Part 1: Knowledge Graphs

Part 2: Knowledge Extraction

Part 3: Graph Construction

Part 4: Critical Analysis
Tutorial Outline

1. Knowledge Graph Primer [Jay]
2. Knowledge Extraction from Text
   a. NLP Fundamentals [Sameer]
   b. Information Extraction [Bhavana]
   Coffee Break
3. Knowledge Graph Construction
   a. Probabilistic Models [Jay]
   b. Embedding Techniques [Sameer]
4. Critical Overview and Conclusion [Bhavana]
John was born in Liverpool, to Julia and Alfred Lennon.
Information Extraction

3 IMPORTANT SUB-PROBLEMS

CATEGORIES OF IE TECHNIQUES

KNOWLEDGE FUSION

IE SYSTEMS IN PRACTICE
Information Extraction

3 CONCRETE SUB-PROBLEMS

- Defining domain
- Learning extractors
- Scoring the facts

3 LEVELS OF SUPERVISION

- Supervised
- Semi-supervised
- Unsupervised
Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain
Learning extractors
Scoring the facts

3 LEVELS OF SUPERVISION

Supervised
Semi-supervised
Unsupervised
Defining Domain: Manual

Subsets:
- Mammals
- Reptiles

Disjoint sets:
- Animals
- Food

Everything

[ Toward an Architecture for Never-Ending Language Learning, Carlson et al. AAAI 2010 ]
Defining Domain: Manual

- Highly semantic ontology
- Leads to high precision extractions
- Expensive to create
- Requires domain experts

[For an Architecture for Never-Ending Language Learning, Carlson et al. AAAI 2010]
Defining Domain: Semi-automatic

- Subset of types are manually defined
- SSL methods discover new types from unlabeled data

/Exploratory Learning, Dalvi et al., ECML 2013/
/Hierarchical Semi-supervised Classification with Incomplete Class Hierarchies, Dalvi et al., WSDM 2016/
Defining Domain: Semi-automatic

- Assume: Types and type hierarchy is manually defined
  E.g. River, City, Food, Chemical, Disease, Bacteria

- Relations are automatically discovered using clustering methods

<table>
<thead>
<tr>
<th>Discovered relation</th>
<th>Patterns</th>
<th>Seed instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>River -in heart of-City</td>
<td>“in heart of”</td>
<td>“Seine, Paris”, “Nile, Cairo”</td>
</tr>
<tr>
<td></td>
<td>“in the center of”</td>
<td>“Tiber river, Rome”</td>
</tr>
<tr>
<td></td>
<td>“which flows through”</td>
<td>“River arno, Florence”</td>
</tr>
<tr>
<td>Food -to produce-Chemical</td>
<td>“to produce”</td>
<td>“Salt, Chlorine”</td>
</tr>
<tr>
<td></td>
<td>“to make”</td>
<td>“Sugar, Carbon dioxide”</td>
</tr>
<tr>
<td></td>
<td>“to form”</td>
<td>“Protein, Serotonin”</td>
</tr>
<tr>
<td>Disease -caused by-Bacteria</td>
<td>“caused by”</td>
<td>“pneumonia, legionella”</td>
</tr>
<tr>
<td></td>
<td>“is the causative agent of”</td>
<td>“mastitis, staphylococcus aureus”</td>
</tr>
<tr>
<td></td>
<td>“is the cause of”</td>
<td>“gonorrhea, neisseria gonorrhoeae”</td>
</tr>
</tbody>
</table>

- Easier to derive types using existing resources
- Relations are discovered from the corpus
- Leads to moderate precision extractions
- Partially semantic ontology

[Discovering Relations between Noun Categories, Mohamed et al., EMNLP 2011]
Defining Domain: Automatic

- Any noun phrase is a candidate entity
- Any verb phrase is a candidate relation

- Cheapest way to induce types/relations from corpus
- Little expert annotations needed
- Limited semantics
- Leads to noisy extractions

[Open Information Extraction from the Web, Banko et al., IJCAI 2007]
Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring candidate facts

