Automated knowledge base construction

7. Commonsense Knowledge (CSK)

Simon Razniewski
Summer term 2022
Interactive pattern matching with spaCy

• Interactive tool for exploring pattern matches
  • https://explosion.ai/demos/matcher

• General documentation:
  • https://spacy.io/usage/rule-based-matching
Background

• So far largely (implicitly): Small data
  • First sentence from an entity description, one Wikipedia article, one Fandom infobox, ...

• More often than not:
  • Huge set of possible sources
  • Precision-orientation may suggest filtering, selecting
    • Semistructured content >> Wikipedia >> General web
  • If no premium resources exist/high recall is desired
→ Leads to the topic of extraction consolidation
Consolidation need

- `bornIn(John, Paris) + bornIn(John, London)`
  (single-value conflict)

- `hasParent(Mary, John) + hasParent(John, Mary)`
  (antisymmetric relation violated)

- `hasParent(Mary, {John1, John2, ... John377})`
  (implausible counts)

- `bornIn/raisedIn/livedIn/diedIn(John, Sydney) + positionHeld(John, Prime Minister of Canada)`
  (spatial implausibility)
  ... temporal implausibility
  ... topical implausibility
  ...

...
Extraction consolidation

• **Family of approaches**
  - Tresholding
  - Constraint reasoning
  - Multi-source validation

• **Details in**
  - Section 8.5 of course textbook (see website)
  - Slides of lecture 8 of old course (link)

• Today: **Application domain** of AKBC where multi-source extraction and consolidation is essential: **Commonsense knowledge**
Outline

1. **Introduction to CSK**
   1. What is CSK?
   2. Why is it important?
   3. How to represent it?
   4. What makes it challenging?

2. Crowdsourced CSKB construction

3. Text-extraction for ACSKB construction

4. CSKBs: Summary and Outlook
What is commonsense knowledge?

- Possible qualifications
  - Across cultures
  - From early in life (=children)

- E.g., elementary school exam questions
  - [http://data.allenai.org/ai2-science-questions](http://data.allenai.org/ai2-science-questions)

**Definition 1 (by commonality):**
Knowledge shared by most humans
What is commonsense knowledge?

Definition 2 (by knowledge type):

Knowledge about concepts and events

- Concepts: *City, footballer, organization*
- Events: *Football match, birthday party*

- Differentiation from encyclopedic knowledge on instances
  - Instances: *Saarbrücken, Ronaldo, Manchester United*
Definition Pro/Con

• Definition 1 (by commonality):
  • *Popsicle, is, frozen – only known in North America*
  • *Lion is dangerous/cute - depends whom you ask*
    → Inclusion/exclusion decision challenging

• Definition 2 (by knowledge type):
  • *Apple MacBook, Ford Model T*
    → Class-instance not trivial to separate
  • *USA borders Pacific Ocean – excluded as instance knowledge*
  • *Mitochondria, hasPart, inner membrane – not common knowledge*
    → Open-ended
      → Can be somewhat mitigated by ranking-based evaluation
## Definition: Merger

### Knowledge

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>Class knowledge</th>
<th>Instance knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shared by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Virtually</td>
<td><em>Fire is hot</em></td>
<td><em>USA borders Pacific</em></td>
</tr>
<tr>
<td>everyone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Many</td>
<td><em>Elephants have tusks</em></td>
<td></td>
</tr>
<tr>
<td>- Some</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Few</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Basic CSK</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Advanced CSK</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Fire is hot
- Elephants have tusks
- Mitochondria have inner membrane
- Newton born in Woolsthorpe
Examples of CSK

• Taxonomical
  • *Elephant, isA, mammal*

• Properties
  • *Elephant, lives in, Savanna*

• Parts
  • *Elephants, hasPart, trunk*

• Measures
  • *Adult elephant, weight, ~2..5 tons*
  • *Elephant, lifespan, ~60 years*

• Activities
  • *Seeing elephant, requires, go to zoo*
  • *Go to zoo, subevent, buy ticket*
  • *Go to zoo, typicalDuration, 2 hours*
Why CSK? Amazing progress without

- 2018: Human-level performance on reading comprehension on SQuAD (Stanford QA dataset)
- 2017: Super-human performance on speech recognition
- 2016: Google neural machine translation
- 2015: Super-human performance on image captioning
  
[From Yejin Choi, ACL 2020]
Solving only a “dataset” without solving the underlying “task”!

