Understand Relations in Knowledge Base Construction

Han Wang

December 2nd, 2014
Outline

- Knowledge Base Construction
- Relation Extraction
- Understand Relations
What is a Knowledge Base (KB)?

“Comprehensive and semantically organized machine-readable collection of universally relevant or domain-specific entities, classes, and SPO facts (attributes, relations).”

“plus spatial and temporal dimensions plus commonsense properties and rules plus contexts of entities and facts (textual & visual witnesses, descriptors, statistics) plus …..”

— Gerhard Weikum
A Sample KB

HanWang type GraduateStudent
HanWang type PhDStudent
PhDStudent subClassOf GraduateStudent

HanWang hasAdvisor PeterFox
HanWang memberOf TetherlessWorldConstellation
TetherlessWorldConstellation subOrganizationOf RensselaerPolytechnicInstitute
RensselaerPolytechnicInstitute hasAbbreviation RPI
RensselaerPolytechnicInstitute locatedAt Troy

HanWang passed ResearchQualifier
HanWang researchQualifierDate “2014-12-02”
What can a KB do?

- Disambiguation
  - “Han is pursuing a Ph.D. with Prof. Fox after he got an MSIT degree at RPI.”
  - “Later on, Han, accompanied by Chewbacca, Leia, and Deena Shan, traveled to the planet of Aguarl 3.

- Question answering
  - Where does Han live in 2014?

- Semantic search
  - Entities and relations
KB History

WordNet
A lexical database for English
155,287 words
117,659 synsets

Cyc

Wikipedia
Over 4.6 million articles
Over 73,000 active editors


DBpedia
4.58 million things
583 million facts

Freebase
46 million things
2.7 billion facts

yago
10 million things
120 million facts

WolframAlpha

Facebook
Entity Graph

Microsoft Satori

Google Knowledge Graph

Google Knowledge Vault
Research Tasks around KB – KB Growth

- Taxonomy extraction/construction
  - Define KB backbone - classes
  - Human work: Wordnet
    - Establish mapping between Wikipedia categories and Wordnet

- Entity extraction
  - Find instances of KB classes (unary “is-a” relation)
  - Use lexical patterns in free text (such as X and Y, including X and Y)
  - Also extract from web tables, lists, etc.

- Relation extraction
  - Find facts about entities

- KB alignment
  - Connect KBs
  - Establish mapping between classes
  - Measure (string, lexical, semantic, etc.) similarities

- Cold start
  - Construct KBs without external KBs
Research Tasks around KB — KB Validation

- Entity resolution
  - Merge duplicated entities
  - Entity linking

- Relation resolution
  - Merge duplicated relations
  - Relation clustering/paraphrasing

- Temporal and spatial validation
  - Add temporal and/or spatial constraints to facts

- Error detection
  - Remove false statements
Research Tasks around KB — KB Intelligence

• Knowledge representation

• Semantic search
  • Search entities instead of strings

• Question answering
  • Search for answers based on KBs

• Automatic reasoning
Related Fields

• Information Retrieval

• Natural Language Processing

• Semantic Web

• Machine Learning
Relation Extraction

• Extracts semantic relations between entities from text

• Numerous foci:
  • Supervision: full vs. semi vs. un
  • Unary relations vs. binary relations vs. n-ary relations
  • open extraction vs. fixed set extraction
  • Extraction source: semi-structured data vs. free text
  • Feature used: only lexical vs. full linguistic features
  • Extract temporal information
Distant Supervised Relation Extraction

• Input: an existing KB + unlabeled text
  • For a pair entities participating in a relation in the KB, find them in some sentences in the text.
  • Use these sentences as the training data for extracting this relation.

