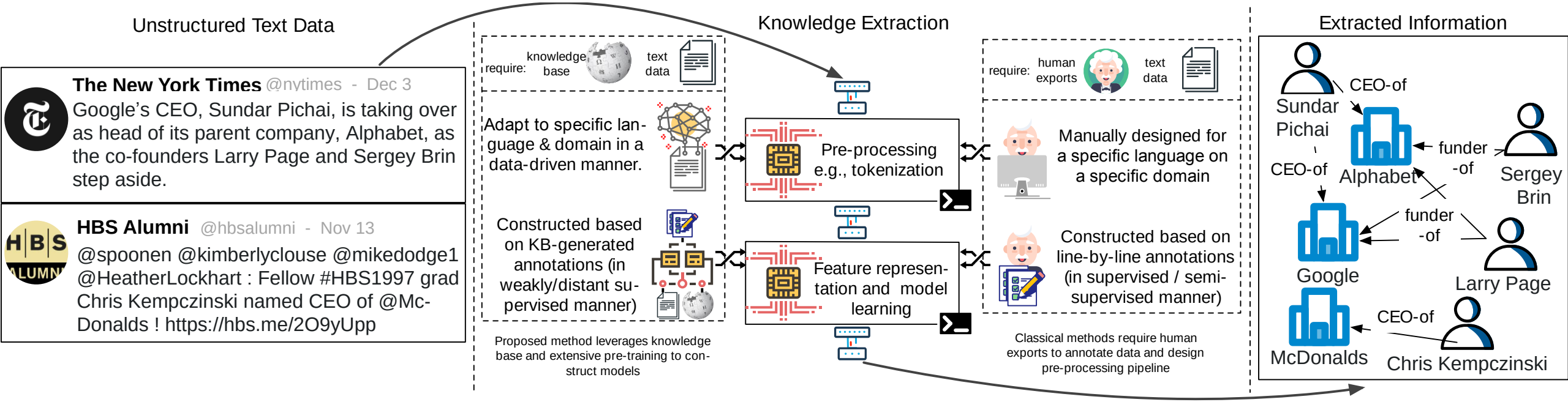
Human

Automated Models

Structured Knowledge & Insights

AutoIE: Texts → Structured Information

- **Structuring** is the key to turn raw texts into knowledge
- We propose **automated information extraction** (**AutoIE**) to convert unstructured text into structured information
  - Automated named entity recognition
  - Automated relation extraction



Tutorial Outline

- Introduction
- Part I: Supervised Methods (Handcrafted Features)
- Part II: Supervised Methods (Deep Learning Models)
- Part III: Weakly/Distantly Supervised Methods
- Part IV: Unsupervised Methods
- Summary and Future Directions

Part I: Supervised Methods (Handcrafted Features)

IEEE-BigData 2019 Tutorial

Taming Unstructured Big Data: Automated Information Extraction from Massive Text

Xuan Wang, Yu Zhang, Qi Li, Jiawei Han
Computer Science, University of Illinois at Urbana-Champaign

December 11, 2019

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
    - $[\text{Jim}]_{\text{Person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{Organization}}$ in $[\text{2006}]_{\text{Time}}$.

Classical Supervised Methods

- NER as a sequence labeling problem
- Handcrafted Features + Conditional Random Fields
- Popular Tools
  - Stanford NER (Finkel et al., ACL'05)
    - https://nlp.stanford.edu/software/CRF-NER.shtml
  - Ohio State University Twitter NER (Ritter et al., EMNLP'11, KDD'12)
    - https://github.com/aritter/twitter_nlp

Classical Supervised Methods: Training Data

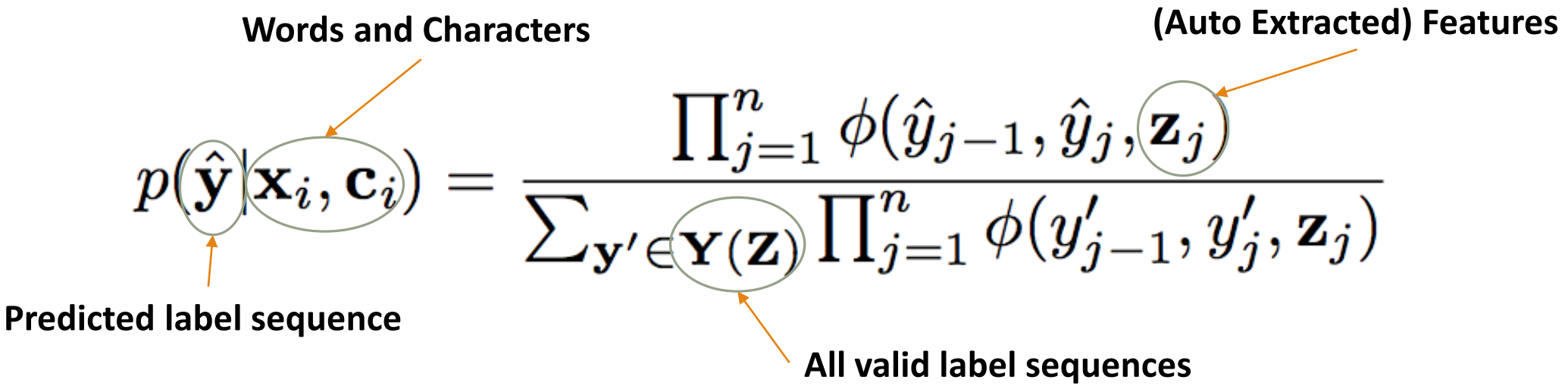
- Two popular labeling schemes
  - BIO: **B**egin, **I**n, **O**ut
  - BIOES: **B**egin, **I**n, **O**ut, **E**nd, **S**ingleton
  - BIOES is arguably better than BIO (Ratinov and Roth, ACL 09)
- Example:
  - LABELS: $[\text{Jim}]_{\text{Person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{Organization}}$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

Classical Supervised Methods: Feature Engineering

- Bag-of-words Features
  - Unigrams, Bigrams, …
  - Lowercase
  - Lemmazation
  - Stemming
  - Part-of-speech Tags
  - Word Shape
- Context Features
  - E.g., Left/right 4 words
- Character n-grams
  - Useful for unknown words
- Brown Clusters
  - Data-driven similar words
  - Alleviate sparsity issues
- WordNet Clusters
  - Semanticly similar words
- Dictionaries
  - Domain-specific dictionaries

Classical Supervised Methods: CRF

- Conditional Random Field (CRF) model
  - Capture the dependencies between labels



Classical Supervised Methods: Drawbacks

- They require
  - Human Expert Effort = Domain + Linguistic Expert Effort
  - Every single sentence must be fully annotated

- Therefore, it is difficult to adapt them to new domain from the original training domain
  - New training data → Expensive in specific domains (e.g., biomedical)

Two Types of Relation Extraction Tasks

- Sentence-level RE
  - Given two entities in a sentence (e.g., the result of entity recognition), predict the relation between these two entities.
  - Relations between the same pair may vary with sentences.
  - There may be "*no_relation*".

Example	Entity Types & Label
Carey will succeed **Cathleen P. Black**, who held the position for 15 years and will take on a new role as **chairwoman** of Hearst Magazines, the company said.	**Types:** PERSON/TITLE **Relation:** *per:title*
Irene Morgan Kirkaldy, who was born and reared in **Baltimore**, lived on Long Island and ran a child-care center in Queens with her second husband, Stanley Kirkaldy.	**Types:** PERSON/CITY **Relation:** *per:city_of_birth*
Pandit worked at the brokerage Morgan Stanley for about 11 years until 2005, when he and some Morgan Stanley colleagues quit and later founded the hedge fund **Old Lane Partners**.	**Types:** ORGANIZATION/PERSON **Relation:** *org:founded_by*
Baldwin declined further comment, and said JetBlue chief **executive** Dave Barger was unavailable.	**Types:** PERSON/TITLE **Relation:** *no_relation*

Examples in the TACRED dataset [Zhang et al., 2017]

Two Types of Relation Extraction Tasks

- Corpus-level RE
  - Given two entities (which may not co-occur in one sentence), predict their relations reflected by the whole corpus.
  - Need to utilize signals from the whole corpus

Text Corpus

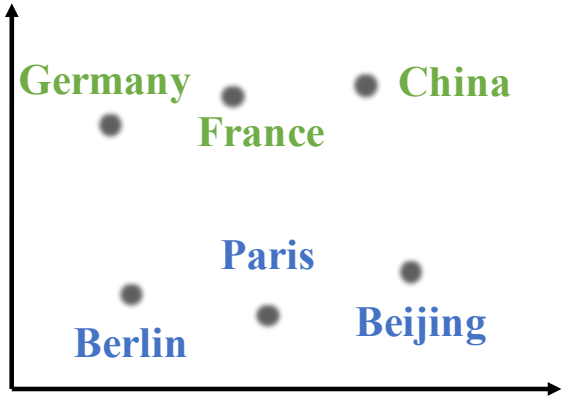
ID	Sentence
1	**Beijing** , the capital of **China**, is a megacity rich in history.
2	**Tokyo** , **Japan** 's capital , was originally a small village .
3	**Bill Gates** is a co-founder of the **Microsoft** Corporation .
4	**Apple** is working closely with **MS** to fix an issue .
5	**Steve Jobs** was an American businessman, inventor .

Extracted Relation Instances

Bill Gates — Founder of — Microsoft

Microsoft

Steve Jobs — Founder of — Apple



Classical Supervised Methods

- Relation extraction as a classification task
- Representation:
  - Features:
    - Words between and including entities
    - Types of entities (person, location, etc)
    - # of entities between the two entities
    - # words separating the two entities
    - Path between the two entities on a parse tree
  - Structured representations (Kernels)
- Discriminative classifier: SVM, Log-linear model, etc.

Classical Supervised Methods: Drawbacks

- They require
  - Human effort for training data annotation
  - Feature engineering such as POS tagging, shallow parsing and dependency parsing are prone to errors

- Difficult to extend to new relation types for want of labeled data
- Difficult to extend to higher order relations

Part II: Supervised Methods (Deep Learning Models)

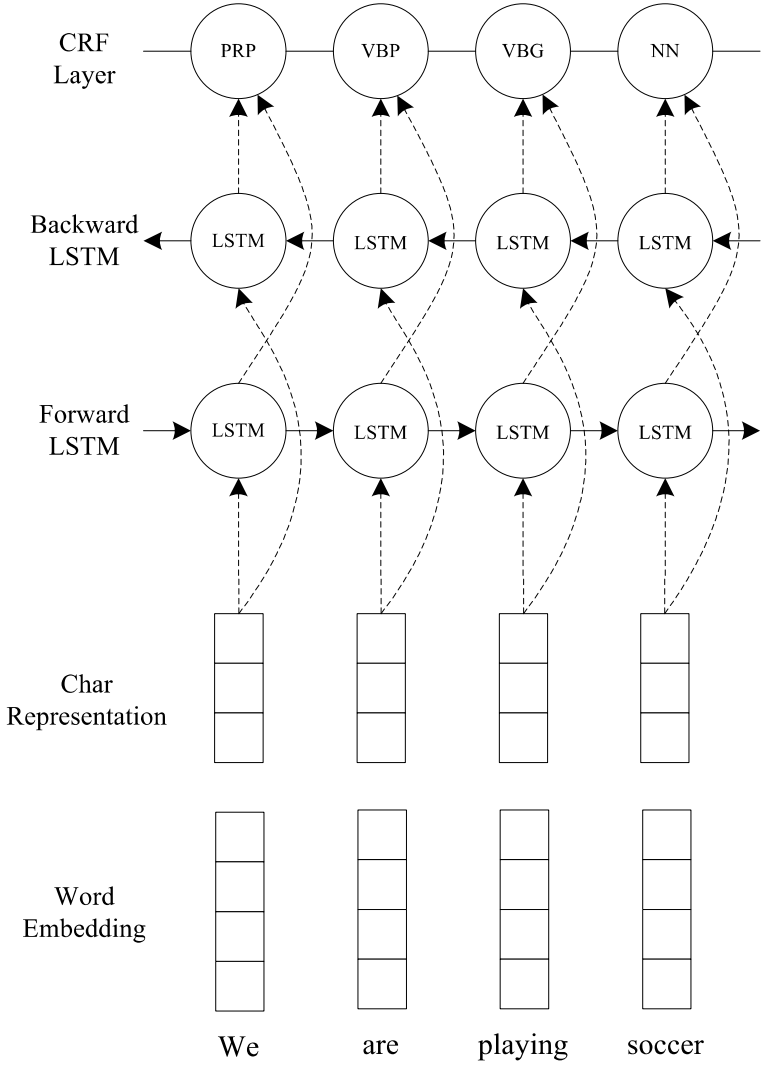
IEEE-BigData 2019 Tutorial

Taming Unstructured Big Data: Automated Information Extraction from Massive Text

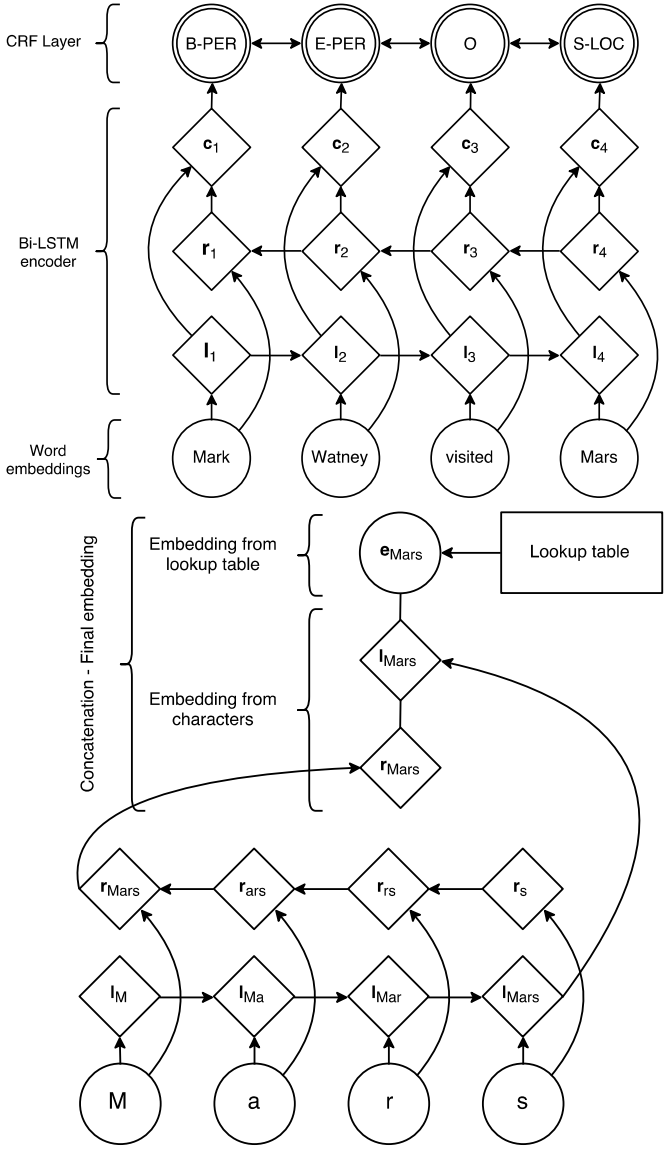
Xuan Wang, Yu Zhang, Qi Li, Jiawei Han
Computer Science, University of Illinois at Urbana-Champaign

December 11, 2019

LSTM-CNN-CRF and LSTM-CRF



Ma and Hovy, 2016

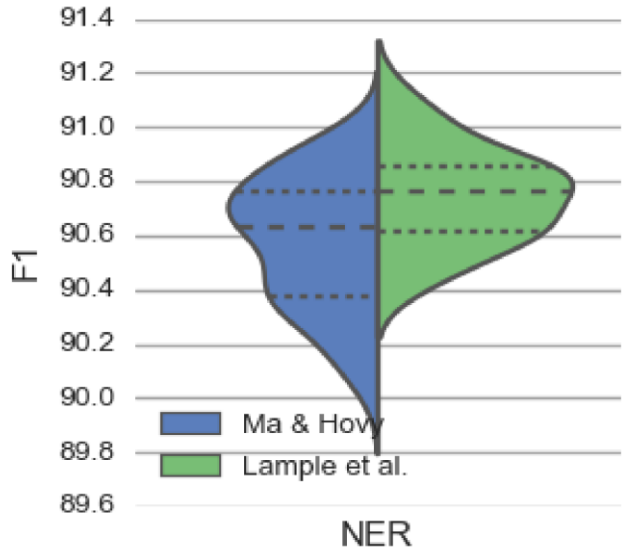


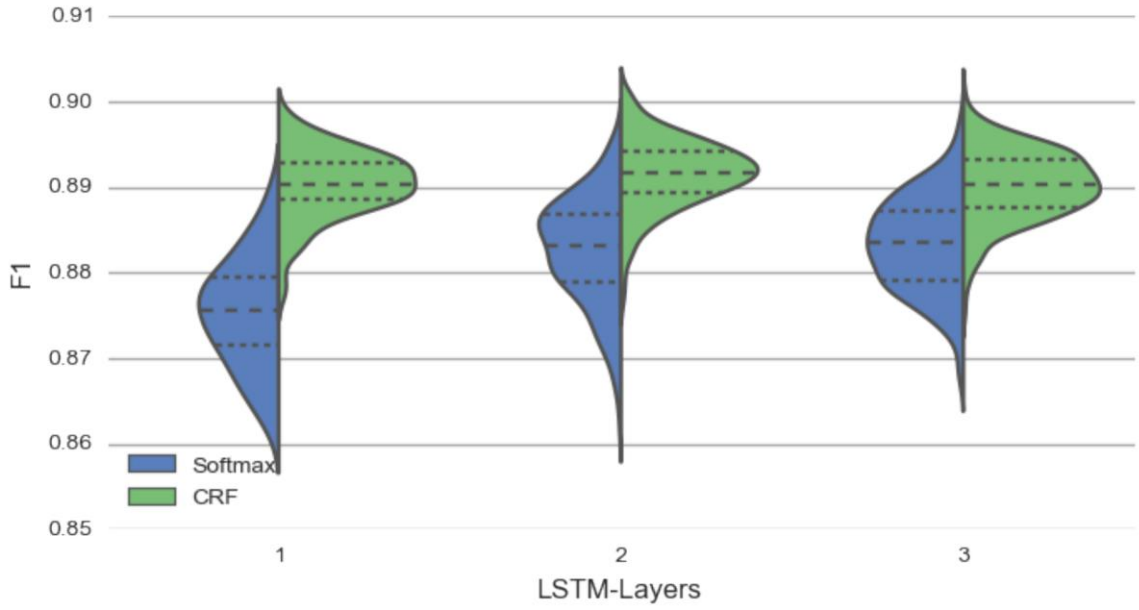
Lample *et al.*, 2016

LSTM-CRF and LSTM-CNN-CRF: Comparison

- Reimers and Gurevych, EMNLP 2017

- Question 1: Is LSTM-CNN-CRF really better than LSTM-CRF?
  - LSTM-CRF works better.
- Question 2: Does the CRF layer really help? How many layers in LSTM are the best?
  - CRF layer helps!
  - 2 layers seem enough.



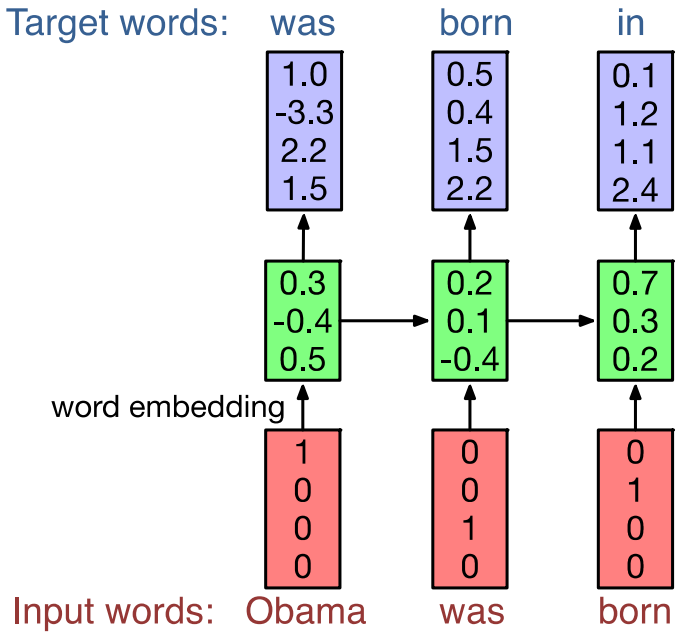


Going Beyond Word Embedding: Language Model

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

What is Language Model?

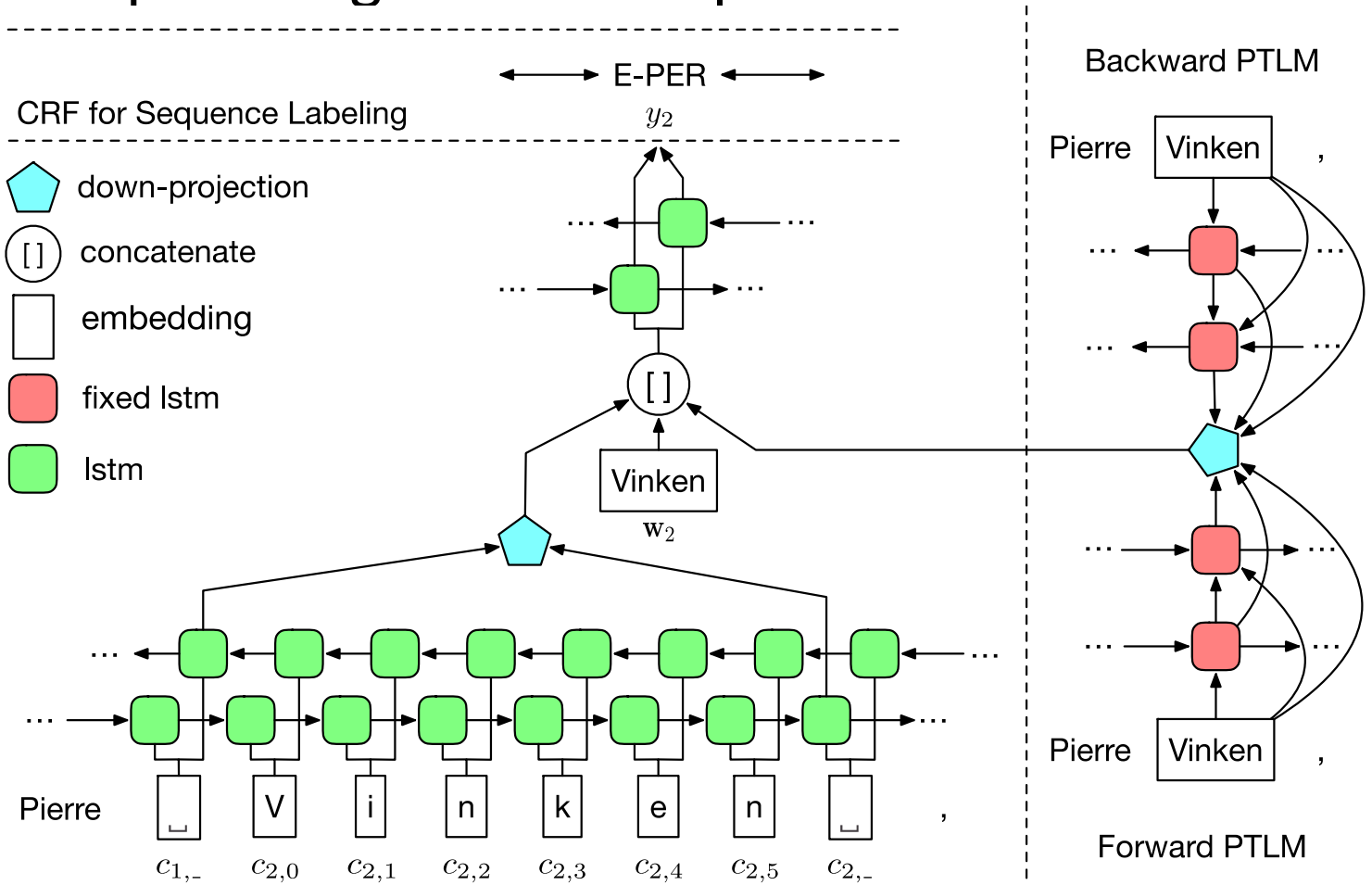
- Describing the generation of text:
  - Predicting the next word based on previous contexts



