Scalable Construction and Reasoning of Massive Knowledge Bases

Xiang Ren¹  Nanyun Peng¹  William Yang Wang²

University of Southern California¹
University of California, Santa Barbara²
Turning Unstructured Text Data into Structures

Unstructured Text Data
(account for ~80% of all data in organizations)

Knowledge & Insights

(Chakraborty, 2016)
This **hotel** is my favorite **Hilton property** in **NYC**! It is located right on 42nd street near **Times Square**, it is close to all subways, **Broadways shows**, and next to great restaurants like **Junior’s Cheesecake**, **Virgil’s BBQ** and many others.

--- *TripAdvisor*

**Structured Facts**

1. “**Typed**” entities
2. “**Typed**” relationships
Why Text to Structures?

Structured Search & Exploration

SQL

City contains "ist" ×
Category equals "Friends" ×
Birthday on 09/04/2000 × Age = 30 ×
Lastname equals "plugins" ×
Is active ▼ Yes □ No □

SparQL

Dialog Systems

Question Answering

Scientific Inference
A Product Use Case: Finding “Interesting Hotel Collections”

Technology Transfer to TripAdvisor

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

Grouping hotels based on structured facts extracted from the review text

A Scientific Use Case: Precision Medicine

Molecular tumor board

Problem: Hard to scale

U.S. 2016: 1.7 million new cases, 600K deaths

902 cancer hospitals

Memorial Sloan Kettering

Sequence: Tens of thousand
Board can review: A few hundred

Better Structured Search with Reasoning Capabilities

who was the president of usa when churchill died

About 16,400,000 results (0.68 seconds)

United States of America / President (1965)

Lyndon B. Johnson
Text to Structures: Applications

Technology Transfer
- ARL
- Microsoft
- NIH
- Mayo Clinic
- bing
- Yelp
- TripAdvisor

Intelligent Personal Assistant
- Cortana
- Siri
- Amazon Echo
- Google Now
- Facebook

Online Education
- Coursera
- Udacity
- edX
- Canvas
- NovoED
- EdX Network
- Open2Study
- FutureLearn

Medical records
Scientific papers
Clinical reports
...

Healthcare

Social media posts
Web blogs
News articles
...

Computational Social Sciences

Corporate reports
News streams
Customer reviews
...

Business Intelligence
This hotel is my favorite Hilton property in NYC! It is located right on 42nd street near Times Square, it is close to all subways, Broadways shows, and next to many great …
Effort-Light Structure Extraction

- Enables *quick* development of applications over various corpora
- Extracts *complex* structures without introducing human error
Effort–Light Structure Extraction: Where Are We?

Human labeling effort

- Supervised learning methods
  - Stanford CoreNLP, 2005 - present
  - UT Austin Dependency Kernel, 2005
  - IBM Watson Language APIs

- Weakly-supervised learning methods
  - CMU NELL, 2009 - present
  - UW KnowItAll, Open IE, 2005 - present
  - Max-Planck YAGO, 2008 - present

- Distantly-supervised learning methods
  - Stanford: Snorkel, MIML-RE 2012 - present

Hand-crafted methods

- UCB Hearst Pattern, 1992
- NYU Proteus, 1997

Effort–Light Structure Extraction (EMNLP’16, 17, ACL’18, KDD’15, 16, 17, WWW’17, 18…)

Feature engineering effort
“Distant” Supervision: What Is It?

“Matchable” structures: entity names, entity types, typed relationships ...

Freely available!
- Common knowledge
- Life sciences
- Art …

Number of Wikipedia articles

Rapidly growing!

Text corpus

Knowledge Bases

“Un-matchable”

(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), …

Learning with Distant Supervision: Challenges

1. Sparsity of “Matchable”
   - Incomplete knowledge bases
   - Low-confidence matching

2. Accuracy of “Expansion”
   - For “matchable”: Are all the labels assigned accurately?
   - For “un-matchable”: How to perform inference accurately?

(Ren et al., KDD’15)

It is my favorite city in the United States

The United States needs a new strategy to meet this challenge
Effort-Light StructMine: Methodology

- Text corpus
- Data-driven text segmentation (SIGMOD’15, WWW’16)
- Entity names & context units
- Structures from the remaining unlabeled data
- Learning Corpus-specific Model (KDD’15, 16, EMNLP’16, WWW’17)
- Partially-labeled corpus
- Knowledge bases
Structured Knowledge

<table>
<thead>
<tr>
<th>Entity</th>
<th>Entity</th>
<th>Entity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>T790M</td>
<td>EGFR</td>
<td>gefitinib</td>
<td>Resist</td>
</tr>
<tr>
<td>Obama</td>
<td>U.S.</td>
<td></td>
<td>President_of</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Challenges of Obtaining Training Data

• Constructing data sets is labor intensive
• Many different
  • Languages
  • Domains
  • Modalities
  • …
Joint representation learning

• Learning comprehensive representations from *heterogeneous sources.*
  • *unlabeled data*
  • annotations for *related tasks, domains, languages.*

• Encoding structured knowledge to learn robust representations and make *holistic decisions.*
  • *linguistic structures*

T790M is present as a minor clone in NSCLC.

present
Low-resource IE: Another Way to Reduce Human Effort

A Review of Previous Efforts

- **Human labeling effort**
  - Supervised learning systems
    - Stanford CoreNLP, 2005 - present
    - UT Austin Dependency Kernel, 2005
    - IBM Watson Language APIs
  - Weakly-supervised learning systems
    - CMU NELL, 2009 - present
    - UW KnowItAll, Open IE, 2005 - present
    - Max-Planck YAGO, 2008 - present
  - Heterogeneous learning signals
    - CMU, Yang et. al. 2017
    - RPI, Lin et. al. 2017
  - Joint Representation Learning
    - (EMNLP’15, ACL’16, TACL’17, ACL-Repl4NL’17, IJCNLP’17, …)

- **Hand-crafted Systems**
  - UCB Hearst Pattern, 1992
  - NYU Proteus, 1997

- **Feature engineering effort**
Knowledge Bases are Highly Incomplete

Query Start Node: “United States”  Query End Node: “English”
Query: ?(United States, English)
Knowledge Base Reasoning

• Question: can we infer missing links based on background KB?

• Path-based methods
  • Path-Ranking Algorithm (PRA), Lao et al. 2011
  • RNN + PRA, Neelakantan et al, 2015
  • Chains of Reasoning, Das et al, 2017

• Embedding-based methods
  • RESCAL, Nickel et al., 2011
  • TransE, Bordes et al, 2013
  • TransR/CTransR, Lin et al, 2015

• Integrating Path and Embedding-Based Methods
  • DeepPath, Xiong et al, 2017
  • MINERVA, Das et al, 2018
  • DIVA, Chen et al., 2018
Tutorial Outline

• Introduction

• Part I: Effort–Light Structure Extraction
  • Tea break at 10:00am

• Part II: Low-resource IE

• Part III: Knowledge Base Reasoning

• Summary & Future Directions
Scalable Construction and Reasoning of Massive Knowledge Bases

Part I: Effort-Light Structure Extraction
Framework Overview

Text corpus

Data-driven text segmentation (SIGMOD’15, WWW’16)

Entity names & context units

Learning Corpus-specific Model (KDD’16, EMNLP’16, 17, WWW’17)

Partially-labeled corpus

Knowledge bases

Entity Recognition and Coarse-grained Typing (KDD’15)

Fine-grained Entity Typing (KDD’16, EMNLP’16)

Joint Entity and Relation Extraction (WWW’17)

Corpus to Structured Network: The Roadmap
Corpus to Structured Network: The Roadmap

Text corpus

- Structures from the remaining *unlabeled* data
- Data-driven text segmentation (SIGMOD’15, WWW’16)
- Learning Corpus-specific Model (KDD’16, EMNLP’16, 17, WWW’17)

- Partially-labeled corpus
- Knowledge bases

Entity Recognition and Coarse-grained Typing (KDD’15)

- Fine-grained Entity Typing (KDD’16, EMNLP’16)

- Joint Entity and Relation Extraction (WWW’17)
Recognizing Entities of Target Types in Text

The best BBQ I’ve tasted in Phoenix! I had the pulled pork sandwich with coleslaw and baked beans for lunch. The owner is very nice. …

The best BBQ I’ve tasted in Phoenix! I had the pulled pork sandwich with coleslaw and baked beans for lunch. The owner is very nice. …
Traditional Named Entity Recognition (NER) Systems

- Heavy reliance on corpus-specific human labeling
- Training sequence models is slow

A manual annotation interface

NER Systems:
- Stanford NER
- Illinois Name Tagger
- IBM Alchemy APIs

...
Weak-Supervision Systems: Pattern-Based Bootstrapping

- Requires manual seed selection & mid-point checking

Seeds for Food
- Pizza
- French Fries
- Hot Dog
- Pancake
  ...

