Constructing and Mining Web-scale Knowledge Graphs

Evgeniy Gabrilovich (Google)
gabr@google.com

Nicolas Usunier (Facebook)
usunier@fb.com
The opinions expressed herein are the sole responsibility of the tutorial instructors and do not necessarily reflect the opinion of Facebook Inc. or Google Inc.

Technologies described might or might not be in actual use.
Acknowledgements

Thanks to Luna Dong, Matthew Gardner, Ni Lao, Kevin Murphy, Antoine Bordes, and Jason Weston for fruitful discussions that helped improve this tutorial.

Special thanks to Philip Bohannon and Rahul Gupta for letting us use their slides on entity deduplication and relation extraction.
Outline of the tutorial

PART 1: Knowledge graphs
1. Applications of knowledge graphs
2. Freebase as an example of a large scale knowledge repository
3. Research challenges
4. Knowledge acquisition from text

PART 2: Methods and techniques
1. Relation extraction
2. Entity resolution
3. Link prediction
PART 1: KNOWLEDGE GRAPHS
The role of knowledge

• “Knowledge is Power” Hypothesis (the Knowledge Principle): “If a program is to perform a complex task well, it must know a great deal about the world in which it operates.”

• The Breadth Hypothesis: “To behave intelligently in unexpected situations, an agent must be capable of falling back on increasingly general knowledge.”

Lenat & Feigenbaum
Artificial Intelligence 47 (1991)
“On the Threshold of Knowledge”
Why (knowledge) graphs?

- We’re surrounded by **entities**, which are connected by **relations**
- We need to store them somehow, e.g., using a **DB** or a **graph**
- **Graphs** can be processed **efficiently** and offer a convenient abstraction
Knowledge graphs

Microsoft’s Satori

Facebook’s Entity Graph

Google’s Knowledge Graph

OpenIE (Reverb, OLLIE)

Freebase, Wikidata, Yago, DBpedia, NELL: Never-Ending Language Learning
A sampler of research problems

- **Growth**: knowledge graphs are incomplete!
  - *Link prediction*: add relations
  - *Ontology matching*: connect graphs
  - *Knowledge extraction*: extract new entities and relations from web/text

- **Validation**: knowledge graphs are not always correct!
  - *Entity resolution*: merge duplicate entities, split wrongly merged ones
  - *Error detection*: remove false assertions, find stale facts

- **Interface**: how to make it easier to access knowledge?
  - *Semantic parsing*: interpret the meaning of queries
  - *Question answering*: compute answers using the knowledge graph

- **Intelligence**: can AI emerge from knowledge graphs?
  - *Automatic reasoning* and planning
  - Generalization and abstraction
A sampler of research problems

• **Growth**: knowledge graphs are incomplete!
  - *Link prediction*: add relations
  - *Ontology matching*: connect graphs
  - *Knowledge extraction*: extract new entities and relations from web/text

• **Validation**: knowledge graphs are not always correct!
  - *Entity resolution*: merge duplicate entities, split wrongly merged ones
  - *Error detection*: remove false assertions

• **Interface**: how to make it easier to access knowledge?
  - *Semantic parsing*: interpret the meaning of queries
  - *Question answering*: compute answers using the knowledge graph

• **Intelligence**: can AI emerge from knowledge graphs?
  - *Automatic reasoning* and planning
  - Generalization and abstraction
A SAMPLER OF APPLICATIONS OF KNOWLEDGE GRAPHS
Surfacing structured results in web search

Augmenting the presentation with relevant facts
Surfacing facts proactively

Google search for "Italy population"

Google search for "Italy"

Results for Italy population:
- Italy Population (2016): 59.83 million
- Italian Tourism Official Website
- Italy - Wikipedia, the free encyclopedia

Results for Italy:
- Country in Europe
- Capital: Rome
- Currency: Euro
- Points of Interest: Colosseum, Lake Garda, Leaning Tower of Pisa, Amalfi Coast, Lake Como
Exploratory search

Google search results for "pisa sightseeing"
Question answering

Google

EVI (Amazon)

Siri (Apple)
Answering implicit questions

Here is a condition – what does it mean?

Here are some symptoms – what do they indicate?

Talk at the WebQA workshop
“Cura Te Ipsum: answering symptom queries with question intent”
14:00 on Thursday, July 21
Patellar tendinitis

- Cause:
  - Frequent jumping

- Symptom:
  - Knee pain
  - Swollen knee

- Treatment:
  - NSAID
  - Analgesic
Forthcoming afternoon tutorial today

“Question Answering with Knowledge Base, Web and Beyond”

Scott Wen-tau Yih (Microsoft Research)

Hao Ma (Microsoft Research)

TODAY: Sunday, July 17, 14:00 – 17:30

Room C
Connecting people, places and things
Connecting people, places and things

Structured search within the graph
Towards a knowledge-powered digital assistant

- Natural way of accessing/storing knowledge
- Dialogue system
- Personalization
- Emotion
FREEBASE AS AN EXAMPLE OF A LARGE SCALE KNOWLEDGE REPOSITORY
- **Freebase** is an open, [Creative Commons](https://creativecommons.org) licensed repository of **structured data**
- **Typed entities** rather than **strings**

Relations are typed too!
The world changes, but we don’t retract facts

We just add more facts!

Marriage Mediator Type

Key: /people/marriage

'Marriage' defines a relationship between two people. The person type uses it to store the two people in the relationship as well as a beginning and end date.

More

Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>ID</th>
<th>Expected Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spouse</td>
<td>/people/marriage/spouse</td>
<td>/people/person</td>
</tr>
<tr>
<td>From</td>
<td>/people/marriage/from</td>
<td>/type/datetime</td>
</tr>
<tr>
<td>To</td>
<td>/people/marriage/to</td>
<td>/type/datetime</td>
</tr>
<tr>
<td>Type of union</td>
<td>/people/marriage/type_of_union</td>
<td>/people/marriage_union_type</td>
</tr>
<tr>
<td>Location of ceremony</td>
<td>/people/marriage/location_of_ceremony</td>
<td>/location/location</td>
</tr>
</tbody>
</table>
A graph of inter-related objects
Schema limitations

What is the capital of South Africa
Schema limitations (cont’d)

Who is the father of Lil Wayne

About 25,800,000 results (0.32 seconds)

Lil Wayne
Rapper
Dwayne Michael Carter, Jr., known by his stage name Lil Wayne, is an American rapper from New Orleans, Louisiana. Wikipedia
Subject-Predicate-Object (SPO) triples in Freebase

/m/04m8, /m/019xz9>

/people/person/place_of_birth

Place of birth

/en/albert_einstein
Albert Einstein

/en/ulm
Ulm

YAGO2 uses SPOTL tuples (SPO + Time and Location)
Freebase transition to WikiData

- Freebase is read-only since March ‘15, data in transition to WikiData
  
  http://dumps.wikimedia.org/wikidatawiki/

- WikiData: 18.5M data items as of June 30, 2016
- Freebase: 47.5M entities, 2.9B facts

- Different requirements for citation and sourcing of facts
  - **Unique** (each item represents a clearly identifiable)
  - **Notable** (corresponding page on a Wikimedia site such as Wikipedia)
  - **Statements** do not necessarily encode true facts, but rather **claims**, which have verifiable **sources**
“From Freebase to WikiData: The Great Migration”,
Tanon et al., WWW 2016

Using Google Search and the Knowledge Vault to match Freebase assertions to potential citations

Slides:
AUTOMATIC KNOWLEDGE ACQUISITION FROM TEXT USING AN EXISTING KNOWLEDGE BASE
Knowledge discovery: Man vs. Machine

Relying on humans
- Volunteer contributions at Freebase.com
- Import of large datasets
  (e.g., EDGAR, MusicBrainz)

Automatic extraction
- Extraction from web pages
  Using an existing knowledge base!