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

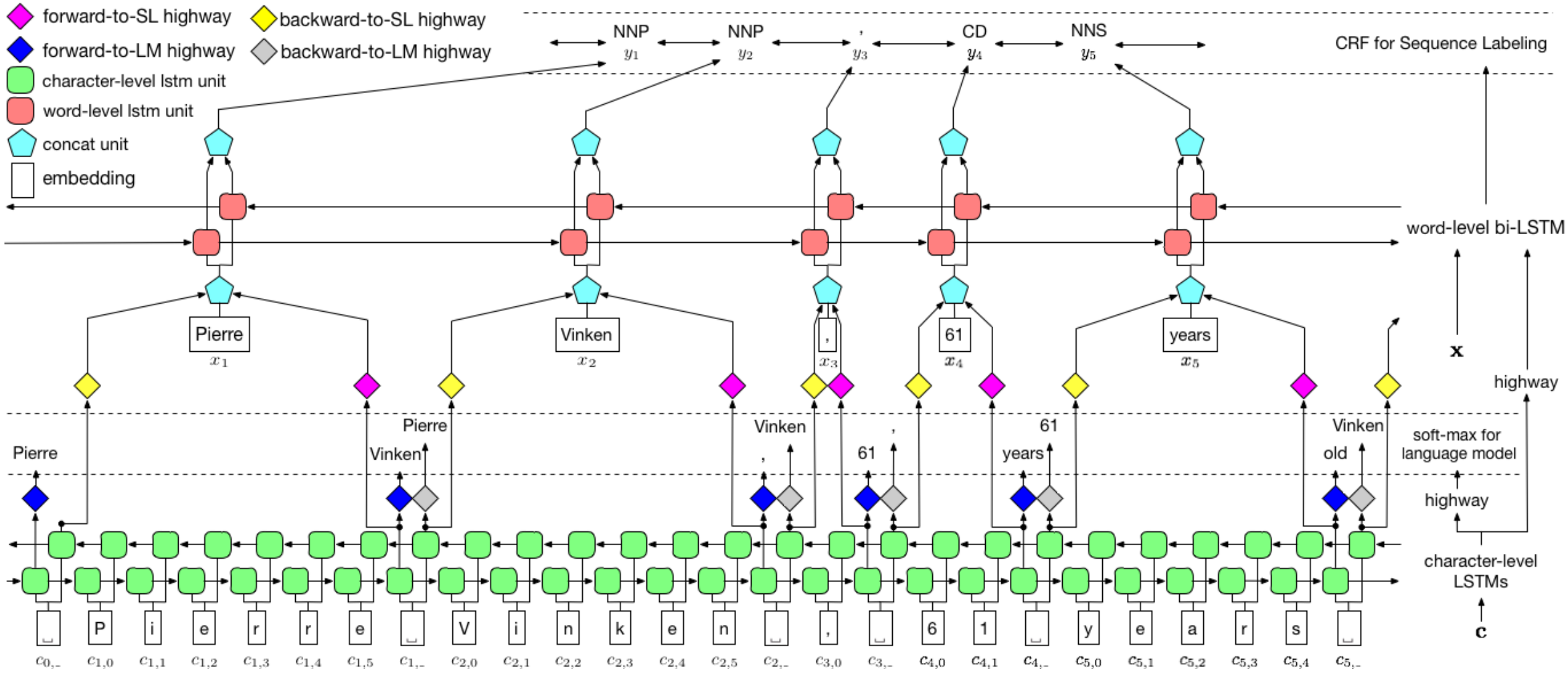
ELMo (Peters et al., NAACL'18, best paper)

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



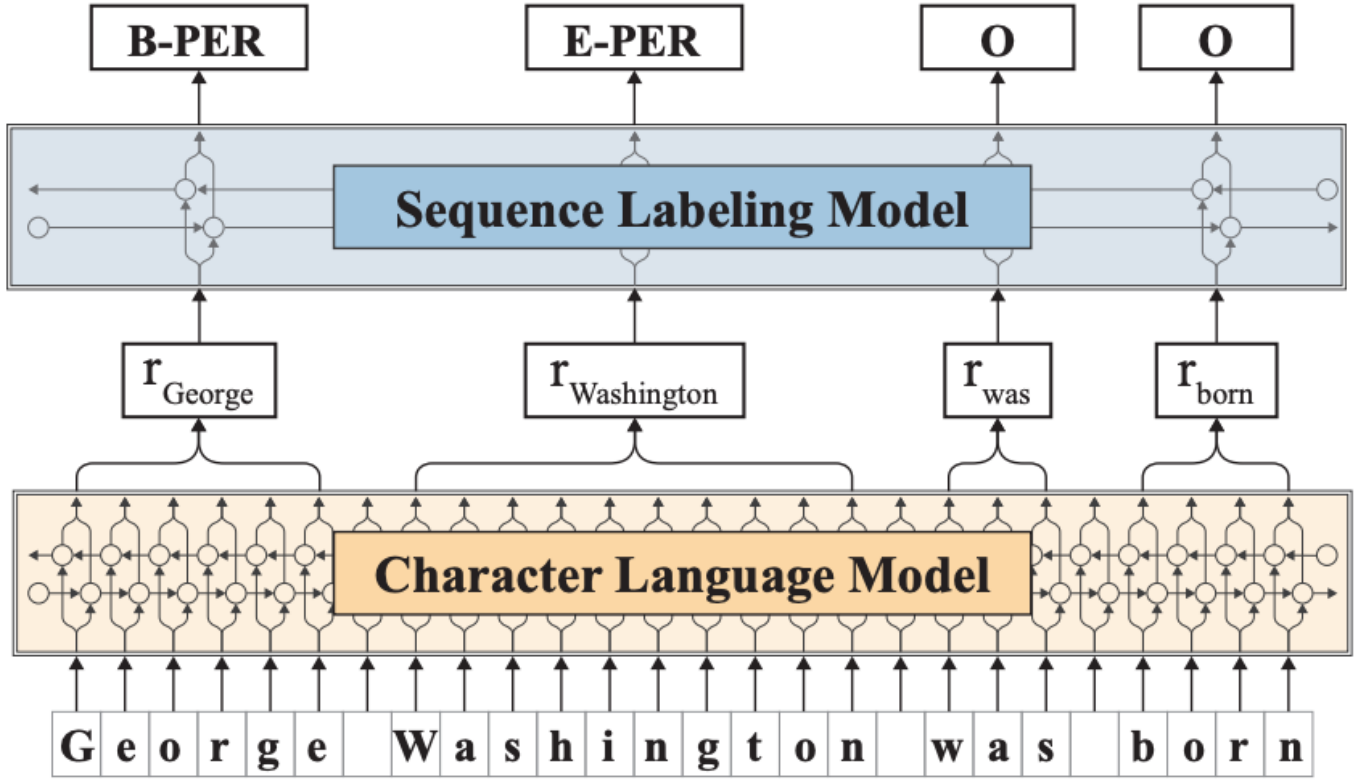
LM-LSTM-CRF (Liu et al., AAAI'18)

- Propose to use character-level language model as a *Co-Training* objective:



Flair (Akbik et al., COLING'18)

- It has a similar architecture with LM-LSTM-CRF.
- The character-level language model is pre-trained on additional training corpus.

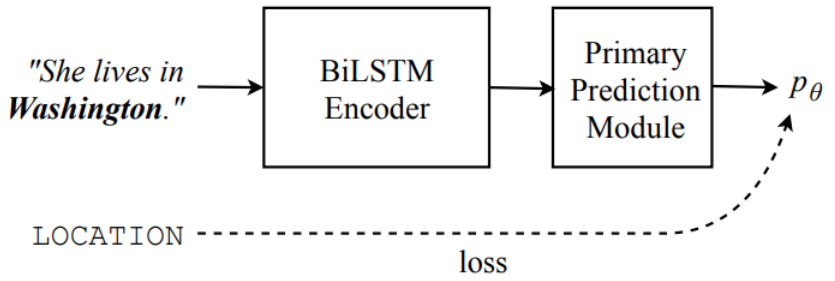


Cross-view Training (Clark et al., 2018)

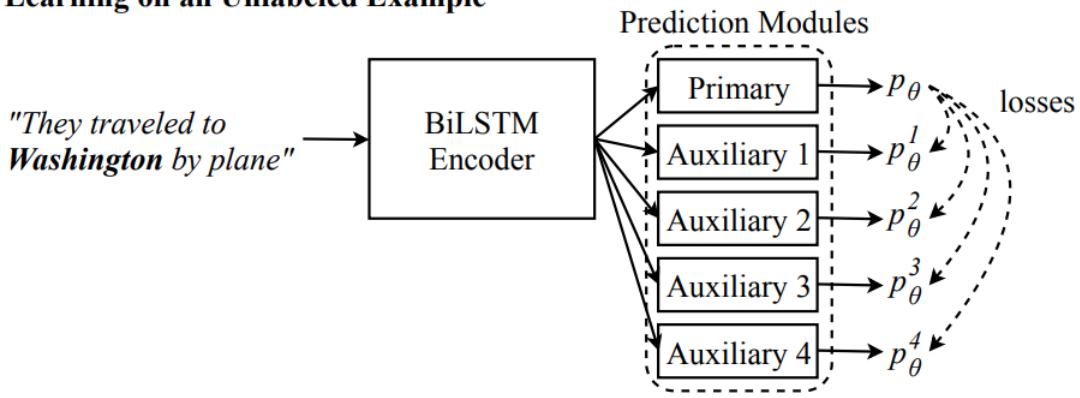
- Using more auxiliary task as supervision
- Can be combined with multi-tasking

Method	CCG Acc.	Chunk F1	NER F1	FGN F1	POS Acc.	Dep. Parse UAS	Dep. Parse LAS	Translate BLEU
Shortcut LSTM (Wu et al., 2017)	95.1				97.53			
ID-CNN-CRF (Strubell et al., 2017)			90.7	86.8				
JMT† (Hashimoto et al., 2017)		95.8			97.55	94.7	92.9	
TagLM* (Peters et al., 2017)		96.4	91.9					
ELMo* (Peters et al., 2018)			92.2					
Biaffine (Dozat and Manning, 2017)						95.7	94.1	
Stack Pointer (Ma et al., 2018)						95.9	94.2	
Stanford (Luong and Manning, 2015)								23.3
Google (Luong et al., 2017)								26.1
Supervised	94.9	95.1	91.2	87.5	97.60	95.1	93.3	28.9
Virtual Adversarial Training*	95.1	95.1	91.8	87.9	97.64	95.4	93.7	–
Word Dropout*	95.2	95.8	92.1	88.1	97.66	95.6	93.8	29.3
ELMo (our implementation)*	95.8	96.5	92.2	88.5	97.72	96.2	94.4	29.3
ELMo + Multi-task*†	95.9	96.8	92.3	88.4	**97.79**	96.4	94.8	–
CVT*	95.7	96.6	92.3	88.7	97.70	95.9	94.1	**29.6**
CVT + Multi-task*†	96.0	96.9	92.4	88.4	97.76	96.4	94.8	–
CVT + Multi-task + Large*†	**96.1**	**97.0**	**92.6**	**88.8**	97.74	**96.6**	**95.0**	–

Learning on a Labeled Example



Learning on an Unlabeled Example



Inputs Seen by Auxiliary Prediction Modules

Auxiliary 1: *They traveled to* _____

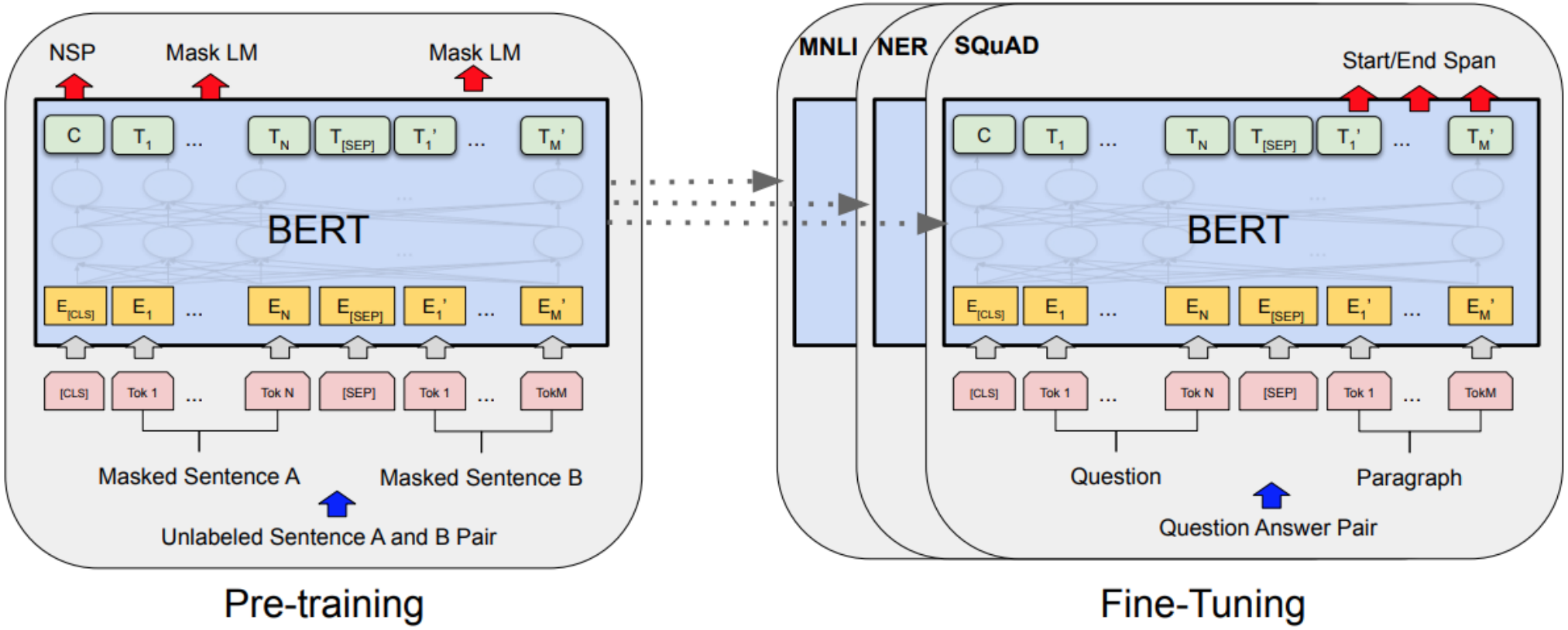
Auxiliary 2: *They traveled to* ***Washington*** _____

Auxiliary 3: _____ ***Washington*** *by plane*

Auxiliary 4: _____ *by plane*

BERT (Devlin et al., NAACL'19, best paper)

- Multi-layer bidirectional Transformer encoder
- Use masked language model + next sentence prediction
- Conduct fine-tuning after pre-training on each task.

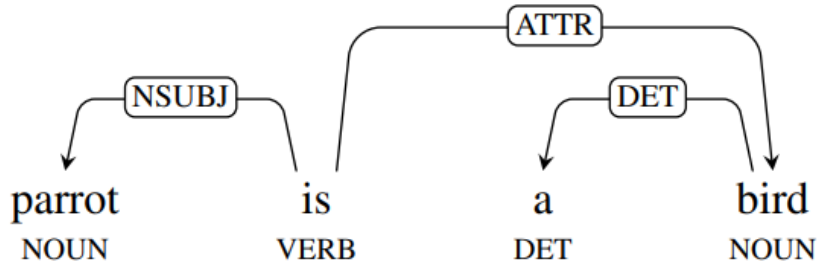


Going Beyond Word Embedding: New STOAs

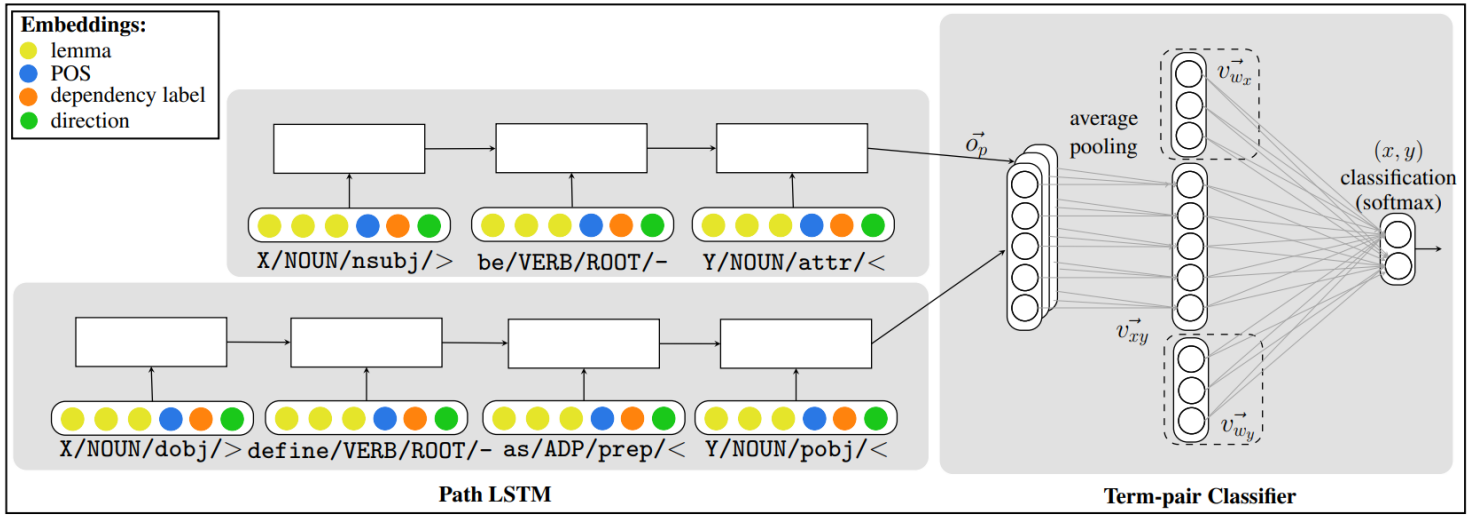
- Using **Language Model** for better representations: F1 on CoNLL03
  - Word-level Language Model:
    - ELMo (Peters et al., NAACL'18, **best paper**) 92.2
    - LD-Net (Liu et al., EMNLP'18) 92.0, ~5X faster
  - Char-level Language Model:
    - LM-LSTM-CRF (Liu et al., AAAI' 18) 91.4
    - Flair (Akbik 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

Recurrent Networks along Shortest Dependency Paths

- [Shwartz et al., 2016] Aggregating sentences for corpus-level RE
  - How to denoise the word sequence between two entities?
    - Semantic parsing -> Shortest dependency path
  - Encode sequences along the SDP
    - Four channels: Word, POS tag, dependency label, and direction (of the dependency edge)



- One pair of entities may appear in multiple sentences: aggregate sentence representations using average pooling



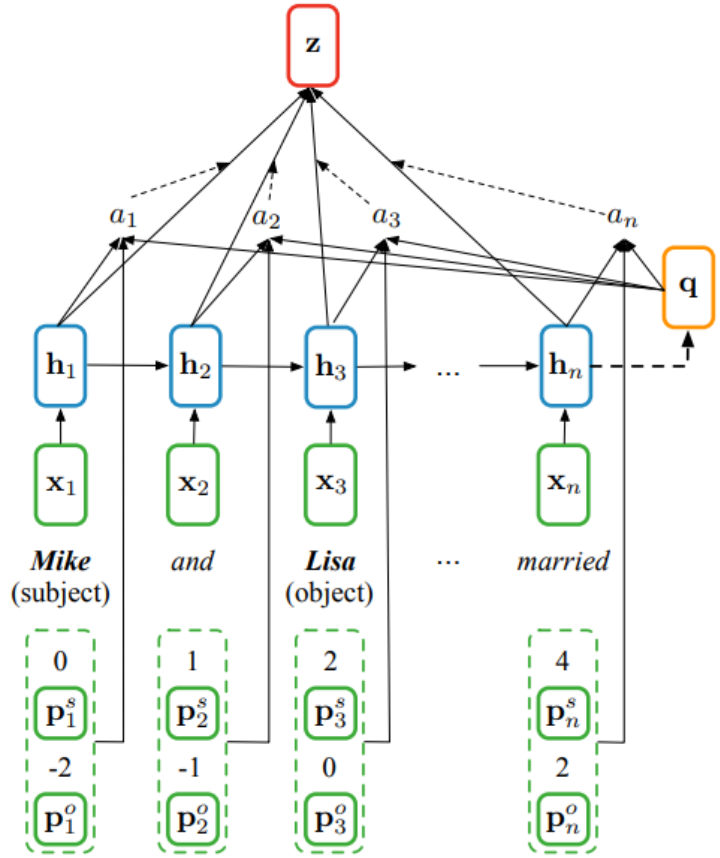
Recurrent Networks with Position-Aware Attention

- [Zhang et al., 2018] TACRED: A large (106,264) sentence-level RE dataset
  - Decide which entity is more important in predicting the relation
  - q: output state of the LSTM ("summary" of the whole sentence)
  - p^s: position feature w.r.t the subject entity
  - p^o: position feature w.r.t the object entity

$$u_i = \mathbf{v}^\top \tanh(\mathbf{W}_h \mathbf{h}_i + \mathbf{W}_q \mathbf{q} + \mathbf{W}_s \mathbf{p}_i^s + \mathbf{W}_o \mathbf{p}_i^o)$$

$$a_i = \frac{\exp(u_i)}{\sum_{j=1}^{n} \exp(u_j)}$$

$$\mathbf{z} = \sum_{i=1}^{n} a_i \mathbf{h}_i$$



Mike and Lisa got married .