Annotate corpus using entities

Apply patterns to find new entities

Select Top patterns

Generate candidate patterns

Score candidate patterns

Patterns for Food
- the best <X> I’ve tried in their <X> tastes amazing
  ...

Systems:
- CMU NELL
- UW KnowItAll
- Stanford DeepDive
- Max-Planck
- PROSPERA
  ...

e.g., (Etzioni et al., 2005), (Talukdar et al., 2010), (Gupta et al., 2014), (Mitchell et al., 2015), ...

Leveraging Distant Supervision

1. **Detect** entity names from text
2. **Match** name strings to KB entities
3. **Propagate** types to the un-matchable names

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Phoenix</em> is my all-time favorite dive bar in <em>New York City</em>.</td>
</tr>
<tr>
<td>S2</td>
<td>The best <em>BBQ</em> I’ve tasted in <em>Phoenix</em>.</td>
</tr>
<tr>
<td>S3</td>
<td><em>Phoenix</em> has become one of my favorite bars in <em>NY</em>.</td>
</tr>
</tbody>
</table>

(Lin et al., 2012), (Ling et al., 2012), (Nakashole et al., 2013)
Current Distant Supervision: Limitation

1. Context-agnostic type prediction
   • Predict types for each mention regardless of context

2. Sparsity of contextual bridges

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Phoenix</em> is my all-time favorite dive bar in <em>New York City</em>.</td>
</tr>
<tr>
<td>S2</td>
<td>The best <em>BBQ</em> I’ve tasted in <em>Phoenix</em>.</td>
</tr>
<tr>
<td>S3</td>
<td><em>Phoenix</em> has become one of my favorite bars in <em>NY</em>.</td>
</tr>
</tbody>
</table>
Current Distant Supervision: Limitation

1. Context-agnostic type prediction

2. Sparsity of contextual bridges
   - Some relational phrases are infrequent in the corpus → ineffective type propagation

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Phoenix</em> is my all-time favorite dive bar in <em>New York City</em>.</td>
</tr>
<tr>
<td>S3</td>
<td><em>Phoenix</em> has become one of my favorite bars in <em>NY</em>.</td>
</tr>
</tbody>
</table>
The ClusType Approach (KDD’15)

<table>
<thead>
<tr>
<th>ID</th>
<th>Segmented Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Phoenix is my all-time favorite dive bar in New York City.</td>
</tr>
<tr>
<td>S2</td>
<td>The best BBQ I’ve tasted in Phoenix.</td>
</tr>
<tr>
<td>S3</td>
<td>Phoenix has become one of my favorite bars in NY.</td>
</tr>
</tbody>
</table>

Putting two sub-tasks together:
1. Type label propagation
2. Relation phrase clustering

Correlated mentions

Similar relation phrases
Type Propagation in ClusType

Smoothness Assumption
If two objects are similar according to the graph, then their type labels should be also similar.

Edge weight / object similarity
\[ f^T L f = \sum_{i,j} W_{ij} (f_i - f_j)^2 \]

(Belkin & Partha, NIPS’01), (Ren et al., KDD’15)
Relation Phrase Clustering in ClusType

• Two relation phrases should be grouped together if:
  1. Similar string
  2. Similar context
  3. Similar types for entity arguments

“Multi-view” clustering

Two subtasks mutually enhance each other

(Ren et al., KDD’15)
ClusType: Comparing with State-of-the-Art Systems (F1 Score)

<table>
<thead>
<tr>
<th>Methods</th>
<th>NYT</th>
<th>Yelp</th>
<th>Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern (Stanford, CONLL’14)</td>
<td>0.301</td>
<td>0.199</td>
<td>0.223</td>
</tr>
<tr>
<td>SemTagger (U Utah, ACL’10)</td>
<td>0.407</td>
<td>0.296</td>
<td>0.236</td>
</tr>
<tr>
<td>NNPLB (UW, EMNLP’12)</td>
<td>0.637</td>
<td>0.511</td>
<td>0.246</td>
</tr>
<tr>
<td>APOLLO (THU, CIKM’12)</td>
<td>0.795</td>
<td>0.283</td>
<td>0.188</td>
</tr>
<tr>
<td>FIGER (UW, AAAI’12)</td>
<td>0.881</td>
<td>0.198</td>
<td>0.308</td>
</tr>
<tr>
<td>ClusType (KDD’15)</td>
<td>0.939</td>
<td>0.808</td>
<td>0.451</td>
</tr>
</tbody>
</table>

- vs. bootstrapping: context-aware prediction on “un-matchable”
- vs. label propagation: group similar relation phrases
- vs. FIGER: no reliance on complex feature engineering

**NYT**: 118k news articles (1k manually labeled for evaluation); **Yelp**: 230k business reviews (2.5k reviews are manually labeled for evaluation); **Tweet**: 302 tweets (3k tweets are manually labeled for evaluation)

Precision $(P) = \frac{\text{#Correctly-typed mentions}}{\text{#System-recognized mentions}}$, Recall $(R) = \frac{\text{#Correctly-typed mentions}}{\text{#ground-truth mentions}}$, F1 score $= \frac{2(PR)}{(P+R)}$
Corpus to Structured Network: The Roadmap

Entity Recognition and Coarse-grained Typing (KDD’15)

细粒度实体类型化 (KDD’16, EMNLP’16)

联合实体和关系抽取 (WWW’17)

Text corpus

Data-driven text segmentation (SIGMOD’15, WWW’16)

Learning Corpus-specific Model (KDD’16, EMNLP’16, 17, WWW’17)

entity names & context units

Partially-labeled corpus

Knowledge bases

结构化网络：道路图

数据驱动的文本分割

学习专有模型

部分有标签的语料库

实体名称及语境单位

从剩余的未标注数据中学习结构

从语料库到结构化网络
From Coarse-Grained Typing to Fine-Grained Entity Typing

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Donald Trump</em> spent 14 television seasons presiding over a game show, NBC’s The Apprentice.</td>
</tr>
</tbody>
</table>

A few common types

Location  Person

Organization

A type hierarchy with 100+ types (from knowledge base)

(Ling et al., 2012), (Nakashole et al., 2013), (Yogatama et al., 2015)
Current Distant Supervision: Context-Agnostic Labeling

- Inaccurate labels in training data
- Prior work: all labels are “perfect”

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Donald Trump spent 14 television seasons presiding over a game show, NBC’s The Apprentice</td>
</tr>
</tbody>
</table>

Entity: Donald Trump

Entity Types: person, artist, actor, author, businessman, politician

Entity types from knowledge base
Modeling **Clean** and **Noisy** Mentions Separately

For a **clean mention**, its “positive types” should be **ranked higher** than all its “negative types”

<table>
<thead>
<tr>
<th>ID</th>
<th>Noisy Entity Mention</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S1</strong></td>
<td><em>Donald Trump</em> spent 14 television seasons presiding over a game show, NBC’s The Apprentice</td>
</tr>
</tbody>
</table>

