Knowledge Graphs and Information Retrieval
A Symbiotic Relationship

Sumit Bhatia

IBM Research AI, New Delhi, India

Email: sumitbhatia@in.ibm.com
Website: http://sumitbhatia.net
Twitter: @sbhatia_

December 6, 2018
Introduction
Have you used Knowledge Graphs?
Knowledge Graphs

Have you used Knowledge Graphs?

Most probably, YES!!!
Have you used Knowledge Graphs?

Most probably, YES!!!
What is a Knowledge Graph?

No clear definition...
What is a Knowledge Graph?

No clear definition...

Structured representation of knowledge
What is a Knowledge Graph?

No clear definition...

Structured representation of knowledge

Collection of facts about real-world concepts
What is a Knowledge Graph?

No clear definition...

Structured representation of knowledge

Collection of facts about real-world concepts

Represented as \(< h, r, t >\) triples
What is a Knowledge Graph?

No clear definition...

Structured representation of knowledge

Collection of facts about real-world concepts

Represented as $< h, r, t >$ triples

$<\text{IIT Delhi}, \text{locatedAt}, \text{New Delhi}>$
What is a Knowledge Graph?

No clear definition...

Structured representation of knowledge

Collection of facts about real-world concepts

Represented as $< h, r, t >$ triples

$\langle$IIT Delhi, locatedAt, New Delhi$\rangle$

$h,t$ – entities, represented as nodes in the graph
What is a Knowledge Graph?

No clear definition...

Structured representation of knowledge

Collection of facts about real-world concepts

Represented as $<h, r, t>$ triples

$<\text{IIT Delhi}, \text{locatedAt}, \text{New Delhi}>$

$h,t$ – entities, represented as nodes in the graph

$r$ – relation between entities, represented as an edge in the graph
Knowledge Graph

Google

Search results for "steve jobs birthday"

About 1,84,00,000 results (0.59 seconds)

Steve Jobs / Date of birth

24 February 1955
Let us consider this graph:

![Knowledge Graph Diagram]

- Steve Jobs
- Apple
- iPhone
- Palo Alto
- Steve Wozniak
- USA
- Steve Balmer
- Seattle
- Bill Gates
- Windows
- Microsoft
Let us consider this graph:

Open World Assumption:
Let us consider this graph:

![Knowledge Graph Diagram](image)

**Open World Assumption:** What is not known to be true is unknown
Knowledge Graph

What makes a Knowledge Graph
What makes a Knowledge Graph

- Not every data in graph format is a Knowledge Graph
Knowledge Graph

What makes a Knowledge Graph

- Not every data in graph format is a Knowledge Graph
- Citation network, web graph, road network, etc. are not knowledge graphs
Knowledge Graph

What makes a Knowledge Graph

- Not every data in graph format is a Knowledge Graph
- Citation network, web graph, road network, etc. are not knowledge graphs
- **Recall**: A Knowledge Graph consists of entities and their relationships
What makes a Knowledge Graph

- Not every data in graph format is a Knowledge Graph
- Citation network, web graph, road network, etc. are not knowledge graphs
- Recall: A Knowledge Graph consists of entities and their relationships

T-Box (terminology box): Defines the real world concepts and relations between them. Also known as Ontology/schema.
Knowledge Graph

What makes a Knowledge Graph

- Not every data in graph format is a Knowledge Graph
- Citation network, web graph, road network, etc. are not knowledge graphs
- Recall: A Knowledge Graph consists of entities and their relationships

T-Box (terminology box): Defines the real world concepts and relations between them. Also known as Ontology/schema.

A-Box (axiom box): Represents the facts, statements, or axioms
Let us consider an example
Let us consider an example

Our T-BOX defines these classes of entities – PERSON, MALE, FEMALE
Let us consider an example

Our T-BOX defines these classes of entities – PERSON, MALE, FEMALE

Relations could be defined as: MALE isA PERSON; FEMALE isA PERSON ;PERSON hasMother FEMALE
Let us consider an example

Our T-BOX defines these classes of entities – PERSON, MALE, FEMALE

Relations could be defined as: MALE isA PERSON; FEMALE isA PERSON ;PERSON hasMother FEMALE

Consider this A-Box

- Rahul isA MALE
- Sonia isA FEMALE
- Priyanka isA FEMALE
Let us consider an example

Our T-BOX defines these classes of entities – PERSON, MALE, FEMALE

Relations could be defined as: MALE \textit{isA} PERSON; FEMALE \textit{isA} PERSON ;PERSON \textit{hasMother} FEMALE

Consider this A-Box

- Rahul \textit{isA} MALE
- Sonia \textit{isA} FEMALE
- Priyanka \textit{isA} FEMALE
- Rahul \textit{hasMother} Sonia
Let us consider an example

Our T-BOX defines these classes of entities – PERSON, MALE, FEMALE

Relations could be defined as: MALE isA PERSON; FEMALE isA PERSON; PERSON hasMother FEMALE

Consider this A-Box

- Rahul isA MALE
- Sonia isA FEMALE
- Priyanka isA FEMALE
- Rahul hasMother Sonia
- Sonia hasMother Rahul
Let us consider an example

Our T-BOX defines these classes of entities – PERSON, MALE, FEMALE

Relations could be defined as: MALE $isA$ PERSON; FEMALE $isA$ PERSON ;PERSON $hasMother$ FEMALE

Consider this A-Box

- Rahul $isA$ MALE
- Sonia $isA$ FEMALE
- Priyanka $isA$ FEMALE
- Rahul $hasMother$ Sonia
- Sonia $hasMother$ Rahul
- Sonia $hasMother$ Priyanka
A Typical Knowledge Graph Entity

Here is the Wikipedia page for Narendra Modi
A Typical Knowledge Graph Entity

Here is the Wikidata page for Narendra Modi

Narendra Modi (Q1058)