Part III: Weakly/Distantly Supervised Methods

IEEE-BigData 2019 Tutorial

Taming Unstructured Big Data: Automated Information Extraction from Massive Text

Xuan Wang, Yu Zhang, Qi Li, Jiawei Han
Computer Science, University of Illinois at Urbana-Champaign

December 11, 2019

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

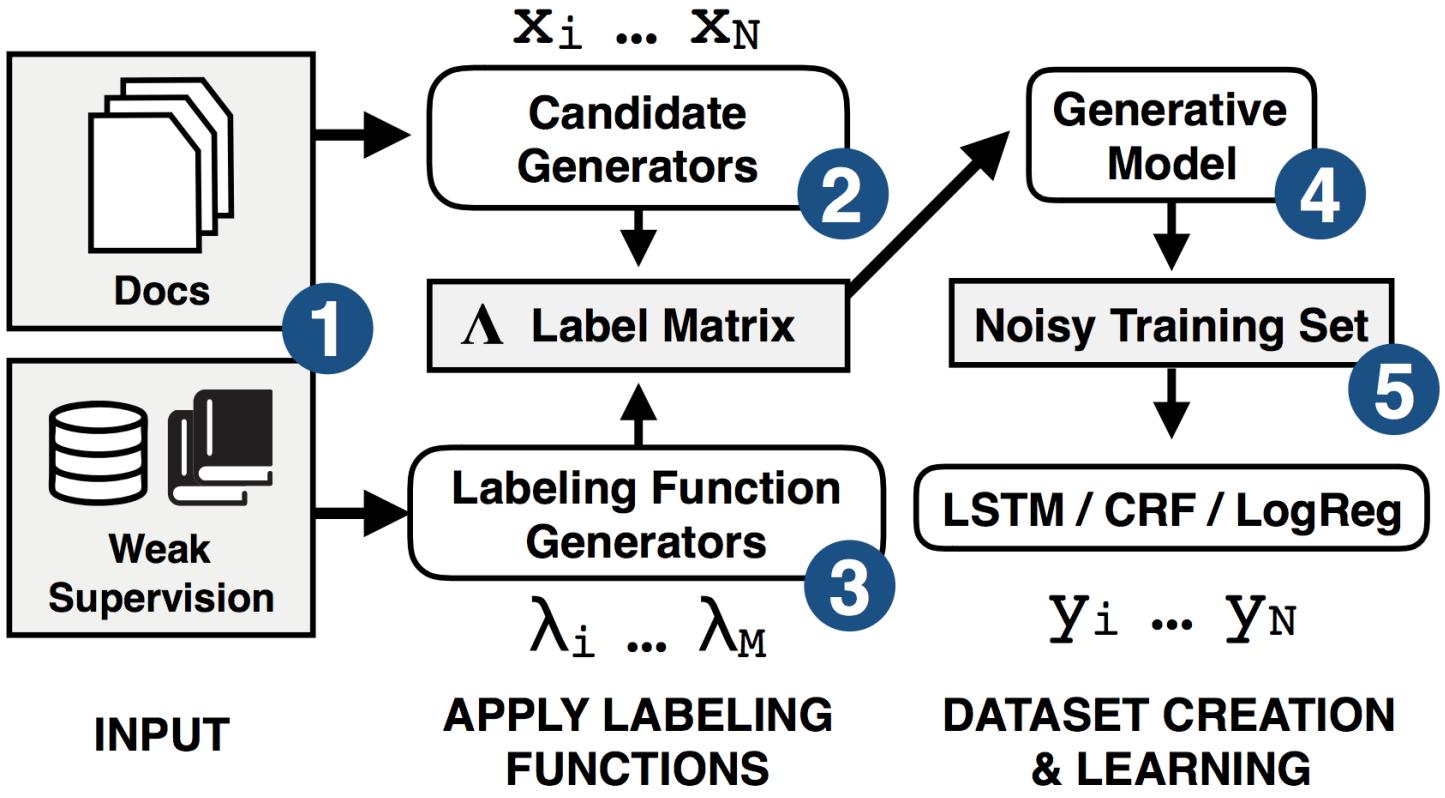
- String-match / rule-based distant supervision generation

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

- Distant-LSTM-CRF
  - Leverage AutoPhrase to extract “aspect terms”
- AutoNER
  - A novel “Tie-or-Break” labeling scheme + tailored neural model

SwellShark (Fries et al, 2017)

- Data Programming for Entity Typing
  - Entity detection: Regular expressions based on part-of-speech tags
  - Requires expert efforts
    - Candidate Generators



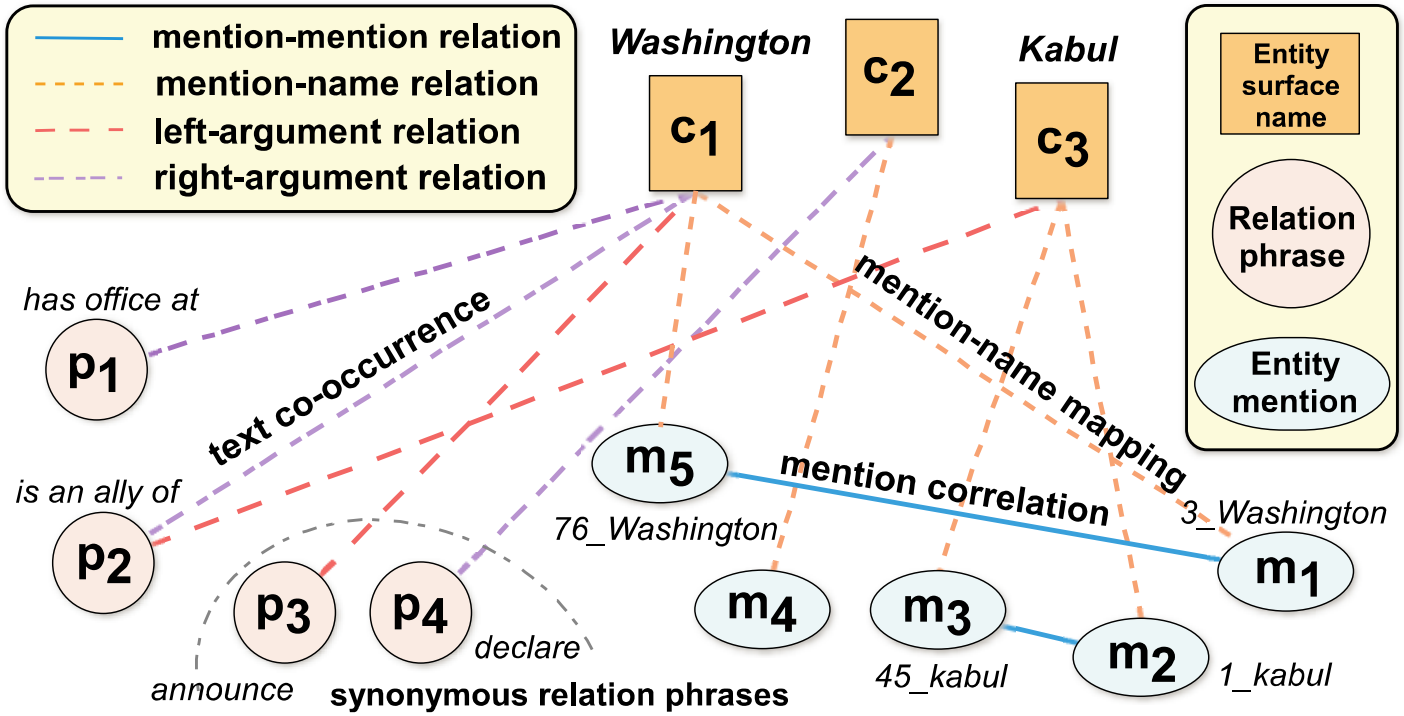
ClusType (Xiang et al, KDD'15)

- POS-constrained phrase segmentation for mining candidate entity mentions and relation phrases, simultaneously
- Construct a heterogeneous graph to represent available information in a unified form

Entity mentions are kept as individual objects **to be disambiguated**

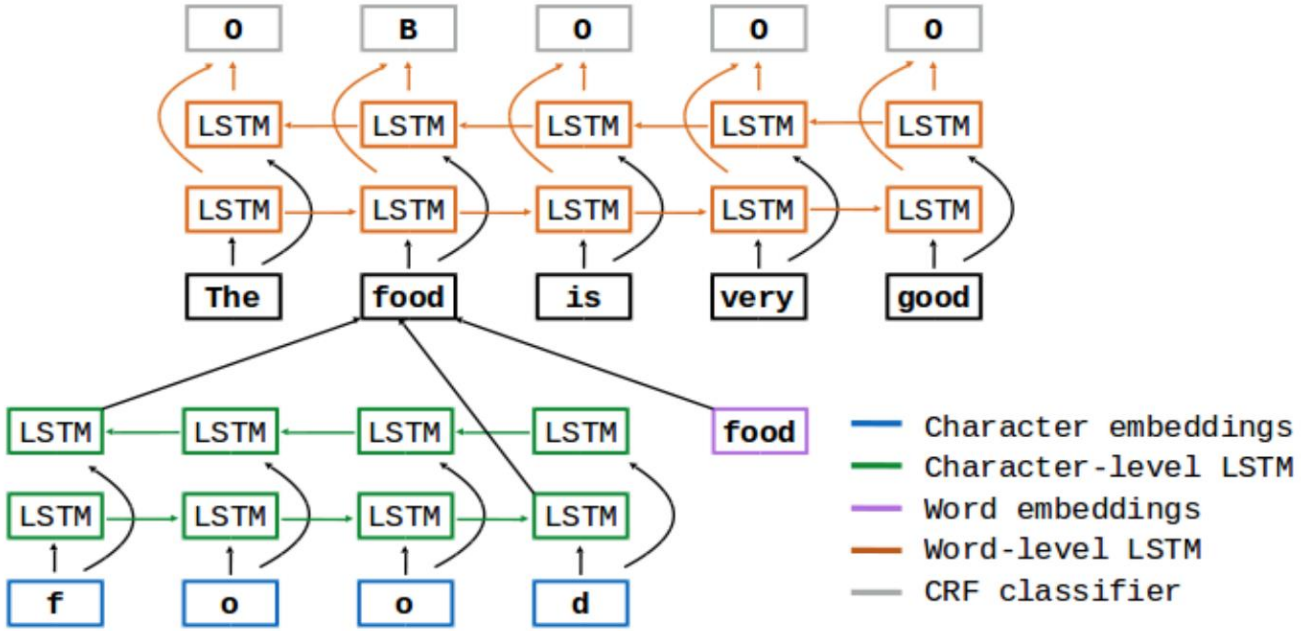
Linked to entity surface names & relation phrases

Weight assignment: The more two objects are likely to share the same label, the larger the weight will be associated with their connecting edge



Distant-LSTM-CRF (Giannakopoulos et al., 2018)

- AutoPhrase + LSTM-CRF
  - AutoPhrase generates candidate entities
  - Syntactic rules to label aspect terms
  - LSTM-CRF builds models
    - Both word & char info are used
- Problem
  - Many false-negative labels



AutoNER (Shang et al., 2018): 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: "Tie or Break" Labeling Scheme

- 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

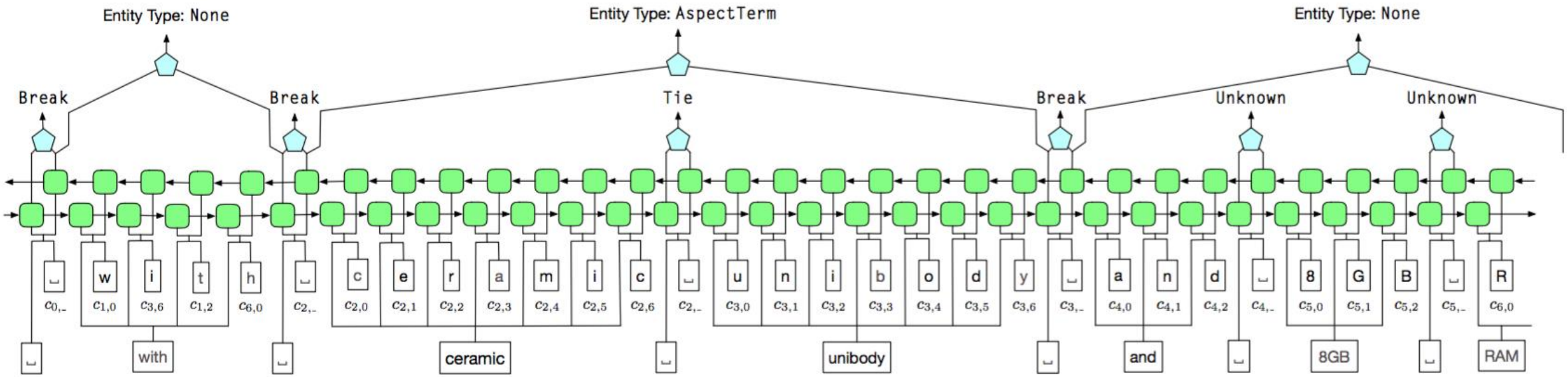


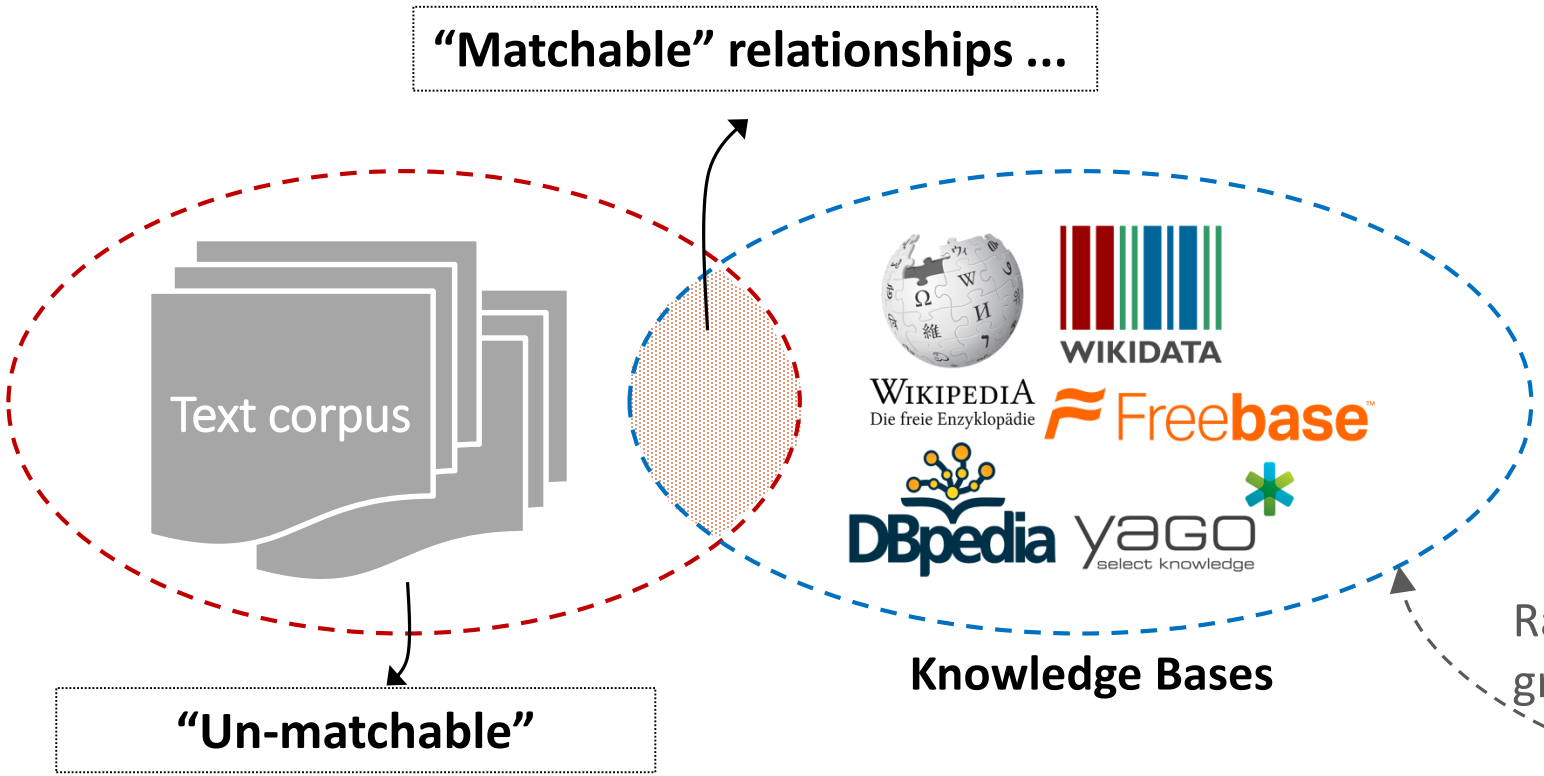
Figure 2: The illustration of AutoNER with `Tie or Break` tagging scheme. The named entity type is {`AspectTerm`}. "ceramic unibody" is a matched `AspectTerm` entity and "8GB RAM" is an unknown-typed high-quality phrase. `Unknown` labels will be skipped during the model training.

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.

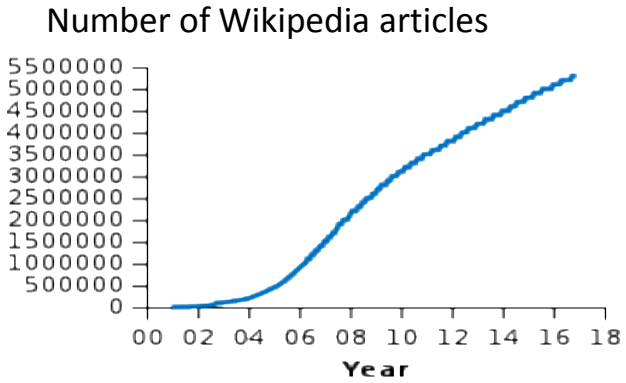
Method	Human Effort other than Dictionary	BC5CDR			NCBI-Disease		
		Pre	Rec	F1	Pre	Rec	F1
Supervised Benchmark	Gold Annotations	88.84	85.16	**86.96**	86.11	85.49	**85.80**
SwellShark	Regex Design + Special Case Tuning	86.11	82.39	84.21	81.6	80.1	**80.8**
	Regex Design	84.98	83.49	**84.23**	64.7	69.7	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**

Distantly Supervised Relation Extraction



Freely available!

- Common knowledge
- Life sciences
- Art …



Rapidly growing!

Human crowds

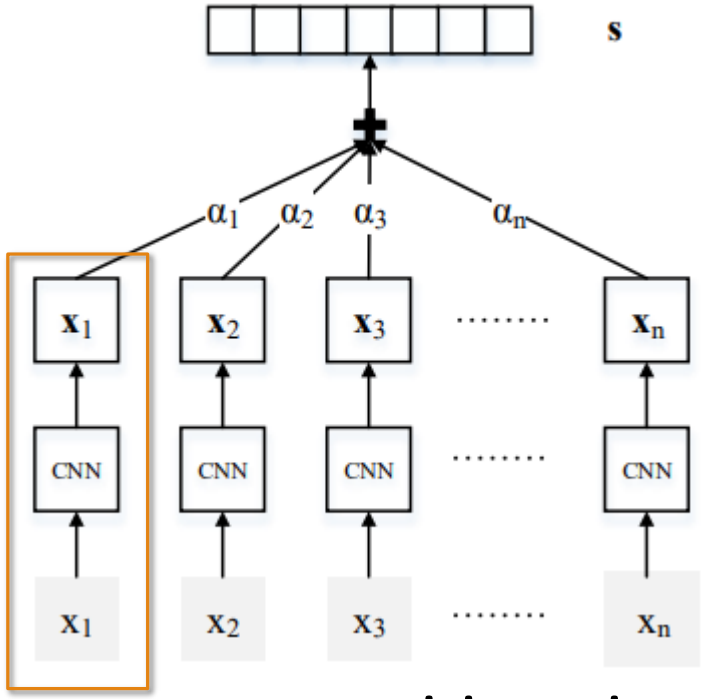
(Mintz et al., 2009) , (Riedek et al., 2010), (Lin et al., 2012), (Ling et al., 2012), (Surdeanu et al., 2012), (Xu et al., 2013), (Nagesh et al., 2014), …

https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Position-Aware CNN with Attention Over Instances

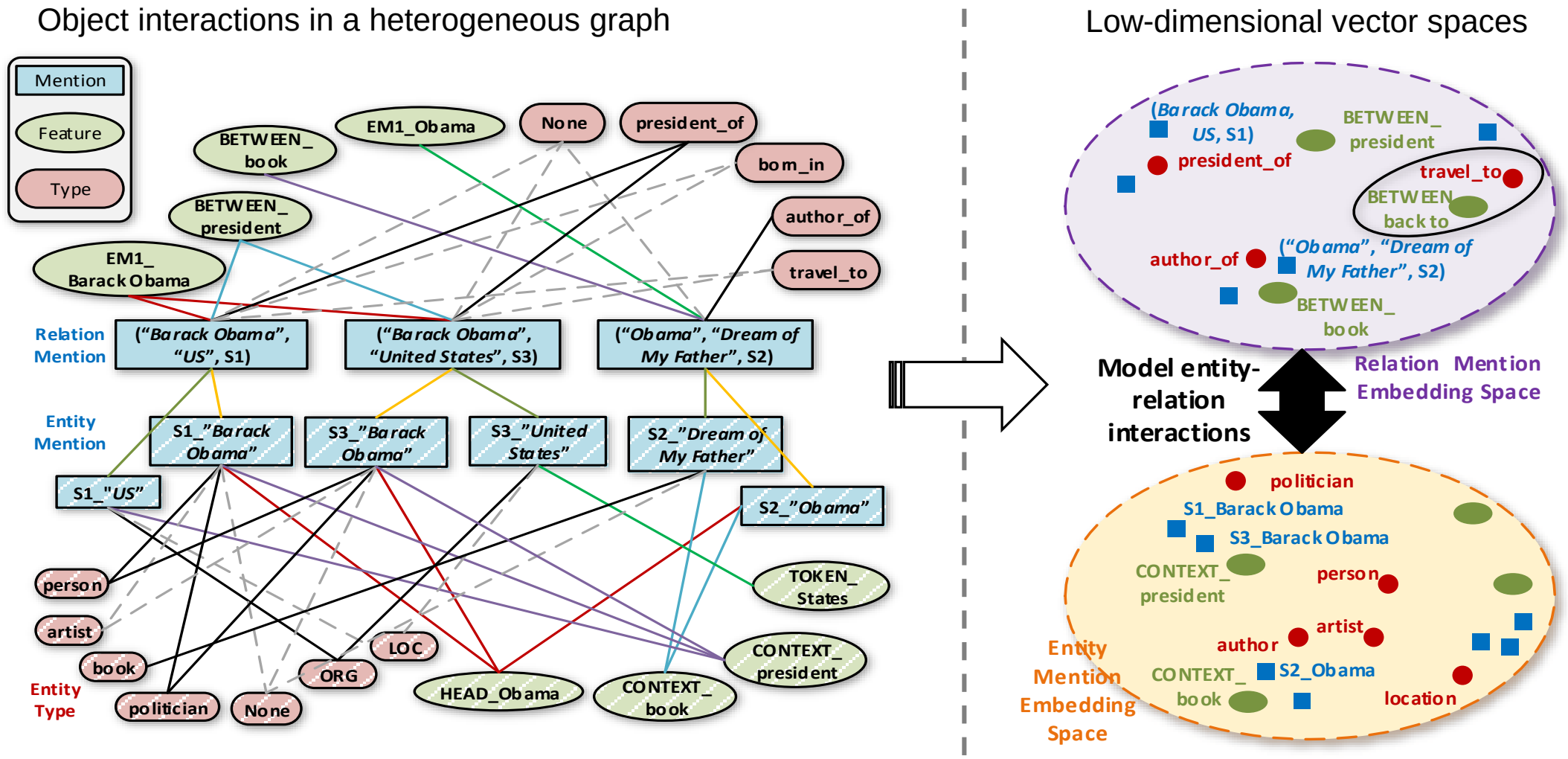
- [Lin et al., 2016] Attention over matched instances
  - One tuple in KB can be matched multiple times in the corpus.
  - Not all matched instances indicate the relation in KB.
    - KB: (**Beijing**, *capital of*, **China**)
    - Matched sentence: **Beijing** is a city in **China**.
  - Entity pair representation is a weighted sum of sentence representation.

$$\mathbf{s} = \sum_i \alpha_i \mathbf{x}_i, \quad \alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}, \quad e_i = \mathbf{x}_i \mathbf{A} \mathbf{r},$$



***n* sentences containing entity pair (Beijing, China)**

CoType: Co-Embedding for Typing Entities and Relations



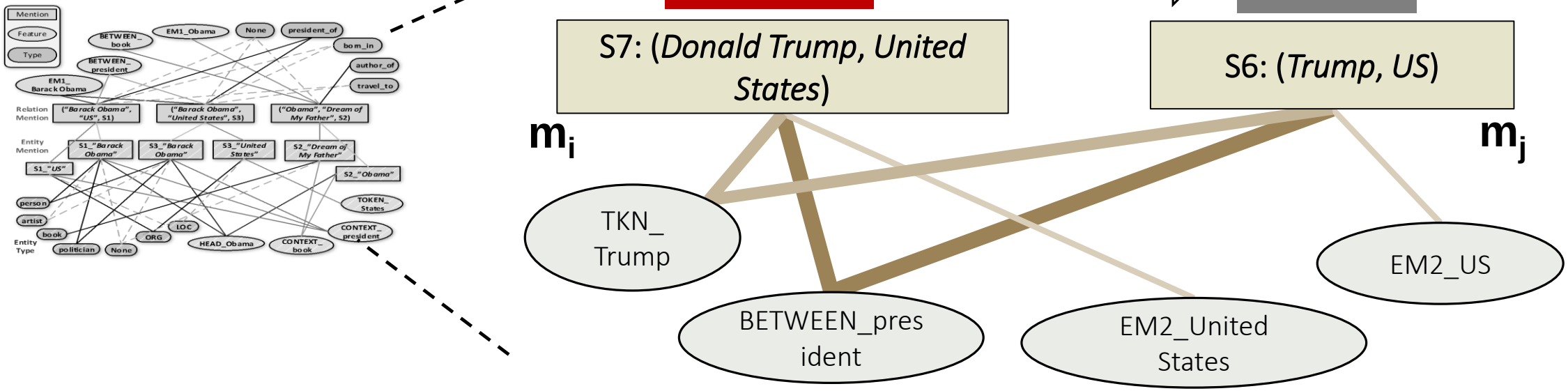
(Ren et al. WWW'17)