3 LEVELS OF SUPERVISION

Supervised

Semi-supervised

Unsupervised
Information Extraction

3 Concrete Sub-Problems

- Defining domain
- Learning extractors
- Scoring candidate facts

3 Levels of Supervision

- Supervised
- Semi-supervised
- Unsupervised
Learning Extractors: Manual

- Human defined high-precision extraction patterns for each relation

Person-member of-Band

<Person> works for <Band>
<Person> is part of <Band>

Extract relation instances
(John Lennon, The Beatles)
(Brian Jones, The Rolling Stones)
Information Extraction

3 CONCRETE SUB-PROBLEMS

- Defining domain
- **Learning extractors**
- Scoring candidate facts

3 LEVELS OF SUPERVISION

- Supervised
- Semi-supervised
- Unsupervised
Learning Extractors: Semi-supervised

Bootstrapping

- Seed instances
- Set of relation instances (I)
- Set of extraction patterns (P)
- Extract patterns that occur around relation instances in I
- Apply patterns in P to extract more relation instances
Learning Extractors: Semi-supervised

Person-member of-Band

Seed instances

Relation instances
(John Lennon, Beatles)
(Brian Jones, The Rolling Stones)

Learn patterns

<CễPERSON> works for <CễBAND>
<CễPERSON> is part of <CễBAND>
<CễBAND> includes <CễPERSON>

<CễBAND> was admired by <CễPERSON>

Semantic Drift!

Add top-k instances

Candidate facts
(Ringo Starr, The Beatles)
(Nick Mason, Pink Floyd)

Apply patterns

[Toward an Architecture for Never-Ending Language Learning, Carlson et al. AAAI 2010]
Learning Extractors: Interactive

1. Person-member of-Band
   - Seed instances

2. Relation instances
   - (John Lennon, Beatles)
   - (Brian Jones, The Rolling Stones)
   - Learn patterns
   - <PERSON> works for <BAND>
   - <PERSON> is part of <BAND>
   - <BAND> was invited by <PERSON>
   - <BAND>’s manager <PERSON>

3. Positive instances
   - (Nick Mason, Pink Floyd)
   - (Allen Klein, The Beatles)

4. Candidate facts
   - Apply correct patterns
   - Helps reduce semantic drift!

[Open information extraction to KBP relations in 3 hours, Soderland et al., TAC KBP 2013]
Information Extraction

3 Concrete Sub-Problems

- Defining domain
- Learning extractors
- Scoring candidate facts

3 Levels of Supervision

- Supervised
- Semi-supervised
- Unsupervised
Learning Extractors: Unsupervised

• Identify candidate relations:
  for each verb find the longest sequence of words
  s.t. syntactic and lexical constraints are satisfied

• Identify arguments for each relation:
  For each identified relation phrase r,
  find the closest noun-phrases on the left and right of r
  satisfying certain syntactic constraints

- Syntactic constraint
  Regular expressions of POS tags

- Lexical constraint
  |distinct arguments|
a relation phrase takes
Hudson was born in Hampstead, which is a suburb of London.

e1: (Hudson, was born in, Hampstead)
e2: (Hampstead, is a suburb of, London)
Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain
Learning extractors

Scoring candidate facts

3 LEVELS OF SUPERVISION

Supervised
Semi-supervised
Unsupervised
Scoring the candidate facts

- Human defined scoring function or Scoring function learnt using supervised ML with large amount of training data {expensive, high precision}

- Small amount of training data is available
  scoring refined over multiple iterations using both labeled and unlabeled data

- Completely automatic (Self-training)
  Confidence(extraction pattern) $\propto$ (#unique instances it could extract)
  Score(candidate fact) $\propto$ (#distinct extraction patterns that support it)
  {cheap, leads to semantic drift}
Impact of early supervision

Defining domain
- Extractors for each relation of interest
- Puts constraints on the space of possibly true extractions

Scoring the candidate facts

Domain expertise needed

- Enables inheritance and mutual exclusion at extractor level
- Early removal of noisy extraction pattern can avoid semantic drift in later stages
Effect of supervision on extractions

