Automated knowledge base construction

1. Introduction

Simon Razniewski
Summer term 2022
Outline

1. Introducing each other
2. Course organization
3. What, Why, How
4. Lab 1
Simon Razniewski

• Senior Researcher at MPII, Department 5
  • Heading “Knowledge Base Construction and Quality” area

• Background
  • Assistant professor FU Bozen-Bolzano, Italy, 2014-2017
  • PhD FU Bozen-Bolzano, 2014
  • Diplom at TU Dresden, 2010

• Research areas:
  • Logics, databases, Semantic Web
  • More recently IR, (applied) NLP, ML, ...

• Research focus: Knowledge base construction and quality
  • Analyzing what knowledge bases know, and what they don’t
  • Commonsense knowledge base construction
Tutorial teachers

• Hiba Arnaout
• Shrestha Ghosh
• Sneha Singhania
• Tuan-Phong Nguyen

• Doctoral researchers at D5, MPII
• Knowledge base construction, question answering, knowledge coverage, commonsense knowledge, ...
Department 5

- Department 5: Database and information systems

- Knowledge discovery: *extracting, organizing, searching, exploring and ranking facts* from *structured, semi-structured, textual* and *multimodal information sources*

- **yago**
  - Knowledge Base
    - Earliest prominent machine-generated knowledge base (2007)
    - Contains **more than 10 million entities** and **more than 120 million facts**
  
- Gerhard Weikum 259th most cited computer scientist worldwide
And you?

- Course of study
- Preknowledge
- ...
- Comments?

- https://tinyurl.com/4xpk8enh
Outline

1. Introducing each other
2. Course organization
3. What, Why, How
4. Lab 1
Learning outcomes

• **Knowledge**
  • What AKBC is about (“What”)
  • What AKBC is good for (“Why”)
  • What main tasks and challenges in AKBC are
  • What common approaches to problems in AKBC are (“How”)

• **Skills**
  • Analyze potentials and limitations of AKBC approaches
  • Learn to choose right source and method for right task
  • Implement simple solutions for main problems in AKBC
    • Scraping, typing, linking, …

• **Abilities**
  • Build your own AKBC pipeline for a problem

→ Very practical focus!
Prerequisites

• Basic concepts of ML
  • We won’t go deep

• Python programming
  • Essential
  • Still time to learn

• Helpful but not required
  • Basic notions of information retrieval (IRDM?)
  • Computational linguistics (SNLP?)
Formal organization

• Credit points: 6, **hours: 180 (!)**

• Registration
  ▪ Subscribe to the mailing list https://groups.google.com/g/akbc2022/
  ▪ Register in HISPOS until 4.7. for the exam

• When?
  • Lecture: Wednesday 12:15-13:45
  • Lab: Wednesday 16:15-17:45

• How to pass this course?
  ▪ **8 small practical assignments**
    ▪ Pass/fail
    ▪ To be admitted to exam, pass at least 6
  ▪ Oral exam
Assignments

• Published on lecture day (Wednesday)
• Due Monday 23:59 the week after

• Labs are there to start solving the assignments

• Discussing assignments together is allowed, but each student must write their own solution
  • No sharing of code!
  • Plagiarism = course failed for both
  • Avoid triangular plagiarism = cite sources
    • “Approach for NER adapted from stackoverflow.com/how-to-...”

• Libraries that solve core tasks not allowed
  • In doubt ask..