• Han Wang, a graduate student in RPI, …
• Han received an MSIT degree from RPI.
• Han is currently working as a research assistant in RPI.
• Han watched a hockey game between RPI and Yale.
Learning Features (Sentence Level)

• Lexical: words between and around entity mentions and their POS tags

• Syntactic: dependency parse paths between entity mentions and other words

• Named entity tags

• Conjunction of the above features
Distant Supervision Hypotheses

• All sentences containing the entity pair express the relation. [Mintz et al., 2009]
Distant Supervision Hypotheses

- All sentences containing the entity pair express the relation. [Mintz et al., 2009]

- At least one sentence containing the entity pair expresses the relation. [Riedel et al., 2010]
Distant Supervision Hypotheses

- All sentences containing the entity pair express the relation. [Mintz et al., 2009]

- At least one sentence containing the entity pair express the relation. [Riedel et al., 2010]

- At least one sentence containing the entity pair expresses the relation, and these these two entities can have multiple relations. [Hoffmann et al., 2011; Surdeanu et al., 2012]
Relaxing hypotheses Works

[Weston et al., 2013]
Never-Ending Language Learning (NELL) [Carlson et al., 2010]

![Recently-Learned Facts](image)

<table>
<thead>
<tr>
<th>instance</th>
<th>iteration</th>
<th>date learned</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>honda_sabre is a model of automobile</td>
<td>881</td>
<td>24-oct-2014</td>
<td>92.6</td>
</tr>
<tr>
<td>tower_185 is a scraper</td>
<td>883</td>
<td>02-nov-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>nucleus_tenaiae is a kind of brain tissue</td>
<td>883</td>
<td>02-nov-2014</td>
<td>99.9</td>
</tr>
<tr>
<td>disorders_4 is a physiological condition</td>
<td>881</td>
<td>24-oct-2014</td>
<td>94.8</td>
</tr>
<tr>
<td>n1_02 is a term used by physicists</td>
<td>883</td>
<td>02-nov-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>tie is a clothing item to go with tux</td>
<td>886</td>
<td>21-nov-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>caterpillar is a company also known as cat</td>
<td>886</td>
<td>21-nov-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>philippines is a country also known as tanzania</td>
<td>884</td>
<td>08-nov-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>arizona_diamondbacks is a sports team that won the world series</td>
<td>881</td>
<td>24-oct-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>clinton is a U.S. politician endorsed by the U.S. politician richard_nixon</td>
<td>883</td>
<td>02-nov-2014</td>
<td>96.9</td>
</tr>
</tbody>
</table>

http://rtw.ml.cmu.edu/rtw/
Never-Ending Language Learning (NELL) [Carlson et al., 2010]

- Extracts both entities and relations

- Bootstrapping semi-supervised learning
  - Starts with a small amount of labeled data as seeds
  - Grows the seeds with high-confidence instances found by the current model
  - Suffers from “semantic drift”

- Overcomes the semantic drift by constraining the semi-supervised learning
  - Mutual exclusion. E.g. a “PhDStudent” cannot also be a “Professor”
  - Relation argument type checking. E.g. arguments of “employeeOf” should be “Person” and “Organization”
  - Unstructured and semi-structured text features. E.g. extractions from free text and web tables should agree with each other
Knowledge Vault [Dong et al., 2014]

- Automatically extracts triples from the Web
  - Various extraction sources: free text, HTML DOM trees, HTML tables, and human annotations (schema.org, RDFa)
  - Uses distant supervision to get triples as the training data.
  - Uses local closed world assumption (a.k.a. partial completeness assumption) to assign labels to training data
  - Extraction results are fused together from different sources with a confidence score
- Uses prior knowledge (Freebase) to evaluate the extracted triples
  - Uses Freebase triples to predict the probabilities of the extracted triples
- Combines extraction and prior together
### Knowledge Vault [Dong et al., 2014]

<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>271M</td>
</tr>
<tr>
<td>DeepDive [32]</td>
<td>4</td>
<td>2.7M</td>
<td>34</td>
<td>7M(^a)</td>
</tr>
<tr>
<td>NELL [8]</td>
<td>271</td>
<td>5.19M</td>
<td>306</td>
<td>0.435M(^b)</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(^c)</td>
</tr>
<tr>
<td>Freebase [4]</td>
<td>1,500</td>
<td>40M</td>
<td>35,000</td>
<td>637M(^d)</td>
</tr>
<tr>
<td>Knowledge Graph (KG)</td>
<td>1,500</td>
<td>570M</td>
<td>35,000</td>
<td>18,000M(^e)</td>
</tr>
</tbody>
</table>

Table 1: Comparison of knowledge bases. KV, DeepDive, NELL, and PROSPERA rely solely on extraction, Freebase and KG rely on human curation and structured sources, and YAGO2 uses both strategies. Confident facts means with a probability of being true at or above 0.9.

