Tutorial Overview

Part 1: Knowledge Graphs

Part 2: Knowledge Extraction

Part 3: Graph Construction

Part 4: Critical Analysis

https://kgtutorial.github.io
Tutorial Outline

1. Knowledge Graph Primer [Jay]

2. Knowledge Extraction Primer [Jay]

Coffee Break

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

4. Critical Overview and Conclusion [Sameer]
Critical Overview

SUMMARY
SUCCESS STORIES
DATASETS, TASKS, SOFTWARES
EXCITING RESEARCH DIRECTIONS
Critical Overview

SUMMARY
SUCCESS STORIES
DATASETS, TASKS, SOFTWARES
EXCITING RESEARCH DIRECTIONS
Why do we need Knowledge graphs?

• Humans can explore large database in intuitive ways

• AI agents get access to human common sense knowledge
Knowledge graph construction

• **Who** are the entities (nodes) in the graph?

• **What** are their attributes and types (labels)?

• **How** are they related (edges)?
Knowledge Graph Construction

1. Text
2. Knowledge Extraction
3. Extraction graph
4. Graph Construction
5. Knowledge graph
## Two perspectives

<table>
<thead>
<tr>
<th></th>
<th>Extraction graph</th>
<th>Knowledge graph</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Who are the entities?</strong></td>
<td>• Named Entity Recognition</td>
<td>• Entity Linking</td>
</tr>
<tr>
<td><strong>(nodes)</strong></td>
<td>• Entity Coreference</td>
<td>• Entity Resolution</td>
</tr>
<tr>
<td><strong>What are their attributes?</strong></td>
<td>• Entity Typing</td>
<td>• Collective classification</td>
</tr>
<tr>
<td><strong>(labels)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>How are they related?</strong></td>
<td>• Semantic role labeling</td>
<td>• Link prediction</td>
</tr>
<tr>
<td><strong>(edges)</strong></td>
<td>• Relation Extraction</td>
<td></td>
</tr>
</tbody>
</table>
John was born in Liverpool, to Julia and Alfred Lennon.
Information Extraction

Single extractor

Defining domain
Learning extractors
Scoring candidate facts

Supervised
Semi-supervised
Unsupervised

Fusing multiple extractors
Knowledge Graph Construction

Part 2: Knowledge Extraction → Extraction graph → Part 3: Graph Construction

Knowledge graph
Issues with Extraction Graph

Extracted knowledge could be:

• ambiguous

• incomplete

• inconsistent
Two approaches for KG construction

PROBABILISTIC MODELS

EMBEDDING BASED MODELS
Two approaches for KG construction

PROBABILISTIC MODELS
EMBEDDING BASED MODELS
# Two classes of Probabilistic Models

## Graphical Model Based
- Possible facts in KG are variables
- Logical rules relate facts
- Probability $\propto$ satisfied rules
- Universal-quantification

## Random Walk Based
- Possible facts posed as queries
- Random walks of the KG constitute “proofs”
- Probability $\propto$ path lengths/transitions
- Local grounding
Illustration of KG Identification

**Uncertain Extractions:**
- .5: Lbl(Fab Four, novel)
- .7: Lbl(Fab Four, musician)
- .9: Lbl(Beatles, musician)
- .8: Rel(Beatles, AlbumArtist, Abbey Road)

**Ontology:**
- Dom(albumArtist, musician)
- Mut(novel, musician)

**Entity Resolution:**
- SameEnt(Fab Four, Beatles)

(Annotated) Extraction Graph

After Knowledge Graph Identification
Random Walk Illustration

Query: R(Lennon, PlaysInstrument, ?)
Two approaches for KG construction

PROBABILISTIC MODELS

EMBEDDING BASED MODELS
Why embeddings?