Modeling Mention-Feature Co-occurrences

- Second-order Proximity

  Mentions with similar distributions over text features should have similar types

Vertex $\mathbf{m_i}$ and $\mathbf{m_j}$ have a large second-order proximity



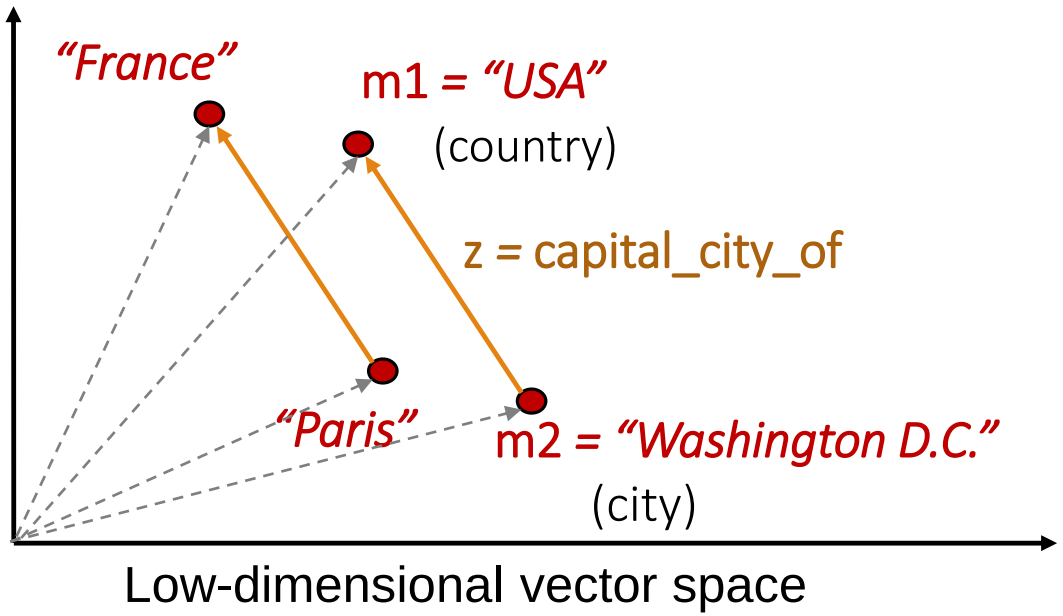
(Tang et al., WWW'15), (Ren et al. WWW'17)

Modeling Entity-Relation Interactions

Object "Translating" Assumption

For a relation mention **z** between entity arguments **m1** and **m2**:

vec(m1) ≈ **vec**(m2) + **vec**(z)



Error on a relation triple (z, m1, m2):

$$\tau(z) = \|\mathbf{m}_1 + \mathbf{z} - \mathbf{m}_2\|_2^2$$

$$\sum_{z_i \in \mathcal{Z}_L} \sum_{v=1}^{V} \max\{0, 1 + \tau(z_i) - \tau(z_v)\}$$

positive relation triple

negative relation triple

(Bordes, NIPS'13), (Ren et al., WWW'17)

Reducing Error Propagation: A Joint Optimization Framework

Modeling **entity-relation interactions**

$$O_{ZM} = \sum_{z_i \in \mathcal{Z}_L} \sum_{v=1}^{V} \max\{0, 1 + \tau(z_i) - \tau(z_v)\}$$

$$\min \quad \mathcal{O} = \mathcal{O}_M + \mathcal{O}_Z + \mathcal{O}_{ZM}$$

$$O_Z = \mathcal{L}_{ZF} + \sum_{i=1}^{N_L} \ell_i + \frac{\lambda}{2} \sum_{i=1}^{N_L} \|\mathbf{z}_i\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_r} \|\mathbf{r}_k\|_2^2$$

Modeling types of **relation mentions**

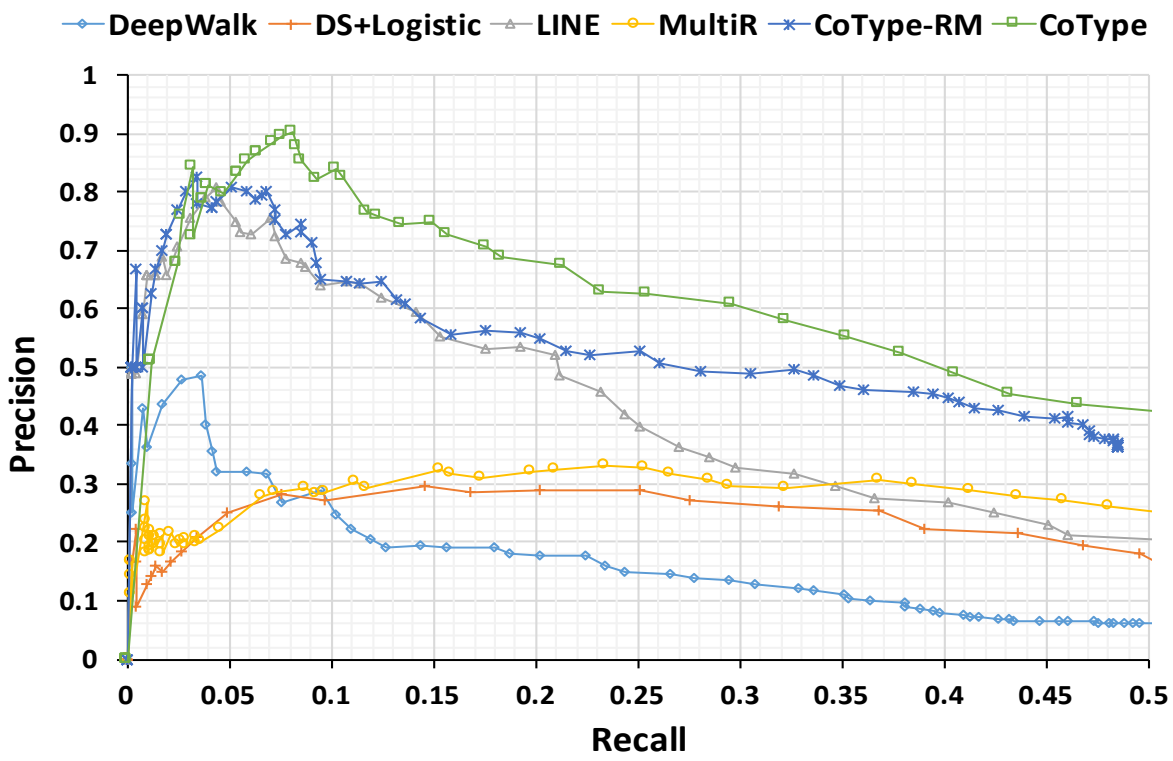
$$O_M = \mathcal{L}_{MF} + \sum_{i=1}^{N'_L} \ell'_i + \frac{\lambda}{2} \sum_{i=1}^{N'_L} \|\mathbf{m}_i\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_y} \|\mathbf{y}_k\|_2^2$$

Modeling types of **entity mentions**

(Ren et al., WWW'17)

CoType: Comparing with State-of-the-Arts RE Systems

- Given candidate relation mentions, predict its relation type if it expresses a relation of interest; otherwise, output "None"



- DS+Logistic (Stanford, ACL'09): logistic classifier on DS
- MultiR (UW, ACL'11): handles inappropriate labels in DS
- DeepWalk (StonyBrook, KDD'14): homogeneous graph embedding
- LINE (MSR, WWW'15): joint feature & type embedding
- CoType-RM (WWW'17): only models relation mentions
- CoType (WWW'17): models entity-relation interactions

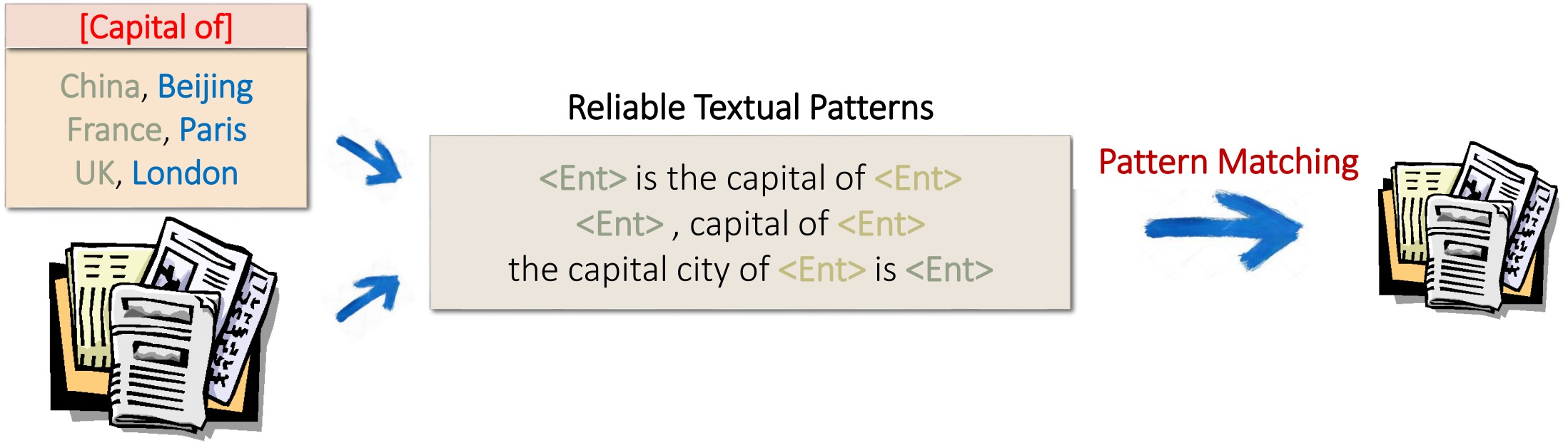
NYT public dataset (Riedel et al. 2010, Hoffmann et al., 2011): 1.18M sentences in the corpus, 395 manually annotated sentences for evaluation, 24 relation types

Weakly Supervised Relation Extraction

- Given a text corpus D and some target relations R, where each target relation r is characterized by a few seed instances $\{(e_{h_k}, e_{t_k}, r)\}$, weakly supervised RE aims at discovering more entity pairs $\{(e_{h_i}, e_{t_i}, r)\}$.

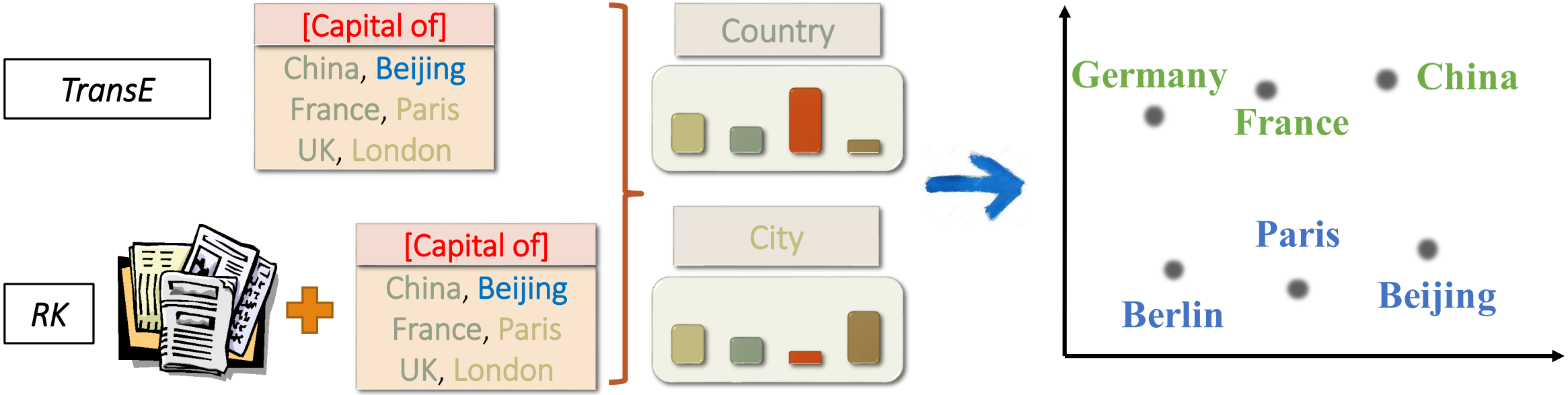
Pattern-Based Approaches

- Snowball [Agichtein et al., 2000] and PATTY [Nakashole et al., 2012]
  - Extract textual patterns (e.g., tokens between a pair of entities) and new relation instances in a bootstrapping manner
  - Assumption: For an entity pair, their relation is described in sentences mentioning both of them.



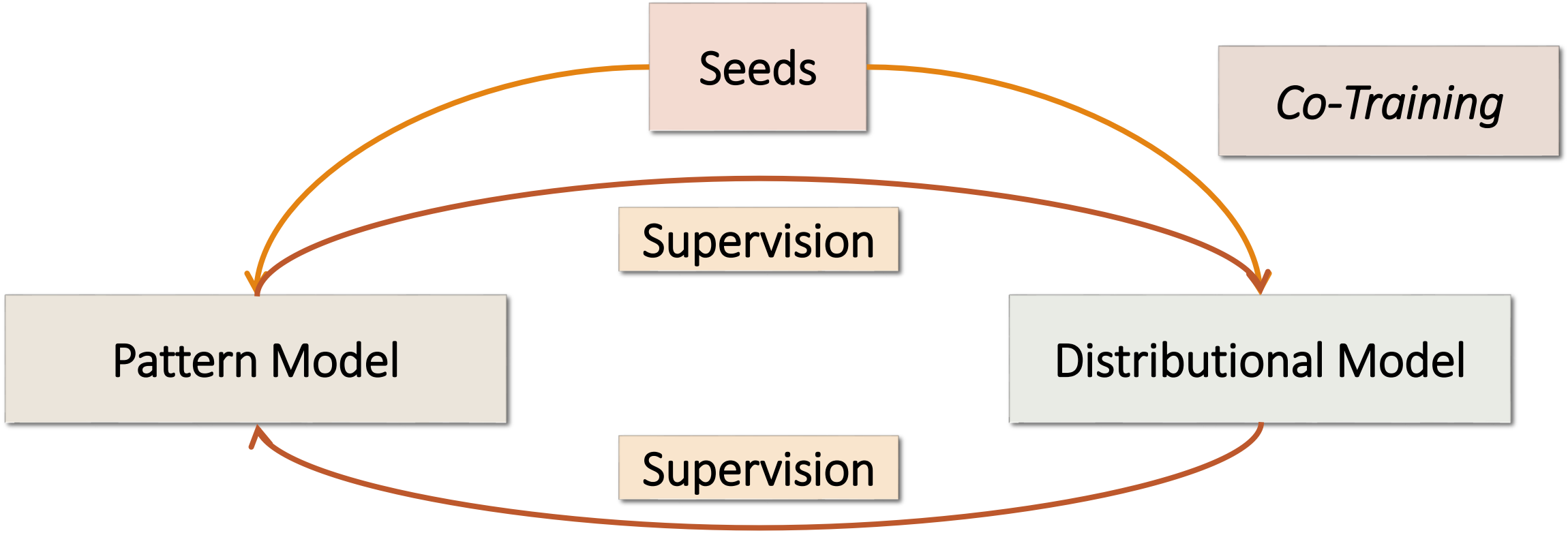
Distributional Approaches

- TransE [Bordes et al., 2014] and RK [Wang et la., 2016]
  - Learn distributed representations for entities based on their co-occurrence
  - Similar entities will have similar representations
  - Train a classifier for relation prediction



Pattern-Enhanced Embedding Learning

- REPEL [Qu et al., 2018]
  - Encourage both models to generate extra supervision for each other to complement the limited supervision from the given seeds



Co-Training Framework

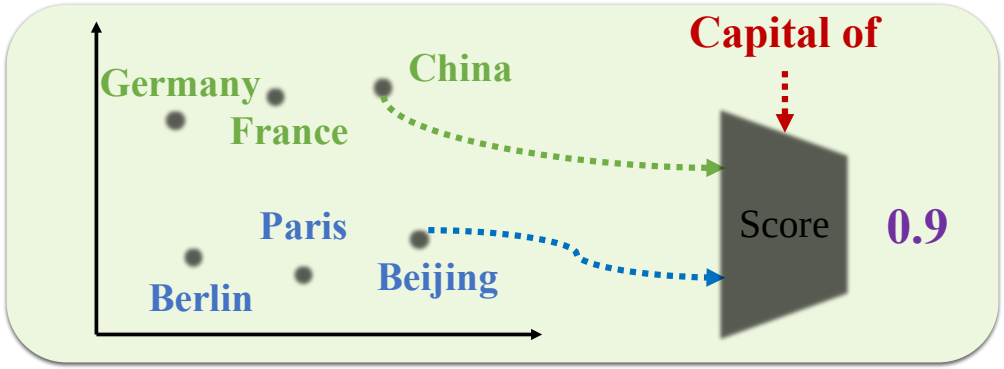
- **Objective function:**

$$Max_{P,D}\{O_P + O_D + O_I\}$$

Reliable Textual Patterns

<Ent> is the capital of <Ent>
<Ent> , capital of <Ent>
the capital city of <Ent> is <Ent>

Relation Classifier

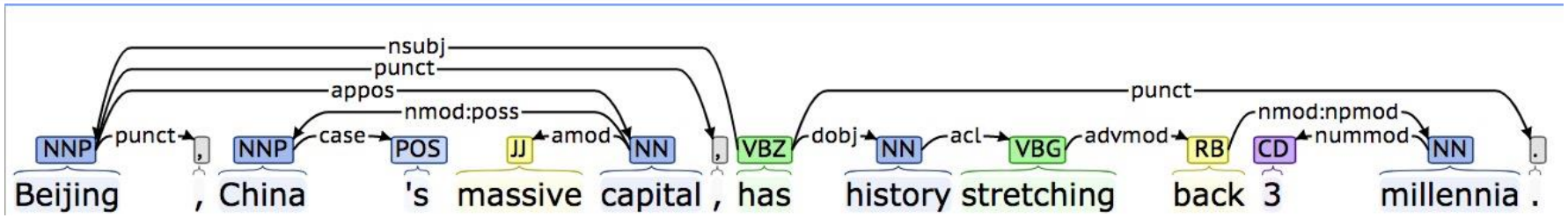


Pattern Model (O_P)

- Patterns along the shortest dependency path

Beijing , China 's massive capital , has history stretching back 3 millennia .

<Ent> <Ent> capital



- Use the given seeds to measure the pattern reliability

$$R(p|r) = \frac{|E_p \bigcap S_r|}{|E_p|}$$

China, Beijing

China, Shanghai

Germany, Berlin

USA, NYC

1 / 4 = 0.25

Distributional Model (O_D)

- Learn entity representations by considering both the text corpus and the given seeds
  - Text: use the objective function of LINE

$$P(w|e) = \frac{\exp(\mathbf{x}_e \cdot \mathbf{c}_w)}{Z}, \qquad O_{text} = \sum_{w,e} n_{w,e} \log P(w|e),$$

  - Seed: use the objective function of TransE
    - It will learn a score function L_D to measure the possibility of each instance

$$L_D(f|r) = 1 - ||x_{e_h} + y_r - x_{e_t}||_2^2,$$

L_D(Beijing, China, Capital) = 0.9
L_D(Chicago, USA, Capital) = 0.3

Modeling the Interaction (O_I)

$$O_I = \max_{P,D} E_{f \sim G_P}[L_D(f)]$$

Generator

Discriminator

Candidate Pairs:

Capital of	
Japan	Tokyo

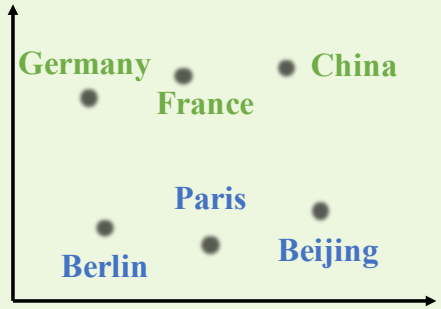
Pattern Model

[ENT] [ENT] 's capital

[ENT] capital [ENT]

capital [ENT] [ENT]

Distributional Model

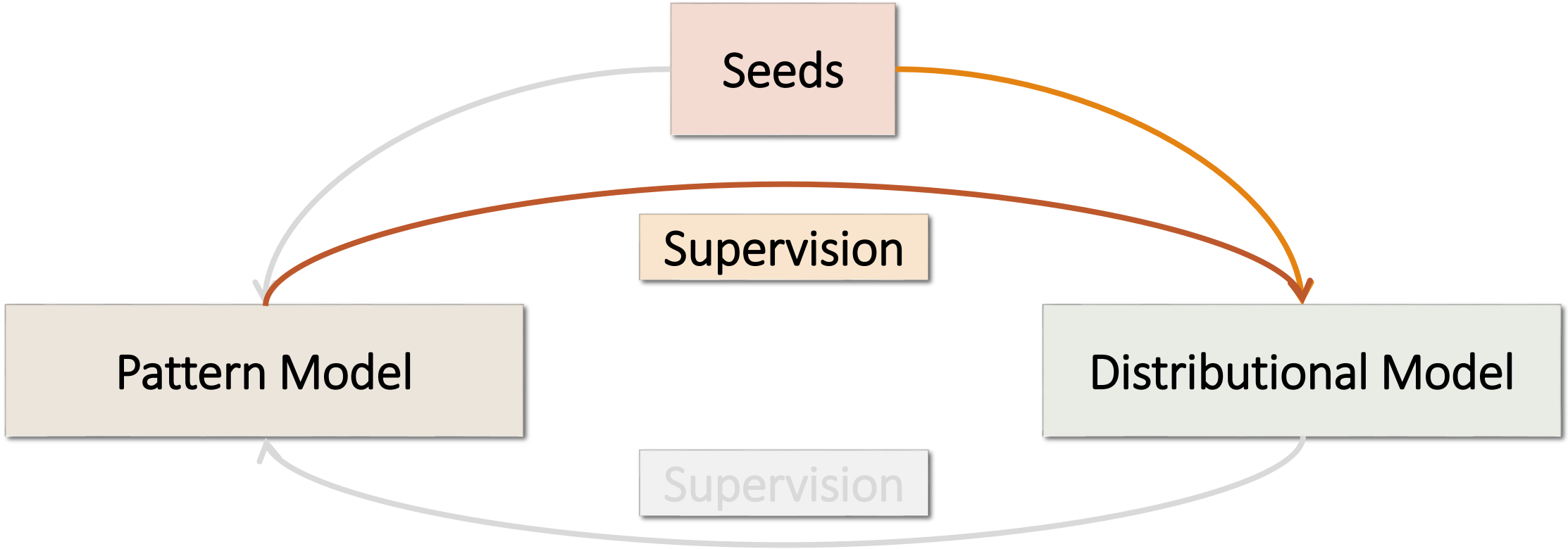


Pattern Reliability:

[ENT] [ENT] 's capital **0.9**

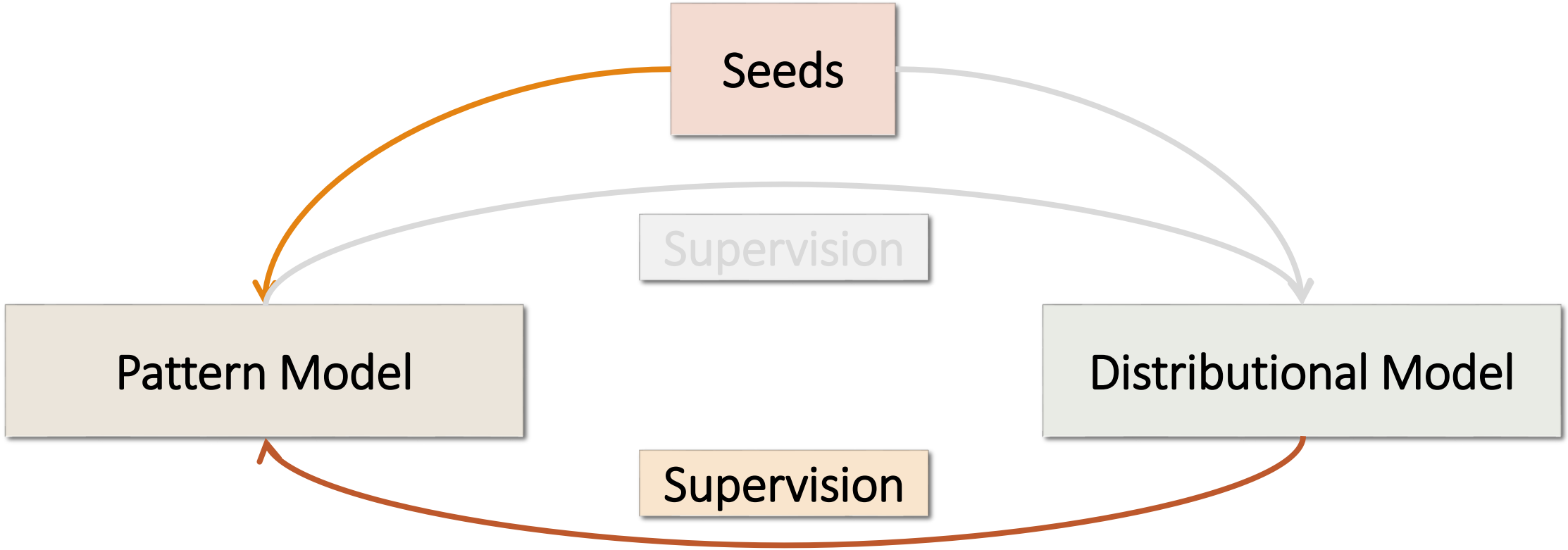
Optimization

- The objective can be optimized with the coordinate gradient descent algorithm.
- Step 1: Fix P and optimize D