http://www.flickr.com/photos/sandreli/4691045841/
Output (representation)

Structured (schema-ful)

Semi-structured (fixed lexicon + free text)

Unstructured (schema-less)

Input

Knowledge Vault

DeepDive

NELL / NEIL

Open IE

Unstructured

Semi-structured

Structured

Humans

Knowledge Graph Construction

Structured Unstructured Semi-structured

Open IE

Semi-structured (fixed lexicon + free text)

DeepDive

Human

Structured (schema-ful)

Knowledge Vault

The diagram illustrates the conversion of input data into structured and semi-structured output, with various tools and methods used for different types of input data.
Open IE – extracting **unstructured** facts from **unstructured** sources (text)

- **TextRunner** [Banko et al., IJCAI ’07], **WOE** [Wu & Weld, ACL ‘10]

- Limitations
  1. Incoherent extractions – the system makes independent decisions whether to include each word in the relation phrase, possibly gluing together unrelated words
  2. Uninformative extractions – those omitting critical information (e.g., “has” instead of “has a population of” or “has a Ph.D. in”)

- **ReVerb** [Fader et al., EMNLP ‘11] solves these problems by adding syntactic constraints
  - Every multi-word relation phrase must begin with a verb, end with a preposition and be a contiguous sequence of words
  - Relation phrases should not omit nouns
  - Minimal number of distinct argument pairs in a large corpus
OLLIE: Open Language Learning for Information Extraction

[Mausam et al., EMNLP ‘12]

Limitations of ReVerb

• Only extracts relations mediated by verbs
• Ignores context, potentially extracting facts that are not asserted

1. “After winning the Superbowl, the Saints are now the top dogs of the NFL.”
   O: (the Saints; win; the Superbowl)
2. “There are plenty of taxis available at Bali airport.”
   O: (taxis; be available at; Bali airport)
3. “Microsoft co-founder Bill Gates spoke at ...”
   O: (Bill Gates; be co-founder of; Microsoft)
4. “Early astronomers believed that the earth is the center of the universe.”
   R: (the earth; be the center of; the universe)
   W: (the earth; be; the center of the universe)
   O: ((the earth; be the center of; the universe) AttributedTo believe; Early astronomers)
5. “If he wins five key states, Romney will be elected President.”
   R,W: (Romney; will be elected; President)
   O: ((Romney; will be elected; President) ClausalModifier if; he wins five key states)
OLLIE (cont’d)

Uses ReVerb to build a seed set.

Assumption: every relation can also be expressed via a verb-based expression

Uses dependency parse structure
Extracting **structured** facts from **unstructured** sources (text)

**/en/tom_cruise**

Thomas Cruise Mapother IV (born July 3, 1962), widely known as Tom Cruise,

is an American film actor and producer.

**/en/united_states**

**/people/person/nationality**

**/people/person/profession**

**/common/topic/alias**

**/people/person/date_of_birth**

**/en/actor**

**/en/film_producer**
Knowledge Fusion

Extractors

Graph priors

Fusion

Cf. Knowledge Graph Identification [Pujara et al., ISWC ‘13]

[with Dong et al., KDD 2014]
Distant supervision [Mintz et al. ‘09]

“Boy rescued from **Thames** after **jumping** from **Westminster Bridge**”

“Man’s life saved after he **plunged** into **Thames** from **Westminster Bridge**”

“Man **jumps** into **Thames** to save woman in front of London Eye tourists, near **Westminster Bridge**”

“Two people evading police **jump** into **Connecticut River** from **North End Bridge**; search ongoing”

<en/westminster_bridge, /.../body_of_water_spanned, /en/river_thames>

< North End Bridge, /.../body_of_water_spanned, Connecticut River >
Knowledge Fusion

Extractors

Graph priors

Fusion

[with Dong et al., KDD 2014]
Link prediction / matrix completion

Compute $\Pr(x_{ijk} = 1 \mid \text{Graph})$:

“A Review of Relational Machine Learning for Knowledge Graphs”

with Nickel, Murphy, and Tresp,

Proceedings of the IEEE, 104(1):11-33, 2016

Special issue on Big Data
Knowledge Fusion

- Webmaster annotations
- NL text
- Page structure
- Tables
- Deep neural network
- Path Ranking Algorithm (PRA)

Extractors

Graph priors

Fusion

[with Dong et al., KDD 2014]
Knowledge Fusion

Extractors

Graph priors

Fusion

(http://) Schema.org

[Mintz et al., RANLP ‘09]

[Magarelli et al., VLDB ‘12]

[Cafarella et al., CACM ‘11]

[Franz et al., ISWC ‘09;
Bordes et al., AI/Stats ‘12;
Drumond et al., SAC ‘12;
Socher et al., NIPS ‘13]

[Lao et al.,
EMNLP ‘11]
(PRA)

[Mintz et al.,
RANLP ‘09]

[Magarelli et al.,
VLDB ‘12]

[Cafarella et al.,
CACM ‘11]

[Franz et al.,
ISWC ‘09;
Bordes et al.,
AI/Stats ‘12;
Drumond et al.,
SAC ‘12;
Socher et al.,
NIPS ‘13]

[Lao et al.,
EMNLP ‘11]
(PRA)

[with Dong et al., KDD 2014]
Fusing multiple extractors

![Graph showing true positive rate against false positive rate for different extractors, with ANO (0.73), TBL (0.76), TXT (0.85), DOM (0.87), and FUSED-EX (0.92).]
The importance of adding more evidence
The yield from different extraction systems

![Graph showing yield from different extraction systems](image)
Fusing extractors with priors
Example:

<Barry Richter, studiedAt, UW-Madison>

“In the fall of 1989, Richter accepted a scholarship to the University of Wisconsin, where he played for four years and earned numerous individual accolades ...”

“The Polar Caps' cause has been helped by the impact of knowledgeable coaches such as Andringa, Byce and former UW teammates Chris Tancill and Barry Richter.”

- Fused extraction confidence: 0.14

- Freebase

<Barry Richter, born in, Madison>
<Barry Richter, lived in, Madison>

- Final belief (fused with prior): 0.61
The importance of prior modeling

![Graph showing the importance of prior modeling with a 2x increase in probability towards higher predicted probabilities.](image-url)
Comparison of knowledge repositories

Total # facts in > 2.5B

<table>
<thead>
<tr>
<th>Name</th>
<th># Entity types</th>
<th># Entity instances</th>
<th># Relation types</th>
<th># Confident facts (relation instances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Vault (KV)</td>
<td>1100</td>
<td>45M</td>
<td>4469</td>
<td>302M</td>
</tr>
<tr>
<td>DeepDive [32]</td>
<td>4</td>
<td>2.7M</td>
<td>34</td>
<td>7M&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>NELL [8]</td>
<td>271</td>
<td>5.19M</td>
<td>306</td>
<td>0.435M&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>PROSPERA [30]</td>
<td>11</td>
<td>N/A</td>
<td>14</td>
<td>0.1M</td>
</tr>
<tr>
<td>YAGO2 [19]</td>
<td>350,000</td>
<td>9.8M</td>
<td>100</td>
<td>4M&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Freebase [4]</td>
<td>1,500</td>
<td>40M</td>
<td>35,000</td>
<td>637M&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>Knowledge Graph (KG)</td>
<td>1,500</td>
<td>570M</td>
<td>35,000</td>
<td>18,000M&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Cf. Open IE (e.g., [Mausam et al. ‘12])

5B assertions (Mausam, Michael Schmitz, personal communication, Oct. ‘13)

302M with Prob > 0.9
381M with Prob > 0.7
Knowledge acquisition from the Web

Unfocused ("comb through the Web")

• Start from a collection of Web pages
  ➔ Non-targeted (blanket) extraction

Focused

• Formulate specific questions or queries, looking for missing data
  • Identify (a small set of) relevant Web pages
  ➔ Targeted extraction
**Automatic knowledge base completion (focused extraction)**