Precision, Human efforts

Recall, Speed
Information Extraction

3 IMPORTANT SUB-PROBLEMS

CATEGORIES OF IE TECHNIQUES

KNOWLEDGE FUSION

IE SYSTEMS IN PRACTICE
Categories of IE Techniques

1. Narrow domain patterns
2. Ontology based extraction
3. Interactive extraction
4. Open domain IE
5. Hybrid approach (Adding structure to OpenIE KB)
(1) Narrow domain patterns

Defining domain

Person headOf Organization

Arg1, DT CEO of Arg2

appos, det, nmod, case

Implies

Arg1 headOf Arg2

Learning extractors

Scoring candidate facts
(1) Narrow domain patterns

<table>
<thead>
<tr>
<th>Defining domain</th>
<th>Learning extractors</th>
<th>Scoring candidate facts</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Icon" /></td>
<td><img src="image2" alt="Icon" /></td>
<td><img src="image3" alt="Icon" /></td>
</tr>
</tbody>
</table>
(2) Ontology based extraction

Defining domain

Everything

Animals
- Disjoint
  - Mammals
  - Reptiles

Food
- Subset
  - Fruits
  - Vegetables

Animal-eats-Food
(2) Ontology based extraction

Bootstrapping

Ontological constraints

[Bootstrapping]

Instances (I) → Extract patterns → Patterns (P) → Apply patterns

[Ontological constraints]

Everything

Animals
  - Mammals
  - Reptiles

Food
  - Fruits
  - Vegetables

Disjoint
(2) Ontology based extraction

Coupled Bootstrap learning

[ Toward an Architecture for Never-Ending Language Learning, Carlson et al. AAAI 2010 ]
(2) Ontology based extraction

**Coupled Bootstrap learning**

- **Arg1 ISA Animal**
  - **Animal**
    - instances (I)
    - patterns (P)
    - Extract patterns
  - Apply patterns
- **Arg2 ISA Food**
  - **Food**
    - instances (I)
    - patterns (P)
    - Extract patterns
  - Apply patterns

**Animal eats Food**

[Toward an Architecture for Never-Ending Language Learning, Carlson et al. AAAI 2010]
(2) Ontology based extraction

• Self-training for scoring candidate facts
  • Confidence(extraction pattern) $\propto$ (#unique instances it could extract)
  • Score(candidate fact) $\propto$ (#distinct extraction patterns that support it)
(2) Ontology based extraction

<table>
<thead>
<tr>
<th>Defining domain</th>
<th>Learning extractors</th>
<th>Scoring candidate facts</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
</tbody>
</table>
(3) Interactive Extraction

Defining domain

Scoring candidate facts

Seed instances

Person-member of-Band

Relation instances
(John Lennon, Beatles)
(Brian Jones, The Rolling Stones)

Learn patterns

<PERSON> works for <BAND>
<PERSON> is part of <BAND>
<BAND> was invited by <PERSON>
<BAND>’s manager <PERSON>

Positive instances

Candidate instances
(Nick Mason, Pink Floyd)
(Allen Klein, The Beatles)

Apply correct patterns

Learning extractors

[ IKE - An Interactive Tool for Knowledge Extraction, Dalvi et al, AKBC 2015 ]
(3) Interactive Extraction

<table>
<thead>
<tr>
<th>Defining domain</th>
<th>Learning extractors</th>
<th>Scoring candidate facts</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Defining Domain" /></td>
<td><img src="image" alt="Learning Extractors" /></td>
<td><img src="image" alt="Scoring Candidate Facts" /></td>
</tr>
</tbody>
</table>
Can we do Web-scale IE?

1. Narrow domain patterns
2. Ontology based extraction
3. Interactive extraction

4. Open domain IE
5. Hybrid approach
   (Adding structure to OpenIE KB)

Assume expert input
Biased towards high precision
High costs
(4) Open domain IE

Open domain
any NP is a candidate entity
Any VP is a candidate relation

Defining domain

Learning extractors

Hudson was born in Hampstead, which is a suburb of London.

Scoring based on classifier
(features: POS tags, dependency parse ...)