Optimization

- The objective can be optimized with the coordinate gradient descent algorithm.
- Step 2: Fix D and optimize P

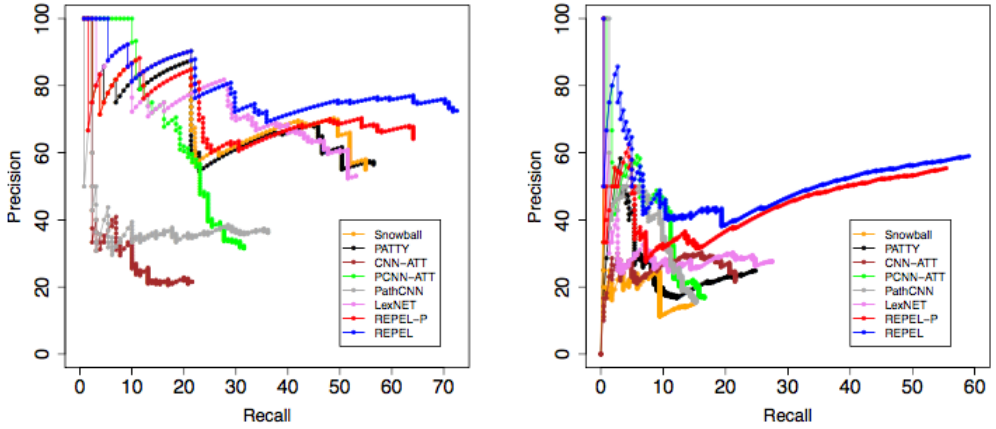


Experiments

- REPEL is comparable with the neural methods in the weakly-supervised setting
- REPEL outperforms the joint training methods
- The distributional module helps improve the pattern module (REPEL vs REPEL-P)

Dataset	Wiki + Freebase	NYT + Freebase
# Documents	150,000	118,664
# Entities	92,443	23,120
# Candidate Patterns	621,782	232,892
# Seed Instances per Relation	50	50
# Relations in KBC	10	10
# Test Instances in KBC	10,734	6,094
# Relations in RE	5	6
# Test Entity Pairs in RE	131	222

Algorithm	Wiki + Freebase				NYT + Freebase			
	P@50	R@50	P@100	R@100	P@50	R@50	P@100	R@100
Snowball	58.00	22.14	65.00	49.62	20.00	4.50	21.00	9.46
CNN-ATT	26.00	9.92	22.00	16.79	24.00	5.41	29.00	13.06
PCNN-ATT	58.00	22.14	36.00	27.48	46.00	10.36	26.00	11.71
PathCNN	36.00	13.74	38.00	29.01	42.00	9.46	26.00	11.71
LexNET	74.00	28.24	61.00	46.56	32.00	7.21	26.00	11.71
REPEL-D	14.00	5.34	17.00	12.98	6.00	1.35	7.00	3.15
REPEL-P	64.00	24.43	70.00	53.44	32.00	7.21	33.00	14.86
REPEL	**78.00**	**29.77**	**76.00**	**58.02**	**48.00**	**10.81**	**43.00**	**19.37**



(a) Wiki (b) NYT

Part IV: Unsupervised Methods

IEEE-BigData 2019 Tutorial

Taming Unstructured Big Data: Automated Information Extraction from Massive Text

Xuan Wang, Yu Zhang, Qi Li, Jiawei Han
Computer Science, University of Illinois at Urbana-Champaign

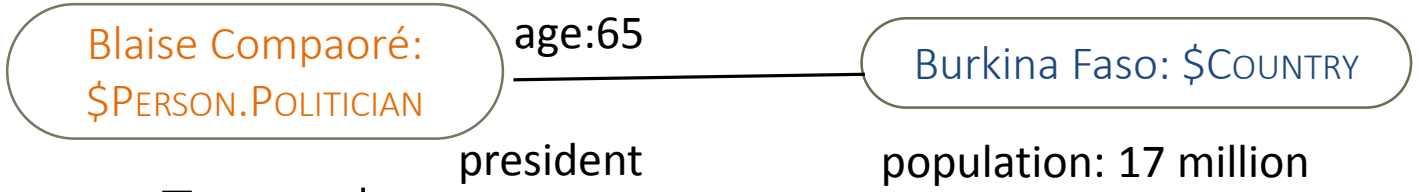
December 11, 2019

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

- [Jiang et al., KDD 2017]
- Motivation:

Given a sentence in a large corpus, "President Blaise Compaoré's government of **Burkina Faso** was founded…", …

We may find:



- 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

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]:

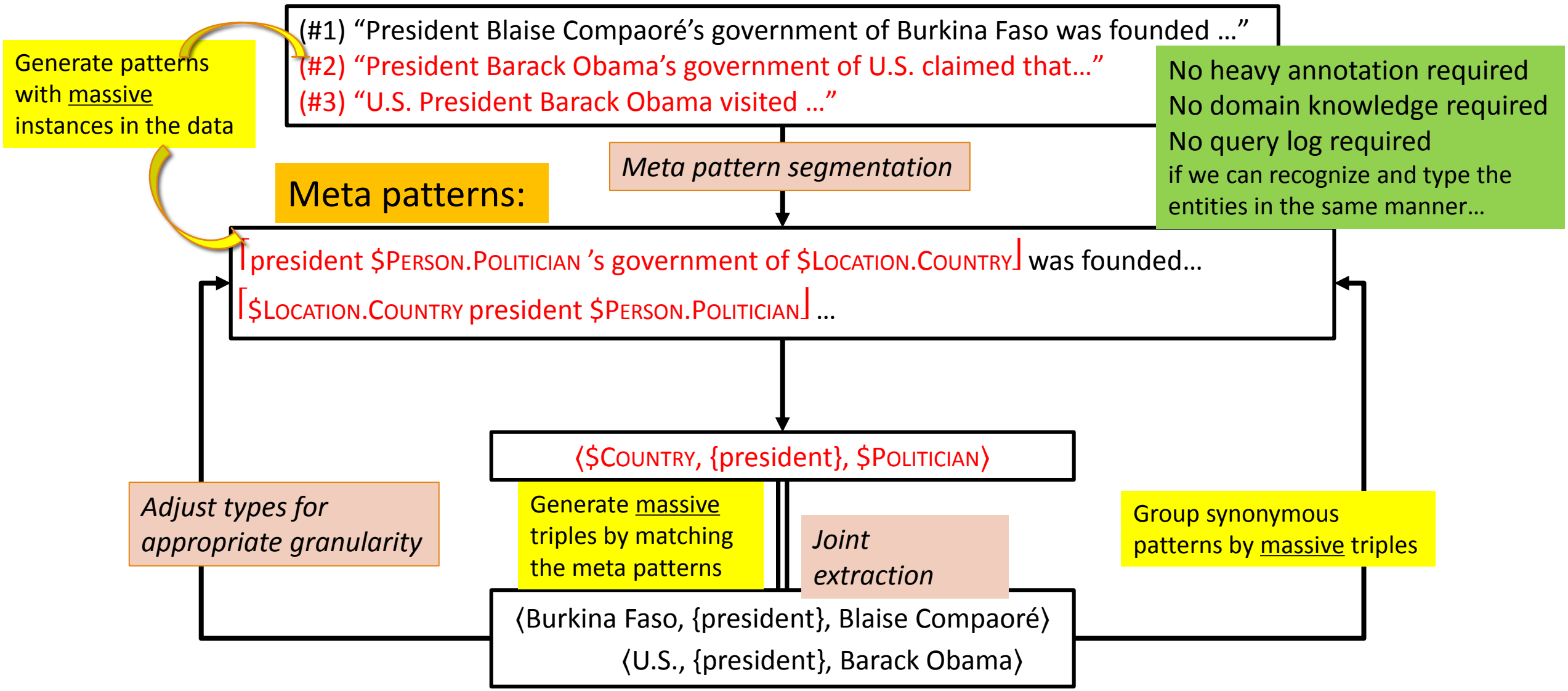
"president of united states" ➡ "A of E", "E 's A", "E A", "A, E"

"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⟩ ☹

Meta-Pattern Methodology

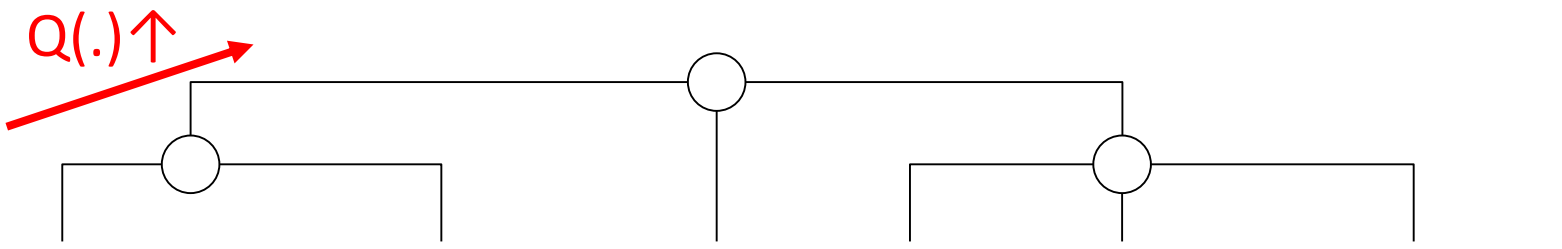


Meta-Pattern Quality Assessment and Segmentation

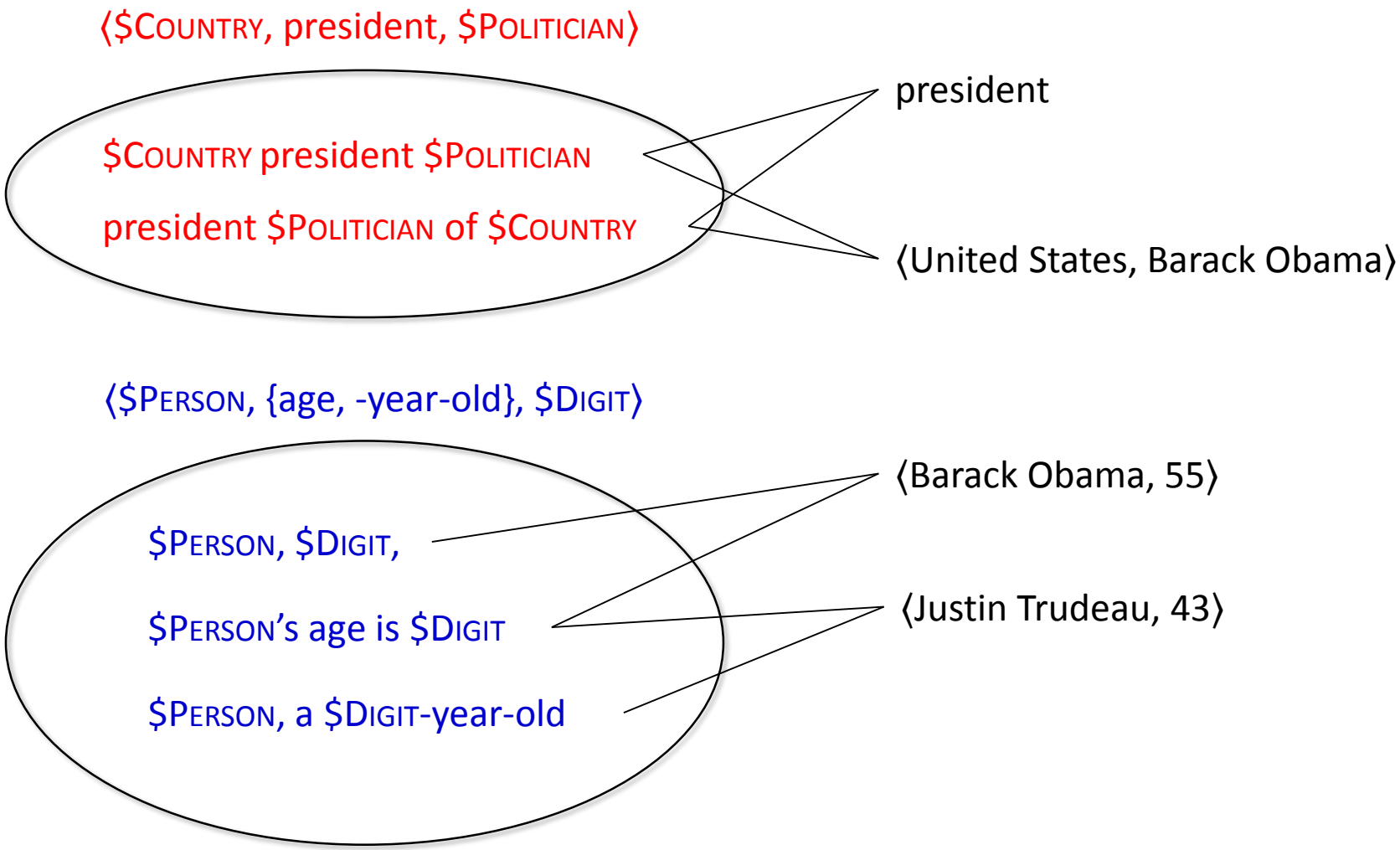
A rich set of features:

- ✓ 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



Grouping Synonymous Patterns



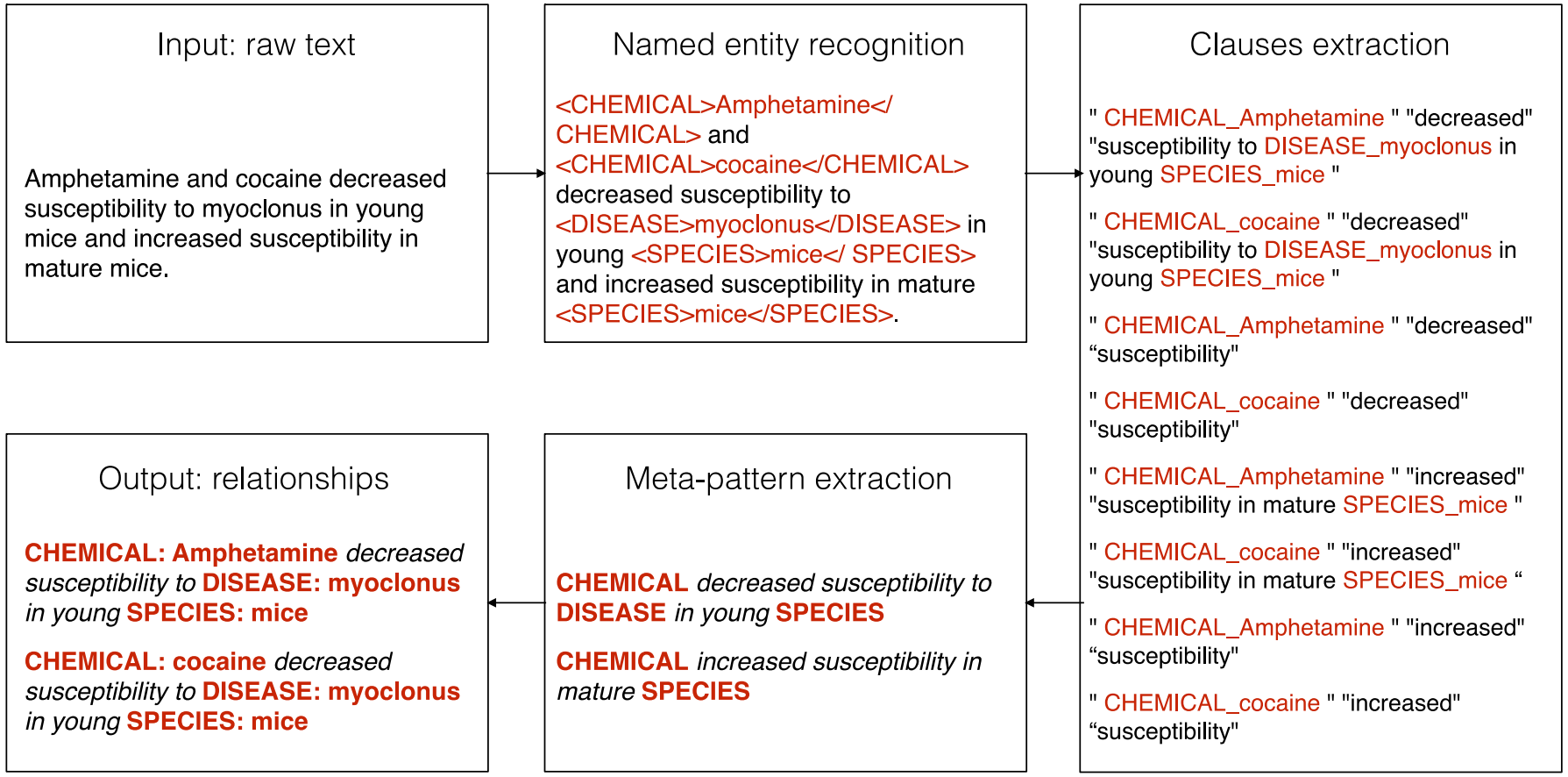
Results: Patterns, Entities and Attribute Values in News Corpus

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

CPIE: Clause+Pattern-guided Information Extraction

- [Wang et al., 2018] Incorporate clause analysis in meta-pattern discovery
- Motivation: Long and complicated sentences in biomedical research articles



Framework

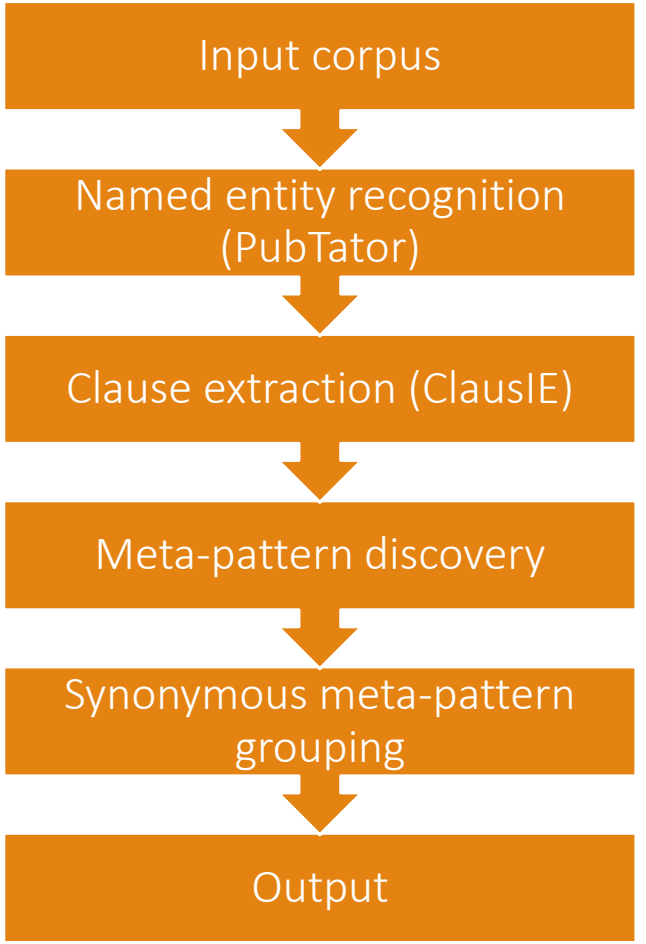
- **Named Entity Recognition (NER):**
  - PubTator (Wei *et al*., 2013)
  - Fully annotated PubMed abstracts: gene, chemical, disease, species and SNP
- **Clause Extraction:**
  - ClausIE (Corro and Gemulla, 2013)
  - Extract all possible clauses from an input sentence.
- **Candidate Meta-pattern Generation:**
  - Take the segment between the first and last typed entity in each input clause as a candidate meta-pattern.

Experiments: Dataset

- Dataset: 28,007 PubMed abstracts

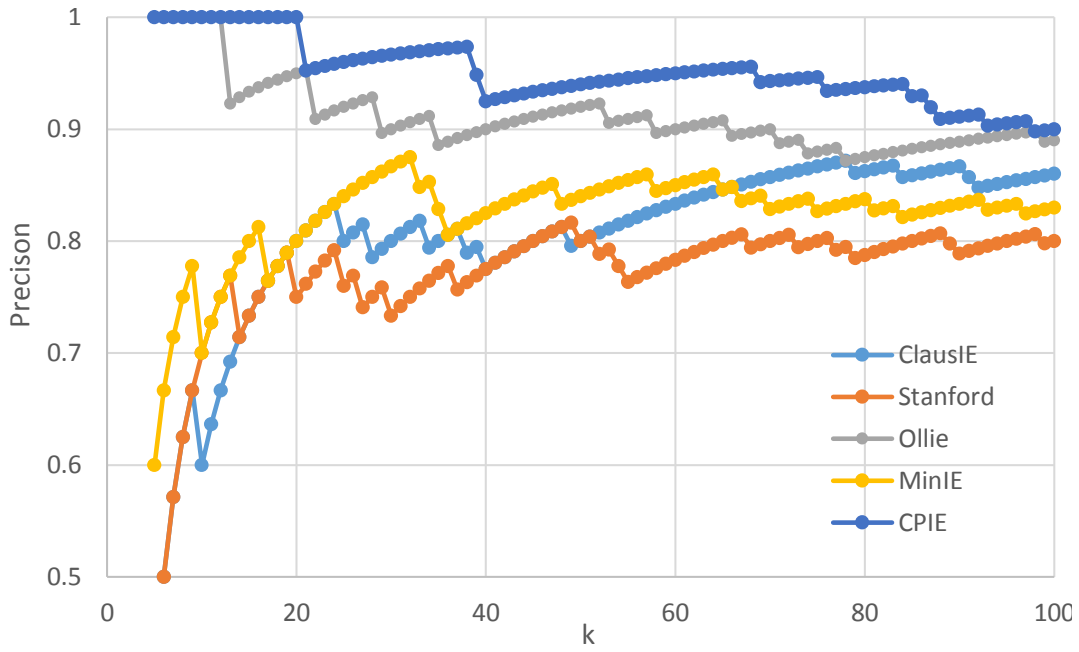
Corpus statistics from Comparative Toxicogenomics Database (CTD)

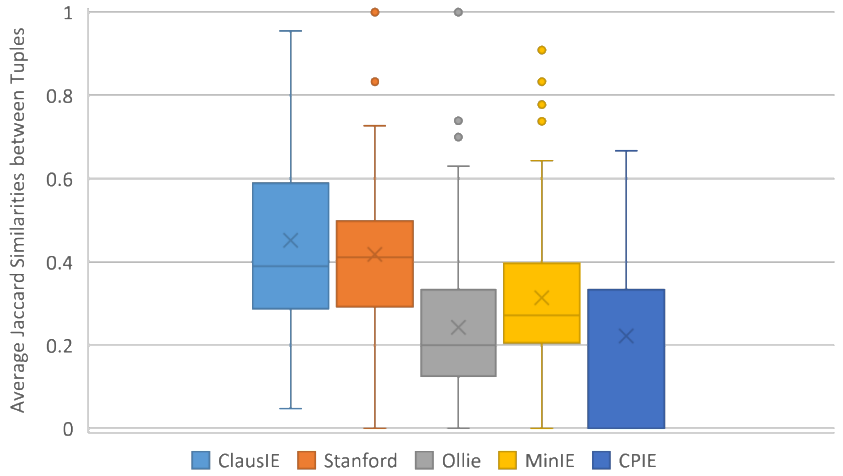
	# of Entity	# of Relation	# of PMID
Chemical-Gene	Chemical: 7,187 Gene: 667	163,126	26,786
Chemical-Disease	Chemical: 7,187 Disease: 598	84,104	1,619
Disease-Gene	Gene: 669 Disease: 599	834	1,622
Total	**Chemical: 7,187 Gene: 669 Disease: 599**	**248,064**	**28,007**