<table>
<thead>
<tr>
<th>Relation</th>
<th>% unknown in Freebase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profession</td>
<td>68%</td>
</tr>
<tr>
<td>Place of birth</td>
<td>71%</td>
</tr>
<tr>
<td>Nationality</td>
<td>75%</td>
</tr>
<tr>
<td>Education</td>
<td>91%</td>
</tr>
<tr>
<td>Spouse</td>
<td>92%</td>
</tr>
<tr>
<td>Parents</td>
<td>94%</td>
</tr>
</tbody>
</table>

8 And the LORD God **planted a garden eastward in Eden; and there he put the man whom he had formed.**

15 Then the LORD God **took the man and put him in the garden of Eden to tend and keep it.**

19 And out of the ground the LORD God formed every beast of the field, and every fowl of the air; and **brought them unto Adam to see what he would call them:** and whatsoever Adam called every living creature, that was the name thereof.
Proactively searching for missing values
[with West et al., WWW ’14]

- Mine search logs for best query templates (per relation)
- Augment queries with disambiguating information
- Thou shalt ask in moderation (asking too much may be harmful!)
The importance of query augmentation

Who is the mother of Frank Zappa

The Mothers of Invention - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/The_Mothers_of_Invention

The Mothers of Invention were an American rock band from California that served as the backing musicians for Frank Zappa, a self-taught composer and ...

History - Personnel - Discography - References

Frank Zappa - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Frank_Zappa


Discography - Moon Zappa - Diva Zappa - Gail Zappa

Who is the mother of Frank Zappa Baltimore Maryland

Frank Zappa - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Frank_Zappa

Frank Vincent Zappa was born in Baltimore, Maryland, on December 21, 1940. His mother, Rose Marie (née Colimore) was of Italian and French ancestry; his ...
Learning to query

/people/person/parents

Color = mean reciprocal rank of true answer

GOOD  BAD
Asking the right (number of) questions

(a) SPOUSES

(b) PLACE OF BIRTH

(c) NATIONALITY
To trust, or not to trust, that is the question
Motivation for Knowledge-Based Trust (KBT) [with Dong et al., VLDB ’15]

Today

Popularity-based trust (links / searches / clicks)

- Exogenous signals

Key idea

Evaluate trustworthiness of sources by the correctness of their factual information

- Endogenous signals
Knowledge-Based Trust vs. PageRank

Often sources with low accuracy

Similar scores

Often tail sources w. high trustworthiness
Assessing fact correctness

<table>
<thead>
<tr>
<th>Triple 1</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triple 2</td>
<td>0.9</td>
</tr>
<tr>
<td>Triple 3</td>
<td>0.3</td>
</tr>
<tr>
<td>Triple 4</td>
<td>0.8</td>
</tr>
<tr>
<td>Triple 5</td>
<td>0.4</td>
</tr>
<tr>
<td>Triple 6</td>
<td>0.8</td>
</tr>
<tr>
<td>Triple 7</td>
<td>0.9</td>
</tr>
<tr>
<td>Triple 8</td>
<td>1.0</td>
</tr>
<tr>
<td>Triple 9</td>
<td>0.7</td>
</tr>
<tr>
<td>Triple 10</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Accuracy 0.7
Extraction
What if the triple was extracted incorrectly?

Surely we should not penalize a source for extraction errors!

And we don’t know the truth 😞

Solution: We will jointly model

1. Extraction correctness
2. Triple correctness
3. Source accuracy
4. Extractor precision/recall

Latent variable model + EM
Whom do you trust?

We may trust facts found in reputable sources.

But how do we know a source is reputable?

If it contains correct facts, of course!
Algorithm: Knowledge-Based Trust [with Dong et al., VLDB ’15]

We use Bayes rule as follows:

\[
p(C_{wdv} = 1 | X_{wdv}) = \frac{\alpha p(X_{wdv} | C_{wdv} = 1)}{\alpha p(X_{wdv} | C_{wdv} = 1) + (1 - \alpha)p(X_{wdv} | C_{wdv} = 0)}
\]

\[
= \frac{1}{1 + \frac{p(X_{wdv} | C_{wdv} = 1)}{p(X_{wdv} | C_{wdv} = 0)} \frac{\alpha}{1 - \alpha}}
\]

\[
= \sigma \left( \log \frac{p(X_{wdv} | C_{wdv} = 1)}{p(X_{wdv} | C_{wdv} = 0)} + \log \frac{\alpha}{1 - \alpha} \right)
\]

where \( \sigma(x) = \frac{1}{1 + e^{-x}} \) is the sigmoid function.

Assuming independence of the extractors, and using Equation (6), we can compute the likelihood ratio as follows:

\[
p(X_{wdv} | C_{wdv} = 1) = \prod_{e : X_{ewdv} = 1} R_e \prod_{e : X_{ewdv} = 0} \frac{1 - R_e}{1 - Q_e}
\]

In other words, for each extractor we can compute a presence vote \( \text{Pre}_e \) for a triple that it extracts, and an absence vote of \( \text{Abs}_e \) for a triple that it does not extract:

\[
\text{Pre}_e = \log R_e - \log Q_e
\]

\[
\text{Abs}_e = \log (1 - R_e) - \log (1 - Q_e).
\]

For each triple \((w, d, v)\) we can compute its vote count as the sum of the presence votes and the absence votes:

\[ 3.3.2 \quad \text{Estimating true value of the data item} \]

In this step, we compute \( p(V_d = v | \hat{C}_d) \), following the “single truth” model of [8]. By Bayes rule we have

\[
p(V_d = v | \hat{C}_d) = \frac{p(\hat{C}_d | V_d = v)p(V_d = v)}{\sum_{v' \in \text{dom}(d)} p(\hat{C}_d | V_d = v')p(V_d = v')}
\]

(16)

We will assume a uniform prior \( p(V_d = v) \), so we just need to focus on the likelihood. Using Equation (5), we have

\[
p(\hat{C}_d | V_d = v)
\]

\[= \prod_{w : \hat{C}_{wdv} = 1} A_w \prod_{w : \hat{C}_{wdv} = 0} \frac{1 - A_w}{n}
\]

(17)

\[
= \prod_{w : \hat{C}_{wdv} = 1} \frac{nA_w}{1 - A_w} \prod_{w : \hat{C}_{wdv} \in \{0,1\}} \frac{1 - A_w}{n}
\]

(18)

Since the latter term \( \prod_{w : \hat{C}_{wdv} \in \{0,1\}} \frac{1 - A_w}{n} \) is constant with respect to \( v \), we can drop it.

Now let us define the vote count as follows:

\[ VCV(w) \triangleq \log \frac{nA_w}{1 - A_w}
\]

(19)

Aggregating over web sources that provide this triple, we define

\[ VCV(d, v) \triangleq \sum_w \mathbb{I}(\hat{C}_{wdv} = 1) VCV(w)
\]

(20)

With this notation, we can rewrite Equation (16) as

\[ p(V_d = v | \hat{C}_d) = \frac{\exp(VCV(d, v))}{\sum_{v' \in \text{dom}(d)} \exp(VCV(d, v'))}
\]

(21)
Limitations

• Extraction is sparse: 74% URLs yield < 5 triples
• KBT is reliable for < 20% websites and << 5% webpages
How many facts are there?
And how many can we represent?

How many facts can be potentially extracted from text?

RESEARCH CHALLENGES

“For with much wisdom comes much sorrow; the more knowledge, the more grief.”