Limitations of probabilistic models

- Representation restricted by manual design
  - Clustering? Asymmetric implications?
  - Information flows through these relations
  - Difficult to generalize to unseen entities/relations

Limitation to Logical Relations

- Representation restricted by manual design
  - Clustering? Asymmetric implications?
  - Information flows through these relations
  - Difficult to generalize to unseen entities/relations

Computational Complexity of Algorithms

- Learning is NP-Hard, difficult to approximate
- Query-time inference is also NP-Hard
- Not easy to parallelize, or use GPUs
- Scalability is badly affected by representation

Embedding based models

- Can generalize to unseen entities and relations
- Efficient inference at large scale
Relation Embeddings
Part 1: Knowledge Graphs

Part 2: Knowledge Extraction

Part 3: Graph Construction
Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING RESEARCH DIRECTIONS
Success stories

Open Information Extraction

ConceptNet
An open, multilingual knowledge graph

YAGO

NELL Knowledge Base Browser
CMU Read the Web Project

DeepDive
v0.8.0
Think about features, not algorithms.
Success story: OpenIE (ReVerb)

Open Information Extraction

Argument 1: entity: The Beatles
Argument 2: 

Relation: 

All 

Search

were bigger than Jesus (100)
came to America (95)
appeared on The Ed Sullivan Show (88)
broke up in 1970 (56)
Here Comes the Sun (46)
came to America (45)
is for the future (44)
are a great band (42)

perform on The Ed Sullivan Show (39)

were Musical ensemble (36)

are a great band »

Extracted Synonyms:

were
is
was

Extracted from these sentences:

are The Beatles are the best band, hands down but Oasis did make a great cover. (via ClueWeb12)
The Beatles are a great band. (via ClueWeb12)
The Beatles are the best band. (via ClueWeb12)
The Beatles are the greatest band ... Started 1 month ago by georgedcc Yeah, Songs in the Key of Life is a bit much for 1 listen. (via ClueWeb12)
The Beatles, arguably, are the greatest band, and may or may not have the greatest name. (via ClueWeb12)
The point is, from my view, The Beatles are a good band, but way behind the greatest artists to ever grace rock. (via ClueWeb12)
Success story: **NELL**

**beatles (musicartist)**

**literal strings:** BEATLES, Beatles, beatles

**Help NELL Learn!**

NELL wants to know if these beliefs are correct.
If they are or ever were, click thumbs-up. Otherwise, click thumbs-down.

- **beatles** is a musical artist
- **beatles** is a musician in the genre classic pop (musicgenre)
- **beatles** is a musician in the genre pop (musicgenre)
- **beatles** is a musician in the genre rock (musicgenre)

**categories**

- musicartist(100.0%)
Success story: **YAGO**

- **Input:** Wikipedia infoboxes, WordNet and GeoNames
- **Output:** KG with 350K entity types, 10M entities, 120M facts
- Temporal and spatial information
beatles
An English term in ConceptNet 5.5

Derived terms
- en beatle
- en beatledom
- en beatlemania
- en beatlesque
- en fourth beatle

beatles is a type of...
- a British band
- man (n)
- band (n)
- musician (n)
- album (n)

Links to other sites
- dbpedia.org The Beatles
- sw.opencyc.org Beatle
- umbel.org Beatle
- wordnet-rdf.princeton.edu 400520405-N
- wordnet-rdf.princeton.edu 108386847-n
- wikidata.dbpedia.org Q1299
- en.wiktionary.org Beatles
- dbpedia.org The Beatles (No. 1)
- wikidata.dbpedia.org Q738260
- fr.wiktionary.org Beatles
- dbpedia.org The Beatles (The Original Studio Recordings)
- wikidata.dbpedia.org Q603122
Success story

- **DBPedia** is automatically extracted structured data from Wikipedia
  - 17M canonical entities
  - 88M type statements
  - 72M infobox statements
DeepDive

- Best Precision/recall/F1 in KBP-slot filling task 2014 evaluations (31 teams participated)
## IE systems in practice

<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>ConceptNet</td>
<td><img src="image" alt="ConceptNet" /></td>
<td><img src="image" alt="ConceptNet" /></td>
<td></td>
</tr>
<tr>
<td>NELL</td>
<td><img src="image" alt="NELL" /></td>
<td><img src="image" alt="NELL" /></td>
<td>Heuristic rules</td>
</tr>
<tr>
<td>Knowledge Vault</td>
<td><img src="image" alt="Knowledge Vault" /></td>
<td><img src="image" alt="Knowledge Vault" /></td>
<td>Classifier</td>
</tr>
<tr>
<td>OpenIE</td>
<td><img src="image" alt="OpenIE" /></td>
<td><img src="image" alt="OpenIE" /></td>
<td></td>
</tr>
</tbody>
</table>
Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING RESEARCH DIRECTIONS
Datasets