• Weekly assignments are evil!? 
  • Established psychological “trick” to help you learn and pass!
Assignment content

- Coding
- 4 assignments are in competition format
  - Crisp input/output problem specification
    - “From the first sentence of Wikipedia, extract the type of an entity”
  - Labelled training/test data set
  - Unseen (hidden) evaluation dataset
    - To avoid overfitting
  → Ranked list by a standard metric, e.g., precision or F1-score
    - But pass/fail does not depend on rank
## Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Lecture</th>
<th>Tutorial (tutor)</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.4.</td>
<td>1. Introduction (pdf)</td>
<td>Data familiarization (Sneha)</td>
</tr>
<tr>
<td>4.5.</td>
<td>2. Crawling and Scraping (pdf)</td>
<td>Scraping (Phong)</td>
</tr>
<tr>
<td>11.5.</td>
<td>3. Entity typing (pdf)</td>
<td>Typing from first WP sentence (Hiba)</td>
</tr>
<tr>
<td>18.5.</td>
<td>4. Taxonomy induction, coreference and disambiguation (pdf)</td>
<td>Taxonomy induction (Hiba)</td>
</tr>
<tr>
<td>25.5.</td>
<td>5. Relation extraction</td>
<td>Relation extraction (Shrestha)</td>
</tr>
<tr>
<td>1.6.</td>
<td>6. Relation extraction II</td>
<td>Open information extraction (Shrestha)</td>
</tr>
<tr>
<td>8.6.</td>
<td>7. Commonsense knowledge</td>
<td>Commonsense (Phong)</td>
</tr>
<tr>
<td>15.6.</td>
<td>8. Language models and knowledge bases</td>
<td>KBC from LMs (Sneha)</td>
</tr>
<tr>
<td>22.6.</td>
<td>9. Applications</td>
<td>Exam preparation (Simon)</td>
</tr>
<tr>
<td>29.6.</td>
<td>10. TBD / Backup slot</td>
<td>TBD / Backup slot</td>
</tr>
<tr>
<td>11.7.+12.7.</td>
<td>Oral exam (register till 4.7. in LSF)</td>
<td>-</td>
</tr>
<tr>
<td>12.9.</td>
<td>Re-exam</td>
<td>-</td>
</tr>
</tbody>
</table>
Outline

1. Introducing each other
2. Course organization
3. What, Why, How
4. Lab 1
3. Introduction to AKBC

I. Motivation
II. Terminology
III. Topics
IV. Construction techniques
V. Applications
VI. Past, present and future
I. Motivation
• https://en.wikipedia.org/wiki/Max_Planck_Institute_for_Informatics

• https://www.wikidata.org/wiki/Q565400
What for?

• One central hub for *interlanguage interlinking* of 100+ Wikipedia editions

• Your AI chatbot wants to know where MPII, MIT and KAIST are located? → structured query

• A library wants to *distinguish* which of the 100+ literary John Smiths *wrote* “A *description of New England*”? → Wikidata ID
Samples of advanced queries

• Who discovered the most planets:
  http://tinyurl.com/y7rldyqc

• Distribution of places ending with “-weiler” in Germany:
  https://w.wiki/67o

• Living relatives of Louis XIV of France:
  https://w.wiki/549E
The Semantic Web

• Term coined by Tim Berners-Lee for a machine-readable Web

• Web content originally from humans for humans

→ Make machines read human language, or make humans write machine-readable structured data?
3. Introduction to AKBC

I. Motivation

II. Terminology

III. Topics

IV. Construction techniques

V. Applications

VI. Past, present and future
Facts (triples) and their constituents

- **Entities**: Objects about which statements can be made
  
  *Paris; Trump; Irony*

- **Property/predicate/relation/attribute**: What can be said
  
  *locatedIn(entity, location), worksAt(person, organization), antonymOf(term, term)*

- **Fact/statement/claim/triple**: Core building block of KBs
  
  *<Paris, locatedIn, France>*

  ➔ General form:

  `<subject, predicate, object>
  
  <s, p, o>`
Subjects and objects

- Machine-generated identifiers
  - Wikidata: Q4262, Q67245

- Canonical name strings
  - DBpedia, YAGO: “John_Smith_(politician)”

- Internationalized resource identifier (IRI)
  - Semantic web: http://dbpedia.org/resource/Max_Planck