Types ranked:

- (+) actor 0.88
- (+) artist 0.74
- (+) person 0.55
- (+) author 0.41
- (+) politician 0.33
- (+) business 0.31

“Best” candidate type: (+) actor

*Types in KB: person, artist, actor, author, businessman, politician*

For a **noisy mention**, its “**best candidate type**” should be **ranked higher** than all its “non-candidate types”

*Ren et al., EMNLP’16, KDD’16*
Hierarchical Type Inference

• Top-down nearest neighbor search in the given type hierarchy

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_i$</td>
<td>President Trump gave an all-hands address to troops at the U.S. Central Command headquarters</td>
</tr>
</tbody>
</table>

Vectors for text features

Test mention: $S_i$ Trump

(Ren et al., EMNLP’16)
Partial Label Embedding (KDD’16)

Extract Text Features

“Label Noise Reduction” with PLE

Train Classifiers on De-noised Data

Prediction on New Data

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Donald Trump</em> spent 14 television seasons presiding over a game show, NBC’s <em>The Apprentice</em></td>
</tr>
</tbody>
</table>

Text features: TOKEN_Donald, CONTEXT: television, CONTEXT: season, TOKEN_trump, SHAPE: AA

S1: *Donald Trump*

Entity Types: person, artist, actor, author, businessman, politician

“De-noised” labeled data

More effective classifiers

(Ren et al., KDD’16)
Performance of Fine-Grained Entity Typing

Accuracy = \frac{\text{# mentions with all types correctly predicted}}{\text{# mentions in the test set}}

Accuracy on different type levels

- **Raw**: candidate types from distant supervision
- **WSABIE** (Google, ACL’15): joint feature and type embedding
- **Predictive Text Embedding** (MSR, WWW’15): joint mention, feature and type embedding
  - Both WSABIE and PTE suffer from “noisy” training labels
- **PLE** (KDD’16): partial-label loss for context-aware labeling

OntoNotes public dataset (Weischedel et al. 2011, Gillick et al., 2014):
13,109 news articles, 77 annotated documents, 89 entity types
Corpus to Structured Network: The Roadmap

- Text corpus
- Data-driven text segmentation (SIGMOD’15, WWW’16)
- Structures from the remaining unlabeled data
- Learning Corpus-specific Model (KDD’16, EMNLP’16, 17, WWW’17)
- Partially-labeled corpus
- entity names & context units
- Knowledge bases

Entity Recognition and Coarse-grained Typing (KDD’15) → Fine-grained Entity Typing (KDD’16, EMNLP’16) → Joint Entity and Relation Extraction (WWW’17)
The Women’s March was a worldwide protest on January 21, 2017. The protest was aimed at Donald Trump, the recently inaugurated president of the United States. The first protest was planned in Washington, D.C., and was known as the Women’s March on Washington.
Prior Work: Relation Extraction (RE)

**Substantial task-specific human annotation**

- Supervised RE systems
  - Hard to be ported to deal with different kinds of corpora

**Pattern-based bootstrapping RE systems**

- Focus on “explicit” relation mentions
  - “Semantic drift”

**No task-specific human annotation**

- Distantly-supervised RE systems (cont.)
  - Error propagation
  - Noisy candidate type labels

---

Prior Work: An “Incremental” System Pipeline

Error propagation cascading down the pipeline

Entity mention detection

Context-aware entity typing

Relation mention detection

Context-aware relation typing

Entity boundary errors:
The Women’s March was a worldwide protest on January 21, 2017.

Entity type errors:
The Women’s March was a worldwide protest on January 21, 2017. → person

Relation mention errors:
(women, protest) X
(protest, January 21, 2017)

Relation type errors
(women, protest) → is a X
(protest, January 21, 2017)

Error propagation cascading down the pipeline

Entity mention detection

Context-aware entity typing

Relation mention detection

Context-aware relation typing

Entity boundary errors:
The Women’s March was a worldwide protest on January 21, 2017.

Entity type errors:
The Women’s March was a worldwide protest on January 21, 2017. → person

Relation mention errors:
(women, protest) X
(protest, January 21, 2017)

Relation type errors
(women, protest) → is a X
(protest, January 21, 2017)

(Mintz et al., 2009), (Riedel et al., 2010), (Hoffmann et al., 2011), (Surdeanu et al., 2012), (Nagesh et al., 2014), ...
The **CoType** Approach (WWW’17)

1. Data-driven detection of entity and relation mentions
   - Data-driven text segmentation
   - Syntactic pattern learning from KBs

2. Joint typing of entity and relation mentions
   - Context-aware type modeling
   - Model **entity-relation interactions**

(Ren et al. WWW’17)
**CoType: Co-Embedding for Typing Entities and Relations**

Object interactions in a heterogeneous graph

Entity Type
- politician
- person
- artist
- book
- LOC
- ORG
- book

Relation Mention
- CONTEXT_president
- EM1_President
- EM1_Book
- S1_"Barack Obama"
- S3_"Barack Obama"
- S3_"United States"
- S2_"Dream of My Father"
- S2_"Obama"

Entity Mention
- S1_"US"
- "Barack Obama"

Low-dimensional vector spaces

Model entity-relation interactions

(Ren et al. WWW’17)
Object “Translating” Assumption

For a relation mention \( z \) between entity arguments \( m_1 \) and \( m_2 \):

\[
\text{vec}(m_1) \approx \text{vec}(m_2) + \text{vec}(z)
\]

Error on a relation triple (\( z, m_1, m_2 \)):

\[
\tau(z) = \|m_1 + z - m_2\|^2
\]

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

(Bordes, NIPS’13), (Ren et al., WWW’17)
Reducing Error Propagation: A Joint Optimization Framework

Modeling entity-relation interactions

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

\[ \min \mathcal{O} = \mathcal{O}_M + \mathcal{O}_Z + \mathcal{O}_{ZM} \]

Modeling types of relation mentions

\[ O_Z = \mathcal{L}_{ZF} + \sum_{i=1}^{N_L} \ell_i + \frac{\lambda}{2} \sum_{i=1}^{N_L} \|z_i\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_r} \|r_k\|_2^2 \]

Modeling types of entity mentions

\[ O_M = \mathcal{L}_{MF} + \sum_{i=1}^{N_L'} \ell_i' + \frac{\lambda}{2} \sum_{i=1}^{N_L'} \|m_i\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_y} \|y_k\|_2^2 \]

(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”

- **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
An Application to Life Sciences

LifeNet: A Knowledge Exploration and Analytics System for Life Sciences

BioInfer Network by human labeling (Pyysalo et al., 2007)
- Human-created
- 1,100 sentences
- 94 protein-protein interactions
- 2,500 man-hours
- 2,662 facts

BioInfer Network by human labeling (Pyysalo et al., 2007)
- Human-created
- 1,100 sentences
- 94 protein-protein interactions
- 2,500 man-hours
- 2,662 facts

LifeNet by Effort-Light StructMine
- Machine-created
- 4 Million+ PubMed papers
- 1,000+ entity types
- 400+ relation types
- <1 hour, single machine
- 10,000x more facts

Performance evaluation on BioInfer:
Relation Classification Accuracy = 61.7%
(11%↑ over the best-performing baseline)

(Pyysalo et al., BMC Bioinformatics’07)
(Ren et al., ACL’17 demo)
Towards Automated Structure Extraction

End-to-end extraction models
(Zheng et al., ACL’17), (Xu et al., 2017), (Liu et al., 2017)

Unifying heterogeneous forms of weak supervisions
(Liu et al., EMNLP’17)

Indirection supervision from auxiliary tasks
(Wu et al., WSDM’18)

Leveraging rich language patterns to facilitate NLU
(Qu et al., 2018), (Liu et al., AAAI’18)
Biomedical Named Entity Recognition by Multi-tasking different datasets