Scoring candidate facts

Scoring (Hudson, was born in, Hampstead) : 0.88
Scoring (Hampstead, is a suburb of, London) : 0.9

[Identifying Relations for Open Information Extraction, Fader et al, EMNLP 2011]
(4) Open domain IE

<table>
<thead>
<tr>
<th>Defining domain</th>
<th>Learning extractors</th>
<th>Scoring candidate facts</th>
</tr>
</thead>
</table>
Pros and Cons of Open domain IE

• Open domain IE paradigm can be easily applied
  • on a large scale corpus
  • in a new domain (no training data)

• Main disadvantages
  • Poor aggregation
    Doesn’t detect different surface forms for same entity or relation
  • Lack of semantics
    OpenIE merely tells us how many times the lexical fact occurred in a corpus
(5) Hybrid approach (adding structure to Open IE KB)

[ Canonicalizing Open Knowledge Bases, Galárraga at al., CIKM 2014 ]
(5) Hybrid approach

- **Clustering entities**

- **Clustering relations**

Verb phrases: be an abbreviation-for, be known as, stand for, be an acronym for, be spoken in, be the official language of, be the national language of, be bought, acquire

Freebase relation: location.country.official_language, organization.organization.acquired_by
(5) Hybrid approach

[Discovering Semantic Relations from the Web and Organizing them with PATTY, SIGMOD 2013]
(5) Hybrid approach

<table>
<thead>
<tr>
<th>Defining domain</th>
<th>Learning extractors</th>
<th>Scoring candidate facts</th>
</tr>
</thead>
</table>

**Open domain IE** + **Distant supervision to add structure**
Categories of IE Techniques

1. Narrow domain patterns
   - Assume expert input
   - Biased towards high precision
   - High cost

2. Ontology based extraction
3. Interactive extraction
   - No expert annotations
   - Biased towards high recall
   - Low cost

4. Open domain IE
5. Hybrid approach
   - (Adding structure to OpenIE KB)
Information Extraction

3 IMPORTANT SUB-PROBLEMS
CATEGORIES OF IE TECHNIQUES
 KNOWLEDGE FUSION
IE SYSTEMS IN PRACTICE
Knowledge fusion

**Single extractor**

- Defining domain: Manual
- Learning extractors: Semi-automatic
- Scoring candidate facts: Automatic

**Fusing multiple extractors**
Multiple extractors

• **Extractor 1:** text patterns to extract ISA relations
e.g. coupled pattern learner

• **Extractor 2:** learning wrappers for HTML pages to extract ISA relations
  from structured text
Knowledge fusion schemes

- Voting (AND vs OR of extractors)
- Co-training (multiple extraction methods)
- Multi-view learning (multiple data sources)
- Classification
(1) Voting Schemes

• **AND of two extractors:**
  • For a candidate extraction to be promoted to a fact in KB, both the extractors should support the fact
  • \( \text{score}(\text{fact}) = \text{Min}(\text{score}_{\text{extractor1}}(\text{fact}), \text{score}_{\text{extractor2}}(\text{fact})) \)

• **OR of two extractors**
  • For a candidate extraction to be promoted to a fact in KB, both the extractors should support the fact
  • \( \text{score}(\text{fact}) = \text{Max}(\text{score}_{\text{extractor1}}(\text{fact}), \text{score}_{\text{extractor2}}(\text{fact})) \)

• **Hand-coded heuristic rules**
  • E.g. (at least one extractor has confidence > 0.9) or (two extractors support the fact with confidence > 0.6)
    .....
(2) Co-training

[ Combining Labeled and Unlabeled Data with Co-Training, Blum and Mitchell, CoLT 1998 ]
(3) Multi-view learning

- Task: Entity typing
- Each entity can be represented using two independent data views

**Entity:** Carnegie Mellon University

View-1: Text-patterns
- _arg1_ located in, 100
- are offered by _arg1_, 56
- _arg1_ has branches in, 23
- _arg1_ is renowned for, 15

View-2: HTML-tables
- doc04::2:1, 1
- doc10::1:2, 1
- doc17::1:1, 1

Text sentences:
- Carnegie Mellon University is located in Pittsburgh,
- Various masters and PhD programs in computer science are offered by Carnegie Mellon University,
- Online courses in machine learning are offered by Stanford University,
- Carnegie Mellon University has branches in Pittsburgh, Qatar and Sillicon Valley.