-- Ecclesiastes 1:18
Limits of automatic extraction

637M facts (non-redundant)

223M facts

35% of Freebase

302M facts with Prob(true) > 0.9

12% of Freebase
Relations that are rarely expressed in text

<table>
<thead>
<tr>
<th>Relation</th>
<th>% entity pairs not found</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>/people/person/gender</td>
<td>30%</td>
<td>Pronouns</td>
</tr>
</tbody>
</table>

**Sample object:**

*Biaxin 250 film coated tablet* (/m/0jxc5vb)

**Sample subject:**

*Fluocinolone Acetonide 0.25 cream* (/m/0jxlbx9)

**Sample book title:**

*The birth day: a brief narrative of Eliza Reynolds, who died on Sunday, Oct 19, 1834* (/m/0ydpbtq)

Albert Einstein  male
## Relations that are rarely expressed in text

<table>
<thead>
<tr>
<th>Relation</th>
<th>% entity pairs not found</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>/people/person/gender</code></td>
<td>30%</td>
<td>Pronouns</td>
</tr>
<tr>
<td><code>/people/person/profession</code></td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td><code>/people/person/children and /people/person/parents</code></td>
<td>36%</td>
<td></td>
</tr>
<tr>
<td><code>/medicine/drug_formulation/manufactured_forms</code></td>
<td>99.9%</td>
<td><strong>Sample object:</strong> &quot;Biaxin 250 film coated tablet&quot; (/m/0jxc5vb)</td>
</tr>
<tr>
<td><code>/medicine/manufactured_drug_form/available_in</code></td>
<td>99.4%</td>
<td><strong>Sample subject:</strong> “Fluocinolone Acetonide 0.25 cream” (/m/0jxlbx9)</td>
</tr>
<tr>
<td><code>/book/author/works_written and /book/written_work/author</code></td>
<td>37%</td>
<td><strong>Sample book title:</strong> “The birth day: a brief narrative of Eliza Reynolds, who died on Sunday, Oct 19, 1834” (/m/0ydpbtq)</td>
</tr>
</tbody>
</table>
1 In the beginning God created the heaven and the earth.
2 And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters.
3 And God said, Let there be light: and there was light. ...

(Genesis 2)

7 And the LORD God formed man of the dust of the ground, and breathed into his nostrils the breath of life; and man became a living soul.
8 And the LORD God planted a garden eastward in Eden; and there he put the man whom he had formed.

19 And out of the ground the LORD God formed every beast of the field, and every fowl of the air; and brought them unto Adam to see what he would call them: and whatsoever Adam called every living creature, that was the name thereof.

Implicitly stated information: Adam
Challenge: Disparity of opinion

Who invented the radio?
Challenge: Fictional contexts

Should we assert

</en/abraham_lincoln, /people/person/profession, /en/vampire_hunter>?
Challenge: negative examples

• We already know a lot … but those are only positive examples!

• Many ways to get negative examples … none of them perfect 😞
  
  • Deleted assertions in Freebase
    • Was the deletion justified?
  
  • Inconsistencies identified with manually specified rules
    • Poor coverage
  
  • Examples judged by humans
    • Optimized for accuracy on the positive class
  
  • Automatically create negative examples using the closed world assumption
    • Noisy, unless applied to functional relations
  
  • Feedback from Web users
    • Difficult to judge automatically

Released! See goo.gl/MJb3A

Crowdsourcing
Negative examples (cont’d): feedback from Web users
What is the name of the animal on the Quizz.us logo?

Your answer: 

Send

I don’t know

Quoll
Animal

The quoll is a carnivorous marsupial native to mainland Australia, New Guinea, and Tasmania. It is primarily nocturnal and spends most of the day in its den. Of the six species of quoll, four are found in Australia and two in New Guinea. [Wikipedia]

Scientific name: Dasyurus
Rank: Genus
Higher classification: Dasyuridae
Lower classifications: Tiger quoll, New Guinean quoll, Bronze quoll, Northern quoll, Western quoll, Eastern quoll
Entity resolution / deduplication

- Multiple mentions of the same entity is incorrect and confusing.
Commonsense knowledge

“Bananas are yellow.”

“Jasmine flowers smell good.”

“Balls bounce.”

- Commonsense information is hard to collect (*too obvious*)
- Yet commonsense reasoning is often crucial
Commonsense knowledge

ConceptNet

- Nodes represent concepts (words or short NL phrases)
- Labeled relationships connecting them
  saxophone → UsedFor → jazz
  learn → MotivatedByGoal → knowledge
ConceptNet (cont’d)

- ConceptNet is a *(hyper)*graph
  - Edges about the edges

- Each statement has *justifications*
  - Provenance + reliability assessment

- The graph is **ID-less**
  - Every node has all the information necessary to identify it
  - Multiple branches can be developed in parallel and later merged
    - Take the union of the nodes and edges
    - No reconciliation

[Havasi et al., RANLP ‘07; Speer and Havasi, LREC ‘12]

http://conceptnet5.media.mit.edu/
Commonsense knowledge in YAGO

- WebChild [Tandon et al., WSDM ‘14]
  (strawberry, hasTaste, sweet), (apple, hasColor, green)
- Acquired from the web using semi-supervised learning
- Uses WordNet senses and web statistics to construct seeds

- Acquiring comparative commonsense knowledge from the web [Tandon et al., AAAI ‘14]
  (car, faster, bike), (lemon, more-sour, apple)
  - Uses Open IE

- Earlier work: [Tandon et al., AAAI ‘11]
  CapableOf(dog, bark), PartOf(roof, house)
  - Uses web n-gram data with seeds from ConceptNet
CYC


• OpenCYC
  • 239K terms, 2M triples

• ResearchCYC
  • 500K concepts, 5M assertions, 26K relations
Multiple modalities

How to jointly acquire knowledge from all these sources?

Text

Video

Images

Artificial worlds?

Speech/sounds

Multiple modalities

How to jointly acquire knowledge from all these sources?

Text

Video

Images

Artificial worlds?
Natural interfaces to knowledge

“Where is New York City?”
Natural interfaces to knowledge (cont’d)

“Where did Kobe Bryant play in high school?”

Kobe Bryant

From Wikipedia, the free encyclopedia

Kobe Bean Bryant (born August 23, 1978) is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA). He entered the NBA directly from high school, and has played for the Lakers his entire career, winning five NBA championship. Bryant is a 15-time All-Star, 15-time member of the All-NBA Team, and 12-time member of the All-Defensive team. As of March 2013, he ranks third and fourth[3] on the league's all-time post-season scoring and all-time regular season scoring lists, respectively.

Bryant enjoyed a successful high school basketball career at Lower Merion High School in Pennsylvania, where he was recognized as the top high school basketball player in the country. He declared his eligibility for the NBA Draft upon graduation, and was selected with the 13th overall pick in the 1996 NBA Draft by the Charlotte Hornets, then traded to the Los Angeles Lakers. As a rookie, Bryant earned himself a reputation as a high-flyer and a fan favorite by winning the 1997 Slam Dunk Contest.
PART 2: METHODS AND TECHNIQUES
Methods and techniques

1. Relation extraction:
   - Supervised models
   - Semi-supervised models
   - Distant supervision

2. Entity resolution
   - Single entity methods
   - Relational methods

3. Link prediction
   - Rule-based methods
   - Probabilistic models
   - Factorization methods
   - Embedding models

Not in this tutorial:
   - Entity classification
   - Group/expert detection
   - Ontology alignment
   - Object ranking
RELATION EXTRACTION
Relation Extraction

Extracting semantic relations between sets of [grounded] entities

Numerous variants:
- Undefined vs pre-determined set of relations
- Binary vs n-ary relations, facet discovery
- Extracting temporal information
- Supervision: \{fully, un, semi, distant\}-supervision
- Cues used: only lexical vs full linguistic features
Supervised relation extraction

• Sentence-level labels of relation mentions
  - "Apple CEO Steve Jobs said.." => (SteveJobs, CEO, Apple)
  - "Steve Jobs said that Apple will.." => NIL