Prime Minister of India

<table>
<thead>
<tr>
<th>Language</th>
<th>Label</th>
<th>Description</th>
<th>Also known as</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Narendra Modi</td>
<td>Prime Minister of India</td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>Narendra Modi</td>
<td>भारत के प्रधानमंत्री</td>
<td>नरेंद्र मोदी</td>
</tr>
<tr>
<td>Bangla</td>
<td>নরেন্দ্র মোদী</td>
<td>ভারতের সর্বাধিকার্য প্রধানমন্ত্রী।</td>
<td>নরেন্দ্র মোদী</td>
</tr>
<tr>
<td>Telugu</td>
<td>నరేంద్ర మోదీ</td>
<td>భారతదేశంలో శ్రేష్ఠం ప్రధానమంత్రి</td>
<td>నరేంద్ర మోదీ</td>
</tr>
</tbody>
</table>

Statements

instance of  
- human

2 references

image

edit
A Typical Knowledge Graph Entity

Here is the **Wikidata** page for Narendra Modi

| educated at | | | | |
|-------------|-------------|------------|--------|
| | Gujarat University | master’s degree | | |
| | academic degree | | | edit |
| | academic major | political science | | |
| | end time | 1983 | | |
| | academic major | | | |
| | end time | 1978 | | |
| | | | add reference | |

| residence | | | | |
|-----------|-------------|------------|--------|
| | New Delhi | | | |
| | | | add reference | |
| | 7, Lok Kalyan Marg | | | |
| | | | add reference | |

| member of political party | | | | |
|---------------------------|-------------|------------|--------|
| | Bharatiya Janata Party | | | |
| | | | add reference | |
A Typical Knowledge Graph Entity

Here is the *Wikidata* page for Narendra Modi

<table>
<thead>
<tr>
<th>family name</th>
<th>Modi</th>
<th>edit</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 0 references</td>
<td>+ add reference</td>
<td></td>
</tr>
<tr>
<td>+ add value</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>date of birth</th>
<th>17 September 1950</th>
<th>edit</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 1 reference</td>
<td>+ add value</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>place of birth</th>
<th>Vadnagar, located in the administrative territorial entity Bombay State, India</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 0 references</td>
<td>+ add reference</td>
</tr>
<tr>
<td>+ add value</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>spouse</th>
<th>Jashodaben Narendrabhai Modi</th>
<th>edit</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 1 reference</td>
<td>+ add value</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>native language</th>
<th>Gujarati</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 1 reference</td>
<td>+ add value</td>
</tr>
</tbody>
</table>
Some Knowledge Graphs

- Google, Bing, Yahoo, Yandex - all have one
- dbPedia (https://wiki.dbpedia.org/) – created out of Wikipedia infoboxes ($\approx 800M$ triples)
- Wikidata (https://www.wikidata.org/) – Comprehensive, structured knowledge base from Wikipedia ($\approx 7B$ triples)
- YAGO [25] – developed and maintained by Max-Planck Institute ($\approx 200M$ triples)
- Many proprietary, custom knowledge graphs in different domains
Knowledge Graphs and Information Retrieval

• KGs provide a structured view of your data
• Ontologies need to be there, highly precise and sensitive to errors/noise
• Information Retrieval is mostly over unstructured data, probabilistic in nature, comparatively more robust to noise

Focus of this tutorial: How can these two benefit or complement each other?
Knowledge Graphs and Information Retrieval

- KGs provide a structured view of your data
• KGs provide a structured view of your data
• Ontologies need to be there, highly precise and sensitive to errors/noise
• KGs provide a structured view of your data
• Ontologies need to be there, highly precise and sensitive to errors/noise
• Information Retrieval is mostly over unstructured data, probabilistic in nature, comparatively more robust to noise
Knowledge Graphs and Information Retrieval

- KGs provide a structured view of your data
- Ontologies need to be there, highly precise and sensitive to errors/noise
- Information Retrieval is mostly over unstructured data, probabilistic in nature, comparatively more robust to noise

Focus of this tutorial: How can these two benefit or complement each other?
Overview of Tutorial

- Finding Entities of Interest
  - Entity Search and Recommendation
  - Entity Linking and Disambiguation
  - Knowledge Graphs for Query Expansion
- Entity exploration: Knowing more about the entities
  - Relationship Search
  - Path Ranking
- Upcoming Challenges
Finding the Right Entities

Google search for "steve jobs birthday"

About 1,84,00,000 results (0.59 seconds)

Steve Jobs / Date of birth

24 February 1955
Finding the Right Entities

Google search for "Singapore telephone code"

About 26,10,000 results (0.52 seconds)

Singapore / Dialing code

+65
Entities are the *fundamental units* of a Knowledge graph. How to get to the right entities in the graph?
Entities are the *fundamental units* of a Knowledge graph. How to get to the right entities in the graph?

Given a Knowledge Base, $K = \{\mathcal{E}, \mathcal{R}\}$, a document corpus $\mathcal{D}$, and a named entity mention $m$, map/link the mention $m$ to its corresponding entity $e \in \mathcal{E}$. 

Steve Jobs

Apple

iPhone

Palo Alto

Steve Wozniak

Steve Balmer

Seattle

Bill Gates

Windows

Microsoft

USA

Web Queries: steve jobs birthday

NL Questions: When did Steve resign from Microsoft?

NL Text: ....Jobs and Wozniak started Apple Computers from their garage...
Entities are the *fundamental units* of a Knowledge graph. How to get to the right entities in the graph?

Given a Knowledge Base, \( K = \{ \mathcal{E}, \mathcal{R} \} \), a document corpus \( D \), and a named entity mention \( m \), map/link the mention \( m \) to its corresponding entity \( e \in \mathcal{E} \).