- General phrases
  - TupleKB: <industry, grow over, past few decade>

- Literals: Attribute values that are no entities
  - www.mpi-inf.mpg.de
  - Often with units: 1.63m; 54.85° N

- Same for predicates, sometimes canonicalized, sometimes just text
Classes and class hierarchies

• **Classes/types**: Allow to group similar entities

  Presidents, nouns, Greek gods

• **Type/property hierarchy**: Tree-like hierarchy among types/properties (cf. inheritance in object-oriented programming)

  <Town, subclassOf, Administrative_unit>
Classes

- owl:Thing
- foaf:Person
- dbo:Person
- yago:WikicatAgnostics
- yago:WikicatAmericanAcademics
- yago:WikicatAmericanAgnostics
- yago:WikicatAmericanHumanitarians
- yago:WikicatAmericanInventors
- yago:WikicatAmericanNobelLaureates
- yago:WikicatAmericanPeople
Taxonomies

https://angryloki.github.io/wikidata-graph-builder/?property=P279&item=Q5
Knowledge base: Definition

A knowledge base (KB) is a collection of structured data about entities and relations with the following characteristics:

- **Content**: The KB contains entities and their semantic types for a given domain of interest. Additionally, attributes of entities (including numeric and string literals) and relationships between entities are captured.

- **Schema and Scale**: Unlike a conventional database, there is often no pre-determined relational schema where all knowledge has to fit into a static set of relations. If fixed, longitudinal evolution must allow ad-hoc additions where the set of types and relations may grow to ten or hundred thousands.

- **Open Coverage**: New entities and facts emerge and get covered in new web sources at high rate. Therefore, we have to view KB construction and maintenance as a “never-ending” task, following an open world assumption and acknowledging the high pace of real-world changes.

[Weikum et al., FnT 2021]
3. Introduction to AKBC

I. Motivation
II. Terminology
III. Topics
IV. Construction techniques
V. Applications
VI. Past, present and future
Common topics of knowledge bases

• **Lexical knowledge**
  - <shout, isA, verb>
  - <shout, subformOf, communicate>

• **Instance knowledge** (“Encyclopedic KBs”):
  - <Paris, capitalOf, France>
  - <MPII, foundedIn, 1988>
  - <Angela Merkel, major, Physics>

• **Class knowledge** (“Commonsense”):
  - <Pizza, is, tasty>
  - <Elephant, color, grey>
  - <turnOnPC, requires, power>
Lexical KBs

- FrameNet (1998)
- (Wiktionary (2002))
- SenticNet (2010)
- ...
WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: shout [Search WordNet]

Display Options: (Select option to change) ▼ Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- **S:** (n) cry, outcry, call, yell, shout, vociferation (a loud utterance; often in protest or opposition) "the speaker was interrupted by loud cries from the rear of the audience"

Verb

- **S:** (v) shout (utter in a loud voice; talk in a loud voice (usually denoting characteristic manner of speaking)) "My grandmother is hard of hearing--you'll have to shout"

  - **S:** (v) shout, shout out, cry, call, yell, scream, holler, hollo, squall (utter a sudden loud cry) "she cried with pain when the doctor inserted the needle"; "I yelled to her from the window but she couldn't hear me"
    - direct troponym / full troponym
    - verb group
    - direct hypernym / inherited hypernym / sister term
    - derivationally related form
    - phrasal verb
    - sentence frame

- **S:** (v) exclaim, cry, cry out, outcry, call out, shout (utter aloud; often with surprise, horror, or joy) "I won! he exclaimed"; "Help! she cried"; "I'm here, the mother shouted when she saw her child looking lost"

- **S:** (v) abuse, clapperclaw, blackguard, shout (use foul or abusive language towards) "The actress abused the policeman who gave her a parking ticket"; "The angry mother shouted at the teacher"
FrameNet

• **Example Frame – “Revenge”:** Because of some *injury* to something-or-someone important to an *avenger* (maybe himself), the *avenger* inflicts a *punishment* on the *offender*. The *offender* is the person responsible for the *injury*.