- **KG as datasets**
  - FB15K-237: Knowledge base completion dataset based on Freebase
  - DBPedia: Structured data extracted from Wikipedia
  - NELL: Read the web datasets
  - AristoKB: Tuple knowledge base for Science domain

- **Text datasets**
  - Clueweb09: 1 billion webpages (sample of Web)
  - FACC1: Freebase Annotations of the Clueweb09 Corpora
  - Gigaword: automatically-generated syntactic and discourse structure
  - NYTimes: The New York Times Annotated Corpus

- **Datasets related to Semi-supervised learning for information extraction**
  Link: entity typing, concept discovery, aligning glosses to KB, multi-view learning

---

Shared tasks

- Text Analysis Conference on Knowledge Base Population (TAC KBP)
  - Slot filling task
  - Cold Start KBP Track
  - Tri-Lingual Entity Discovery and Linking Track (EDL)
  - Event Track
  - Validation/Ensembling Track
Software: NLP

• Stanford CoreNLP: a suite of core NLP tools [link] (Java code)

• FIGER: fine-grained entity recognizer assigns over 100 semantic types [link] (Java code)

• FACTORIE: out-of-the-box tools for NLP and information integration [link] (Scala code)

• EasySRL: Semantic role labeling [link] (Java code)
Software: Extracting and Reasoning

- **Open IE** (University of Washington)
  Open IE 4.2 [link](#) (Scala code)
  Stanford Open IE [link](#) (Java code)

- **Interactive Knowledge Extraction (IKE)** (Allen Institute for Artificial Intelligence)
  [link](#) (Scala code)

- **PSL**: Probabilistic soft logic
  [link](#) (Java code)

- **ProPPR**: Programming with Personalized PageRank
  [link](#) (Java code)
Critical Overview

SUMMARY
SUCCESS STORIES
DATASETS, TASKS, SOFTWARES
EXCITING RESEARCH DIRECTIONS
Exciting Active Research

• INTERESTING APPLICATIONS OF KG
• MULTI-MODAL INFORMATION EXTRACTION
• KNOWLEDGE AS SUPERVISION
• COMMON KNOWLEDGE
Exciting Active Research

• INTERESTING APPLICATIONS OF KG
• MULTI-MODAL INFORMATION EXTRACTION
• KNOWLEDGE AS SUPERVISION
• COMMON KNOWLEDGE
Interesting application of Knowledge Graphs

The Literome Project [link]

- Automatic system for extracting genomic knowledge from PubMed articles
- Web-accessible knowledge base
Interesting application of Knowledge Graphs

Chronic disease management:
develop AI technology for predictive and preventive personalized medicine to reduce the national healthcare expenditure on chronic diseases (90% of total cost)
Exciting Active Research

• INTERESTING APPLICATIONS OF KG
• MULTI-MODAL INFORMATION EXTRACTION
• KNOWLEDGE AS SUPERVISION
• COMMON KNOWLEDGE
Knowledge Base Completion

\[ \phi(s, r, o) \]

<table>
<thead>
<tr>
<th>Model</th>
<th>Score ( \psi_r(e_s, e_o) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESCAL [21]</td>
<td>( e_s^T W_r e_o )</td>
</tr>
<tr>
<td>SE [3]</td>
<td>(</td>
</tr>
<tr>
<td>TransE [1]</td>
<td>(</td>
</tr>
<tr>
<td>DistMult [34]</td>
<td>( \langle e_s, r, e_o \rangle )</td>
</tr>
<tr>
<td>ComplEx [33]</td>
<td>( \langle e_s, r, e_o \rangle )</td>
</tr>
<tr>
<td>ConvE</td>
<td>( f(\text{vec}(f([e_s; r] \ast \omega)) W) e_o )</td>
</tr>
</tbody>
</table>

Table from Dettmers, et al. (2017)
Multimodal KB Embeddings

Object \rightarrow \text{Encoder} \rightarrow \Phi(s, R, o)

Entity \rightarrow \text{Lookup}
Images \rightarrow \text{CNN}
Text \rightarrow \text{LSTM}
Numbers, etc. \rightarrow \text{FeedFwd}
Exciting Active Research

• INTERESTING APPLICATIONS OF KG
• MULTI-MODAL INFORMATION EXTRACTION
• KNOWLEDGE AS SUPERVISION
• COMMON KNOWLEDGE
Knowledge as Supervision