![Diagram of entities and connections]

- Steve Jobs
- Apple
- iPhone
- Palo Alto
- Steve Wozniak
- USA
- Microsoft
- Bill Gates
- Windows
- Seattle
- Steve Balmer

**Web Queries:**
- steve jobs birthday

**NL Questions:**
- When did Steve resign from Microsoft?

**NL Text:**
...
...Jobs and Wozniak started Apple Computers from their garage...

Finding Right Entities

Entities are the *fundamental units* of a Knowledge graph. How to get to the right entities in the graph?

Given a Knowledge Base, $K = \{\mathcal{E}, \mathcal{R}\}$, a document corpus $\mathcal{D}$, and a named entity mention $m$, map/link the mention $m$ to its corresponding entity $e \in \mathcal{E}$.

Web Queries:
Steve jobs birthday
Finding Right Entities

Entities are the *fundamental units* of a Knowledge graph. How to get to the right entities in the graph?

Given a Knowledge Base, $K = \{\mathcal{E}, \mathcal{R}\}$, a document corpus $D$, and a named entity mention $m$, map/link the mention $m$ to its corresponding entity $e \in \mathcal{E}$.

Web Queries:
steve jobs birthday

NL Questions:
When did Steve resign from Microsoft?
Finding Right Entities

Entities are the *fundamental units* of a Knowledge graph. How to get to the right entities in the graph?

Given a Knowledge Base, \( K = \{\mathcal{E}, \mathcal{R}\} \), a document corpus \( \mathcal{D} \), and a named entity mention \( m \), map/link the mention \( m \) to its corresponding entity \( e \in \mathcal{E} \).

Web Queries:
steve jobs birthday

NL Questions:
When did Steve resign from Microsoft?

NL Text:
....Jobs and Wozniak started Apple Computers from their garage...
• Same entity can be represented by multiple surface forms
Challenges

- **Same entity can be represented by multiple surface forms**
  Barack Obama, Barack H. Obama, President Obama, Senator Obama
Challenges

- Same entity can be represented by multiple surface forms
  Barack Obama, Barack H. Obama, President Obama, Senator Obama
  President of the United States

when did steve leave apple?
Challenges

- Same entity can be represented by multiple surface forms
  Barack Obama, Barack H. Obama, President Obama, Senator Obama
  President of the United States
- Same surface form could refer to multiple entities
Challenges

- **Same entity can be represented by multiple surface forms**
  Barack Obama, Barack H. Obama, President Obama, Senator Obama
  President of the United States

- **Same surface form could refer to multiple entities**
  Michael Jordan – Basketball player or Berkeley professor
Challenges

- **Same entity can be represented by multiple surface forms**
  Barack Obama, Barack H. Obama, President Obama, Senator Obama
  President of the United States

- **Same surface form could refer to multiple entities**
  Michael Jordan – Basketball player or Berkeley professor
  when did steve leave apple?
Challenges

- **Same entity can be represented by multiple surface forms**
  Barack Obama, Barack H. Obama, President Obama, Senator Obama
  President of the United States

- **Same surface form could refer to multiple entities**
  Michael Jordan – Basketball player or Berkeley professor
  when did steve leave apple?

- **Out of KG mentions**
Entity Linking

Related problems:

- Record linkage/de-duplication in databases
- Entity Resolution/name matching
- Co-reference resolution, Word Sense disambiguation
Entity Linking Process

Well studied in NLP [19] open source software like Stanford NLP toolkit [18]

Use of dictionaries

Ranking target entities based on:

• graph based features
• text/document based features
Entity Linking Process

Entity Recognition → Target List Generation → Ranking

Well studied in NLP [19] open source software like Stanford NLP toolkit [18]

Use of dictionaries

Ranking target entities based on:
- graph based features
- text/document based features
Entity Linking Process

- Entity Recognition
- Target List Generation
- Ranking

Named Entity Recognition
Well studied in NLP [19]
open source software like Stanford NLP toolkit [18]
Entity Linking Process

Named Entity Recognition
Well studied in NLP [19]
open source software like Stanford NLP toolkit [18]

Use of dictionaries
Named Entity Recognition
Well studied in NLP [19]
open source software like Stanford NLP toolkit [18]

Use of dictionaries

Ranking target entities based on:
- graph based features
- text/document based features
Candidate Entity List Generation

Much of the variation between different entity linking algorithms could be explained by quality of candidate search components [14].

Acronym expansions and coreference resolutions lead to significant performance gains [14].

The candidate set should be exhaustive enough but not too big to affect efficiency.
Candidate Entity List Generation

- Much of the variation between different entity linking algorithms could be explained by quality of candidate search components [14]
Candidate Entity List Generation

- Much of the variation between different entity linking algorithms could be explained by quality of candidate search components [14]

- Acronym expansions and coreference resolutions lead to significant performance gains [14]
Candidate Entity List Generation

- Much of the variation between different entity linking algorithms could be explained by quality of candidate search components [14]

- Acronym expansions and coreference resolutions lead to significant performance gains [14]

- The candidate set should be exhaustive enough but not too big to affect efficiency
Dictionary based Methods

An offline dictionary of entity names created out of external sources mapping different possible surface forms of entity names to their corresponding entities in the KG.
Candidate Entity List Generation

Dictionary based Methods
An offline dictionary of entity names created out of external sources mapping different possible surface forms of entity names to their corresponding entities in the KG

- Domain specific sources like Gene name dictionary [20]
Candidate Entity List Generation

Dictionary based Methods
An offline dictionary of entity names created out of external sources mapping different possible surface forms of entity names to their corresponding entities in the KG

- Domain specific sources like Gene name dictionary [20]
- Wikipedia/DBPedia
  - Page Titles
  - Disambiguation/Redirect pages
  - Anchor text of Wikipedia in links
Dictionary based Methods

An offline dictionary of entity names created out of external sources mapping different possible surface forms of entity names to their corresponding entities in the KG