• **Frame elements:**
  • *avenger, offender, injury, injured_party, punishment.*

• **Invoking terms:**
  • Nouns: *revenge, vengeance, reprisal, retaliation*
  • Verbs: *avenge, revenge, retaliate (against), get back (at), get even (with), pay back*
  • Adjectives: *vengeful, vindictive*
Encyclopedic KBs ("Instance-oriented KBs")

- Cyc (1984)
- YAGO (2007)*
- DBpedia (2007)
- Wikidata (2012)

* developed at MPII
<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>activeYearsEndDate</td>
<td>2004-11-04 (xsd:date) 2008-11-16 (xsd:date)</td>
</tr>
<tr>
<td>activeYearsStartDate</td>
<td>1997-01-08 (xsd:date) 2005-01-03 (xsd:date) 2009-01-20 (xsd:date)</td>
</tr>
<tr>
<td>almaMater</td>
<td>db:Occidental_College db:Columbia_College_Columbia_University db:Harvard_Law_School</td>
</tr>
<tr>
<td>award</td>
<td>db:Nobel_Peace_Prize</td>
</tr>
<tr>
<td>birthDate</td>
<td>1961-08-04 (xsd:date) 1961-8-4</td>
</tr>
<tr>
<td>birthPlace</td>
<td>db:Hawaii db:Honolulu db:Kapiolani_Medical_Center_for_Women_and_Children</td>
</tr>
<tr>
<td>orderInOffice</td>
<td>44th President of the United States</td>
</tr>
<tr>
<td>party</td>
<td>db:Democratic_Party_(United_States)</td>
</tr>
<tr>
<td>region</td>
<td>db:Illinois</td>
</tr>
</tbody>
</table>
Commonsense KBs (class-oriented)

• Cyc (1984)
• ConceptNet (1999)
• TupleKB (2017)
• Quasimodo (2019)*
• Ascent (2021)*

* Developed at MPII
ConceptNet
Elephant

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Largest land animals</td>
<td>44</td>
</tr>
<tr>
<td>Herbivore</td>
<td>34</td>
</tr>
<tr>
<td>Intelligent</td>
<td>32</td>
</tr>
<tr>
<td>Endangered</td>
<td>22</td>
</tr>
<tr>
<td>Social</td>
<td>14</td>
</tr>
<tr>
<td>Trunk</td>
<td>333</td>
</tr>
<tr>
<td>Tusk</td>
<td>167</td>
</tr>
<tr>
<td>Ear</td>
<td>166</td>
</tr>
<tr>
<td>Foot</td>
<td>65</td>
</tr>
<tr>
<td>Skin</td>
<td>62</td>
</tr>
<tr>
<td>Mouth</td>
<td>62</td>
</tr>
<tr>
<td>Teeth</td>
<td>43</td>
</tr>
<tr>
<td>Body</td>
<td>43</td>
</tr>
<tr>
<td>Size</td>
<td>40</td>
</tr>
<tr>
<td>Brain</td>
<td>40</td>
</tr>
<tr>
<td>War</td>
<td>87</td>
</tr>
<tr>
<td>Wild</td>
<td>67</td>
</tr>
</tbody>
</table>

Elephant uses...
- Their trunks
- Their tusks
- Mud
- Their ears
- Their long trunks

Elephant lives...
- In group
- On land
- In the wild
- In grassland
- Up to 70 years

Elephant is used...
- In war
- For warfare
- As beast of burden
- For safari tourism
- In ceremony

Elephant is able...
- To find
- To track one another
- To spend substantial time
- To recognize their friends
- To eat a wide...
3. Introduction to AKBC

I. Motivation
II. Terminology
III. Topics
IV. Construction techniques
V. Applications
VI. Past, present and future
How to build KBs?
Possible approaches