• Traditional relation extraction datasets
  - ACE 2004
  - MUC-7
  - Biomedical datasets (e.g. BioNLP challenges)

• Learn classifiers from +/- examples

• Typical features: context words + POS, dependency path between entities, named entity tags, token/parse-path/entity distance
Examples of features

X was born on DDDD in Y

- **DEP**: X <nsubjpass / born prep> on pobj> DATE prep> in pobj> Y
- **NER**: X = PER, Y = LOC
- **POS**: X = NOUN, NNP; Y = NOUN, NNP
- **Context**: born, on, in, "born on"
Supervised relation extraction

- Used to be the “traditional” setting [Riloff et al., 06; Soderland et al., 99]

- **Pros**
  - High quality supervision
  - Explicit negative examples

- **Cons**
  - Very expensive to generate supervision
  - Not easy to add more relations
  - Cannot generalize to text from different domains
Semi-supervised relation extraction

**Generic algorithm**
1. Start with seed triples / golden seed patterns
2. Extract patterns that match seed triples/patterns
3. Take the top-k extracted patterns/triples
4. Add to seed patterns/triples
5. Go to 2

**Many published approaches in this category:**
- Dual Iterative Pattern Relation Extractor [Brin, 98]
- Snowball [Agichtein & Gravano, 00]
- TextRunner [Banko et al., 07] – almost unsupervised

**Differ in pattern definition and selection**
TextRunner [Banko et al., 07]

- **Almost unsupervised**
  - Relations not fixed: does not follow Knowledge Graph schema (growing)
  - No labeled data
  - Mostly unlabeled text
  - Uses heuristics to **self-label** a starting corpora (using a parser), such as
    - Path length < k
    - Path does not cross sentence-like boundaries like relative clauses
    - Neither entity is a pronoun

- **Self-training**
  - Generate +/- examples $\rightarrow$ learn classifier
  - Extract new relation mentions using this classifier
  - Generate triples from aggregated mentions, assign probabilistic score using [Downey et al., 2005]

- Later improved in Reverb [Fader et al., 11]
Distantly-supervised relation extraction

- **Existing knowledge base + unlabeled text** → generate examples
  - Locate pairs of related entities in text
  - Hypothesizes that the relation is expressed

Google CEO Larry Page announced that...

George Lucas, founder of LucasFilms...

Star Wars was directed by George Lucas...

I went to Paris, France for the summer...
Distant supervision: modeling hypotheses

Typical architecture:

1. Collect many pairs of entities co-occurring in sentences from text corpus

2. If 2 entities participate in a relation, several hypotheses:
   1. All sentences mentioning them express it [Mintz et al., 09]

“Barack Obama is the 44th and current President of the US.” → (BO, employedBy, USA)
Knowledge Graph

U.S. president Barack Obama told reporters...

Obama’s speech at the U.S. congress...

Barack Obama, the president of the USA, said that...

Barack Obama announced that the U.S. will not...

\( F(BO, USA) = \sum F_i(BO, USA) \)

(Aggregate features)

Multiclass logistic regressor

Label

[Obama, president_of, USA]
[Mintz et al., 09]

- Classifier: multiclass logistic regressor

(Barack Obama, USA, \textit{AggFeatures})

- Negative examples
  - Randomly sample unrelated entity pairs occurring in the same sentence
  - > 98% such pairs actually unrelated

- President
- CEO
- Spokesman
- CapitalOf
- NotRelated
Sentence-level features

- **Lexical**: words in between and around mentions and their parts-of-speech tags (conjunctive form)
- **Syntactic**: dependency parse path between mentions along with side nodes
- **Named Entity Tags**: for the mentions
- **Conjunctions** of the above features
  - Distant supervision is used on to lots of data → sparsity of conjunctive forms not an issue
## Sentence-level features

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Left window</th>
<th>NE1</th>
<th>Middle</th>
<th>NE2</th>
<th>Right window</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td>[]</td>
<td>PER</td>
<td>[was/VERB born/VERB in/CLOSED]</td>
<td>LOC</td>
<td>[]</td>
</tr>
<tr>
<td>Lexical</td>
<td>[Astronomer]</td>
<td>PER</td>
<td>[was/VERB born/VERB in/CLOSED]</td>
<td>LOC</td>
<td>[, Missouri]</td>
</tr>
<tr>
<td>Lexical</td>
<td>[#PAD#, Astronomer]</td>
<td>PER</td>
<td>[was/VERB born/VERB in/CLOSED]</td>
<td>LOC</td>
<td>[]</td>
</tr>
<tr>
<td>Syntactic</td>
<td>[]</td>
<td>PER</td>
<td>[↑s was ↓pred born ↓mod in ↓pcomp-p-n]</td>
<td>LOC</td>
<td>[]</td>
</tr>
<tr>
<td>Syntactic</td>
<td>[Edwin Hubble ↓lex-mod]</td>
<td>PER</td>
<td>[↑s was ↓pred born ↓mod in ↓pcomp-p-n]</td>
<td>LOC</td>
<td>[]</td>
</tr>
<tr>
<td>Syntactic</td>
<td>[Astronomer ↓lex-mod]</td>
<td>PER</td>
<td>[↑s was ↓pred born ↓mod in ↓pcomp-p-n]</td>
<td>LOC</td>
<td>[]</td>
</tr>
<tr>
<td>Syntactic</td>
<td>[]</td>
<td>PER</td>
<td>[↑s was ↓pred born ↓mod in ↓pcomp-p-n]</td>
<td>LOC</td>
<td>[]</td>
</tr>
<tr>
<td>Syntactic</td>
<td>[Edwin Hubble ↓lex-mod]</td>
<td>PER</td>
<td>[↑s was ↓pred born ↓mod in ↓pcomp-p-n]</td>
<td>LOC</td>
<td>[↓lex-mod]</td>
</tr>
<tr>
<td>Syntactic</td>
<td>[Astronomer ↓lex-mod]</td>
<td>PER</td>
<td>[↑s was ↓pred born ↓mod in ↓pcomp-p-n]</td>
<td>LOC</td>
<td>[↓lex-mod]</td>
</tr>
<tr>
<td>Syntactic</td>
<td>[]</td>
<td>PER</td>
<td>[↑s was ↓pred born ↓mod in ↓pcomp-p-n]</td>
<td>LOC</td>
<td>[↓inside Missouri]</td>
</tr>
<tr>
<td>Syntactic</td>
<td>[Edwin Hubble ↓lex-mod]</td>
<td>PER</td>
<td>[↑s was ↓pred born ↓mod in ↓pcomp-p-n]</td>
<td>LOC</td>
<td>[↓inside Missouri]</td>
</tr>
<tr>
<td>Syntactic</td>
<td>[Astronomer ↓lex-mod]</td>
<td>PER</td>
<td>[↑s was ↓pred born ↓mod in ↓pcomp-p-n]</td>
<td>LOC</td>
<td>[↓inside Missouri]</td>
</tr>
</tbody>
</table>

Table 3: Features for ‘Astronomer Edwin Hubble was born in Marshfield, Missouri’.
### Examples of top features