Single-task/dataset learning

- CRF
- Word BiLSTM
- Word emb
- Concat
- Char BiLSTM
- Char emb
- LM

Multi-task/dataset learning

- CRF1
- CRF2
- Word BiLSTM
- Word emb
- Concat
- Char BiLSTM
- Char emb

(Liu et al., AAAI’18)

(Wang et al., 2018)
State-of-the-art Biomed Entity Tagger

- **One** tagger for many biomed entity types (gene, disease, chemical, etc.)
- State-of-the-art performance on several benchmark datasets

(Wang et al., 2018)
Heterogeneous Supervision for Relation Extraction

- A principled framework to unify KB-supervision, manual rules, crowd-sourced labels, etc.
- Multiple “labeling functions” annotate one instance $\rightarrow$ resolve conflicts & redundancy $\rightarrow$ “expertise” of each labeling function

(Liu et al., EMNLP’17)
Indirect Supervision for Relation Extraction – using QA Pairs

• Questions → positive / negative answers
• pos pairs → similar relation; neg pairs → distinct relations

(Wu et al., WSDM’18)
Pattern-enhanced Distributional Representation Learning

Pattern Module
- [ENT] [ENT] ’s capital
- [ENT] capital [ENT]
- capital [ENT] [ENT]

Distributional Module
- Capital of
- Germany
- France
- China
- Paris
- Berlin
- Beijing
- Score 0.9

Existing Integration Frameworks
- Seeds
- Pattern Module
- Distributional Module

Our Co-training Framework
- Seeds
- Pattern Module
- Distributional Module

(Qu et al., WWW’18)
Corpus to Structured Network: The Roadmap

Text corpus

Data-driven text segmentation (SIGMOD’15, WWW’16)

Learning Corpus-specific Model (KDD’15, KDD’16, EMNLP’16, WWW’17)

entity names & context units

Partially-labeled corpus

Knowledge bases

Structures from the remaining unlabeled data

Entity Recognition and Coarse-grained Typing (KDD’15)

Fine-grained Entity Typing (KDD’16)

Joint Entity and Relation Extraction (WWW’17)
References I

• **Xiang Ren**, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, Jiawei Han. CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases. WWW, 2017.

• **Xiang Ren**, Ahmed El-Kishky, Heng Ji, and Jiawei Han. Automatic Entity Recognition and Typing in Massive Text Data (Conference Tutorial). SIGMOD, 2016.

• **Xiang Ren***, Wenqi He*, Meng Qu, Lifu Huang, Heng Ji, Jiawei Han. AFET: Automatic Fine-Grained Entity Typing by Hierarchical Partial-Label Embedding. EMNLP, 2016.

• **Xiang Ren***, Wenqi He*, Meng Qu, Heng Ji, Clare R. Voss, Jiawei Han. Label Noise Reduction in Entity Typing by Heterogeneous Partial-Label Embedding. KDD, 2016.

• **Xiang Ren**, Wenqi He, Ahmed El-Kishky, Clare R. Voss, Heng Ji, Meng Qu, Jiawei Han. Entity Typing: A Critical Step for Mining Structures from Massive Unstructured Text (Invited Paper). MLG, 2016.


• Meng Qu, **Xiang Ren**, Yu Zhang, Jiawei Han. Weakly-supervised Relation Extraction by Pattern-enhanced Embedding Learning. WWW, 2018.

• Ellen Wu, **Xiang Ren**, Frank Xu, Ji Li, Jiawei Han. Indirect Supervision for Relation Extraction using Question-Answer Pairs WSDM, 2018.

• Liyuan Liu*, **Xiang Ren***, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji, Jiawei Han. Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach. EMNLP, 2017.

• Tarique A. Siddiqui*, **Xiang Ren***, Aditya Parameswaran, Jiawei Han. FacetGist: Collective Extraction of Document Facets in Large Technical Corpora. CIKM, 2016.
References II

- Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, Jiawei Han. Mining Quality Phrases from Massive Text Corpora. SIGMOD, 2015.

- Marina Danilevsky, Chi Wang, Nihit Desai, Xiang Ren, Jingyi Guo, and Jiawei Han. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents. SDM, 2014.

- Xiang Ren, Yuanhua Lv, Kuansan Wang, Jiawei Han. Comparative Document Analysis for Large Text Corpora. WSDM, 2017.

- Jialu Liu, Xiang Ren, Jingbo Shang, Taylor Cassidy, Clare R. Voss, Jiawei Han. Representing Documents via Latent Keyphrase Inference. WWW, 2016.

- Hyungsul Kim, Xiang Ren, Yizhou Sun, Chi Wang, and Jiawei Han. Semantic Frame-Based Document Representation for Comparable Corpora. ICDM, 2013.

- Xiang Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han. ClusCite: Effective Citation Recommendation by Information Network-Based Clustering. KDD, 2014.


- Xiang Ren, Yujing Wang, Xiao Yu, Jun Yan, Zheng Chen, Jiawei Han. Heterogeneous Graph-Based Intent Learning from Queries, Web Pages and Wikipedia Concepts. WSDM 2014b.


- Xiao Yu, Xiang Ren, Quanquan Gu, Yizhou Sun and Jiawei Han. Collaborative Filtering with Entity Similarity Regularization in Heterogeneous Information Networks. IJCAI-HINA, 2013.
Scalable Construction and Reasoning of Massive Knowledge Bases

Part II: Joint Representation Learning for Low-resource Information Extraction
### Structured Knowledge

<table>
<thead>
<tr>
<th>Entity</th>
<th>Entity</th>
<th>Entity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>T790M</td>
<td>EGFR</td>
<td>gefitinib</td>
<td>Resist</td>
</tr>
<tr>
<td>Obama</td>
<td>U.S.</td>
<td></td>
<td>President_of</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Sources:**
- PubMed
- Twitter
- Facebook
- Medical Records
- Yelp
- Amazon
- SEC

**IE:** Information Extraction
Challenges of Obtaining Training Data

• Constructing data sets is labor intensive
• Many different
  • Languages
  • Domains
  • Modalities
  • …
Joint representation learning models for *low-resource IE*.

- Learning comprehensive representations from *heterogeneous sources*.
  - *unlabeled data*
  - annotations for *related tasks, domains and languages*.

- Encoding structured knowledge to learn robust representations and make *holistic decisions*.
  - *linguistic structures*
Named Entity Recognition (NER)

- Identifying entities (in social media domain, usually person, organization, location and GPE) boundaries and their type from the plain text.

```
成都(GPE.NAM) 电信(ORG.NAM) 到底有没的时间观念
哦，一托再托，日妈(PER.NOM) 我们时间就不是时间哇，等了你两天啥子速度。
 Chengdu(GPE.NAM) Telecom(ORG.NAM) do you have no concept of time, delay again and again, mother(PER.NOM) (curse word) our time is not time, waited for you for two days what a speed.
```
Structured Model for NER

- **Sequence Tagging Models:**

  西班牙(LOC), 语言(ORG), 数据(ORG)

  开始、内部、外部

  \[
  P(y | x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \exp \left\{ \sum_{k=1}^{K} \theta_k f_k (y_t, y_{t-1}, x_t) \right\}
  \]

  make joint decisions over a sequence

  McCallum and Li, 2003
Representation Learning for NER

- Recurrent Neural Networks (RNNs) for Representation:
  - Automatically learns data representations for features
  - Model input dependencies.

Huang et. al., 2015  Lample et. al., 2016  Ma and Hovy., 2016
Representation Learning for NER

Recurrent Neural Networks (RNNs)

\[ A_t = g(Vx_t + UA_{t-1} + c) \]

Very deep neural network, back propagation training
Long-Short Term Memory Networks (LSTMs)

LSTMs are special RNNs that use gates to control the information flow and essentially capture *long-term dependencies* of the input.