Doc04, table 2
- Column 1
  - Carnegie Mellon University
  - Stanford University
  - UMass Amherst
- Column 2
  - Pittsburgh, PA
  - Stanford, CA
  - Amherst, MA

[Multi-View Hierarchical Semi-supervised Learning by Optimal Assignment of Sets of Labels to Instances, Dalvi et al. in preparation, link]
(3) Multi-view learning

Maximize score of label assignment, Minimize disagreement between views

Update parameters per view

Extractor for View A

Extractor for View B

Instance labels

Multi-View Hierarchical Semi-supervised Learning by Optimal Assignment of Sets of Labels to Instances, Dalvi et al. in preparation, [link]
(4) Classification

Text documents (TXT)

HTML trees (DOM)

HTML Tables (TBL)

Per candidate fact per extractor features: # sources, Avg score ...

Classifier

$P(\text{candidate fact} = \text{true})$

[Dong, Xin et al. “Knowledge vault: a web-scale approach to probabilistic knowledge fusion.” KDD (2014)]
Knowledge fusion schemes

• Voting (AND vs OR of extractors)
• Co-training (multiple extraction methods)
• Multi-view learning (multiple data sources)
• Classification
Information Extraction

3 IMPORTANT SUB-PROBLEMS
CATEGORIES OF IE TECHNIQUES
KNOWLEDGE FUSION

IE SYSTEMS IN PRACTICE
IE systems in practice

• Conceptnet
• NELL
• Knowledge vault
• Open IE
ConceptNet is a freely-available semantic network, designed to help computers understand the meanings of words that people use.

This knowledge was derived from thousands of human contributors.
Never Ending Language Learning (NELL)

Read the Web
Research Project at Carnegie Mellon University

Knowledge Base (latent variables)
- Beliefs
- Candidate Beliefs

Knowledge Integrator

Text Context patterns (CPL)
- Actively search for web text (OpenEval)

Orthographic classifier (CML)
- Infer new beliefs from old (PRA)

URL specific HTML patterns (SEAL)
- Image classifier (NEIL)

Human advice
- Ontology extender (OntExt)

[Never-Ending Learning, Mitchell et al., AAAI 2015]
Knowledge Vault

Web

Extractors

Fusion

Priors

[Architecture diagram taken from Kevin Murphy’s slides]
Open IE (KnowItAll)

[Architecture diagram taken from Oren Etzioni's slides]
IE systems at a glance

<table>
<thead>
<tr>
<th>Defining domain</th>
<th>Learning extractors</th>
<th>Scoring candidate facts</th>
<th>Fusing extractors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# IE systems at a glance

<table>
<thead>
<tr>
<th></th>
<th>Defining domain</th>
<th>Learning extractors</th>
<th>Scoring candidate facts</th>
<th>Fusing extractors</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConceptNet</td>
<td><img src="icon1" alt="Icon" /></td>
<td><img src="icon2" alt="Icon" /></td>
<td><img src="icon3" alt="Icon" /></td>
<td></td>
</tr>
<tr>
<td>NELL</td>
<td><img src="icon4" alt="Icon" /></td>
<td><img src="icon5" alt="Icon" /></td>
<td><img src="icon6" alt="Icon" /></td>
<td>Heuristic rules</td>
</tr>
<tr>
<td>Knowledge Vault</td>
<td><img src="icon7" alt="Icon" /></td>
<td><img src="icon8" alt="Icon" /></td>
<td><img src="icon9" alt="Icon" /></td>
<td>Classifier</td>
</tr>
<tr>
<td>OpenIE</td>
<td><img src="icon10" alt="Icon" /></td>
<td><img src="icon11" alt="Icon" /></td>
<td><img src="icon12" alt="Icon" /></td>
<td></td>
</tr>
</tbody>
</table>
Tutorial Outline

1. Knowledge Graph Primer [Jay]

2. Knowledge Extraction from Text
   a. NLP Fundamentals [Sameer]
   b. Information Extraction [Bhavana]

   Coffee Break

3. Knowledge Graph Construction
   a. Probabilistic Models [Jay]
   b. Embedding Techniques [Sameer]

4. Critical Overview and Conclusion [Bhavana]
Thank You

SEE YOU AFTER THE COFFEE BREAK!