- Domain specific sources like Gene name dictionary [20]
- Wikipedia/DBPedia
  - Page Titles
  - Disambiguation/Redirect pages
  - Anchor text of Wikipedia in links
- Anchor text from Web pages to Wikipedia articles
Dictionary based Methods

An offline dictionary of entity names created out of external sources mapping different possible surface forms of entity names to their corresponding entities in the KG

- Domain specific sources like Gene name dictionary [20]
- Wikipedia/DBPedia
  - Page Titles
  - Disambiguation/Redirect pages
  - Anchor text of Wikipedia in links
- Anchor text from Web pages to Wikipedia articles
- Acronym expansions
## Candidate Entity List Generation

<table>
<thead>
<tr>
<th>Surface Form</th>
<th>Entity Canonical Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>&lt;Barack Obama, Person&gt;</td>
</tr>
<tr>
<td>Barack H. Obama</td>
<td>&lt;Barack Obama, Person&gt;</td>
</tr>
<tr>
<td>USA</td>
<td>&lt;United States of America, Country&gt;</td>
</tr>
<tr>
<td>America</td>
<td>&lt;United States of America, Country&gt;</td>
</tr>
<tr>
<td>Big Apple</td>
<td>&lt;New York, City&gt;</td>
</tr>
<tr>
<td>NYC</td>
<td>&lt;New York, City&gt;</td>
</tr>
<tr>
<td>NY</td>
<td>&lt;New York, City&gt;</td>
</tr>
</tbody>
</table>
## Candidate Entity List Generation

<table>
<thead>
<tr>
<th>Surface Form</th>
<th>Entity Canonical Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>&lt;Barack Obama, Person&gt;</td>
</tr>
<tr>
<td>Barack H. Obama</td>
<td>&lt;Barack Obama, Person&gt;</td>
</tr>
<tr>
<td>USA</td>
<td>&lt;United States of America, Country&gt;</td>
</tr>
<tr>
<td>America</td>
<td>&lt;United States of America, Country&gt;</td>
</tr>
<tr>
<td>Big Apple</td>
<td>&lt;New York, City&gt;</td>
</tr>
<tr>
<td>NYC</td>
<td>&lt;New York, City&gt;</td>
</tr>
<tr>
<td>NY</td>
<td>&lt;New York, City&gt;</td>
</tr>
<tr>
<td>NY</td>
<td>&lt;New York, State&gt;</td>
</tr>
</tbody>
</table>
## Candidate Entity List Generation

<table>
<thead>
<tr>
<th>Surface Form</th>
<th>Entity Canonical Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>&lt;Barack Obama, Person&gt;</td>
</tr>
<tr>
<td>Barack H. Obama</td>
<td>&lt;Barack Obama, Person&gt;</td>
</tr>
<tr>
<td>USA</td>
<td>&lt;United States of America, Country&gt;</td>
</tr>
<tr>
<td>America</td>
<td>&lt;United States of America, Country&gt;</td>
</tr>
<tr>
<td>Big Apple</td>
<td>&lt;New York, City&gt;</td>
</tr>
<tr>
<td>NYC</td>
<td>&lt;New York, City&gt;</td>
</tr>
<tr>
<td>NY</td>
<td>&lt;New York, City&gt;</td>
</tr>
<tr>
<td>NY</td>
<td>&lt;New York, State&gt;</td>
</tr>
</tbody>
</table>

Simple term match – partial or exact

...Obama visited Singapore in 2016...

Matches: Barack Obama, Mount Obama, Michelle Obama,..., etc.
Candidate Entity Ranking

The candidate entity set can be big!

For KORE50 dataset:
- 631 candidates on an average per mention in YAGO [25]

Approaches for ranking can be clubbed under two broad categories:
- Text based
- Graph structure based
Candidate Entity Ranking

The candidate entity set can be big!
The candidate entity set can be big!

For KORE50 dataset:

- 631 candidates on an average per mention in YAGO [25]
Candidate Entity Ranking

The candidate entity set can be big!

For KORE50 dataset:

- **631** candidates on an average per mention in YAGO [25]
- **2000+** in Watson KG [5]

Approaches for ranking can be clubbed under two broad categories:

- Text based
- Graph structure based
Candidate Entity Ranking

...Obama is in Hawaii this week...

{Barack Obama, Michelle Obama, Mt. Obama}
...Obama is in Hawaii this week...
{Barack Obama, Michelle Obama, Mt. Obama}

- Similarity between entity name and mention
  - Term overlap, edit distance, etc.
Candidate Entity Ranking

...Obama is in Hawaii this week...
{Barack Obama, Michelle Obama, Mt. Obama}

- Similarity between entity name and mention
  - Term overlap, edit distance, etc.
- Entity Popularity – Wikipedia page views [13, 12]
...Obama is in Hawaii this week... {Barack Obama, Michelle Obama, Mt. Obama}

- Similarity between entity name and mention
  - Term overlap, edit distance, etc.
- Entity Popularity – Wikipedia page views [13, 12]
- Wikipedia/web anchor text/ inlinks [22, 15]
Candidate Entity Ranking

...Obama is in Hawaii this week...
{Barack Obama, Michelle Obama, Mt. Obama}

• Similarity between entity name and mention
  • Term overlap, edit distance, etc.
• Entity Popularity – Wikipedia page views [13, 12]
• Wikipedia/web anchor text/ inlinks [22, 15]

...when did Steve leave apple...
{Steve Jobs, Steve Wozniak, Steve Ballmer}
Candidate Entity Ranking

...Obama is in Hawaii this week...
{Barack Obama, Michelle Obama, Mt. Obama}

- Similarity between entity name and mention
  - Term overlap, edit distance, etc.
- Entity Popularity – Wikipedia page views [13, 12]
- Wikipedia/web anchor text/ inlinks [22, 15]

...when did Steve leave apple...
{Steve Jobs, Steve Wozniak, Steve Ballmer}
Context Matters!
Role of Context

...when did Steve leave apple...