Very deep neural network, back propagation training

Picture credit: colah's blog, 2015
Neural Sequence Tagging Models

Linear Chain CRF for NER

BiLSTMs for Feature Learning

Embeddings

Input Texts

End-to-end training

Huang et. al., 2015  Lample et. al., 2016  Ma and Hovy, 2016
Challenges for low-resource settings

• HUGE gap on social media (noisy) v.s news text:
  • informal language and insufficient annotations.

Ma and Hovy, 2016
Cherry and Guo, 2015
Yu et al, 2008

English
Chinese
Ideas

• Leverage existing resources to learn representations that generalize across multiple types of data.
  • Multi-task Learning.
  • Domain Adaptation.
  • Cross-lingual Transfer.
Distributional Similarity of Words

Generalizability

- Rose

- Violet

Peng and Dredze, 2015
Joint Learning of Word Embeddings and Named Entity Recognition

Model for Learning Word Representations

Model for Named Entity Recognition

Peng and Dredze, 2015
Joint Learning of Word Embeddings and Named Entity Recognition

Peng and Dredze, 2015
Joint Learning of Word Embeddings and Named Entity Recognition

Skip-gram model to learn word representations

\[ L_u(X | e_x) = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(x_{t+j} | x_t) \]

\[ p(x_i | x_j) = \frac{\exp(e^T_{xi} e_{xj})}{\sum_{i'} \exp(e^T_{xi'} e_{xj})} \]

Log-bilinear CRF model for named entity recognition

\[ L_s(Y | X; \theta | e_x) = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t | x_t) \]

\[ P(y | x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t, e_x)\} \]

2 millions of unannotated weibo message for training

1350 NER annotated weibo message for training
Chengdu (GPE.NAM) Telecom (ORG.NAM) do you have no concept of time, delay again and again, mother (PER.NOM) (curse word) our time is not time, waited for you for two days what a speed.

Peng and Dredze, 2016
Multi-task Learning of Word Segmentation and Named Entity Recognition

Model for Chinese Word Segmentation

Model for Named Entity Recognition

http://www.cs.jhu.edu/~npe
Multi-task Learning of Word Segmentation and Named Entity Recognition

Peng and Dredze, 2016
McDonald’s Seeks Its Fast-Food Soul
- NYTimes 3/7/2015

Nivre and McDonald (2008) used the output of one dependency parser to provide features for another.
- Stacking Dependency Parsers, Martins+ (EMNLP 2008)

Peng and Dredze, 2017
Multi-task Multi-domain Learning

Task Specific Models

Domain Projections

Shared Representation Learner

Input data \( w^{(1)} \) \( \ldots \) \( w^{(n-1)} \) \( w^{(n)} \)

Peng and Dredze, 2017
Multi-task Multi-domain learning for sequence tagging

- Domains: news and social media
- Tasks: Chinese word segmentation and NER
- Datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Train</th>
<th>#Dev</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>News CWS</td>
<td>39,567</td>
<td>4,396</td>
<td>4,278</td>
</tr>
<tr>
<td>News NER</td>
<td>16,814</td>
<td>1,868</td>
<td>4,636</td>
</tr>
<tr>
<td>Social CWS</td>
<td>1,600</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Social NER</td>
<td>1,350</td>
<td>270</td>
<td>270</td>
</tr>
</tbody>
</table>

2 millions of unannotated weibo message for training
NER on Chinese Social Media

Named Entity Recognition on Chinese social media
Closing The Gap

Ours new Chinese Social NER system leveraging heterogeneous sources for representation learning

Peng and Dredze, 2017
How to build NER for a new language using
(1) Comparable Corpora
(2) English NER tagger

Wang, Peng, and Duh, 2017
Idea

Bilingual Word Embeddings

Cross-lingual NER Tagger

Johns Hopkins University

Johns Hopkins University is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins. [5]

Wang, Peng, and Duh, 2017
Training Bilingual Word Embeddings

Word2Vec

Mixed-Language Pseudo-Document

约翰·霍普金斯大學
約翰·霍普金斯大學 是一所主校區
位於美國馬里蘭州巴爾的摩市的研究型私立大學。截止至2012年，
共有36名校友獲諾貝爾獎[3]。

Johns Hopkins University
Johns Hopkins University is an
American private research university in
Baltimore, Maryland. Founded in 1876, the
university was named for its benefactor,
the entrepreneur Johns Hopkins.[5]
Training Cross-lingual NER Tagger

1. Fixed Embeddings
2. Multi-task training

Word2Vec Objective

Bilingual Word Embeddings

Cross-lingual NER Tagger

English NER Tagger

Johns Hopkins University
is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins. [5]
Multi-task training (update Chinese, fix English) is effective!

---

**Results (F1 score)**

![Graph showing F1 score vs embedding dimension for different training methods]
Joint representation learning models for *low-resource IE*.

- Learning comprehensive representations from *heterogeneous sources*.
  - *unlabeled data*
  - annotations for *related tasks, domains and languages*.

- Encoding structured knowledge to learn robust representations and make *holistic decisions*.
  - *linguistic structures*
T790M is present as a minor clone in NSCLC, and may be selected for during therapy. This mutation has been shown to prevent the activation of BIM in response to gefitinib but can be overcome by an irreversible inhibitor of EGFR.
Knowledge Bases for Drug-Gene-Mutation Interaction

• People manually curate drug-gene-mutation interaction databases for precision medicine:
  • Gene Drug Knowledge Database (GDKD) (Dienstmann et al., 2015)
  • Clinical Interpretations of Variants in Cancer (CiViC) (Washington University School of Medicine)
Special Challenges

• **N-ary relations:**
  • Traditional feature-based classification method usually use features defined on the *shortest syntactic dependency paths* between two entities.
  • Such features are hard to define in the N-ary case.

• **Cross sentence relations:**
  • Traditional features become sparser and learning becomes harder.
A Representation Learning Framework

Peng et. al., 2017
Contextual Entity Representation

Word Input Text $w^{(1)} \ldots w^{(n-1)} w^{(n)}$

concatenation

Relation Classifier

Graph LSTM

Representation Learner
Long-Short Term Memory Networks (LSTMs)

Capture *long-term dependencies* of the input.

*However*, it still only explicitly models the dependencies between the adjacent inputs.
T790M is present as a minor clone in NSCLC,

This mutation has been shown to prevent the activation of BIM in response to getinib

Peng et. al., 2017
Directed Cyclic Graph

\[ h_0 \rightarrow A \rightarrow A \rightarrow A \rightarrow A \rightarrow A \rightarrow A \rightarrow h_t \]

\[ x_0 \rightarrow A \rightarrow A \rightarrow A \rightarrow A \rightarrow A \rightarrow A \rightarrow x_t \]

Peng et. al., 2017
Goals:

- different types of dependencies: adjacency, syntactic dependencies, coreferences, and discourse relations.
- long-distance dependencies: entities span sentences.

Challenges: how to define a neural architecture over a cyclic graph?
Work beyond Linear-Chain

• NLP: Tree LSTM (Tai et. al. 2015, Miwa and Bansal, 2016)

• Programming verification: Gated Graph Neural Network (Li et. al. 2016)

• Graph Convolutional Networks (Kipf and Welling, 2017)
Challenge in Training

• Existing approach
  • Unroll recurrence for a number of steps
  • Analogous to loopy belief propagation (LBP)

• Problems
  • Expensive: Many steps per epoch
  • Information does not propagate from distant nodes
Training Graph LSTMs

• *Provably,* all directed cyclic graph without self-loop can be decomposed into two DAGs.

T790M→ is → present → as → a → minor → clone → in → NSCLC

Peng et. al., 2017
Training Graph LSTMs

• Approximate a cyclic graph by two directed acyclic graphs (DAGs), and stack the DAGs.