A. Humans (CYC, ConceptNet, Wikidata)

B. Structured extraction (YAGO, DBpedia)

C. Text extraction (NELL, Textrunner)

D. Constraints and pattern mining

Our focus
A. Humans: Experts

- Potentially best quality

- Difficult to scale
  - CYC: “In 1986, Doug Lenat estimated the effort to complete the KB to be 250,000 rules and 350 man-years of effort.”
Humans: Crowdsourcing/Gamification

- Make work fun (?)
Humans: Volunteers

• Wikidata: 18k active users

• Intrinsic motivation achieves great things

• Broad expertise, compared with selected experts or paid crowdsourcing

• [https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all](https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all)
Humans: Challenges

• ConceptNet:
  • Common knowledge, normalization

• Crowdsourcing: Quality assurance

• Wikidata: Modelling and agreement
  • E.g., ethnicity, notable_work, ...
  • Multilingual concept alignment
B. Structured extraction

- Wikipedia already provides structured data

- All we need to do is harvest...
Work done?

- Noise
- Canonicalization of entities and predicates
- Usage of category system

Examples: YAGO, DBpedia
C. Text extraction

• In principle **most powerful**
  • No need for humans
  • No restriction to Wikipedia existence

• In practice **big noise challenges**
  • Many pipeline steps
    • Named-entity recognition, named-entity disambiguation, relation extraction, relation canonicalization, extraction consolidation, ..

• Examples: NELL, Textrunner
Merkel is of German and Polish descent. Her paternal grandfather, Ludwik Kasner, was a German policeman of Polish ethnicity, who had taken part in Poland's struggle for independence in the early 20th century.[22] He married Merkel's grandmother Margarethe, a German from Berlin, and relocated to her hometown where he worked in the police. In 1930, they Germanized the Polish name Kaźmierczak to Kasner.[23][24][25][26] Merkel's maternal grandparents were the Danzig politician Willi Jentzsch, and Gertrud Alma née Drange, a daughter of the city clerk of Elbing (now Elbląg, Poland) Emil Drange. Since the mid 1990s, Merkel has publicly mentioned her Polish heritage on several occasions and described herself as a quarter Polish, but her Polish roots became better known as a result of a 2013 biography.

In 1968, Merkel joined the Free German Youth (FDJ), the official communist youth movement sponsored by the ruling Marxist–Leninist Socialist Unity Party of Germany.[30][31][32] Membership was nominally voluntary, but those who did not join found it difficult to gain admission to higher education.[33] She did not participate in the secular coming of age ceremony Jugendweihe, however, which was common in East Germany. Instead, she was confirmed.[34] During this time, she participated in several compulsory courses on Marxism-Leninism with her grades only being regarded as "sufficient".
D. Constraints

Databases
  • Key, foreign key, range, ...

Knowledge bases:
  • Events start earlier than they end
  • Every human must have two parents
  • Mayors of cities must be humans
  • The parent of a person’s sibling is the person’s parent

• Can be used to...
  ... reject KB modifications
  ... indicate missing information
  ... infer new facts

• But reality is messy..
3. Introduction to AKBC

I. Motivation

II. Terminology

III. Topics

IV. Construction techniques

V. Applications

VI. Past, present and future
What KBs are good for

• Master data
• Data mining
• Search enhancements
• Question answering
• Language generation
• Entity linking
• Learning more knowledge
• …
Master data (1)

| wd:Q6258248 | John Smith |
| wd:Q6258251 | John Smith |
| wd:Q6258255 | John Smith |
| wd:Q6258259 | John Smith |
| wd:Q6258261 | John Smith |
| wd:Q6258263 | John Smith |
| wd:Q6258265 | John Smith |
| wd:Q6258267 | John Smith |
| wd:Q6258270 | John Smith |
| wd:Q6258271 | John Smith |
| wd:Q6258276 | John Smith |
| wd:Q6258278 | John Smith |
| wd:Q6258281 | John Smith |
| wd:Q6258284 | John Smith |
| wd:Q6258286 | John Smith |
| wd:Q6258288 | John Smith |
| wd:Q6258290 | John Smith |
| wd:Q6258293 | John Smith |
| wd:Q6258294 | John Smith |
| wd:Q6258296 | John Smith |