\(^a\)CE Zhang (U Wisconsin), private communication

\(^b\)Bryan Kiesel (CMU), private communication


\(^d\)This is the number of non-redundant base triples, excluding reverse predicates and “lazy” triples derived from flattening CVTs (complex value types).

Open Information Extraction

- Extracts relational tuples without any pre-defined vocabulary or training data

- Extracts verb phrases as the relations [Fader et al., 2011]
  - Longest sequences containing the verb and starting with the verbs

  - Syntactic constraint: matching POS patterns (e.g. VB, VB DET | PREP, VB NP PREP)
  - Lexical constraint: number of distinct argument pairs > threshold

- Extracts relations mediated by nouns, adjectives, and more [Mausam et al., 2012]
  - Open pattern templates: syntactic patterns + semantic/lexical constraints
Open Information Extraction

Over 5 billion extractions from over a billion web pages
Revisit: Relations

- A relation can be expressed in many ways (patterns)
- How to establish the “definition” of a relation?
- What makes two relations different?
- How to measure the similarity of two relations?

- Han Wang, a graduate student in RPI, …
- Han received an MSIT degree from RPI.
**PATTY [Nakashole et al., 2012]**

- A Taxonomy of Relational Patterns with Semantic Types

- Syntactic-Lexical-Ontological (SOL) Pattern Model
  - Syntactic-Lexical: surface words, wildcards, POS tags
  - Ontological: semantic type classes as entity placeholders
  - Support set of pattern: set of entity-pairs for placeholders

- An example:
  - Pattern: `<Person>'s ADJECTIVE voice * in <Song>`
  - Matching sentences:
    - *Amy Winehouse’s soul voice in her song ‘Rehab’*
    - *Jim Morrison’s haunting voice and charisma in ‘The End’*
    - *Joan Baez’s angel-like voice in ‘Farewell Angelina’*
  - Support set:
    - *(Amy Winehouse, Rehab)*
    - *(Jim Morrison, The End)*
    - *(Joan Baez, Farewell Angelina)*
PATTY [Nakashole et al., 2012]

- Synonymous patterns (pattern synsets)
  - The support sets of two patterns are equivalent
  - `<Person>` graduated from `<School>` = `<Person>` obtained degree in * from `<School>`
  - and PRONOUN ADJECTIVE advisor `<Person>` = under the supervision of `<Person>`

- Subsumption patterns
  - The support set of a pattern is the subset of the support set of another pattern
  - `<Person>` nominated for `<Award>` \[\]
  - `<Person>` winner of `<Award>`
PATTY [Nakashole et al., 2012]

350,000 SOL patterns with 4 million instances
VerbNet [Kipper 2008]

- A verb lexicon that groups approximately 5,800 verbs in over 270 classes

- Inspired by Beth Levin’s classification of verb classes and their syntactic alternations [Levin, 1993]
  - Members within a single verb class participate in shared types of alternations because of an underlying shared semantic meaning

- Verbs are classified according to shared syntactic behaviors
# VerbNet [Kipper 2008]

## spray-9.7

<table>
<thead>
<tr>
<th>Members: 1, Frames: 4</th>
</tr>
</thead>
</table>

### Members

**OVERLOAD (WN 3, 1; G 1)**

### Roles

- **Agent [+animate]**
- Theme
- Destination [+location & -region]

### Frames

#### NP V NP PP:destination

**Example:** "Jessica loaded boxes into the wagon."

**Syntax:** Agent V Theme ((+loc | +dest_conf)) Destination

**Semantics:** motion(during(E), Theme) NOT(Prep(start(E), Theme, Destination)) Prep(end(E), Theme, Destination) cause(Agent, E)

#### NP V NP:theme

**Example:** "Jessica loaded the wagon with boxes."

**Syntax:** Agent V Destination {with} Theme

**Semantics:** motion(during(E), Theme) NOT(location(start(E), Theme, Destination)) location(end(E), Theme, Destination) cause(Agent, E)