- True: $\text{spouseOf}(\text{Barack}, \text{Michelle})$

Problem 1: Each annotation takes time
Problem 2: Each annotation is a drop in the ocean

- True: $X$ husband of $Y$ => $\text{spouseOf}(X,Y)$

Many different options
- Generalized Expectation
- Posterior Regularization
- Labeling functions in SNORKEL
Exciting Active Research

• INTERESTING APPLICATIONS OF KG
• MULTI-MODAL INFORMATION EXTRACTION
• KNOWLEDGE AS SUPERVISION
• COMMON KNOWLEDGE
Frogs lay eggs that develop into tadpoles and then into adult frogs. This sequence of changes is an example of how living things _____

(A) go through a life cycle
(B) form a food web
(C) act as a source of food
(D) affect other parts of the ecosystem
Future KG construction system

Corrects its own mistakes

Consume online streams of data

Supports humanity

Represent context beyond facts

Future......
Thank You

Jay Pujara
jaypujara.org
jay@cs.umd.edu
@jay_mlr

Sameer Singh
sameersingh.org
sameer@uci.edu
@sameer_
## Two perspectives

<table>
<thead>
<tr>
<th></th>
<th>Extraction graph</th>
<th>Knowledge graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who are the entities?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(nodes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What are their</td>
<td></td>
<td></td>
</tr>
<tr>
<td>attributes? (labels)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How are they related?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(edges)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
John was born in Liverpool, to Julia and Alfred Lennon.
NLP annotations ➔ features for IE

Combine tokens, dependency paths, and entity types to define rules.

Bill Gates, the CEO of Microsoft, said ...
Mr. Jobs, the brilliant and charming CEO of Apple Inc., said ...
... announced by Steve Jobs, the CEO of Apple.
... announced by Bill Gates, the director and CEO of Microsoft.
... mused Bill, a former CEO of Microsoft.
and many other possible instantiations...
Success story: OpenIE

- **Key contributions:**
  - No need for human defined relation schemas
  - First ever successful open-source open domain IE system

- **ReVerb**
  - Input = [Clueweb09 corpus](#) (1B web pages)
  - Output = 15M high-precision extractions
Open IE Systems

---|---|---|---|---
OpenIE v 1.0 | v 2.0 | v 3.0 | OpenIE 4.0 | OpenIE 5.0
TextRunner | ReVerb | OLLIE |
CRF | POS-tag | Dependency | SRL-based | Supports
Self- | based | parse based | extraction; | compound
training | relation | extraction | temporal, | noun
| extraction | | | spatial | phrases;
| | | | extractions | numbers;
| | | | | lists

Increase in precision, recall, expressiveness

Derived from Prof. Mausam’s slides
Success story: **NELL**

- **Key technical contributions:**
  - “Never ending learning” paradigm
  - “Coupled bootstrap learning” to reduce semantic drift

- Input: [Clueweb09 corpus](#) (1B web pages)

- Ontology: ~2K predicates
  - $O(100K)$ constraints between predicates

- Output: 50 million candidate facts
  - 3 million high-confidence facts
Success story: **YAGO**

**Key contributions:**
- **Rich Ontology:** Linking Wikipedia categories to WordNet
- **High Quality:** High precision extractions (~95%)
Success story: **ConceptNet**

- Commonsense knowledge base

**Key contributions:**
- **Freely available resource:** covers wide range of common sense concepts and relations organized in an easy-to-use semantic network
- **NLP toolkit based on this resource:** supports analogy, text summarization, context dependent inferences

- ConceptNet4 was manually built using inputs from thousands of people
  - 28 million facts expressed in natural language
  - spanning 304 different languages
DeepDive

• Machine learning based extraction system

• Key contributions:
  • scalable, high-performance inference and learning engine
  • Developers contribute features (rules) not algorithms
  • Combines data from variety of sources (webpages, pdf, figures, tables)
Future......
Aristo ScienceKB

- AI2’s TupleKB dataset: [link](#)

- **Open problems**
  - Best KR for Science domain
  - Domain targeted KB completion
  - Measuring recall w.r.t. end task
(1) Future research directions: Going beyond facts

- Most of the existing KGs are designed to represent and extract binary relations → good enough for search engines

- Applications like QA demand in depth knowledge about higher level structures like activities, events, processes
(2) Future research directions: Online KG Construction