[From Yejin Choi, ACL 2020]
Importance of CSK

Reusable and scrutable asset for a range of AI tasks

• Reusable:
  • CSK can be plugged into a range of tasks, e.g., QA, dialogue, object recognition, text generation, ...
  • Contrasts with typical end-to-end learning

• Scrutable:
  • Humans can inspect, add and remove content
    • Relevant in applications where errors are costly
    • Relevant in applications at risk of bias/discrimination
  • Humans can inspect discrete statements used for reasoning
    • Relevant for debugging complex downstream use cases
  • Contrasts with end-to-end learning and pretrained language models
Knowledge representation challenges

- Encyclopedic KBs: Typically binary truth notion
  - Trump, born in, NY
  - House of Cards, producer, Netflix
  - New York, mayor, Bloomberg, [2002-2013]

- CSK: Generalizes across subjects
  - Lions, have, manes - percentage?

- Fuzzy time notion
  - Lions, drink, milk - when?

- Spatial and cultural context
  - Lion, is, cute
  - Elk, usedFor, transport
Linguistics - Generics

• It is complicated
  • *Ducks lay eggs* >> *Ducks are female*
    • Even though former set is a subset of latter
  • *Dinosaurs are extinct/Elephants are biggest land animals*
    • Not applicable to individuals

[Generics oversimplified, Leslie, 2013]
Epistemic logics

• “Zoo visitors believe lions are cute”
• “Rural dwellers believe lions are dangerous”

• Used in CycL
  • Reification
  • Modals for belief and desires

[CYC: Towards programs with common sense, Lenat et al., 1990]

[Kripke 1963], and others
Episodic logics

KNext, Lore projects in early 2000s

(REP. RONNIE FLIPPO (D., ALA.), ONE OF THE MEMBERS OF THE DELEGATION, SAYS 0 HE WAS PARTICULARLY IMPRESSED *-1 BY MR. KRENZ’S READY ADMISSION THAT EAST GERMANY NEEDED *-2 TO CHANGE.)

AN ELECTED-REPRESENTATIVE MAY SAY A PROPOSITION.
A DELEGATION MAY HAVE MEMBERS.
A MALE-INDIVIDUAL MAY BE IMPRESSED -ED BY AN ADMISSION.
AN ADMISSION CAN BE READY.
A COUNTRY MAY NEED TO CHANGE.

((:I (:Q DET ELECTED-REPRESENTATIVE) SAY[V] (:Q DET PROPOS))
 (:I (:Q DET DELEGATION[N]) HAVE[V]
  (:Q DET (:F PLUR MEMBER[N])))
 (:I (:Q DET MALE-INDIVIDUAL) (:F BE[PASV] IMPRESS[V])
  (:F BY[P] (:Q DET ADMISSION[N])))
 (:I (:Q DET ADMISSION[N]) READY[A])
 (:I (:Q DET COUNTRY) NEED[V] (:F KA CHANGE[V])))

Schubert, Lenhart. "Can we derive general world knowledge from texts." *HLT 2002*
Graded formalisms

• Heuristic level in CycL
  • True, default true, unknown, default false and false as statement labels

• Ordinal grades
  [Schubert and Tong, NAACL 2003]

• Simplified
  [Zhang et al., TACL 2017]
  • Very likely
  • Likely
  • Plausible
  • Technically possible
  • Impossible

1. SEEMS LIKE A REASONABLE GENERAL CLAIM (Of course. Yes.)
   A grand-jury may say a proposition. A report can be favorable.

2. SEEMS REASONABLE BUT EXTREMELY SPECIFIC OR OBSCURE
   (I suppose so)
   A surgeon may carry a cage. Gladiator pecs can be Reeves-type.

3. SEEMS VACUOUS (That’s not saying anything)
   A thing can be a hen. A skiff can be nearest.

4. SEEMS FALSE (No. I don’t think so. Hardly)
   A square can be round. Individual -s may have a world.

5. SOMETHING IS OBVIOUSLY MISSING (Give me a complete sentence)
   A person may ask. A male-individual may attach an importance.