Topological order is well-defined, back propagation training
Chain LSTMs v.s. Graph LSTMs

Linear-chain LSTM

\[i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)\]
\[o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)\]
\[\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)\]
\[f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)\]
\[c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1}\]
\[h_t = o_t \odot \tanh(c_t)\]

Graph LSTM (one DAG)

\[i_t = \sigma(W_i x_t + \sum_{j \in P(t)} U_i^{m(t,j)} h_j + b_i)\]
\[o_t = \sigma(W_o x_t + \sum_{j \in P(t)} U_o^{m(t,j)} h_j + b_o)\]
\[\tilde{c}_t = \tanh(W_c x_t + \sum_{j \in P(t)} U_c^{m(t,j)} h_j + b_c)\]
\[f_{t,j} = \sigma(W_f x_t + U_f^{m(t,j)} h_j + b_f)\]
\[c_t = i_t \odot \tilde{c}_t + \sum_{j \in P(t)} f_{t,j} \odot c_j\]
\[h_t = o_t \odot \tanh(c_t)\]
Multi-task Learning

\[
R_1 \quad \ldots \quad R_k
\]

\[
\text{Relation Classifier}
\]

\[
\text{Contextual Entity Representation} \quad \text{concatenation} \quad (w^{(1)}, \ldots, w^{(n)})
\]

Peng et. al., 2017
Multi-task Learning

Peng et. al., 2017
Domain: Molecular Tumor Board

• Ternary interaction: (drug, gene, mutation)

• Distant supervision
  • Knowledge bases: GDKD + CIVIC
  • Text: PubMed Central articles (~ 1 million full-text articles)

• We got 3,462 paragraphs about drug-gene-mutation relations from distant supervision.
## Absolute Recall

<table>
<thead>
<tr>
<th></th>
<th>Drug</th>
<th>Gene</th>
<th>Mutation</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGKD + CiViC</td>
<td>16</td>
<td>12</td>
<td>41</td>
<td>59</td>
</tr>
<tr>
<td>Single-Sent</td>
<td>68</td>
<td>228</td>
<td>221</td>
<td>530</td>
</tr>
<tr>
<td>Cross-Sent</td>
<td>103</td>
<td>512</td>
<td>445</td>
<td>1461</td>
</tr>
</tbody>
</table>

Numbers of *distinct* drugs, genes and mutations and their interactions in the knowledge bases vs. PubMed scale automatic extraction.

Machine reading extracted orders of magnitudes more knowledge.

Cross-sentence extraction triples the yield.
Sample Precision

Precision

- Random
- $P > 0.5$
- $P > 0.9$
Automatic Evaluation

- Logistic Regression
- CNN
- Linear LSTM
- Graph LSTM
Multi-Task Learning

Code and data available at: http://hanover.azurewebsites.net/

<table>
<thead>
<tr>
<th></th>
<th>Drug-Gene-Mutation</th>
<th>Drug-Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph LSTM</td>
<td>80.7</td>
<td>76.7</td>
</tr>
<tr>
<td>+ Multi-task</td>
<td>82.0</td>
<td>78.5</td>
</tr>
</tbody>
</table>
Conclusion

• Jointly learning comprehensive representations from *heterogeneous sources*:
  
  • *Data and code available at:*  
    [https://github.com/hltcoe/golden-horse/](https://github.com/hltcoe/golden-horse/)

• Encoding linguistic structures to learn robust representations:
  
  • *Data and code available at:*  
    [http://hanover.azurewebsites.net/](http://hanover.azurewebsites.net/)
Knowledge Graph Reasoning: Past, Present, and Future

William Wang
Department of Computer Science

UC SANTA BARBARA

NAACL 2018 Tutorial

w. Xiang Ren and Nanyun Peng (USC)
Agenda

• Motivation
• Path-Based Reasoning
• Embedding-Based Reasoning
• Bridging Path-Based and Embedding-Based Reasoning: DeepPath, MINERVA, and DIVA
• Conclusions
Knowledge Graphs are Not Complete
Benefits of Knowledge Graph

• Support various applications
  • Structured Search
  • Question Answering
  • Dialogue Systems
  • Relation Extraction
  • Summarization

• Knowledge Graphs can be constructed via information extraction from text, but...
  • There will be a lot of missing links.
  • Goal: complete the knowledge graph.
Reasoning on Knowledge Graph

Query node: Band of Brothers
Query relation: tvProgramLanguage

- Band of Brothers
  - Mini-Series
  - HBO
  - tvProgramCreator: Graham Yost
  - tvProgramGenre: "Graham Yost"

- Neal McDonough
  - nationality: English
  - writtenBy: Graham Yost
  - castActor: Tom Hanks

- Tom Hanks
  - profession: Actor
  - awardWorkWinner: Band of Brothers

- Caesars Entertainment
  - serviceLocation: United States

- United States
  - countrySpokenIn: United States

- Michael Kamen
  - music: Band of Brothers
  - tvProgramGenre: "Michael Kamen"

- Caesars Entertainment
  - serviceLanguage: English

- Graham Yost
  - personLanguages: English

- Tom Hanks
  - personLanguages: English

- Neal McDonough
  - personLanguages: English

- Michael Kamen
  - personLanguages: English

- United States
  - countryOfOrigin: United States

- Tom Hanks
  - nationality: English

- Neal McDonough
  - nationality: English

- Michael Kamen
  - nationality: English
KB Reasoning Tasks

• Predicting the missing link.
  • Given e1 and e2, predict the relation r.

• Predicting the missing entity.
  • Given e1 and relation r, predict the missing entity e2.

• Fact Prediction.
  • Given a triple, predict whether it is true or false.
Related Work

• **Path-based methods**
  - Path-Ranking Algorithm, Lao et al. 2011
  - ProPPR, Wang et al, 2013 (My PhD thesis)
  - Subgraph Feature Extraction, Gardner et al, 2015
  - RNN + PRA, Neelakantan et al, 2015
  - Chains of Reasoning, Das et al, 2017

Why do we need path-based methods?
It’s accurate and explainable!
Path-Ranking Algorithm (Lao et al., 2011)

1. Run random walk with restarts to derive many paths.

2. Use supervised training to rank different paths.
ProPPR (Wang et al., 2013; 2015)

- ProPPR generalizes PRA with recursive probabilistic logic programs.
- You may use other relations to jointly infer this target relation.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>about(X,Z): - handLabeled(X,Z)</td>
<td># base</td>
</tr>
<tr>
<td>about(X,Z): - sim(X,Y),about(Y,Z)</td>
<td># prop</td>
</tr>
<tr>
<td>sim(X,Y): - link(X,Y)</td>
<td># sim, link</td>
</tr>
<tr>
<td>sim(X,Y): -</td>
<td></td>
</tr>
<tr>
<td>hasWord(X,W), hasWord(Y,W),</td>
<td># sim, word</td>
</tr>
<tr>
<td>linkedBy(X,Y,W)</td>
<td># by(W)</td>
</tr>
<tr>
<td>linkedBy(X,Y,W): - true</td>
<td></td>
</tr>
</tbody>
</table>
Chain of Reasoning (Das et al, 2017)

1. Use PRA to derive the path.
2. Use RNNs to perform reasoning of the target relation.
Related Work

• **Embedding-based method**
  • RESCAL, Nickel et al, 2011
  • TransE, Bordes et al, 2013
  • Neural Tensor Network, Socher et al, 2013
  • TransR/CTransR, Lin et al, 2015
  • Complex Embeddings, Trouillon et al, 2016

Embedding methods allow us to compare, and find similar entities in the vector space.
RESCAL (Nickel et al., 2011)

- Tensor factorization on the
  - (head)entity-(tail)entity-relation tensor.
TransE (Bordes et al., 2013)

- Assumption: in the vector space, when adding the relation to the head entity, we should get close to the target tail entity.