<table>
<thead>
<tr>
<th>Relation</th>
<th>Feature type</th>
<th>Left window</th>
<th>NE1</th>
<th>Middle</th>
<th>NE2</th>
<th>Right window</th>
</tr>
</thead>
<tbody>
<tr>
<td>/architecture/structure/architect</td>
<td>LEX $\nearrow$</td>
<td>ORG</td>
<td>, the designer of the</td>
<td>PER</td>
<td>(\uparrow_s) designed (\downarrow_{by\text{-}subj}) by (\downarrow_{pcn}) PER</td>
<td>(\uparrow_s) designed</td>
</tr>
<tr>
<td>/book/author/works_written</td>
<td>LEX</td>
<td>PER</td>
<td>s novel</td>
<td>ORG</td>
<td>(\uparrow_{pcn}) by (\uparrow_{mod}) story (\uparrow_{pred}) is (\downarrow_s) PER</td>
<td>PER</td>
</tr>
<tr>
<td>/book/book.edition/author.editor</td>
<td>LEX $\nearrow$</td>
<td>ORG</td>
<td>(\uparrow_{pcn}) by (\uparrow_{by\text{-}subj}) (\uparrow_{pred}) is (\downarrow_s) PER</td>
<td>PER</td>
<td>s novel</td>
<td></td>
</tr>
<tr>
<td>/business/company/founders</td>
<td>LEX</td>
<td>PER</td>
<td>(\uparrow_{nn}) series (\downarrow_{gen})</td>
<td>PER</td>
<td>(\uparrow_{nn}) owner (\downarrow_{person})</td>
<td>PER</td>
</tr>
<tr>
<td>/business/company/place.founded</td>
<td>LEX $\nearrow$</td>
<td>ORG</td>
<td>- based</td>
<td>LOC</td>
<td>(\uparrow_s) founded (\downarrow_{mod}) in (\downarrow_{pcn}) LOC</td>
<td></td>
</tr>
<tr>
<td>/film/film/country</td>
<td>LEX</td>
<td>PER</td>
<td>, released in</td>
<td>LOC</td>
<td>(\uparrow_s) opened (\downarrow_{mod}) in (\downarrow_{pcn}) LOC</td>
<td>(\uparrow_s) opened</td>
</tr>
<tr>
<td>/geography/river/mouth</td>
<td>LEX</td>
<td>LOC</td>
<td>, which flows into the</td>
<td>LOC</td>
<td>(\uparrow_s) is (\downarrow_{pred}) tributary (\downarrow_{mod}) of (\downarrow_{pcn}) LOC</td>
<td>(\downarrow_{det}) the</td>
</tr>
</tbody>
</table>
Distant supervision: modeling hypotheses

Typical architecture:

1. Collect many pairs of entities co-occurring in sentences from text corpus

2. If 2 entities participate in a relation, several hypotheses:
   1. **All** sentences mentioning them express it [Mintz et al., 09]
   2. **At least one** sentence mentioning them express it [Riedel et al., 10]

“Barack Obama is the 44th and current President of the **US**.” \(\rightarrow\) (BO, employedBy, USA)

“Obama flew back to the **US** on Wednesday.” \(\rightarrow\) (BO, employedBy, USA)
[Riedel et al., 10]

- Every mention of an entity-pair does not express a relation

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>New York Times</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>nationality</td>
<td>38%</td>
<td>20%</td>
</tr>
<tr>
<td>place_of_birth</td>
<td>35%</td>
<td>20%</td>
</tr>
<tr>
<td>contains</td>
<td>20%</td>
<td>10%</td>
</tr>
</tbody>
</table>

- Violations more in news than encyclopediac articles

- Assert triple from only a few mentions, not all
[Riedel et al., 10]

- Factor graph:

- Multiple-instance setting
Distant supervision: modeling hypotheses

Typical architecture:

1. Collect many pairs of entities co-occurring in sentences from text corpus

2. If 2 entities participate in a relation, several hypotheses:
   1. **All** sentences mentioning them express it [Mintz et al., 09]
   2. **At least one** sentence mentioning them express it [Riedel et al., 10]
   3. **At least one** sentence mentioning them express it and 2 entities can express multiple relations [Hoffmann et al., 11] [Surdeanu et al., 12]

"Barack Obama is the 44th and current President of the US." \(\rightarrow\) (BO, employedBy, USA)

"Obama flew back to the US just Wednesday’s." \(\not\rightarrow\) (BO, flewTo, USA)
[Surdeanu et al., 12]

- Relation extraction is a multi-instance multi-label problem.

“Barack Obama is the 44th and current President of the US.” $\rightarrow$ (BO, `employedBy`, USA)

“Obama was born in the US just as he always said.” $\rightarrow$ (BO, `bornIn`, USA)

“Obama flew back to the US on Wednesday.” $\rightarrow$ `NIL`

- Training via EM with initialization with [Mintz et al., 09]

![Diagram](Image)
Relaxing hypotheses improves precision

Precision-recall curves on extracting from New York Times articles to Freebase [Weston et al., 13]
Distant supervision

- **Pros**
  - Can scale to the web, as no supervision required
  - Generalizes to text from different domains
  - Generates a lot more supervision in one iteration

- **Cons**
  - Needs high quality entity-matching
  - Relation-expression hypothesis can be wrong
    - Can be compensated by the extraction model, redundancy, language model
  - Does not generate negative examples
    - Partially tackled by matching unrelated entities
Plenty of extensions

- **Using language models** [Downey et al., 07]
  - Do two entities seem to express a given relation, given the context?

- **Joint relation extraction + other NLP tasks**
  - Named Entity tagging [Yao et al., 10]
  - Possibly with entity resolution and/or coreference

- Jointly + **repeatedly training** multiple extractors [Carlson et. al., 10]

- **Unsupervised** extraction [Poon & Domingos, 10]

- **Jointly perform relation extraction and link prediction** [Bordes et al., 12; Weston et al., 13; Riedel et al., 13]
ENTITY RESOLUTION
Entity resolution

- Single entity resolution
- Relational entity resolution

Exemplary use case involving Tony Parker and Pau Gasol.
Single-entity entity resolution

- Entity resolution *without using the relational context* of entities

- Many *distances/similarities* for single-entity entity resolution:
  - Edit distance (Levenshtein, etc.)
  - Set similarity (TF-IDF, etc.)
  - Alignment-based
  - Numeric distance between values
  - Phonetic Similarity
  - Equality on a boolean predicate
  - Translation-based
  - Domain-specific
Case study: deduplicating places [Dalvi et al., 14]

- Multiple mentions of the same place is wrong and confusing.
Origin of duplicates

• Duplicates are often created during check-in:
  ▪ Different spellings
  ▪ GPS Errors
  ▪ Wrong checkins

• Frequently, these duplicates have:
  • few attribute values
  • names were typed hurriedly
Effectively matching place names is hard

<table>
<thead>
<tr>
<th>Good Matches (Help Recall)</th>
<th>Bad Matches (Hurt Precision)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guggenheim Art Museum Manhattan</td>
<td>Guggenheim Starbucks</td>
</tr>
<tr>
<td>DishDash</td>
<td>Dish Dash Restaurant</td>
</tr>
<tr>
<td>Ippudo New York</td>
<td>Ipudo</td>
</tr>
<tr>
<td>Central Park Café (Sunnyvale)</td>
<td>Central Park Restaurant (NYC)</td>
</tr>
<tr>
<td>Central Park Café (Sunnyvale)</td>
<td>Glen Canyon Park</td>
</tr>
</tbody>
</table>

- Easy to find cases where the “bad match” pair is more similar than the “good match” pair
- Existing similarity metrics (TF-IDF, Levenshtein, Learned-weight edit distance, etc.) generally fail to handle this level of variability
Idea 1: core words

- A core word = a word a human would use to refer to the place, if only a single word were allowed

- **Goal:** try to identify the core word, use it for comparisons

  Guggenheim Starbucks

  Guggenheim Art Museum, Manhattan
Idea 2: spatial context model

- Tokens vary in importance based on geographic context
  - Central Park is common/meaningless in NYC

- **Goal:** filter out context-specific tokens when matching names
Convert into an edit distance

- We match $N_1$ with $N_2$ given:
  - Core words model
  - Spatial contextual model
- Treat $N_1$, $N_2$ as bag of words, and require:
  - Core words match
  - Any words that match are either core or background in both $N_1$ and $N_2$
- Extend this to Levenshtein-like edit distance
Deduplication results [Dalvi et al., 14]