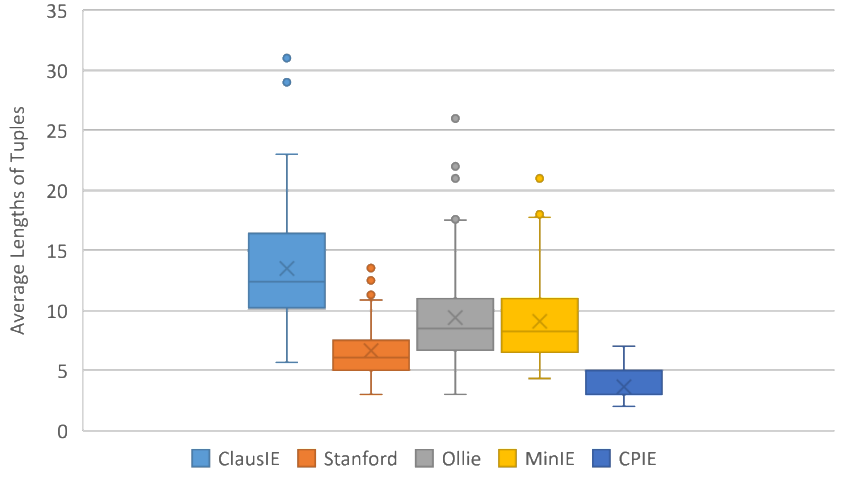
Performance Comparison with OpenIE Methods

	P@50	P@100	MAP	NDCG@50	NDCG@100
ClausIE [10]	0.800	0.860	0.800	0.710	0.875
Stanford [2]	0.800	0.800	0.765	0.714	0.870
Ollie [37]	0.920	0.890	0.917	0.935	0.982
MinIE [13]	0.840	0.830	0.820	0.800	0.918
CPIE	**0.940**	**0.900**	**0.956**	**0.956**	**0.991**



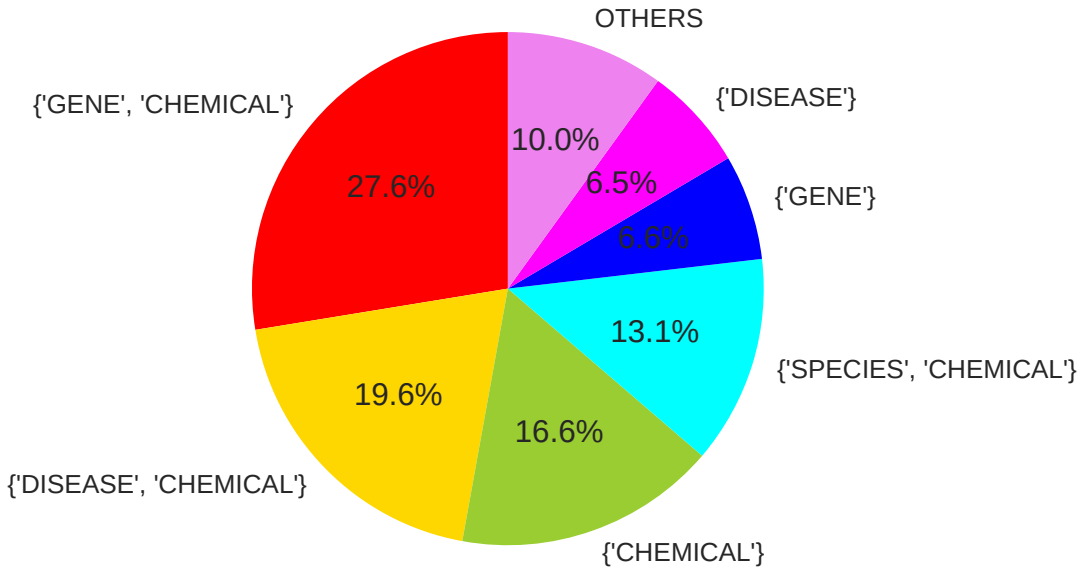


Distinctiveness



Simplicity

Quality Meta-patterns



Quality meta-patterns	Count
CHEMICAL induced DISEASE	306
CHEMICAL inhibited GENE	270
CHEMICAL increased GENE	245
DISEASE be induced by CHEMICAL	233
DISEASE induced by CHEMICAL	181
CHEMICAL inhibits GENE	153
GENE receptor is GENE	145
GENE inhibitor CHEMICAL	138
CHEMICAL inhibited CHEMICAL	126
CHEMICAL inhibited DISEASE	101
DISEASE be induced by CHEMICAL in SPECIES	25
CHEMICAL induces apoptosis in SPECIES DISEASE	25
DISEASE induced by CHEMICAL in SPECIES	20
CHEMICAL against CHEMICAL induced DISEASE	15
SPECIES be treated with CHEMICAL	275

Synonymous meta-pattern groups

Synonymous group	Meta-patterns
CHEMICAL induce DISEASE	CHEMICAL causes DISEASE
	CHEMICAL leads to DISEASE
	CHEMICAL induces DISEASE
	CHEMICAL exposure causes DISEASE
CHEMICAL inhibit GENE	CHEMICAL decreased GENE
	CHEMICAL decreases GENE
	CHEMICAL inhibition of GENE
	CHEMICAL suppressed GENE
CHEMICAL no effect on GENE	CHEMICAL had no effect on GENE
	CHEMICAL did not affect GENE
DISEASE induced by CHEMICAL in SPECIES	DISEASE induced by CHEMICAL in SPECIES
	DISEASE be induced by CHEMICAL in SPECIES
SPECIES treated with CHEMICAL	SPECIES were pretreated with CHEMICAL
	SPECIES were administered with CHEMICAL
	CHEMICAL treated SPECIES
	SPECIES be induced by CHEMICAL
	SPECIES were exposed to CHEMICAL

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

- [Wang et al., 2018]
- 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

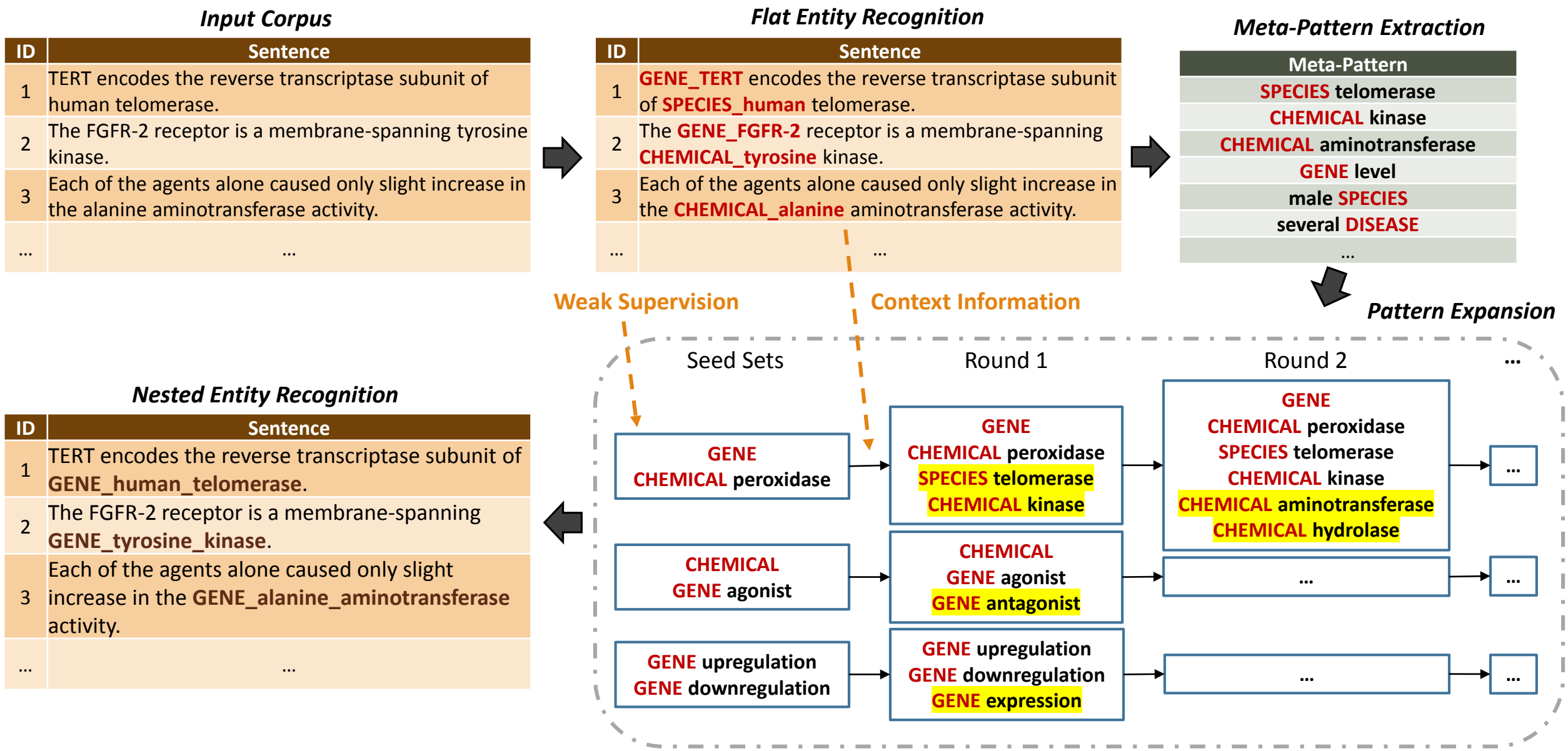
PENNER: Key Ideas of Using Meta-Pattern

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

  CHEMICAL aminotransferase = PROTEIN

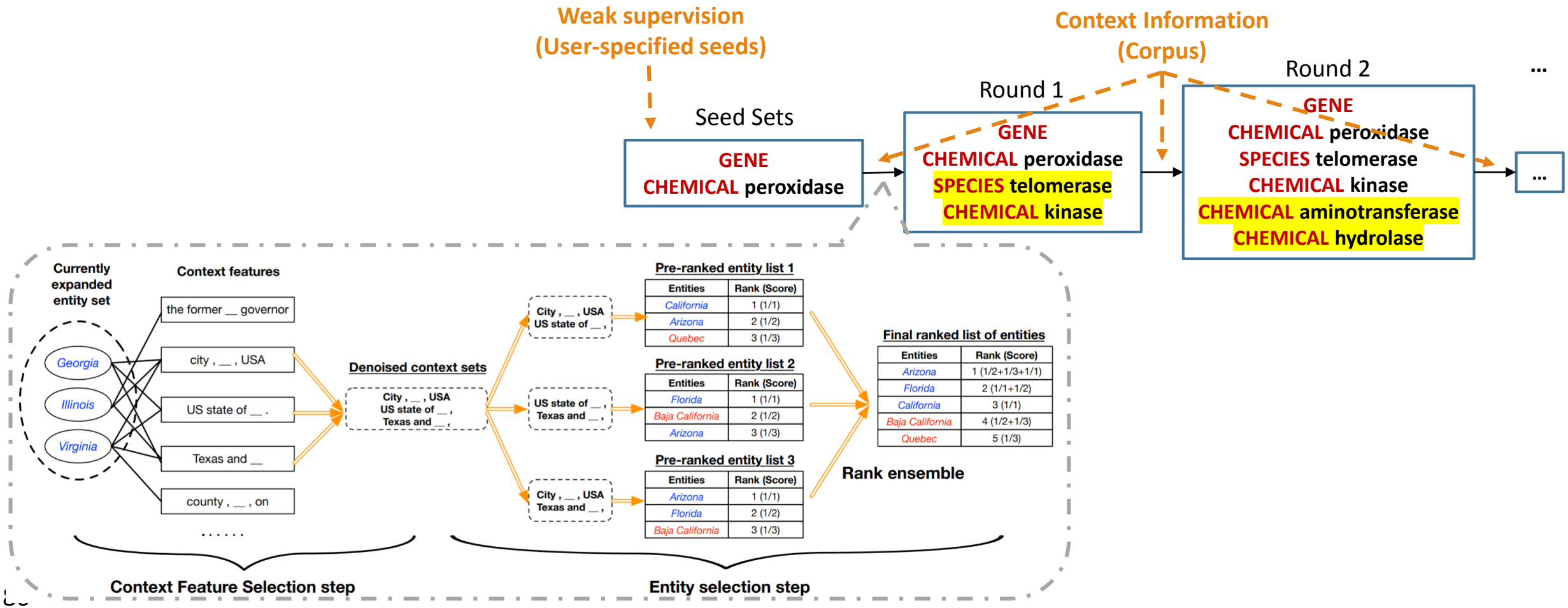
  GENE mRNA release = PROCESS
- Framework
  - Taking a corpus pre-tagged by any flat NER tool as input
  - **Unsupervised** meta-pattern extraction
  - **Few-shot** nested entity recognition for each type
- Evaluation
  - Outperforming baselines in both meta-pattern extraction and nested NER
  - Detecting new entity types with few seeds
  - Improving annotation results over PubTator

Framework Overview



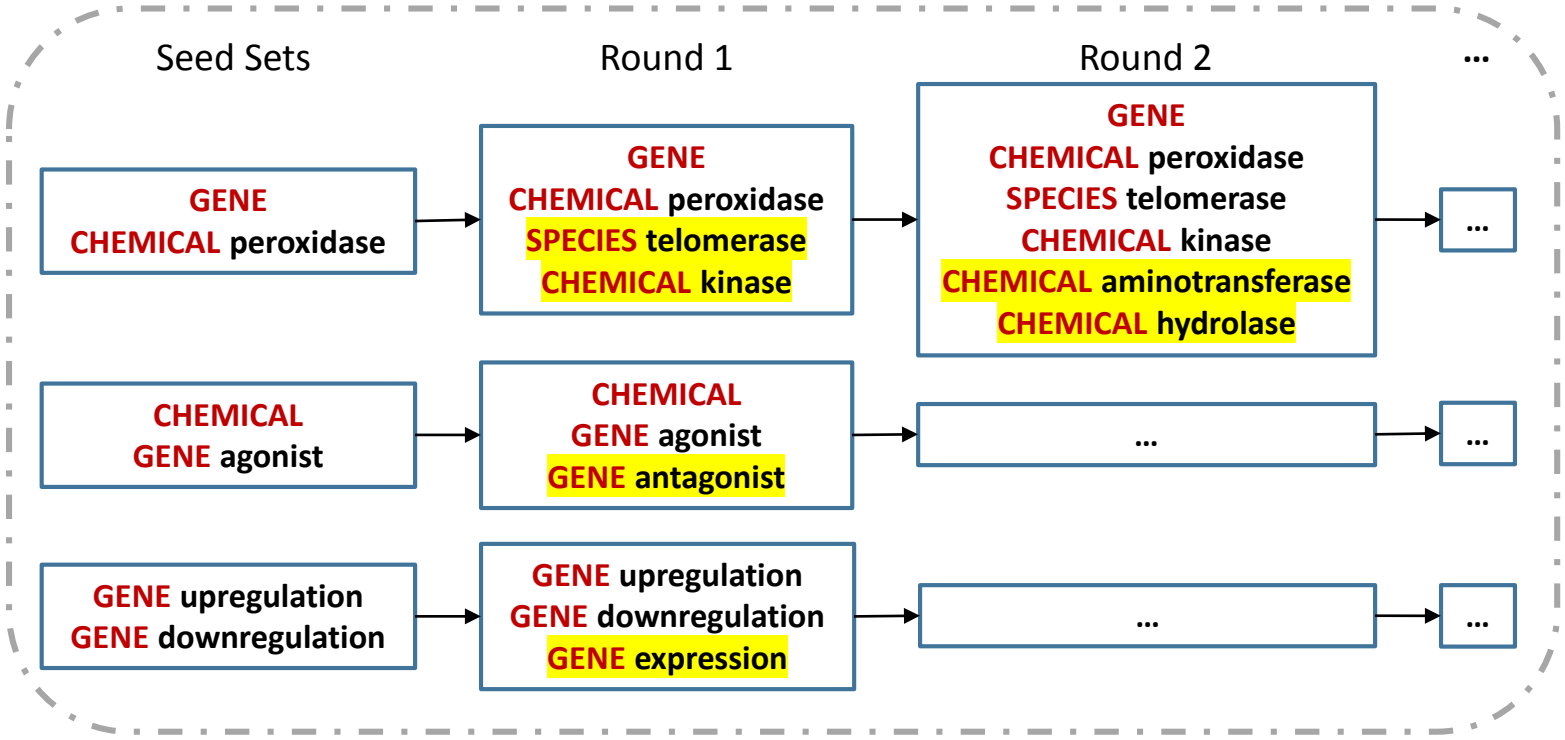
Weakly-supervised Pattern Expansion

- 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

Detecting New Entity Types

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

Seed	{GENE upregulation, GENE downregulation}	{CHEMICAL injection, CHEMICAL inhalation}
1	GENE expression	CHEMICAL treatment
2	GENE phosphorylation	CHEMICAL administration
3	the development of DISEASE	CHEMICAL exposure
4	GENE induction	treatment with CHEMICAL
5	CHEMICAL action	exposure to CHEMICAL
6	identification of GENE	administration of CHEMICAL
7	GENE suppression	pretreatment with CHEMICAL
8	DISEASE reduction	CHEMICAL pretreatment
9	CHEMICAL production	-
10	GENE activity	-

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

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**]$_{\text{CHEMICAL}}$ dismutase (SOD) and aminotransferases like [**alanine**]$_{\text{CHEMICAL}}$ aminotransferase (Ala-AT) and [**aspartate**]$_{\text{CHEMICAL}}$ aminotransferase in different age groups ...
PENNER	The aim of the present study was to determine the effect of HRT on the activities of an antioxidant enzyme [[**superoxide**]$_{\text{CHEMICAL}}$ **dismutase**]$_{\text{GENE}}$ (SOD) and aminotransferases like [[**alanine**]$_{\text{CHEMICAL}}$ **aminotransferase**]$_{\text{GENE}}$ (Ala-AT) and [[**aspartate**]$_{\text{CHEMICAL}}$ **aminotransferase**]$_{\text{GENE}}$ in different age groups ...
	PMID: 10919993
PubTator	Mitogen-activated protein (MAP) kinase [**Erk1/2**]$_{\text{GENE}}$ antagonist mainly inhibited the release of [MCP-1]$_{\text{GENE}}$, whereas MAP kinase [**p38**]$_{\text{GENE}}$ antagonist mainly suppressed the release of [IL-8]$_{\text{GENE}}$ and [RANTES]$_{\text{GENE}}$.
PENNER	Mitogen-activated protein (MAP) kinase [[**Erk1/2**]$_{\text{GENE}}$ **antagonist**]$_{\text{CHEMICAL}}$ mainly inhibited the release of [MCP-1]$_{\text{GENE}}$, whereas MAP kinase [[**p38**]$_{\text{GENE}}$ **antagonist**]$_{\text{CHEMICAL}}$ mainly suppressed the release of [IL-8]$_{\text{GENE}}$ and [RANTES]$_{\text{GENE}}$.
	PMID: 21266192
PubTator	... it suppressed [STAT3]$_{\text{GENE}}$ and [**STAT5**]$_{\text{GENE}}$ phosphorylation in HS-578T cells, whereas it up-regulated [**STAT1**]$_{\text{GENE}}$ phosphorylation and down-regulated [**STAT5**]$_{\text{GENE}}$ phosphorylation in MCF-7 cells.
PENNER	... it suppressed [STAT3]$_{\text{GENE}}$ and [[**STAT5**]$_{\text{GENE}}$ **phosphorylation**]$_{\text{PROCESS}}$ in HS-578T cells, whereas it up-regulated [[**STAT1**]$_{\text{GENE}}$ **phosphorylation**]$_{\text{PROCESS}}$ and down-regulated [[**STAT5**]$_{\text{GENE}}$ **phosphorylation**]$_{\text{PROCESS}}$ in MCF-7 cells.
	PMID: 10498651
PubTator	[**COL1A2**]$_{\text{GENE}}$ expression was decreased by [**vitamin E**]$_{\text{CHEMICAL}}$ treatment or transfection with [**manganese superoxide**]$_{\text{CHEMICAL}}$ dismutase, and was further increased after treatment with [**L-buthionine sulfoximine**]$_{\text{CHEMICAL}}$ ...
PENNER	[[**COL1A2**]$_{\text{GENE}}$ **expression**]$_{\text{PROCESS}}$ was decreased by [[**vitamin E**]$_{\text{CHEMICAL}}$ **treatment**]$_{\text{TREATMENT}}$ or transfection with [[**manganese superoxide**]$_{\text{CHEMICAL}}$ **dismutase**]$_{\text{GENE}}$, and was further increased after [**treatment with** [**L-buthionine sulfoximine**]$_{\text{CHEMICAL}}$]$_{\text{TREATMENT}}$ ...

Automated Information Extraction on PubMed

- PubMed comprises over **29 million** biomedical literature from MEDLINE, life science journals, and online books.
- **BioText Demo Project** is a web-based tool focusing on a more comprehensive annotation of the **entities** and **relations** in PubMed publications.
  - Using the UMLS database as the entity dictionary for distant supervision.
  - Using meta-patterns for open relation extraction.
- Online demo: https://biotext-demo.firebaseapp.com

BioText Demo Project

Important note
Each color only represents a single major type, refer to the list below for more information on sub-types

Labelled Words

Sorted By: None

- Organism
  - Eukaryote
- Fully Formed Anatomical Structure
  - Gene or Genome
- Chemical
- Physiologic Function
  - Organism Function
  - Organ or Tissue Function
- Pathologic Function
  - Disease or Syndrome
  - Experimental Model of Disease

hemodynamic instability 1

Aliskiren-associated acute kidney injury in a patient with pre-existing chronic kidney disease and dilated cardiomyopathy.

PMID: 22127400
Authors: Junji, Yamauchi | Yugo, Shibagaki | Keita, Uehara | Takashi, Yasuda | Kenjiro, Kimura
Journal: *Clinical and experimental nephrology*
Publish Year: 2012

Abstract:

We report a case of acute kidney injury (AKI) caused by a novel direct renin inhibitor, aliskiren. A 43-year-old Japanese man with dilated cardiomyopathy on cardiac resynchronization therapy with defibrillator and chronic kidney disease (CKD) was started on aliskiren in addition to enalapril, carvedilol, furosemide, and spironolactone for worsening cardiac function suggested by the elevation of serum brain natriuretic peptide. After 1 month, he noticed general malaise, loss of appetite and his serum creatinine level increased from 2.0 to 7.24 mg/dL. He had no evidence of exacerbation of hemodynamic instability (heart failure or hypotension) or post-renal cause of AKI. Although a cessation of aliskiren did not ameliorate AKI, renal function returned to baseline after withholding enalapril. Careful monitoring is necessary when aliskiren is used in patients with CKD and/or significant systolic dysfunction since it can cause normotensive ischemic AKI, especially when there is a concomitant use of other renin-angiotensin-aldosterone system inhibitors.

Patterns:

Instances	Meta Pattern	Sentence Extraction
• acute_kidney_injury • aliskiren	DISEASE cause by CHEMICAL	**acute_kidney_injury** AKI cause by {{a novel direct renin inhibitor **aliskiren**}}

Corpus:

- PubMed: 29 million biomedical literature abstracts

Dictionary:

- UMLS: more than 200 biomedical databases

Our demo:

- 5 coarse entity types, 17 fine-grained entity types.
- Meta-pattern discovery for relation extraction.

Demo:

- https://biotext-demo.firebaseapp.com

References

- Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren and Jiawei Han, "Learning Named Entity Tagger using Domain-Specific Dictionary", in Proc. of 2018 Conf. on Empirical Methods in Natural Language Processing (EMNLP'18), Brussels, Belgium, Oct. 2018
- Liyuan Liu, Xiang Ren, Jingbo Shang, Xiaotao Gu, Jian Peng and Jiawei Han, "Efficient Contextualized Representation: Language Model Pruning for Sequence Labeling", in Proc. of 2018 Conf. on Empirical Methods in Natural Language Processing (EMNLP'18), Brussels, Belgium, Oct. 2018
- Liyuan Liu, Jingbo Sahng, Frank Xu, Xiang Ren, Huan Gui, Jian Peng and Jiawei Han, "Empower Sequence Labeling with Task-Aware Neural Language Model", in Proc. of 2018 AAAI Conf. on Artificial Intelligence (AAAI'18), New Orleans, LA, Feb. 2018
- Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. "Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning." arXiv preprint arXiv:1801.09851 (2018).
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.
- Alan Ritter, Oren Etzioni, and Sam Clark. "Open domain event extraction from twitter." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1104-1112. ACM, 2012.
- Alan Ritter, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." In Proceedings of the conference on empirical methods in natural language processing, pp. 1524-1534. Association for Computational Linguistics, 2011.