- Margin based loss function:
  - Minimize the distance between \((h+l)\) and \(t\).
  - Maximize the distance between \((h+l)\) to a randomly sampled tail \(t'\) (negative example).

\[
L = \sum_{(h,l,t) \in S} \sum_{(h',l,t') \in S'} \left[ \gamma + d(h + l, t) - d(h' + l, t') \right]_+
\]
Neural Tensor Networks (Socher et al., 2013)

• Model the bilinear interaction between entity pairs with tensors.

\[
U^T f( e_1^T W^{[1:2]} e_2 + V \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} + b )
\]
Poincaré Embeddings (Nickel and Kiela, 2017)

• Idea: learn hierarchical KB representations by looking at hyperbolic space.

\[ d(u, v) = \text{arcosh} \left( 1 + 2 \frac{\|u - v\|^2}{(1 - \|u\|^2)(1 - \|v\|^2)} \right). \]

Figure 1: (a) Geodesics of the Poincaré disk (b) Embedding of a tree in \( B^2 \) (c) Growth of Poincaré distance

(a) Due to the negative curvature of \( B \), the distance of points increases exponentially (relative to their Euclidean distance) the closer they are to the boundary. (c) Growth of the Poincaré distance \( d(u, v) \) relative to the Euclidean distance and the norm of \( v \) (for fixed \( \|u\| = 0.9 \)). (b) Embedding of a regular tree in \( B^2 \) such that all connected nodes are spaced equally far apart (i.e., all black line segments have identical hyperbolic length).
ConvE (Dettmers et al, 2018)

1. Reshape the head and relation embeddings into “images”.
2. Use CNNs to learn convolutional feature maps.
Bridging Path-Based and Embedding-Based Reasoning with Deep Reinforcement Learning: DeepPath (Xiong et al., 2017)
RL for KB Reasoning: DeepPath (Xiong et al., 2017)

- Learning the paths with RL, instead of using random walks with restart
- Model the path finding as a MDP
- Train a RL agent to find paths
- Represent the KG with pretrained KG embeddings
- Use the learned paths as logical formulas
Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
Supervised v.s. Reinforcement Learning

Supervised Learning
- Training based on supervisor/label/annotation
- Feedback is instantaneous
- Not much temporal aspects

Reinforcement Learning
- Training only based on reward signal
- Feedback is delayed
- Time matters
- Agent actions affect subsequent exploration
Reinforcement Learning

• RL is a general purpose framework for decision making
  • ◦ RL is for an agent with the capacity to act
  • ◦ Each action influences the agent’s future state
  • ◦ Success is measured by a scalar reward signal
  • ◦ Goal: select actions to maximize future reward
Reinforcement Learning

Agent $\psi(s_t)$
Environment

Multi-layer neural nets $\psi(s_t)$
KG modeled as a MDP
DeepPath: RL for KG Reasoning

Query Node: Band of Brothers
Reason Task: tvProgramLanguage

The KG Environment

English

serviceLanguage

Caesars Entertain...

countrySpokenIn

personLanguages

serviceLocation

United States

countryOfOrigin

castActor

Band of Brothers

tvProgramCreator

HBO

writtenBy

Graham Yost

music

Michael Kamen

tvProgramGenre

Mini-Series

Policy Based Agent

Next State

Reward

Reason Step

State

ReLU

ReLU

Softmax

π(а|s)
Components of MDP

• Markov decision process $< S, A, P, R >$
  • $S$: continuous states represented with embeddings
  • $A$: action space (relations)
  • $P(S_{t+1} = s' | S_t = s, A_t = a)$: transition probability
  • $R(s, a)$: reward received for each taken step

• With pretrained KG embeddings
  • $s_t = e_t \oplus (e_{target} - e_t)$
  • $A = \{r_1, r_2, ..., r_n\}$, all relations in the KG
Reward Functions

• Global Accuracy

\[ r_{GLOBAL} = \begin{cases} 
+1, & \text{if the path reaches } e_{target} \\
-1, & \text{otherwise} 
\end{cases} \]

• Path Efficiency

\[ r_{EFFICIENCY} = \frac{1}{length(p)} \]

• Path Diversity

\[ r_{DIVERSITY} = -\frac{1}{|F|} \sum_{i=1}^{\frac{|F|}{N}} cos(p, p_i) \]
Training with Policy Gradient

- Monte-Carlo Policy Gradient (REINFORCE, William, 1992)

\[
\nabla_\theta J(\theta) = \sum_t \sum_{a \in A} \pi(a|s_t; \theta) \nabla_\theta \log \pi(a|s_t; \theta) R(s_t, a_t)
\]

\[
\approx \nabla_\theta \sum_t \log \pi(a = r_t|s_t; \theta) R(s_t, a_t)
\]

\[
R(s_t, a_t) = \lambda_1 r_{global} + \lambda_2 r_{efficiency} + \lambda_3 r_{diversity}
\]
Challenge

- Typical RL problems
  - Atari games (Mnih et al., 2015): 4~18 valid actions
  - AlphaGo (Silver et al. 2016): ~250 valid actions
  - Knowledge Graph reasoning: >= 400 actions

Issue:
- large action (search) space -> poor convergence properties
Supervised (Imitation) Policy Learning

- Use randomized BFS to retrieve a few paths
- Do imitation learning using the retrieved paths
- All the paths are assigned with +1 reward

\[
\nabla_\theta J(\theta) = \sum_t \sum_{a \in A} \pi(a|s_t; \theta) \nabla_\theta \log \pi(a|s_t; \theta)
\]

\[
\approx \nabla_\theta \sum_t \log \pi(a = r_t|s_t; \theta)
\]
Datasets and Preprocessing

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Entities</th>
<th># of Relations</th>
<th># of Triples</th>
<th># of Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB15k-237</td>
<td>14,505</td>
<td>237</td>
<td>310,116</td>
<td>20</td>
</tr>
<tr>
<td>NELL-995</td>
<td>75,492</td>
<td>200</td>
<td>154,213</td>
<td>12</td>
</tr>
</tbody>
</table>

**FB15k-237**: Sampled from FB15k (Bordes et al., 2013), redundant relations removes

**NELL-995**: Sampled from the 995th iteration of NELL system (Carlson et al., 2010b)

- **Dataset processing**
  - Remove useless relations: *haswikipediaurl, generalizations, etc*
  - Add inverse relation links to the knowledge graph
  - Remove the triples with task relations
Effect of Supervised Policy Learning

- **x-axis:** number of training epochs
- **y-axis:** success ratio (probability of reaching the target) on test set

-> Re-train the agent using reward functions
Inference Using Learned Paths

- **Path as logical formula**
  - FilmCountry: actionFilm\(^{-1}\) -> personNationality
  - PersonNationality: placeOfBirth -> locationContains\(^{-1}\)
  - etc ...

- **Bi-directional path-constrained search**
  - Check whether the formulas hold for entity pairs

Uni-directional search  bi-directional search
## Link Prediction Result

<table>
<thead>
<tr>
<th>Tasks</th>
<th>PRA</th>
<th>Ours</th>
<th>TransE</th>
<th>TransR</th>
</tr>
</thead>
<tbody>
<tr>
<td>worksFor</td>
<td>0.681</td>
<td>0.711</td>
<td>0.677</td>
<td>0.692</td>
</tr>
<tr>
<td>athletePlaysForTeam</td>
<td>0.987</td>
<td>0.955</td>
<td>0.896</td>
<td>0.784</td>
</tr>
<tr>
<td>athletePlaysInLeague</td>
<td>0.841</td>
<td>0.960</td>
<td>0.773</td>
<td>0.912</td>
</tr>
<tr>
<td>athleteHomeStadium</td>
<td>0.859</td>
<td>0.890</td>
<td>0.718</td>
<td>0.722</td>
</tr>
<tr>
<td>teamPlaysSports</td>
<td>0.791</td>
<td>0.738</td>
<td>0.761</td>
<td>0.814</td>
</tr>
<tr>
<td>orgHirePerson</td>
<td>0.599</td>
<td>0.742</td>
<td>0.719</td>
<td>0.737</td>
</tr>
<tr>
<td>personLeadsOrg</td>
<td>0.700</td>
<td>0.795</td>
<td>0.751</td>
<td>0.772</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.675</td>
<td>0.796</td>
<td>0.737</td>
<td>0.789</td>
</tr>
</tbody>
</table>