{Steve Jobs, Steve Wozniak, Steve Ballmer}
Role of Context

... when did Steve leave apple...

{Steve Jobs, Steve Wozniak, Steve Ballmer}

- Mention context
  - text of the document/paragraph in which the mention appears
  - a window of terms around the mention
Candidate Entity Ranking

Role of Context

...when did **Steve** leave apple...

\{Steve Jobs,Steve Wozniak,Steve Ballmer\}

- **Mention context**
  - text of the document/paragraph in which the mention appears
  - a window of terms around the mention

- **Entity context representations**
  - Wikipedia article
  - Text around anchors
  - Domain specific models: abstracts of papers containing gene name in titles
Candidate Entity Ranking

Role of Context

...when did Steve leave apple...

\{Steve Jobs,Steve Wozniak,Steve Ballmer\}

- Mention context
  - text of the document/paragraph in which the mention appears
  - a window of terms around the mention

- Entity context representations
  - Wikipedia article
  - Text around anchors
  - Domain specific models: abstracts of papers containing gene name in titles

Compute similarity between mention and entity context representations
Graph Based Features Focus on *strength* between entities, often useful in *collective entity linking*
Candidate Entity Ranking

Graph Based Features Focus on *strength* between entities, often useful in **collective entity linking**

- Simplest graph based measure – Entity Popularity

\[
pop(e) = \frac{nbrCount(e)}{\sum_{e' \in \mathcal{E}} nbrCount(e')}
\]

In Wikipedia graph, inlinks and outlinks can be used to compute popularity
Graph Based Features Focus on *strength* between entities, often useful in **collective entity linking**

- Simplest graph based measure – Entity Popularity

\[
\text{pop}(e) = \frac{\text{nbrCount}(e)}{\sum_{e' \in \mathcal{E}} \text{nbrCount}(e')}
\]  

In Wikipedia graph, inlinks and outlinks can be used to compute popularity

Next we review some measures useful for collective entity linking
Candidate Entity Ranking

Linking/Resolving/Disambiguating Multiple Entities simultaneously

Image Source: [28]
Candidate Entity Ranking

Brad and Angelina were holidaying in Paris.
Brad and Angelina were holidaying in Paris.

- Jaccard Index

\[ J(a, b) = \frac{|A \cap B|}{|A \cup B|} \] (2)

- Adamic Adar

\[ \text{AA}(a, b) = \sum_{n \in A \cup B} \log \left( \frac{1}{\text{degree}(n)} \right) \] (4)
Brad and Angelina were holidaying in Paris.

- **Jaccard Index**
  \[ J(a, b) = \frac{|A \cap B|}{|A \cup B|} \]  

- **Milne-Witten Similarity** [28]
  \[ MW(a, b) = \frac{\log(\max(|A|, |B|)) - \log(|A \cap B|)}{\log(|\mathcal{N}|) - \log(\min(|A|, |B|))} \]  

where, \(A\) and \(B\) are the set of neighbors of entities \(a\) and \(b\), respectively.
Brad and Angelina were holidaying in Paris.

- Jaccard Index

\[
J(a, b) = \frac{|A \cap B|}{|A \cup B|}
\]  

- Milne-Witten Similarity [28]

\[
MW(a, b) = \frac{\log(\max(|A|, |B|)) - \log(|A \cap B|)}{\log(|\mathcal{N}|) - \log(\min(|A|, |B|))}
\]  

where, \( A \) and \( B \) are the set of neighbors of entities \( a \) and \( b \), respectively.

- Adamic Adar [1]

\[
AA(a, b) = \sum_{n \in A \cup B} \log\left(\frac{1}{\text{degree}(n)}\right)
\]
These features can be used in supervised or unsupervised settings.

Choice of features depend on **data/domain** at hand. Many features are specific for Wikipedia, that may not be applicable to other textual data.

Trade off between accuracy and efficiency while designing your systems.
Which search algorithm did Sergey and Larry invent?
Entity Linking as implemented in Watson KG

Which search algorithm did Sergey and Larry invent?

---

Knowledge Graphs for Query Expansion

Even before Knowledge Graphs became popular, query expansion using structured knowledge was shown to be useful [8]. Advancements in entity linking make it more appealing.
Knowledge Graphs for Query Expansion

Even before Knowledge Graphs became popular, query expansion using structured knowledge was shown to be useful [8]. Advancements in entity linking make it more appealing

- All documents in collection are annotated with entity mentions
Knowledge Graphs for Query Expansion

Even before Knowledge Graphs became popular, query expansion using structured knowledge was shown to be useful [8]. Advancements in entity linking make it more appealing

- All documents in collection are annotated with entity mentions
- Perform entity linking for queries
Knowledge Graphs for Query Expansion

Even before Knowledge Graphs became popular, query expansion using structured knowledge was shown to be useful [8]. Advancements in entity linking make it more appealing

- All documents in collection are annotated with entity mentions
- Perform entity linking for queries
- Produce a ranking based on matched terms and knowledge concepts
Knowledge Graphs for Query Expansion

Even before Knowledge Graphs became popular, query expansion using structured knowledge was shown to be useful [8]. Advancements in entity linking make it more appealing

- All documents in collection are annotated with entity mentions
- Perform entity linking for queries
- Produce a ranking based on matched terms and knowledge concepts
- Consider the query \textit{obama in paris}

![Knowledge Graph Example]

Barack Obama visited the French Capital. The US president was in Paris for a visit.
Even before Knowledge Graphs became popular, query expansion using structured knowledge was shown to be useful [8]. Advancements in entity linking make it more appealing

- All documents in collection are annotated with entity mentions
- Perform entity linking for queries
- Produce a ranking based on matched terms and knowledge concepts
- Consider the query *obama in paris*

Additional features and entity properties from KG can also be utilized [29, 29]
Entity Exploration
Entity Exploration

We found the entity of interest.