(300 more)
### Master data (2)

#### Relevant for:
- Museums
- Libraries
- Scientific publications

<table>
<thead>
<tr>
<th>Identifiers</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freebase ID</td>
<td>/m/03mb4s</td>
</tr>
<tr>
<td>GND ID</td>
<td>5066841-9</td>
</tr>
<tr>
<td>VIAF ID</td>
<td>157458492</td>
</tr>
<tr>
<td>ISNI</td>
<td>0000 0004 0491 9823</td>
</tr>
<tr>
<td>GRID ID</td>
<td>grid.419528.3</td>
</tr>
</tbody>
</table>

- 1 reference
- 1 reference
- 1 reference
- 1 reference
- 2 references
Data mining

• Use input facts to extract patterns that allow to predict new facts

\[
\begin{align*}
isCitizenOf(x, y) & \Rightarrow livesIn(x, y) \\
hasAdvisor(x, y) \land graduatedFrom(x, z) & \Rightarrow worksAt(y, z) \\
wasBornIn(x, y) \land isLocatedIn(y, z) & \Rightarrow isCitizenOf(x, z) \\
hasWonPrize(x, G. W. Leibniz) & \Rightarrow livesIn(x, Germany)
\end{align*}
\]

\(\text{isCitizenOf(John, France) } \Rightarrow \text{livesIn(John, France)}\)

• Various approaches based on association rule mining and latent models
Entity linking

https://opentapioca.org/
Search enhancements
Question answering

Try yourself:
- When was Trump born?
- What is the nickname of Ronaldo?
- Who invented the light bulb?
Question answering (2)

- Knowledge bases **key component in question answering systems**
  - E.g., IBM Watson

- **AllenAI science challenge**: Computers currently in 8th grade
  - Knowledge acquisition still major bottleneck
Language generation

- Wikipedia in world’s most spoken language: **1/10** as many articles as English Wikipedia
- World’s fourth most spoken language: **1/100**

→ Wikidata intended to help resource-poor languages

Douglas Adams was a British playwright, screenwriter, novelist, children's book author, and television scriptwriter. He was born in March 1952 in Cambridge, England, and died on March 11, 2001, in Brentwood, California. He was married to Jane Maynard, with whom he had two children. He was buried in myocart, a private crematorium.
3. Introduction to AKBC

I. Motivation
II. Terminology
III. Topics
IV. Construction techniques
V. Applications
VI. Past, present and future
Past

(Cyc)

($relationAllExists$
  $biologicalMother$
  $ChordataPhylum$
  $FemaleAnimal$)

(Wikipedia)

(WordNet)

(Memex (1945))

(Freebase (collaborative))

(WolframAlpha)

(DBpedia)

(Knowledge Graph)

(Alexandria)

(Wikidata)

(Past)

1984 2001 2007 2012 2018
Present

- KBs at most major tech companies and beyond
  - Google, Microsoft, Alibaba, Bloomberg, ...

- Feb 2018: $125 million investment by Microsoft cofounder Paul Allen into non-profit research on common sense knowledge extraction and reasoning

- Research: Major part of NLP conferences taken up by IE/AKBC research
Future

• ?
Outline

1. Introducing each other
2. Course organization
3. What, Why, How
4. Lab 1
Lab 1

• Information extraction where from?
  • Actual web crawling nontrivial
  • Wikipedia a popular high-quality resource

• Learn/practice text manipulation, perform some simple analyses, get to know KB querying
Take home

• Knowledge base construction builds machine-readable structured content from unstructured/semistructured inputs

• Structured data is relevant for a range of knowledge-intensive tasks

• Next week: Crawling and scraping