Mean average precision on NELL-995
Qualitative Analysis

Path length distributions

![Path length distributions diagram]

- NELL-995
- FB15K-237

distribution of reasoning paths

number of paths
Qualitative Analysis

Example Paths

personNationality:
- placeOfBirth -> locationContains\(^{-1}\)
- peoplePlaceLived -> locationContains\(^{-1}\)
- peopleMariage -> locationOfCeremony -> locationContains\(^{-1}\)

tvProgramLanguage:
- tvCountryOfOrigin -> countryOfficialLanguage
- tvCountryOfOrigin -> filmReleaseRegion\(^{-1}\) -> filmLanguage
- tvCastActor -> personLanguage

athletePlaysForTeam:
- athletePlaysSports -> teamPlaysSports\(^{-1}\)
- athleteLedSportsTeam
Bridging Path-Finding and Reasoning w. Variational Inference (teaser): DIVA (Chen et al., NAACL 2018)
DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

• Inferring latent paths connecting entity nodes.

\[
p(r|e_s, e_d) = \arg \max_p \log p(r|e_s, e_d)
\]
DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

- Inferring latent paths connecting entity nodes by parameterizing likelihood (path reasoning) and prior (path finding) with neural network modules.

\[
p = \arg\max_p p(r|e_s, e_d) = \arg\max_p \log \int L p(r|L)p(L|e_s, e_d)
\]
DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

• Marginal likelihood $\log \int_L p(r|L)p(L|e_s, e_d)$ is intractable

• We resort to Variational Bayes by introduce a posterior distribution $q(L|e_s, e_d, r)$

\[
\log p(r|e_s, e_d) \geq ELBO \\
\mathbb{E}_{q(L|e_s, e_d, r)}[\log p(r|L)] \\
- \\
KL(q(L|e_s, e_d, r)||p(L|e_s, e_d))
\]
Parameterization – Path-finder

• Approximate posterior $q_{\phi}(L|e_s, e_d, r)$ and prior $p_\beta(L|e_s, e_d)$: parameterize with RNN

Transition Probability: $p(a_{\tau+1}, e_{\tau+1}|a_{1:\tau}, e_{1:\tau})$
Parameterization – Path Reasoner

• Likelihood $p_\theta (r|L)$ : parameterize with CNN
DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

• Training

\[
\mathbb{E}_{q_\varphi(L|e_s,e_d,r)}[\log p_\theta(r|L)]
\]

\[
KL(q_\varphi(L|e_s,e_d,r) || p_\beta(L|e_s,e_d))
\]

\text{posterior: } q_\varphi, \text{likelihood: } p_\theta (r|L), \text{prior: } p_\beta(L|e_s,e_d)
DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

- Testing

posterior: $q_\varphi$, likelihood: $p_\theta (r|L)$, prior: $p_\beta (L|e_s, e_d)$
KBGAN: Adversarial Learning for Knowledge Graph Completion (NAACL 2018, Monday Morning)

Idea: use adversarial learning to replace random sampling (from a uniform distribution).
Conclusions

• Embedding-based methods are very scalable and robust.
• Path-based methods are more interpretable.
• There are some recent efforts in unifying embedding and path-based approaches.
• DIVA integrates path-finding and reasoning in a principled variational inference framework.
Thanks!

DeepPath Source code:  
https://github.com/xwhan/DeepPath

KBGAN Source code:  
https://github.com/cai-lw/KBGAN

ProPPR Source code:  
https://github.com/TeamCohen/ProPPR
Scalable Construction and Reasoning of Massive Knowledge Bases

Summary
Overall Contributions

• **Effort-Light Structure Extraction**
  ➔ Corpus-specific labeling free, domain/language-independent

• **Joint Models for Low-resource IE:** jointly learning representations from unlabeled data, linguistic structures, annotations from other tasks, domains, and languages. ➔ **Reusable knowledge**

• **Reasoning:** learning to infer missing links from background knowledge.

• A principled approach to manage, explore, and analyze “Big Text Data”
Looking Forward: What’s Next?

Text corpus

Data-driven text segmentation (SIGMOD’15, WWW’16, ...)

Learning Corpus-specific Model (KDD’15, KDD’16, EMNLP’16, WWW’17, ...)

Partially-labeled corpus

Partialy-labeled corpus

entity names & context units

Knowledge bases

Knowledge Base

Network-to-knowledge

Knowledge & Insights

Structures from the remaining unlabeled data

Looking Forward: What’s Next?
Looking Forward: Analyzing Literature to Facilitate Scientific Research

- Literature → Knowledge Base → Scientific Discovery
- More disciplines & More structure analysis functions

Scientific Hypothesis Generation by predicting missing relationships

Gaining insights for various research tasks in different disciplines

Collaborate with life scientists, chemists, physicists, computer scientists, ...
Looking Forward: Engaging with Human Behaviors

User-generated Content (Structured Network) + Structured Behavior Data → Personalized Intelligent Systems

Social media post, Customer review, Chats & messages + Social network, Electronic health record, Transaction record → Smart Health, Business intelligence, Conversational agent

Collaborate with doctors, social scientists, economists, ...
Looking Forward: Integrating with Our Physical World

Textual Signals (Structured Network) + Physical Sensor (Network) Signals

- News streams, Social media post, Organization report
- Geo-sensors
- Audio/video sensors
- Bio-sensors

Data Analytics

Smart-City Operating Systems
Traffic management, Sustainable urban system, Cyber-physical system

Collaborate with network & system researchers, environmental scientists, ...
Application to Vertical Domains

“Which cement stocks go up the most when a Category 3 hurricane hits Florida?”

KENSCHC
One Interface for All

• All domains in a unified knowledge base
• Incrementally learn new domains without forgetting (or instead boosting) existing ones
Learning to Reason for KB Completion

NELL knowledge fragment

Slide from Tom Mitchell.
A Tale of Three Stories

- **Embedding-Based Approaches:**
  - Light-weight, scalable, and robust.

- **Path-Based Approaches:**
  - Explainable and interpretable.

- **Deep Reinforcement Learning Based:**
  - Integrate embedding and path based methods seamlessly.
SOTAs for Reasoning on KBs

• ConvE (Dettmers et al., AAAI 2018)
• Poincare (Nickel and Kiela, NIPS 2017)

• DeepPath (Xiong et al., EMNLP 2017).
• MINERVA (Das et al., ICLR 2018).

• DIVA (Chen et al., NAACL 2018).
Open-sourced Software

• Entity recognition and typing:
  • ClusType: http://shanzhenren.github.io/ClusType/
  • LM-LSTM-CRF: https://github.com/LiyuanLucasLiu/LM-LSTM-CRF
  • CrossType Name Tagger: https://github.com/yuzhimanhua/LM-LSTM-CRF
  • Multi-tasking LSTM-CRF: https://github.com/hltcoe/golden-horse/

• Relation extraction:
  • CoType: https://github.com/shanzhenren/CoType
  • ReQuest: https://github.com/shanzhenren/ReQuest
  • GraphLSTM: http://hanover.azurewebsites.net/

• KB reasoning:
  • DeepPath: https://github.com/xwhan/DeepPath
  • KBGAN: https://github.com/cai-lw/KBGAN
  • ProPPR: https://github.com/TeamCohen/ProPPR
Thank you! Q&A

- **Effort-Light Structure Extraction**
  → Corpus-specific labeling free, domain/language-independent

- **Joint Models for Low-resource IE**: jointly learning representations from unlabeled data, linguistic structures, annotations from other tasks, domains, and languages. → Reusable knowledge

- **Reasoning**: leverage embedding and path based models for discovering new knowledge.

- A principled approach to manage, explore, and analyze “Big Text Data”