- Edit: Levenshtein distance between place names
- TF-IDF: cosine similarity of TF-IDF weighted vector of overlapping names
Entity resolution

Single entity resolution

Relational entity resolution
Relational entity resolution – Simple strategies

- Enrich model with relational features $\rightarrow$ richer context for matching

- Relational features:
  - Value of edge or neighboring attribute
  - Set similarity measures
    - Overlap/Jaccard
    - Average similarity between set members
    - Adamic/Adar: two entities are more similar if they share more items that are overall less frequent
    - SimRank: two entities are similar if they are related to similar objects
    - Katz score: two entities are similar if they are connected by shorter paths
Relational entity resolution – Advanced strategies

- Dependency graph approaches [Dong et al., 05]
- Relational clustering [Bhattacharya & Getoor, 07]
- **Probabilistic Relational Models** [Pasula et al., 03]
- Markov Logic Networks [Singla & Domingos, 06]
- **Probabilistic Soft Logic** [Broecheler & Getoor, 10]
LINK PREDICTION
Link prediction

- Add knowledge from existing graph
- No external source
- Reasoning within the graph

1. Rule-based methods
2. Probabilistic models
3. Factorization models
4. Embedding models
First Order Inductive Learner

- FOIL learns function-free Horn clauses:
  - given positive negative examples of a concept
  - a set of background-knowledge predicates
  - FOIL inductively generates a logical rule for the concept that cover all + and no -

\[
\text{teammate}(x,y) \land \text{playFor}(y,z) \Rightarrow \text{playFor}(x,z)
\]

- Computationally expensive: huge search space large, costly Horn clauses
- Must add constraints → high precision but low recall
- Inductive Logic Programming: deterministic and potentially problematic
Path Ranking Algorithm [Lao et al., 11]

- Random walks on the graph are used to sample paths
- Paths are weighted with probability of reaching target from source
- Paths are used as ranking experts in a scoring function

\[
h(Pa_2(TP,SAS)) = 0.2
\]

\[
h(Pa_1(TP,SAS)) = 0.95
\]

\[
S(TP, playFor, SAS) = \sum_{i \in paths} \theta_{playFor}^i h(pa_i(TP, SAS))
\]
Link prediction with scoring functions

• A scoring function alone does not grant a decision

• **Thresholding**: determine a threshold $\theta$

\[(TP, playFor, SAS) \text{ is True } \text{ iff } S(TP, playFor, SAS) > \theta\]

• **Ranking**:
  • The most likely relation between Kobe Bryant and LA Lakers is:
    \[rel = \arg\max_{r' \in \text{rels}} S(TP, r', SAS)\]
  • The most likely team for Kobe Bryant is:
    \[obj = \arg\max_{e' \in \text{ents}} S(TP, playFor, e')\]

• **As prior** for extraction models (cf. Knowledge Vault [Dong et al., 2014])

• No calibration of scores like probabilities
Random walks boost precision

Precision of generalized facts for three cutoffs [Lao et al., 2011]
Probabilistic Relational Models [Friedman et al., 99]

• **Probabilistic Relational Models** are directed graphical models that can handle link and feature uncertainty

• Probabilistic inference to predict links but also duplicates, classes, clusters, etc. based on conditional probability distributions

• **Limitations:**
  
  • Careful construction: must avoid cycles
  
  • Generative process that models both observations and unknowns
  
  • Tractability issues
Relational Markov Networks [Taskar et al., 02]

- Discriminative model: performs inference over the unknowns only

- Discriminant function: \( P(X = x) = \frac{1}{Z} \exp(\sum_i w_i f_i(x)) \)

- Drawbacks:
  - 1 feature for each state of each clique (large)
  - MAP estimation with belief propagation (slow)
Markov Logic Networks [Richardson & Domingos, 06]

- Knowledge graph = set of hard constraints on the set of possible worlds
  - Markov logic make them soft constraints
  - When a world violates a formula, it becomes less probable but not impossible

**Formulas**
- **Constants:** TP, SAS, NBA
- **Variables:** x, y ranging over their domains (person, team, etc.).
- **Predicates:** teammate(x, y)
- **Atom:** teammate(TP, x)
- **Ground atom:** teammate(TP, PG)

- A Markov Logic Network \((w, F)\) is a set of weighted first-order formulas
  - Probability of a grounding \(x\):
  - **Higher weight** \(\square\) **stronger constraint**

\[
P(X = x) = \frac{1}{Z} \exp\left(\sum_{i \in F} w_i n_i(x)\right)
\]
Probabilistic Soft Logic [Bach et al., 13]

- Framework where rules have continuous truth values
- Atoms like teammate(KB, x) are continuous random variables
- Each predicate has a weight like in MLNs
- Probability of a grounding:

\[ f(I) = \frac{1}{Z} \exp\left[ - \sum_{r \in R} \lambda_r (d_r(I))^p \right] \]

- Inference is very tractable: convex optimization problem.
Multiple Relational Clustering [Kok & Domingos, 07]

- **Hypothesis:** multiple clusterings are necessary to fully capture the interactions between entities.
Multiple Relational Clustering [Kok & Domingos, 07]

- **Markov Logic framework:**
  - Create an *unary predicate for each cluster e.g. cluster22(x)*
  - **Multiple partitions** are learnt together
  - Use connections:
    - Cluster relations by entities they connect and vice versa
  - Use types:
    - Cluster objects of same type
    - Cluster relations with same arity and argument types

- Learning by greedy search and multiple restarts maximizing posterior

- Link prediction is determined by **evaluating truth value of grounded atoms** such as `playFor(TP, SAS)`
Stochastic Blockmodels [Wang & Wong, 87]

- **Blockmodels**: learn partitions of entities and of predicates
  - **Partition entities/relations** into subgroups based on equivalence measure.
  - For each pair of positions **presence or absence of relation**.
  - **Structural equivalence**: entities are structurally equivalent if they have identical relations to and from all the entities of the graph

- **Stochastic blockmodels**:
  - Underlying probabilistic model
  - **Stochastic equivalence**: two entities or predicates are stochastically equivalent if they are “exchangeable” w.r.t. the probability distribution
Infinite Relational Models [Kemp et al., 05]

- **Infinite**: number of clusters increases as we observe more data
- **Relational**: it applies to relational data

Prior assigns a probability to all possible partitions of the entities
- Allow number of clusters to adjust as we observe more data
- **Chinese Restaurant Process**: each new object is assigned to an existing cluster with probability proportional to the cluster size.
Example

- Semantic network with 135 concepts and 49 binary predicates.
- Finds **14 entities clusters and 21 predicate clusters**

- **Scalability issues** with very large knowledge graphs
Variants of SBMs

- Mixed membership stochastic block models [Airoldi et al., 08]
- Nonparametric latent feature relational model [Miller et al., 09]
- Hybrid with tensor factorization [Sutskever et al., 09]
Factorization methods

- **Matrix factorization** is successful: collaborative filtering, recommendation, etc.