Knowing more about the entity

- Finding entities related to entity of interest
- Properties of entities
- Going beyond immediate neighborhood of the entity
Entity Retrieval

- Entity Box in web queries
- Lots of useful information about the query entity
- $\approx 40\%$ of all web queries are entity queries [21]
- Many QA queries can be answered by the underlying Knowledge Base
Related Entity Finding track at TREC [3]
Related Entity Finding track at TREC [3]
Input: Entity Name and Search Intent
Related Entity Finding track at TREC [3]
Input: Entity Name and Search Intent
Output: Ranked list of entity documents – entities embedded in documents
Entity Retrieval

Related Entity Finding track at TREC [3]
Input: Entity Name and Search Intent
Output: Ranked list of entity documents – entities embedded in documents
Example:

Query: Blackberry
Intent: Carriers that carry Blackberry phones
Example Answers: Verizon, AT&T, etc.
Components of Related Entity Ranking [9]^2

Components of Related Entity Ranking [9]

For a given input entity $e_s$, type $T$ of target entity, and a relation description $R$, we wish to rank the target entities as follows:

$$P(e|e_s, T, R) \propto P(R|e_s, e) \times P(e|e_s) \times P(T|e) \quad (5)$$

\[\text{Context Modeling} \quad \text{Co-occurrence} \quad \text{Type Filtering}\]

---

Components of Related Entity Ranking [9]²

For a given input entity $e_s$, type $T$ of target entity, and a relation description $R$, we wish to rank the target entities as follows:

$$P(e|e_s, T, R) \propto P(R|e_s, e) \times P(e|e_s) \times P(T|e) \quad (5)$$

Co-occurrence

\[ P(e|e_s) = \frac{cooc(e, e_s)}{\sum_{e' \in E} cooc(e', e_s)} \]
Entity Retrieval

Co-occurrence

\[ P(e|e_s) = \frac{cooc(e, e_s)}{\sum_{e' \in E} cooc(e', e_s)} \]

Type Filtering

- Wikipedia categories
- Named entity recognizer tools
**Entity Retrieval**

### Co-occurrence

\[
P(e|e_s) = \frac{\text{cooc}(e, e_s)}{\sum_{e' \in E} \text{cooc}(e', e_s)}
\]

### Type Filtering
- Wikipedia categories
- Named entity recognizer tools

### Context Modeling

Co-occurrence language model \( \Theta_{ee_s} \) approximated by documents in which \( e, E_s \) co-occur

\[
P(R|e, e_s) = \prod_{t \in R} P(t|\Theta_{ee_s})
\]
Entity recommendations for web search queries[7]³

Entity recommendations for web search queries[7]³

- Co-occurrence features
  - query logs, user sessions
  - flickr and twitter tags

Entity recommendations for web search queries\cite{blanco2013entity}

- Co-occurrence features
  - query logs, user sessions
  - flickr and twitter tags
- frequency

Entity recommendations for web search queries[7]³

- Co-occurrence features
  - query logs, user sessions
  - flickr and twitter tags
- frequency
- Graph theoretic features
  - Page rank on entity graph
  - Common neighbors between two entities

Learning to rank using text and graph based features[23]^4

---

Learning to rank using text and graph based features\cite{schuhmacher2015}

- Given a web query, retrieve relevant documents,
Learning to rank using text and graph based features[23]\(^4\)

- Given a web query, retrieve relevant documents,
- Identify entities present in them using entity linking methods

---

Learning to rank using text and graph based features\cite{23}^4

- Given a web query, retrieve relevant documents,
- Identify entities present in them using entity linking methods
- Rank these entities using graph theoretic and text based features

Learning to rank using text and graph based features[23]  

- Given a web query, retrieve relevant documents,
- Identify entities present in them using entity linking methods
- Rank these entities using graph theoretic and text based features
- Reformulates entity retrieval/recommendation as ad hoc document retrieval

---

Till now, we have focused on finding entities.
Till now, we have focused on finding entities
Let us focus our attention now on finding about entities
Till now, we have focused on finding entities
Let us focus our attention now on finding *about* entities
Till now, we have focused on finding entities
Let us focus our attention now on finding *about* entities

Relationships of similar types can be clustered and then explored based on user requirements [30]
What are the most important facts about an entity?\textsuperscript{5}

What are the most important facts about an entity? Given a source entity $e_s$, we wish to compute the probability $P(r, e_t | e_s)$.

$$P(r, e_t | e_s) \propto P(e_t) \times P(e_s | e_t) \times P(r | e_s, e_t)$$  \hspace{1cm} (6)
What are the most important facts about an entity? Given a source entity $e_s$, we wish to compute the probability $P(r, e_t | e_s)$

$$P(r, e_t | e_s) \propto P(e_t) \times P(e_s | e_t) \times P(r | e_s, e_t)$$  \hspace{1cm} (6)

**Entity Prior:**

$$P(e_t) \propto relCount(e_t)$$  \hspace{1cm} (7)

---

What are the most important facts about an entity? Given a source entity $e_s$, we wish to compute the probability $P(r, e_t | e_s)$

$$P(r, e_t | e_s) \propto P(e_t) \times P(e_s | e_t) \times P(r | e_s, e_t)$$  \hspace{1cm} (6)

**Entity Prior:**

$$P(e_t) \propto \text{relCount}(e_t)$$  \hspace{1cm} (7)

**Entity Affinity**

$$P(e | e_t) = \frac{\sum_{r_i \in R(e_s, e_t)} w(r_i) \times r_i}{\sum_{r_i \in R(e_t)} w(r_i) \times r_i}$$  \hspace{1cm} (8)