References (Continued)

- Lev Ratinov and Dan Roth. "Design challenges and misconceptions in named entity recognition." In Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pp. 147-155. Association for Computational Linguistics, 2009.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. "Neural Architectures for Named Entity Recognition." In Proceedings of NAACL-HLT, pp. 260-270. 2016.
- Xuezhe Ma and Eduard Hovy. "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1064-1074. 2016.
- Nils Reimers and Iryna Gurevych. "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging." In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 338-348. 2017.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations." In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), vol. 1, pp. 2227-2237. 2018.
- Alan Akbik, Duncan Blythe, and Roland Vollgraf. "Contextual String Embeddings for Sequence Labeling." In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018). 2018.

References (Continued)

- Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. "SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data." arXiv preprint arXiv:1704.06360 (2017).
- Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, and Jiawei Han. "Clustype: Effective entity recognition and typing by relation phrase-based clustering." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 995-1004. ACM, 2015.
- Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. "Unsupervised Aspect Term Extraction with B-LSTM & CRF using Automatically Labelled Datasets." In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 180-188. 2017.
- Xuan Wang, Yu Zhang, Qi Li, Cathy H. Wu, Jiawei Han, “PENNER: Pattern-enhanced Nested Named Entity Recognition in Biomedical Literature”, BIBM’18

References (Continued)

- E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In ACM DL'00, pages 85–94, 2000.
- L. Bing, S. Chaudhari, R. Wang, and W. Cohen. Improving distant supervision for information extraction using label propagation through lists. In EMNLP'15, pages 524–529, 2015.
- A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In NIPS'13, pages 2787–2795, 2013.
- M. Jiang, J. Shang, T. Cassidy, X. Ren, L. M. Kaplan, T. P. Hanratty, and J. Han. Metapad: Meta pattern discovery from massive text corpora. In KDD'17, pages 877–886, 2017.
- Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In AAAI'15, pages 2181–2187. AAAI, 2015.
- Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun. Neural relation extraction with selective attention over instances. In ACL'16, pages 2124–2133, 2016.
- Y. Liu, F. Wei, S. Li, H. Ji, M. Zhou, and H. Wang. A dependency-based neural network for relation classification. In ACL'15, pages 285–290, 2015.
- M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In ACL-IJCNLP'09, pages 1003–1011, 2009.
- N. Nakashole, G. Weikum, and F. Suchanek. Patty: A taxonomy of relational patterns with semantic types. In EMNLP'12, pages 1135–1145, 2012.

References (Continued)

- X. Ren, Z. Wu, W. He, M. Qu, C. R. Voss, H. Ji, T. F. Abdelzaher, and J. Han. Cotype: Joint extraction of typed entities and relations with knowledge bases. In WWW'17, pages 1015–1024, 2017.
- S. Riedel, L. Yao, A. McCallum, and B. M. Marlin. Relation extraction with matrix factorization and universal schemas. In HLT-NAACL'13, pages 74–84, 2013.
- M. Schmitz, R. Bart, S. Soderland, O. Etzioni, et al. Open language learning for information extraction. In EMNLP-CoNLL'12, pages 523–534, 2012.
- V. Shwartz and I. Dagan. Path-based vs. distributional information in recognizing lexical semantic relations. arXiv preprint arXiv:1608.05014, 2016.
- V. Shwartz, Y. Goldberg, and I. Dagan. Improving hypernymy detection with an integrated path-based and distributional method. In ACL'16, pages 2389–2398, 2016.
- Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, Jiawei Han. Open Information Extraction with Meta-pattern Discovery in Biomedical Literature. In ACM BCB'18, 2018.
- C. Xu, Y. Bai, J. Bian, B. Gao, G. Wang, X. Liu, and T.-Y. Liu. Rc-net: A general framework for incorporating knowledge into word representations. In CIKM'14, pages 1219–1228, 2014.
- Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin. Classifying relations via long short-term memory networks along shortest dependency paths. In EMNLP'15, pages 1785–1794, 2015.
- M. Yahya, S. Whang, R. Gupta, and A. Y. Halevy. Renoun: Fact extraction for nominal attributes. In EMNLP'14, pages 325–335. ACL, 2014.

References (Continued)

- B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. In ICLR'15, 2015.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Mannning. Position-aware Attention and Supervised Data Improve Slot Filling. In EMNLP'17, 2017.
- D. Zeng, K. Liu, Y. Chen, and J. Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In EMNLP'15, pages 1753–1762, 2015.
- D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. Relation classication via convolutional deep neural network. In COLING'14, pages 2335–2344, 2014.
- W. Zeng, Y. Lin, Z. Liu, and M. Sun. Incorporating relation paths in neural relation extraction. In EMNLP'17, pages 1769–1778, 2017.

Summary & Future Directions

IEEE-BigData 2019 Tutorial

Taming Unstructured Big Data: Automated Information Extraction from Massive Text

Xuan Wang, Yu Zhang, Qi Li, Jiawei Han
Computer Science, University of Illinois at Urbana-Champaign

December 11, 2019

Summary: from Unstructrured Text to Knowledge

- Power of Structured Heterogeneous Information Networks
- Mining Structures from Massive Unstructured Text (Texts → Structures)
  - Automated Entity Typing (ClusType → AutoNER)
  - Automated Open Relation Extraction?
  - Automated Structured Analysis (Network/Text-based Summarization & Embedding)
- The Path: Unstructured Text to Network to Knowledge

Future Work: Named Entity Recognition

- Improve the distant supervision
  - Can we do better than string match?
  - Can we integrate the phrase mining with NER and let them mutually enhance each other?
- Involve human experts in the loop
  - Given a fixed amount of expert hours, how to build the most reliable NER system?

Future Work: Relation Extraction

- Better utilize unsupervised text corpora
  - Retrieve unlabeled sentences that are similar to labeled sentences for training
  - Fine-tune contextualized language models to generate training samples

- Incorporate embedding techniques into open-domain information extraction
  - Use embeddings to encode patterns and refine extracted tuples

Constructed Heterogeneous Information Networks

Massive Text Data

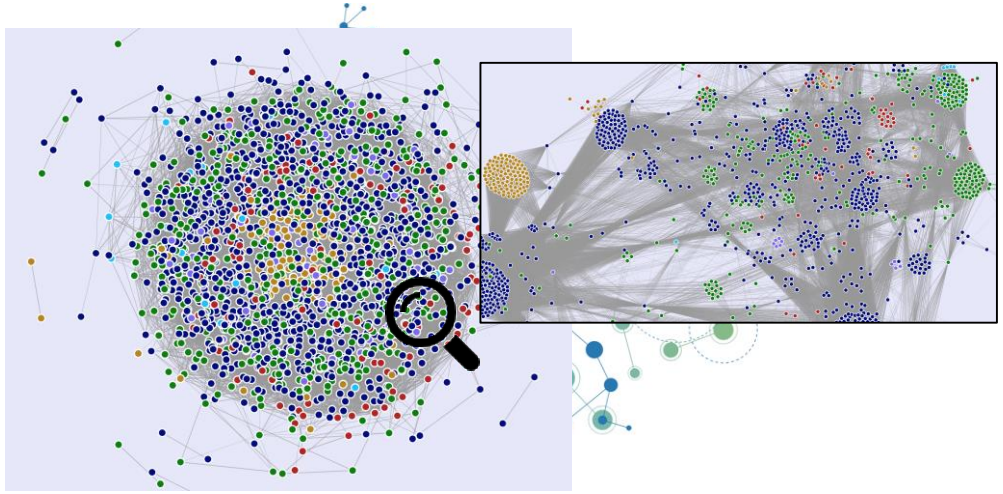
Public Knowledge Bases

Phrase Mining

Entity Recognition

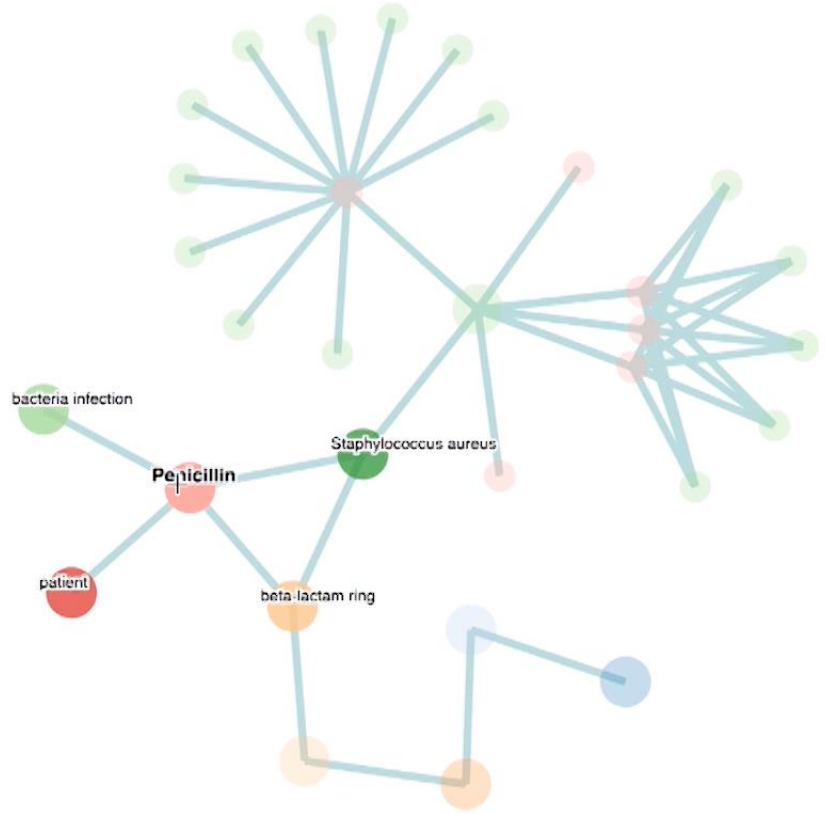
Relation Extraction

- The network constructed from ***massive*** texts is usually ***gigantic***
- How to better explore it?



Heterogeneous Information Network

AutoNet Demo (KDD'18): Basic Structured Analysis



AutoNet System ?

phlebitis

TYPE: Disease

RELATED DOCUMENTS:

Phase II study of the amsacrine analogue CI-921 (...
Remission induction of meningeal leukemia with hig...

RELATED TO:

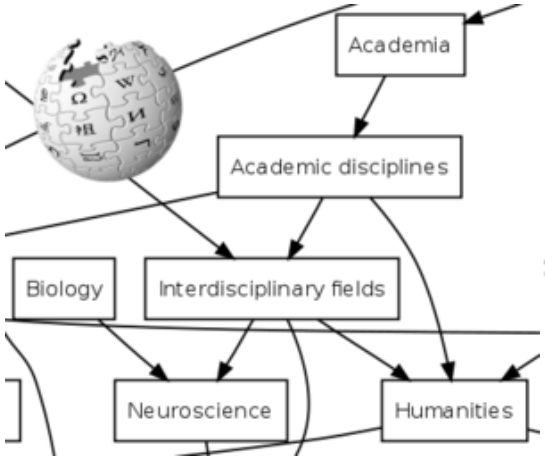
Keyword1;Keyword2 or URL

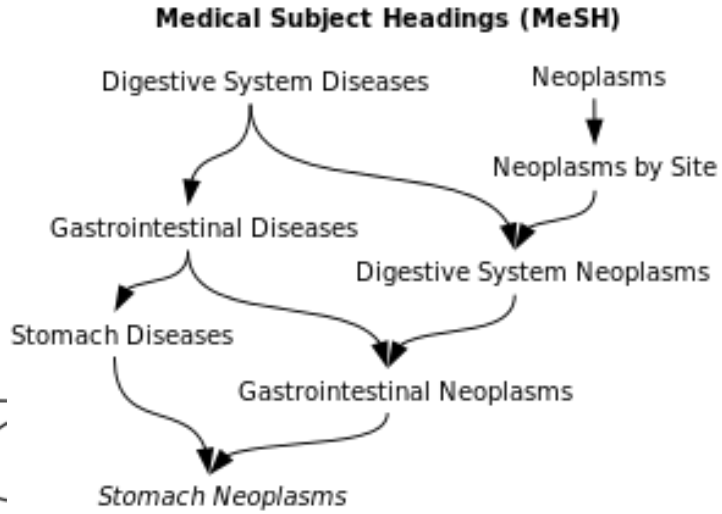
Search | Upload | Expand

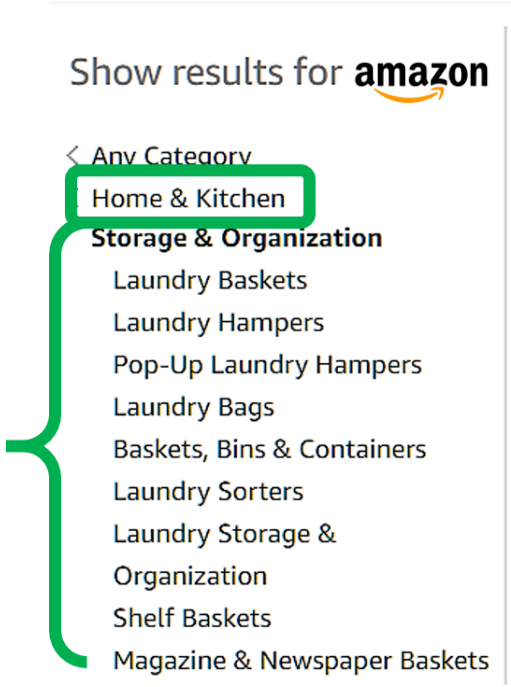
- Retrieves a subnetwork through keyphrase search
- Links back to PubMed articles

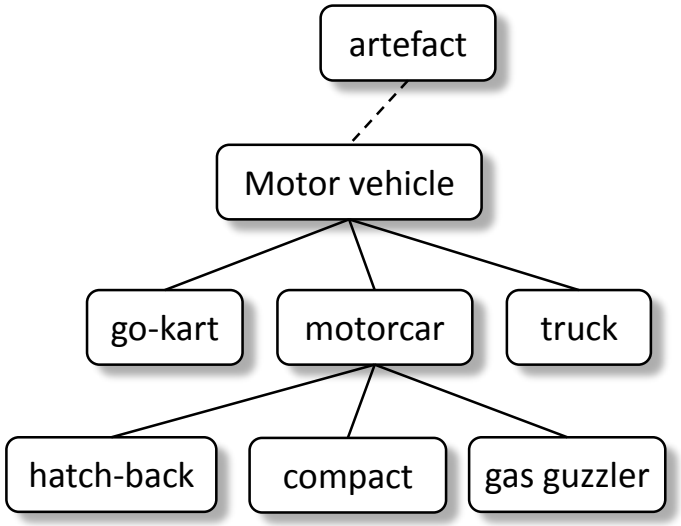
Taxonomy Construction

- Taxonomy is a hierarchical organization of concepts
  - For example: Wikipedia category, ACM CCS Classification System, Medical Subject Heading (MeSH), Amazon Product Category, Yelp Category List, WordNet, and etc.





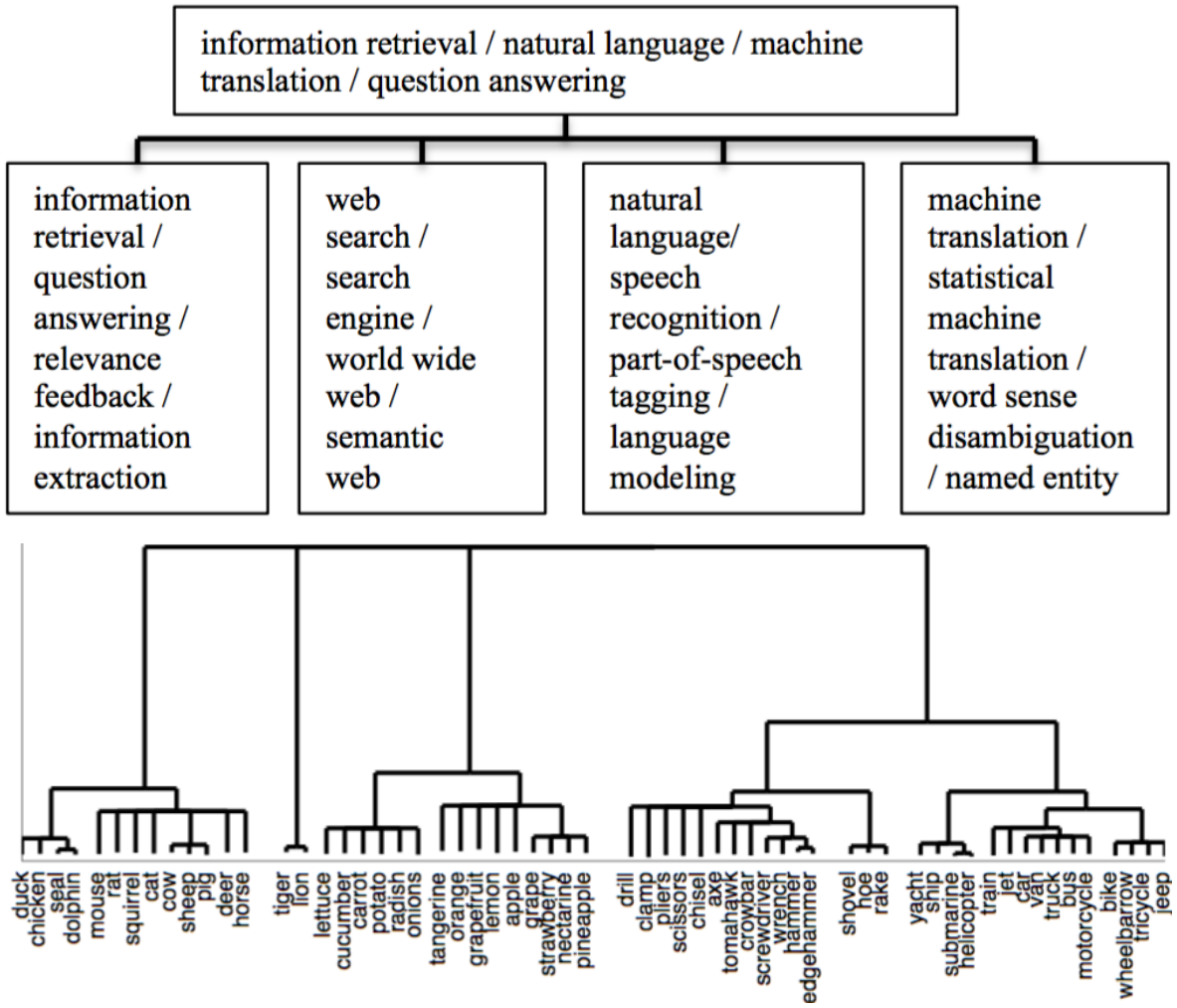




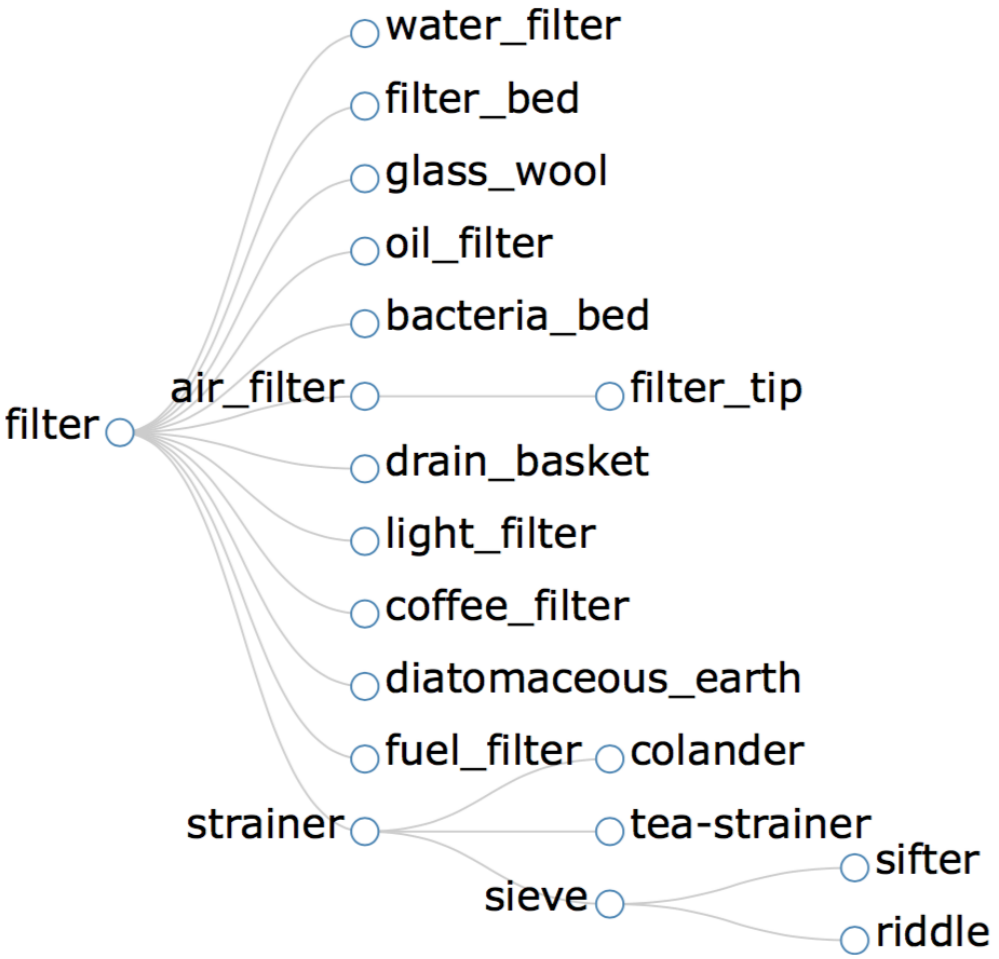
Wikipedia Category | MeSH | Amazon Product Category | WordNet