- Extension to multi-relational case

- **Collective matrix factorization** or tensor factorization
Tensor factorization

- Many methods available: PARAFAC, Tucker, DEDICOM, etc.
- Example of PARAFAC [Harschman, 70]

\[ \mathbf{x} \approx \sum_{i=1}^{R} \mathbf{a}_i^T \mathbf{b}_i \mathbf{c}_i^T \]

- Decomposition as a sum of rank-one tensors

\[ S(KB, playFor, LAL) = \sum_{i=1}^{R} a_{KB}^i b_{LAL}^i c_{playFor}^i \]

- \( A, B \) and \( C \) are learned by alternating least squares
- Does not take advantage of the symmetry of the tensor
RESCAL [Nickel et al., 11]

- Collective matrix factorization inspired by DEDICOM

- A single matrix $A$ stores latent representations of entities (vectors)
- Matrices $R_k$ store latent representations of relations

- Score:

$$S(KB, \text{playFor}, LAL) = a_{KB} R_{\text{playFor}} a_{LAL}^T$$
RESCAL [Nickel et al., 11]

- Training with **reconstruction objective**:

\[
\min_{A,R} \frac{1}{2} \left( \sum_k \| X_k - AR_kA^T \|_F^2 \right) + \lambda_A \| A \|_F^2 + \lambda_R \sum_k \| R_k \|_F^2
\]

- Optimization with **alternating least squares** on \( A \) and \( R_k \)
- Faster than PARAFAC.
Factorization outperforms clustering

Area under the precision-recall curve in link prediction [Nickel et al., 2011]
+ Improved performance when combined with PRA [Nickel et al. 2014]
Embedding models

- Related to Deep Learning methods
- Entities are vectors (low-dimensional sparse)
- Relation types are operators on these vectors

- Embeddings trained to define a similarity score on triples such that:

\[ S(KB, playFor, LAL) > S(KB, playFor, NYK) \]
Training embedding models

- Training by ranking triples from the KG vs negative (generated)

- Optimization via Stochastic Gradient Descent

- For each triple from the training set such as (TP, playFor, SAS):
  1. Unobserved facts (false?) are sub-sampled:
     - (Kobe Bryant, opponent, SA Spurs)
     - (Kobe Bryant, playFor, SA Spurs)
     - (NBA, teammate, SA Spurs)
     - Etc…

  2. It is checked that the similarity score of the true triple is lower:

     \[ S(TP, \text{playFor}, \text{SAS}) > S(TP, \text{playFor}, \text{LAL}) + 1 \]

  3. If not, parameters of the considered triples are updated.
Structured Embeddings [Bordes et al., 11]

- Each entity = 1 vector
- Each relation = 2 matrices
- Score: L1 distance between projected embeddings

$$S(TP, \text{playFor}, SAS) = - \| M_{\text{sub \ playFor}}^{TP} e_{TP} - M_{\text{obj \ playFor}}^{SAS} e_{SAS} \|_1$$
Translating Embeddings [Bordes et al. 13]

- *Simpler model*: relation types are translation vectors

\[ S(john, bornIn, miami) = -\left\| e_{john} + e_{bornIn} - e_{miami} \right\|_2 \]

- Much fewer parameters (1 vector per relation).
The simpler, the better

Ranking object entities on a subset of Freebase [Bordes et al. 13]
Visualization
Using knowledge graph and text together

Why not merging relation extraction and link prediction in the same model?

Extracted facts should agree both with the text and the graph!

“A point guard of once again saved man of the match for the Spurs” his team” San Antonio”
Joint embedding models [Bordes et al., 12; Weston et al., 13]

- Combination of two scores: $S(.) = S_{\text{text}}(.) + S_{\text{FB}}(.)$ (trained separately)

- $S_{\text{text}}(TP, \text{playFor}, SAS) = \langle W^T \Phi(m), e_{\text{playFor}} \rangle$ inspired by WSABIE (Weston et al., 10)

- $S_{\text{FB}}(TP, \text{playFor}, SAS) = \| e_{\text{KB}} + e_{\text{playFor}} - e_{\text{LAL}} \|_2$ (translating embeddings)
Using stored information improves precision even more

Precision-recall curves on extracting from New York Times articles to Freebase [Weston et al., 13]
Universal schemas [Riedel et al., 13]

- Join in a **single learning problem** link prediction and relation extraction

- The same model can score triples made of entities linked with:
  - extracted surface forms from text
  - predicates from a knowledge base
Universal schemas [Riedel et al., 13]

- Combination of three scores:
  \[ S(.) = S_{\text{mention}}(.) + S_{FB}(.) + S_{\text{neighbors}}(.) \]

  \[ S_{\text{mention}}(TP, playFor, SAS) = \langle e_{\text{mention}}, e_{\text{playFor1}} \rangle \]

  \[ S_{FB}(TP, playFor, SAS) = \langle e_{\text{sub}}^{\text{playFor2}}, e_{TP} \rangle + \langle e_{\text{obj}}^{\text{playFor2}}, e_{SAS} \rangle \]

  \[ S_{\text{neighbors}}(TP, playFor, SAS) = \sum_{(TP, rel', SAS) \neq \text{playFor}} w_{\text{rel'}}^{\text{playFor}} \]

- Embeddings for entities, relations and mentions.

- Training by ranking observed facts versus others and making updates using Stochastic Gradient Descent.
Using stored information (still) improves precision

Weighted Mean Averaged Precision on a subset of relations of Freebase [Riedel et al. 13]
Extensions of embedding-based approaches

- Modeling relation paths [Guu et al., 15; Neelakantan et al., 15; Lin et al., 15; Garcia-Duran et al. 15; Toutanova et al., 16]
- Multi-hop inferences such as “where is the U.S. president born”
- Path-based regularization
  e.g., for translations [Lin et al., 15; Garcia-Duran et al. 15]

• Jointly embedding entities and textual description [Xie et al., 16]
RESOURCES
Relevant datasets

- Wikipedia

- Freebase
  - [https://developers.google.com/freebase/data](https://developers.google.com/freebase/data)

- YAGO

- DBpedia

- OpenIE/Reverb
Relevant competitions, evaluations, and workshops

- Knowledge Base Population (KBP) @ TAC
  http://www.nist.gov/tac/2014/KBP/

- Knowledge Base Acceleration (KBA) @ TREC
  http://trec-kba.org/

- Entity Recognition and Disambiguation (ERD) Challenge @ SIGIR 2014

- INEX Link the Wiki track
  http://link.springer.com/chapter/10.1007/978-3-642-23577-1_22

- CLEF eHealth Evaluation Lab
  http://link.springer.com/chapter/10.1007/978-3-642-40802-1_24
Relevant competitions, evaluations, and workshops (cont’d)

• Named Entity Extraction & Linking (NEEL) Challenge (#Microposts2014)
  http://www.scc.lancs.ac.uk/microposts2014/challenge/

• LD4IE 2014 Linked Data for Information Extraction
  http://trec-kba.org/
Tutorials

• Entity linking and retrieval tutorial (Meij, Balog and Odijk)
  • http://ejmeij.github.io/entity-linking-and-retrieval-tutorial/

• Entity resolution tutorials (Getoor and Machanavajjhala)

• Big data integration (Dong and Srivastava)
  • http://lunadong.com/talks/BDI_vldb.pptx

• Tensors and their applications in graphs (Nickel and Tresp)
  • http://www.cip.ifi.lmu.de/~nickel/iswc2012-learning-on-linked-data/

• Probabilistic soft logic (Bach et Getoor)
  • http://psl.umiacs.umd.edu/
Data releases from Google

1. Automatic annotation of ClueWeb09 and ClueWeb12 with Freebase entities (800M documents, 11B entity mentions)

2. Similar annotation of TREC KBA Stream Corpus 2014 (496M documents, 9.4B entity mentions)

3. Human judgments of relations extracted from Wikipedia (50K instances, 250K human judgments)

4. Triples deleted from Freebase over time (63M triples)

Mailing list:

goo.gl/MJb3A
SUMMARY
Price/performance of knowledge acquisition

Noise

Automated

Non-experts

Experts

Cf. Quizz.us

[with Ipeirotis et al., WWW 2014, 2015]
1. Knowledge is **crucial for many AI tasks**

2. **Knowledge brings more knowledge**
   (the more you know, the more you can learn)

   ➡️ Fusion rules!

3. **Knowledge is hard to acquire**
   
   • **A lot** can be found online, but **a lot more** cannot be found
   
   • **A lot** cannot be extracted using today’s methods
Next steps

• We need new extraction methods, from new sources
  • Modalities other than text?

• Going beyond the fixed schema
  • Enforcing structure makes modeling easier yet imposes limitations

• The holy grail: reasoning with knowledge!