---

Entity Exploration – Fact Ranking

What are the most important facts about an entity? Given a source entity $e_s$, we wish to compute the probability $P(r, e_t|e_s)$.

\[
P(r, e_t|e_s) \propto P(e_t) \times P(e_s|e_t) \times P(r|e_s, e_t)
\]  

(6)

Entity Prior:

\[
P(e_t) \propto \text{relCount}(e_t)
\]  

(7)

Entity Affinity

\[
P(e|e_t) = \frac{\sum_{r_i \in R(e_s, e_t)} w(r_i) \times r_i}{\sum_{r_i \in R(e_t)} w(r_i) \times r_i}
\]  

(8)

Relationship Strength

\[
P(r|e_s, e_t) = \frac{\text{mentionCount}(r, e_s, e_t)}{\sum_{r \in R(e_s, e_t)} \text{mentionCount}(r, e_s, e_t)}
\]  

(9)

---

Entity Exploration – Moving Beyond the Neighborhood

Till now, we have limited our attention to relations of the entity and it’s immediate neighborhood.
Till now, we have limited our attention to relations of the entity and it’s immediate neighborhood. What lies after that?
Discovering and Explaining Higher Order Relations Between Entities
Discovering and Explaining Higher Order Relations Between Entities

Charlie Hebdo Shooting Attack
Entity Exploration – Moving Beyond the Neighborhood

Discovering and Explaining Higher Order Relations Between Entities

Charlie Hebdo Shooting Attack
Discovering and Explaining Higher Order Relations Between Entities

Can we tell how are they connected?
Path Ranking

- Thousands of such paths
- Too generic – obvious relations

Abu bakr al-baghdadi

- Nasir Al-Wuhayshi
- AQAP
- Man Haron Monis
- Hizb Ut-Tahrir Australia
- Islamic state of The Caliph
- Terrorist Event
- War
- Gaza
- Mosque
- Attack
- Irani
- London
- U.S.
- Paris

Charlie Hebdo
Path Ranking

- Thousands of such paths
- Too generic – obvious relations
Path Ranking

Three components for ranking possible paths [2]

Specificity: Popular entities given lower scores

\[ \text{spec}(p) = \sum_{e \in p} \text{spec}(e) \]

where:

\[ \text{spec}(e) = \log(1 + 1 / \text{docCount}(e)) \] (10)

Reduces generic paths, but boosts noise entities

Connectivity: A strongly connected path consists of strong edges.

\[ \text{score}(e_a, e_b) = \vec{d}_{e_a} \cdot \vec{d}_{e_b} \] (11)

Cohesiveness:

\[ \text{score}(p) = n - 1 \sum_{i=2}^{n} \text{score}(e_i) = n - 1 \sum_{i=2}^{n} \vec{d}_{e_i} \cdot \vec{d}_{e_{i+1}} \] (12)
Path Ranking

Three components for ranking possible paths [2]

**Specificity**: Popular entities given lower scores

\[ spec(p) = \sum_{e \in p} spec(e); \text{ where: } spec(e) = \log(1 + 1/docCount(e)) \] (10)

Reduces generic paths, but boosts noise entities
Path Ranking

Three components for ranking possible paths [2]

Specificity: Popular entities given lower scores

\[
spec(p) = \sum_{e \in p} spec(e); \text{ where: } spec(e) = \log(1 + 1/docCount(e)) \tag{10}
\]

Reduces generic paths, but boosts noise entities

Connectivity: A strongly connected path consists of strong edges.

\[
score(e_a, e_b) = d_{e_a} \cdot d_{e_b} \tag{11}
\]
Path Ranking

Three components for ranking possible paths [2]

**Specificity:** Popular entities given lower scores

\[ spec(p) = \sum_{e \in p} spec(e); \text{ where: } spec(e) = \log(1 + 1/docCount(e)) \] (10)

Reduces generic paths, but boosts noise entities

**Connectivity:** A strongly connected path consists of strong edges.

\[ score(e_a, e_b) = \vec{d}_{e_a} \cdot \vec{d}_{e_b} \] (11)

**Cohesiveness:**

\[ score(p) = \sum_{i=2}^{n-1} score(e_i) = \sum_{i=2}^{n-1} \vec{d}_{e_i-1} \cdot \vec{d}_{e_{i+1}} \] (12)
Wikipedia:
Aamir Khan’s Bollywood movie Ghajini was the remake of Hollywood movie Memento directed by Christopher Nolan.
Wikipedia:
Shortly after the Syrian uprising began against the Syrian administration headed by Syrian president Bashar al-Assad, al-Julani moved into Syrian territory and, fully supported by al-Baghdadi.... Bashar was supported by major general Qasem Soleimani ....
Application Example from Life Sciences

Predicting Drug-Drug Interactions (DDI)\(^6\)

---

Predicting Drug-Drug Interactions (DDI)\(^6\)

- DDI are a major cause of preventable adverse drug reactions

Application Example from Life Sciences

Predicting Drug-Drug Interactions (DDI)\(^6\)

- DDI are a major cause of preventable adverse drug reactions
- Clinical studies cannot accurately determine all possible DDIs

---

Application Example from Life Sciences

Predicting Drug-Drug Interactions (DDI)\(^6\)

- DDI are a major cause of preventable adverse drug reactions
- Clinical studies can not accurately determine all possible DDIs
- Can we utilize knowledge about drugs to predict possible DDIs?