#### NP V NP:theme

**Example:** "Jessica squirted water."

**Syntax:** Agent V Theme

**Semantics:** motion(during(E), Theme) NOT(location(start(E), Theme, ?Destination)) location(end(E), Theme, ?Destination) cause(Agent, E)

#### NP V NP:destination

**Example:** "Jessica sprayed the wall."

**Syntax:** Agent V Destination

**Semantics:** motion(during(E), ?Theme) NOT(location(start(E), ?Theme, Destination)) location(end(E), ?Theme, Destination) cause(Agent, E)
<table>
<thead>
<tr>
<th>Class Number</th>
<th>Verb Type</th>
<th>Verb Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Verbs of Removing</td>
<td>remove-10.1, banish-10.2, clear-10.3, wipe, manner-10.4.1, wipe, inst-10.4.2, steal-10.5, cheat-10.6, pit-10.7, debunk-10.8, mine-10.9, fire-10.10, resign-10.11</td>
</tr>
<tr>
<td>12</td>
<td>Verbs of Exerting Force: Push/Pull Verbs</td>
<td>drive-11.5, push-12</td>
</tr>
<tr>
<td>14</td>
<td>Learn Verbs</td>
<td>learn-14</td>
</tr>
<tr>
<td>15</td>
<td>Hold and Keep Verbs</td>
<td>hold-15.1, keep-15.2</td>
</tr>
<tr>
<td>16</td>
<td>Verbs of Concealment</td>
<td>concealment-16</td>
</tr>
<tr>
<td>17</td>
<td>Verbs of Throwing</td>
<td>throw-17.1, pelt-17.2</td>
</tr>
<tr>
<td>18</td>
<td>Verbs of Contact by Impact</td>
<td>hit-18.1, swat-18.2, spank-18.3, bump-18.4</td>
</tr>
<tr>
<td>19</td>
<td>Poke Verbs</td>
<td>poke-19</td>
</tr>
<tr>
<td>20</td>
<td>Verbs of Contact; Touch Verbs</td>
<td>touch-20</td>
</tr>
<tr>
<td>21</td>
<td>Verbs of Cutting</td>
<td>cut-21.1, carve-21.2</td>
</tr>
<tr>
<td>23</td>
<td>Verbs of Separating and Disassembling</td>
<td>separate-23.1, split-23.2, disassemble-23.3, differ-23.4</td>
</tr>
<tr>
<td>24</td>
<td>Verbs of Coloring</td>
<td>coloring-24</td>
</tr>
<tr>
<td>25</td>
<td>Image Creation Verbs</td>
<td>image, impression-25.1, scribble-25.2, illustrate-25.3, transcribe-25.4</td>
</tr>
</tbody>
</table>
What else can be done?

• PATTY measures the similarity between relation (patterns) using their argument (entity) instance sets — an extensional way

• Verbnet defines verbs intensionally via shared syntactic behaviors (semantic meanings).

• Pattern attributes
  • Surface words
  • POS tags
  • Dependency paths
  • Semantic roles
  • More deep semantic features (e.g. AMR)
  • Relation (pattern) embedding

• How about combining them together?
Application: Provenance Modeling

- What *calibrations* have been applied to this *image*?

**Solar Science concepts**

**Data Processing concepts**

**Provenance concepts**

Diagram:

- Raw Image
  - Optics Calibration Process
    - Flat-field Calibration
    - Angle of Incidence Calibration
  - Data Calibration Process
    - Junk Data Filter
  - Data Filtering Process
    - Data Product
References


Thank you!

Questions? Comments? Suggestions?
The dependency relation views the (finite) verb as the structural center of all clause structure. All other syntactic units (e.g. words) are either directly or indirectly dependent on the verb.

Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas.
(Back-up) Semantic Role Labeling

Detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles.
(Back-up) Abstract Meaning Representation

AMR captures “who is doing what to whom” in a sentence. Each sentence is represented as a rooted, directed, acyclic graph with labels on edges (relations) and leaves (concepts).

The boy wants the girl to believe him.
The boy wants to be believed by the girl.
The boy has a desire to be believed by the girl.
The boy’s desire is for the girl to believe him.
The boy is desirous of the girl believing him.