- One shot KG construction → Online KG construction
  - Consume online stream of data
  - Temporal scoping of facts
  - Discovering new concepts automatically
  - Self-correcting systems
(2) Future research directions: Online KG Construction

• **Continuously learning and self-correcting systems**
  • [Selecting Actions for Resource-bounded Information Extraction using Reinforcement Learning, Kanani and McCallum, WSDM 2012]
    • Presented a reinforcement learning framework for budget constrained information extraction
  
  • [Never-Ending Learning, Mitchell et al. AAAI 2015]
    • Tom Mitchell says “Self reflection and an explicit agenda of learning subgoals” is an important direction of future research for continuously learning systems.
Existing knowledge graphs

- Too named entity centric (no domain relevance)
- Too noisy (not directly usable by inference systems)
**Upcoming article on “High Precision Knowledge Extraction for Science domain”**
**Upcoming article on ``High Precision Knowledge Extraction for Science domain’’**

AI2’s ScienceKB

Hybrid Approach: Adding structure to Open domain IE

<table>
<thead>
<tr>
<th>Defining domain</th>
<th>Learning extractors</th>
<th>Scoring candidate facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open domain IE</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Distant supervision to add structure

AI2’s TupleKB dataset: [link](#)

- > 300K common-sense and science facts
- > 80% precision
Future research directions: Going beyond facts

- **Fact:** Individual knowledge tuples (plant, take in, CO2)

- **Event frame:** more context how, when, where?

- **Processes:** representing larger structures, sequence of events e.g. Photosynthesis
(3) Exciting active research:  
Ambitious Project

The Allen AI Science Challenge

Is your model smarter than an 8th grader?

$80,000 · a year ago

<table>
<thead>
<tr>
<th>#</th>
<th>Δ1w</th>
<th>Team Name 🌟 in the money</th>
<th>Kernel</th>
<th>Team Members</th>
<th>Score</th>
<th>Entries</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—</td>
<td>Alejandro Mosquera</td>
<td></td>
<td></td>
<td>0.59375</td>
<td>2</td>
<td>1y</td>
</tr>
<tr>
<td>2</td>
<td>new</td>
<td>Cardal</td>
<td></td>
<td></td>
<td>0.59000</td>
<td>2</td>
<td>1y</td>
</tr>
<tr>
<td>3</td>
<td>new</td>
<td>poweredByTalkwalker</td>
<td></td>
<td></td>
<td>0.59000</td>
<td>4</td>
<td>1y</td>
</tr>
</tbody>
</table>
(2) Exciting active research: Multi-modal information extraction

Text + Images → Multi-modal Knowledge Graph
NEIL: Extracting Visual Knowledge from Web Data

[Chen et al., "NEIL: Extracting Visual Knowledge from Web Data," ICCV 2013]
NEIL: Extracting Visual Knowledge from Web Data

Learned facts:
- Monitor is a part of Desktop Computer
- Keyboard is a part of Desktop Computer
- Television looks similar to Monitor

[Chen et al., "NEIL: Extracting Visual Knowledge from Web Data," ICCV 2013]
WebChild: Text + Images

[Tandon et al. “Commonsense in Parts: Mining Part-Whole Relations from the Web and Image Tags.” AAAI ’16]
Knowledge Base Completion

Entity Prediction

Link Prediction
Restrictions in the Model

Each object has a vector representation:
- Limits number of objects
- Large number of parameters
- Is not compositional (doesn’t generalize)

What about other kinds of objects?
- Dates and Numbers: should generalize
- Text: Names and Descriptions
- Images: Portraits, Posters, etc.
Multimodal KB Embeddings

Object → Encoder → 

Entity → Lookup → 

Images → CNN → 

Text → LSTM → 

Numbers, etc. → FeedFwd → 

$v_0$
Augmenting Existing Datasets

<table>
<thead>
<tr>
<th>MovieLens-100k-plus</th>
<th>YAGO3-10-plus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relations</td>
<td>13</td>
</tr>
<tr>
<td>Users</td>
<td>943</td>
</tr>
<tr>
<td>Movies</td>
<td>1682</td>
</tr>
<tr>
<td>Posters</td>
<td>1651</td>
</tr>
<tr>
<td>Ratings</td>
<td>100,000</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>