Create a KG out of existing information about drugs and their interactions with genes, enzymes, molecules, etc.
Application Example from Life Sciences

- Given a pair of drugs, extract features based on physiological effect, side effect, targets, drug targets, chemical structure, etc.
- Perform supervised classification using logistic regression
- Retrospective Analysis: Known DDIs til January 2011 as training.
- Could predict $\approx 68\%$ of DDIs discovered after January 2011 till December 2014.
Upcoming Areas and Future Directions
Future Research Directions

Explaining relations present in a graph [26, 16, 4]
Explaining relations present in a graph [26, 16, 4]

p21 is a potent cyclin-dependent kinase inhibitor (CKI). The p21 (CIP1/WAF1) protein binds to and inhibits the activity of cyclin-CDK2, -CDK1, and -CDK4/6 complexes, and thus functions as a regulator of cell cycle progression at G1 and S phase.
Explaining relations present in a graph [26, 16, 4]

Graph and text joint modeling [27, 31]
Future Research Directions

- Reasoning over Knowledge Graphs
  - KG Completion [10, 24, 17]
  - Complex QA Systems
Future Research Directions

- Reasoning over Knowledge Graphs
  - KG Completion [10, 24, 17]
  - Complex QA Systems

**Question:** Which type of scientist would study the relationship between simple machines and energy?
(A) chemist; (B) biologist; (C) physicist; (D) geologist
Future Research Directions

- Reasoning over Knowledge Graphs
  - KG Completion [10, 24, 17]
  - Complex QA Systems

Question: Which type of scientist would study the relationship between simple machines and energy?
(A) chemist; (B) biologist; (C) physicist; (D) geologist

Question: Which is a physical property of an apple?
(A) what color it is; (B) how pretty it is; (C) how much it costs; (D) when it was picked
Future Research Directions

Complex QA Systems Allen Institute for AI has a complex QA challenge running

- Easy Set: Best system has an accuracy of 68.90%; IR-Solver has 62.55%
- Challenge Set: Best system has an accuracy of 42.32%; IR-Solver has 20.26%

IR-Solver is pretty basic – tf-idf scoring. I know we can certainly do much better (because I have done with a student). Currently, we have around 70% on Easy Set and 38% on challenge set using language modeling, query expansion, and relevance feedback.

Goal is to understand limitations of IR and where advanced reading comprehension techniques should focus rather than questions that IR can answer using term co-occurrences.
Future Research Directions

**Complex QA Systems** Allen Institute for AI has a complex QA challenge running

- **Easy Set:** Best system has an accuracy of 68.90%; IR-Solver has 62.55%
  - **Challenge Set:** Best system has an accuracy of 42.32%; IR-solver has 20.26%
Complex QA Systems Allen Institute for AI has a complex QA challenge running

- Easy Set: Best system has an accuracy of 68.90%; IR-Solver has 62.55%
  - Challenge Set: Best system has an accuracy of 42.32%; IR-solver has 20.26%

IR-Solver is pretty basic – tf-idf scoring
Complex QA Systems Allen Institute for AI has a complex QA challenge running

- Easy Set: Best system has an accuracy of 68.90%; IR-Solver has 62.55%
- Challenge Set: Best system has an accuracy of 42.32%; IR-solver has 20.26%

IR-Solver is pretty basic – tf-idf scoring

I know we can certainly do much better
Complex QA Systems  Allen Institute for AI has a complex QA challenge running

- Easy Set: Best system has an accuracy of 68.90%; IR-Solver has 62.55%
- Challenge Set: Best system has an accuracy of 42.32%; IR-solver has 20.26%

IR-Solver is pretty basic – tf-idf scoring

I know we can certainly do much better (because I have done with a student)
Future Research Directions

Complex QA Systems Allen Institute for AI has a complex QA challenge running

- Easy Set: Best system has an accuracy of 68.90%; IR-Solver has 62.55%
- Challenge Set: Best system has an accuracy of 42.32%; IR-solver has 20.26%

IR-Solver is pretty basic – tf-idf scoring

I know we can certainly do much better (because I have done with a student)

Currently, we have around 70% on Easy Set and 38% on challenge set using language modeling, query expansion, and relevance feedback.
Future Research Directions

Complex QA Systems Allen Institute for AI has a complex QA challenge running

- Easy Set: Best system has an accuracy of 68.90%; IR-Solver has 62.55%
- Challenge Set: Best system has an accuracy of 42.32%; IR-solver has 20.26%

IR-Solver is pretty basic – tf-idf scoring

I know we can certainly do much better (because I have done with a student)

Currently, we have around 70% on Easy Set and 38% on challenge set using language modeling, query expansion, and relevance feedback.

Goal is to understand limitations of IR and where advanced reading...
Conclusions

- KG can provide structure to your unstructured data!
Conclusions

- KG can provide structure to your unstructured data!
- They thus complement traditional IR paradigm of unstructured data
Conclusions

- KG can provide structure to your unstructured data!
- They thus complement traditional IR paradigm of unstructured data
- Many time-tested IR principles can in turn benefit Knowledge Graph exploration
Conclusions

- KG can provide structure to your unstructured data!
- They thus complement traditional IR paradigm of unstructured data
- Many time-tested IR principles can in turn benefit Knowledge Graph exploration
- We wanted to provide an overview of tools/techniques that have worked well in the past, and challenges you may face
Conclusions

- KG can provide structure to your unstructured data!
- They thus complement traditional IR paradigm of unstructured data
- Many time-tested IR principles can in turn benefit Knowledge Graph exploration
- We wanted to provide an overview of tools/techniques that have worked well in the past, and challenges you may face
- Be careful in selecting the KG appropriate for your domain and requirements.
Conclusions

- KG can provide structure to your unstructured data!
- They thus complement traditional IR paradigm of unstructured data
- Many time-tested IR principles can in turn benefit Knowledge Graph exploration
- We wanted to provide an overview of tools/techniques that have worked well in the past, and challenges you may face
- Be careful in selecting the KG appropriate for your domain and requirements.
- Keep in mind the scale and efficiency issues.
Thanks!!!
Suggestions and Questions Welcome!

Slides available at http://sumitbhatia.net/source/knowledge-graph-tutorial.html