Two types of Taxonomy

- Clustering-based Taxonomy

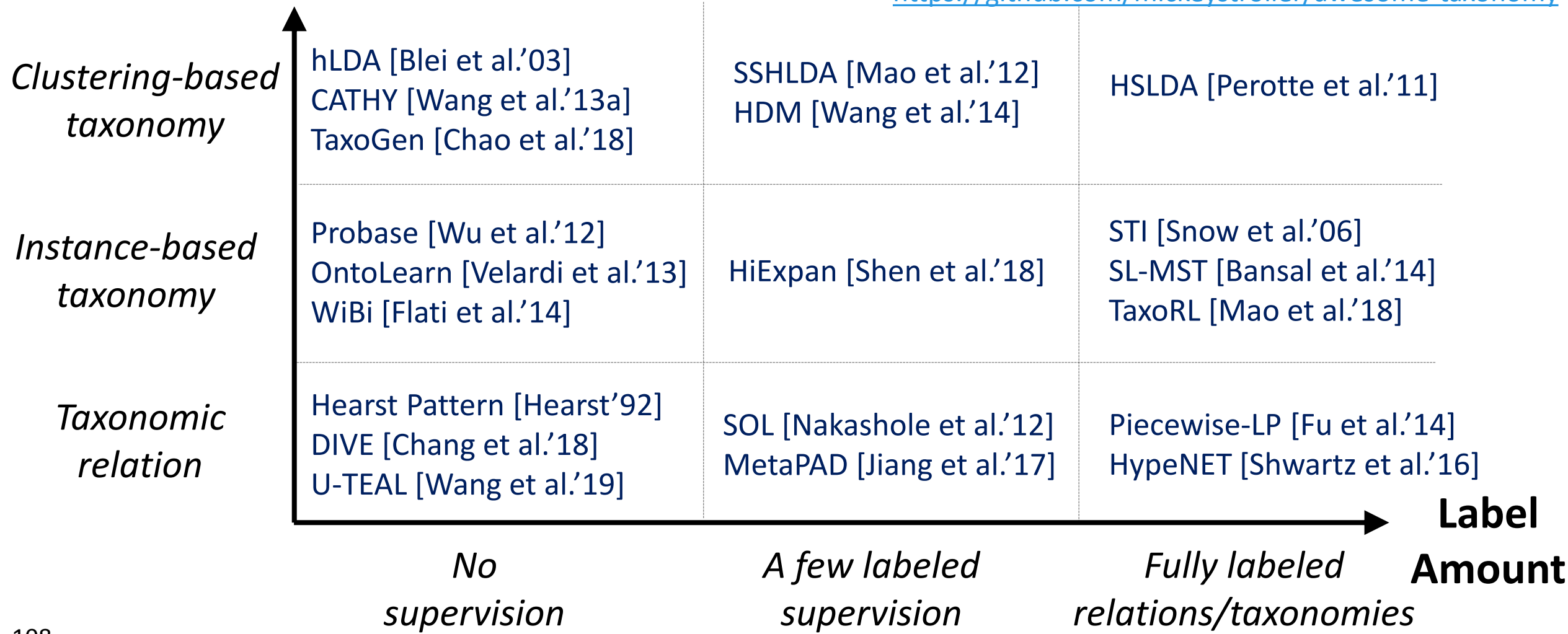


- Instance-based Taxonomy



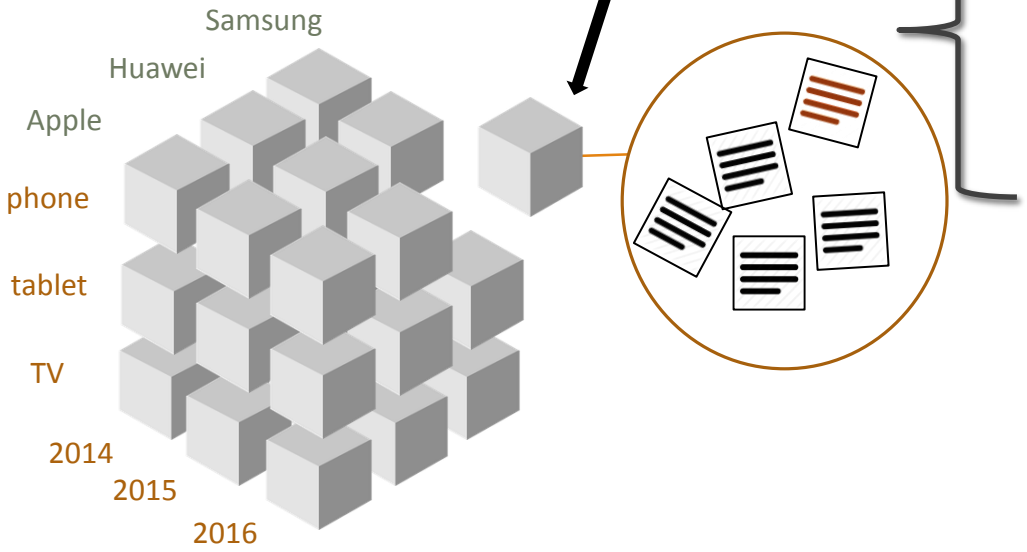
Taxonomy Construction Methods: A Landscape

For a more comprehensive landscape see:
https://github.com/mickeystroller/awesome-taxonomy



Multi-Dimensional Text Cube Construction

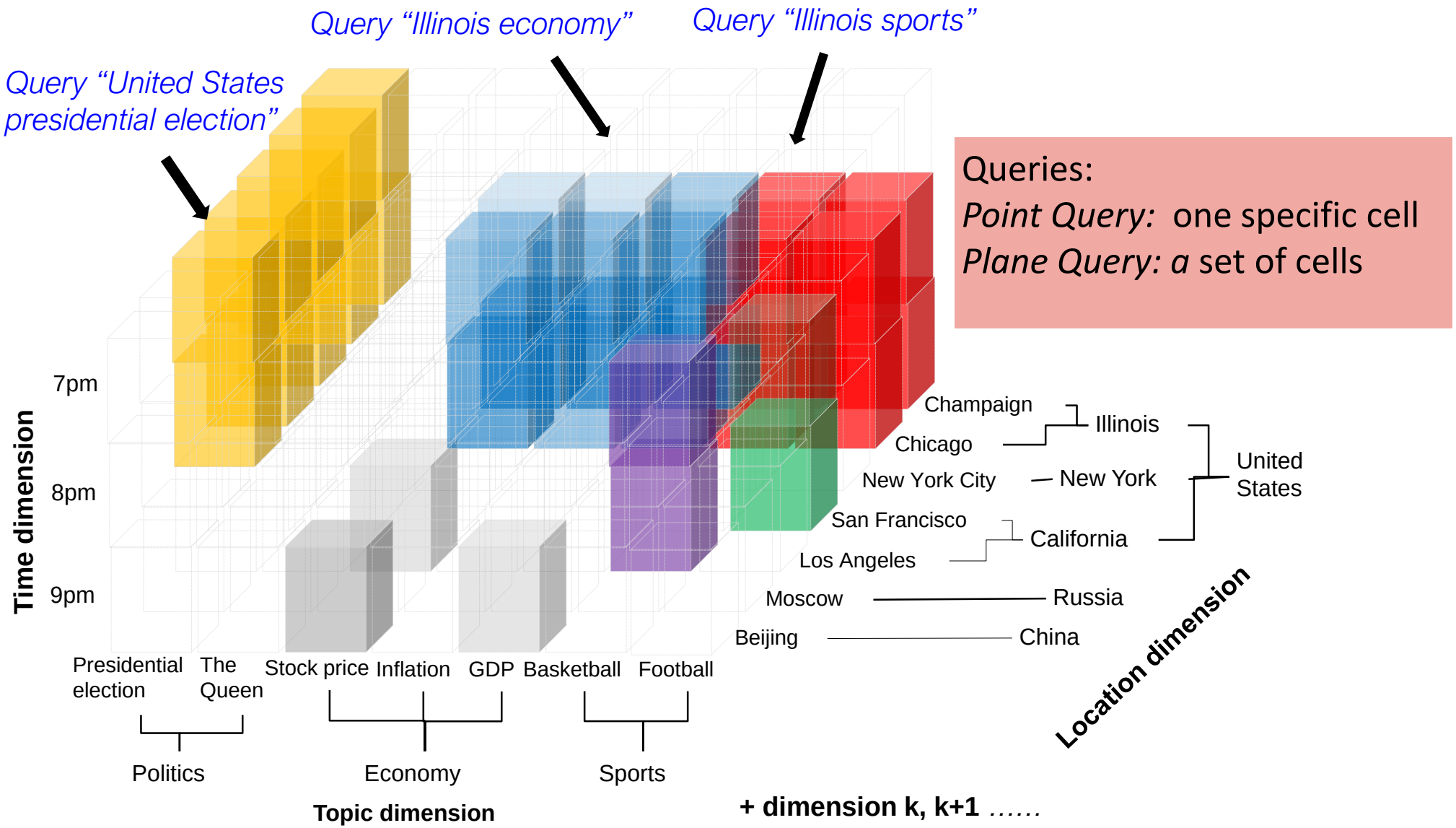
- Numerical data cube (each cell is a numerical value) has been extensively studied
  - Measures: Numerical aggregations as *sum* & *avg*.
- Text cube: Each cell contains a set of documents (e.g., <Apple, TV, 2016>)
  - There is an imminent need to do OLAP analysis on text cubes



Dimensions:
Brand: Apple, Samsung, Huawei…
Product: phone, tablet, TV, laptop…
Field: IR, Machine Learning, NLP…

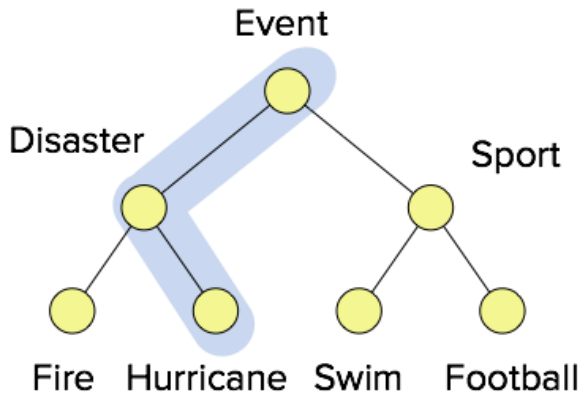
Text:
Aviation Safety Reports from NASA
Product Reviews from Amazon
Research Papers from DBLP

Multi-dimensional Text Cube with Queries & Hierarchies

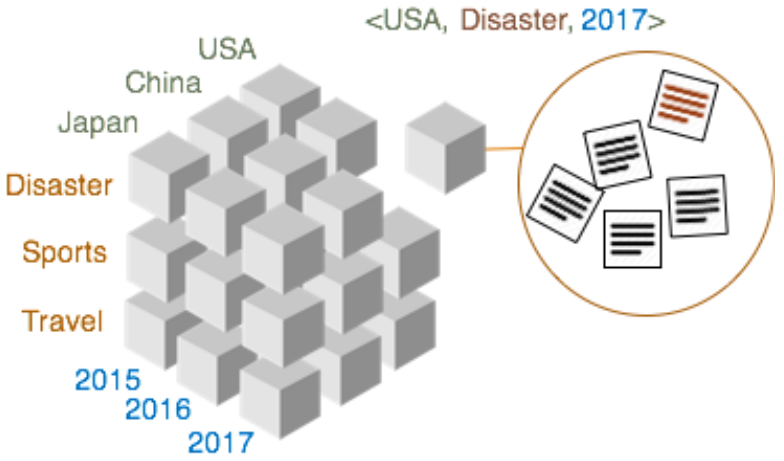


Text Cube Construction: Two Central Tasks

- **1. Taxonomy Construction**
  - How to discover the taxonomy for each dimension?

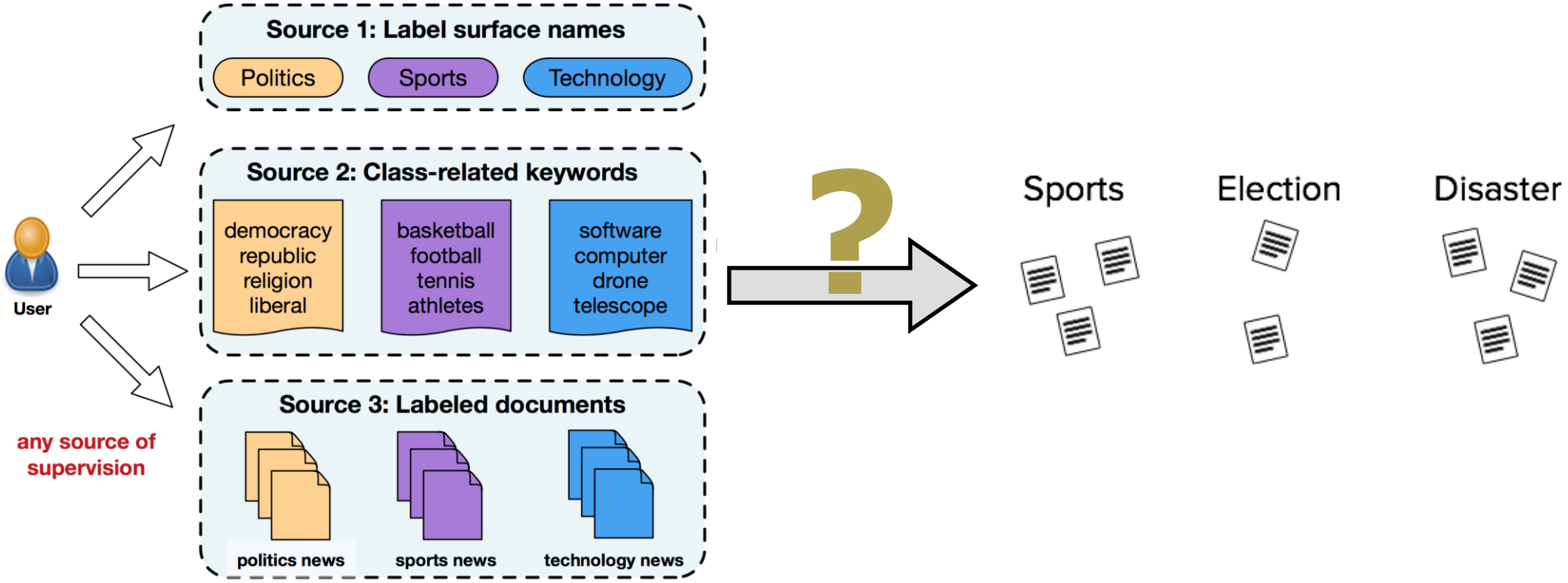


- **2. Document Allocation**
  - How to allocate documents into the cube?



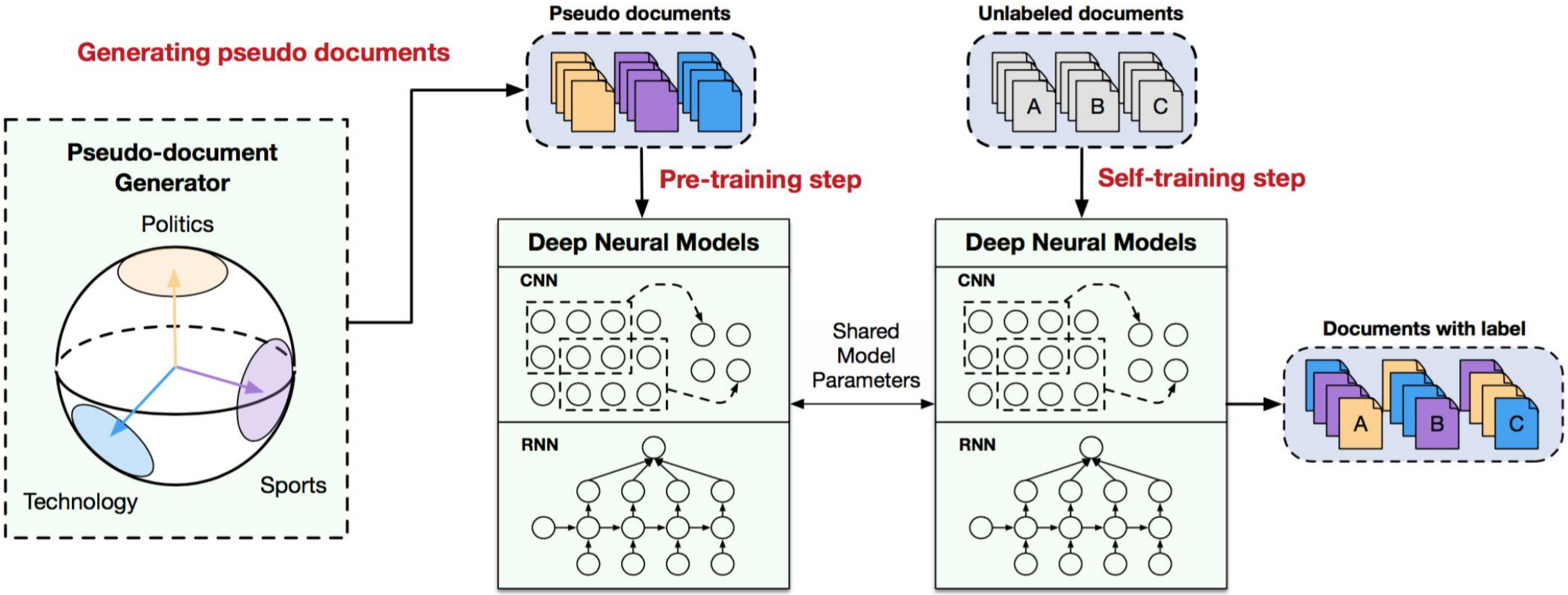
Weakly-Supervised Text Classification

- WeSTClass, WeSHClass [Meng et al., 2018, 2019], HiGitClass [Zhang et al., 2019]
- Require no training data, but a small amount of seed information
  - (1) label names, or (2) relevant keywords, or (3) a few labeled docs



Pseudo Training Data + Self-Training

- Pseudo document generation: generate pseudo documents from seeds.
- Self-training: train deep neural nets (CNN, RNN) with bootstrapping.



References: Heterogeneous Network Construction

- Shang, J., Zhu, Q., Shen, J., Wang, X., Gu, X., Kaplan, L., ... & Han, J. (2018). AutoNet: Automated Network Construction and Exploration System from Domain-Specific Corpora.
- Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. (2016). A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, *29*(1), 17-37.
- Yizhou Sun, Yintao Yu, and Jiawei Han, "Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema", KDD'09
- Liangliang Cao, Andrey Del Pozo, Xin Jin, Jiebo Luo, Jiawei Han, and Thomas S. Huang, "RankCompete: Simultaneous Ranking and Clustering of Web Photos", WWW'10
- M. Ji, J. Han, and M. Danilevsky, "Ranking-based Classification of Heterogeneous Information Networks", KDD'11.
- Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. Proceedings of the VLDB Endowment, 4(11), 992-1003.
- Yizhou Sun, Jiawei Han, Charu C. Aggarwal, and Nitesh V. Chawla, "When Will It Happen? --- Relationship Prediction in Heterogeneous Information Networks", WSDM'12.
- Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han, "Citation Prediction in Heterogeneous Bibliographic Networks", SDM'12.
- Shang, J., Qu, M., Liu, J., Kaplan, L. M., Han, J., & Peng, J. (2016). Meta-path guided embedding for similarity search in large-scale heterogeneous information networks. arXiv preprint arXiv:1610.09769.
- Chen, T., & Sun, Y. "Task-guided and path-augmented heterogeneous network embedding for author identification", WSDM'17.
- Dong, Y., Chawla, N. V., & Swami, A. "metapath2vec: Scalable representation learning for heterogeneous networks", KDD'17.
- Hosseini, A., Chen, T., Wu, W., Sun, Y., & Sarrafzadeh, M. "Heteromed: Heterogeneous information network for medical diagnosis", CIKM'18.
- Shi, Y., Gui, H., Zhu, Q., Kaplan, L., & Han, J. "Aspem: Embedding learning by aspects in heterogeneous information networks", SDM'18.
- Shi, Y., Zhu, Q., Guo, F., Zhang, C., & Han, J. "Easing embedding learning by comprehensive transcription of heterogeneous information networks", KDD'18.

References: Taxonomy Construction

- David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In NIPS.
- David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. Journal of ACM (2010).
- Wei Li and Andrew D McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In ICML.
- D. M. Mimno, W. Li, and A. McCallum. Mixtures of hierarchical topics with pachinko allocation. In ICML, pages 633–640, 2007
- Li, Wei, David M. Blei and Andrew McCallum. "Nonparametric Bayes Pachinko Allocation." *UAI* (2007).
- David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. ICML (2009).
- Su-Jin Shin and Il-Chul Moon. 2017. Guided HTM: Hierarchical Topic Model with Dirichlet Forest Priors. TKDE (2017).
- Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, rivikrama Taula, and Jiawei Han. 2013. A phrase mining framework for recursive construction of a topical hierarchy. In KDD.
- Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, and Jiawei Han. 2013. Constructing topical hierarchies in heterogeneous information networks. ICDM (2013).
- Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. Automatic taxonomy construction from keywords. In KDD.
- Yangqiu Song, Shixia Liu, Xueqing Liu, and Haixun Wang. 2015. Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees. TKDE (2015).
- Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle Vanni and Jiawei Han. "TaxoGen : Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering." (2018).

References: Cube Construction

- D.M.Blei, T.L.Griffiths, M.I.Jordan, and J.B.Tenenbaum. “Hierarchical topic models and the nested Chinese restaurant process”. In NIPS, pages 17–24, 2003.
- F. Tao, H. Zhuang, C. Wang, Q. Wang, T. Cassidy, L. Kaplan, C. Voss, J. Han, "Multi-Dimensional, Phrase-Based Summarization in Text Cubes", Data Eng. Bulletin 39(3):74-84, 2016
- Y. Meng, J. Shen, C. Zhang and J. Han, “Weakly-Supervised Neural Text Classification”, in CIKM, 2018.
- Y. Meng, J. Shen, C. Zhang and J. Han, “Weakly-Supervised Hierarchical Text Classification”, in AAAI, 2019
- Y. Zhang, F. F. Xu, S. Li, Y. Meng, X. Wang, Q. Li, J. Han, “HiGitClass: Keyword-Driven Hierarchical Classification of GitHub Repositories”, in ICDM, 2019.
- D.M.Mimno, W.Li, and A.McCallum. Mixtures of hierarchical topics with pachinko allocation. In ICML, pages 633–640, 2007.
- W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, J. Huang, "STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration over the Twitter Stream", in ICDE, 2015.
- F. Tao, C. Zhang, X. Chen, M. Jiang, T. Hanratty, L. Kaplan, and J. Han. “Doc2cube: Automated document allocation to text cube via dimension-aware joint embedding”. In review, 2018.
- X. Chen, Y. Xia, P. Jin, and J. A. Carroll. “Dataless text classification with descriptive LDA”. In AAAI, pages 2224–2231, 2015.
- F. Ravat, O. Teste, R. Tournier, and G. Zurfluh. “Top_keyword: An aggregation function for textual document OLAP”. In Data Warehousing and Knowledge Discovery, pages 55–64, 2008.
- F. Tao, K. H. Lei, J. Han, C. Zhai, X. Cheng, M. Danilevsky, N. Desai, B. Ding, J. Ge, H. Ji, R. Kanade, A. Kao, Q. Li, Y. Li, C. X. Lin, J. Liu, N. C. Oza, A. N. Srivastava, R. Tjoelker, C. Wang, D. Zhang, and B. Zhao. Eventcube: multi-dimensional search and mining of structured and text data. In KDD, pages 1494–1497, 2013

Our Journey: From Big Data to Big Structures & Knowledge

Han, Kamber and Pei, Data Mining, 3rd ed. 2011

Yu, Han and Faloutsos (eds.), Link Mining, 2010

Sun and Han, Mining Heterogeneous Information Networks, 2012

Y. Sun: SIGKDD'13 Dissertation Award

Wang and Han, Mining Latent Entity Structures, 2015

C. Wang: SIGKDD'15 Dissertation Award

Our Journey: From Big Data to Big Structures & Knowledge

Han, Kamber and Pei, Data Mining, 3rd ed. 2011

Yu, Han and Faloutsos (eds.), Link Mining, 2010

Sun and Han, Mining Heterogeneous Information Networks, 2012

Y. Sun: SIGKDD'13 Dissertation Award

ng and Han, Mining Latent Entity Structures, 2015

Wang: SIGKDD'15 Dissertation Award

Our Journey: From Big Data to Big Structures & Knowledge

Han, Kamber and Pei, Data Mining, 3rd ed. 2011

Yu, Han ... Link ...

... Sun and Han, Mining Heterogeneous ... Information Networks, 2012

Y. Sun: SIGKDD'13 Dissertation Award

... g and Han, Mining Latent Entity Structures, 2015

... Wang: SIGKDD'15 Dissertation Award

Our Journey: From Big Data to Big Structures & Knowledge

Han, Kamber and Pei, Data Mining, 3rd ed. 2011

Yu, Han ... Link ...

Sun and Han, Mining Heterogeneous Information Networks, 2012

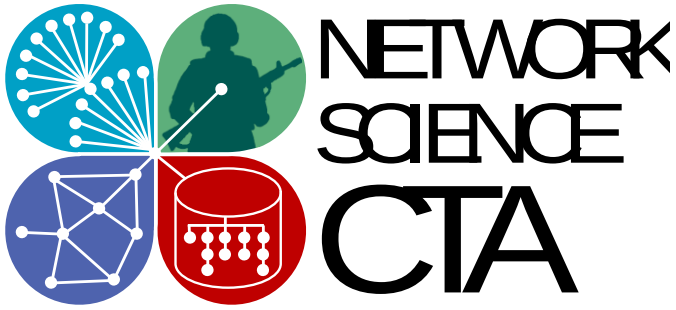
Y. Sun: SIGKDD'13 Dissertation Award

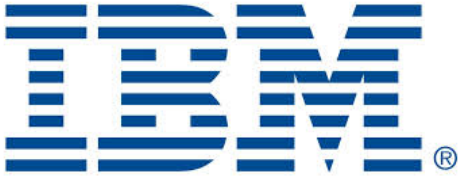
... and Han, Mining Latent Entity Structures, 2015

... Wang: SIGKDD'15 Dissertation Award

Acknowledgements

- Thanks for the research support from: ARL/NSCTA, NIH, NSF, DHS, ARO, DTRA





Thank you !
Q&A