6. HARD TO JUDGE (Huh?? How do you mean that? I don’t know.)
   A female-individual can be psychic. Supervision can be with a company.

| Sam bought a new clock ~> The clock runs |
| Dave found an axe in his garage ~> A car is parked in the garage |
| Tom was accidentally shot by his teammate in the army ~> The teammate dies |
| Two friends were in a heated game of checkers ~> A person shoots the checkers |
| My friends and I decided to go swimming in the ocean ~> The ocean is carbonated |
Graded formalisms (2)

- Dice [Chalier et al., AKBC 2020]
  - 4 dimensions

<table>
<thead>
<tr>
<th></th>
<th>Plausible</th>
<th>Typical</th>
<th>Salient</th>
<th>Remarkable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lions; eat; chicken</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lions; attack; humans</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Lions; drink; water</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
KR - state of the art

• Expressive proposals exist
  • Modal, epistemic, episodic logic

• Instantiation hard
  • Sparse realization in natural language
  • Correct extraction nontrivial

• Most projects:
  Pragmatic choice of (subject, predicate, object) triples with a single score

Lion, hunts, zebra – 0.73
Lion, drinks, milk – 0.45
Triples and done?

• **Still major design decisions left!**
  1. Fixed or open set of predicates
  2. Subject range
  3. Object range

• Fixed vs. open predicates
  • E.g., ConceptNet: ~25 predicates (isCapableOf, requires, isA) vs. TupleKB ~1000 textual phrases

• Subjects: Strings or disambiguated terms?
  • Lynx vs. lynx vs. lynx

• Granularity and modifiers
  • Elephant, Foraging elephant? Newborn elephant?

• Objects: Entities or open phrases?
  • Politician, isCapableOf, promise that impossible things will happen
Outline

1. Introduction to CSK
2. Crowdsourced CSKB construction
3. Text-extraction for ACSKB construction
4. CSKBs: Summary and Outlook
Crowdsourcing Expert annotation

- **WordNet** [Miller and Fellbaum, ~80s]
  - Lexical resource still popular today
  - Is-A, synonym, partOf
  - ~200k word senses
  - 2 expert annotators (=authors)
  - Limitations: Imbalance, idiosyncrasies, lack of scores/ranking

- **Cyc** [Lenat, ~80s]
  - CSK, world knowledge, rules
  - Hired experts on specific domains
  - >1000 person-years of effort estimated

- Next projects: Harness power of laypeople
Open Mind Common Sense / ConceptNet

- Havasi et al., MIT ~1999 [http://conceptnet.io/]
- CSK for ~25 relations
  - Construction statistics
    - ~14k volunteers filled in sentences with blanks
    - ~700 000 English sentences
    - NLP tools: 300 000 concepts and 1.6 million assertions

[ConceptNet — a practical commonsense reasoning tool-kit, Liu and Singh, 2004]
Knowledge about ocean

Similar objects to ocean: sea, water, beaches, aquarium, lake

An inquiring mind wants to know...

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is on the ocean somewhere that people can be?</td>
<td>Yes</td>
</tr>
<tr>
<td>You would find __________ near the ocean.</td>
<td></td>
</tr>
<tr>
<td>Teach OpenMind</td>
<td></td>
</tr>
<tr>
<td>Is on the ocean somewhere that coral reefs can be?</td>
<td>Yes</td>
</tr>
<tr>
<td>ocean is a kind of __________.</td>
<td></td>
</tr>
<tr>
<td>Teach OpenMind</td>
<td></td>
</tr>
<tr>
<td>Would you find an ocean in a pool?</td>
<td>Yes</td>
</tr>
<tr>
<td>an ocean is used for __________.</td>
<td></td>
</tr>
<tr>
<td>Teach OpenMind</td>
<td></td>
</tr>
<tr>
<td>Is on the ocean somewhere that seagulls can be?</td>
<td>Yes</td>
</tr>
<tr>
<td>ocean can be __________.</td>
<td></td>
</tr>
<tr>
<td>Teach OpenMind</td>
<td></td>
</tr>
</tbody>
</table>
Verbosity

- 2-player game inspired by Taboo
- Narrator must describe a word by filling blanks in templates
  - ___ is a kind of ___. Allows for hierarchical categorization.
  - ___ is used for ___. Provides information about the purpose of a word.
  - ___ is typically near/in/on ___ (three templates). Provide spatial data.
  - ___ is the opposite of ___ / ___ is related to ___ (two templates). Provide data about basic relations between words.
- Templates give rise to CSK assertions
- Verification via automated narrator that replays human assertions
- Used to feed ConceptNet
- High-quality: 85% of sentences rated as correct by 6/6 annotators

[Ahn et al., 2006]
Atomic

- Targets event knowledge

https://mosaickg.apps.allenai.org/kg_atomic

[Sap et al., AAAI 2019]
Atomic

• Archetype of large-scale paid crowdsourcing
• Subjects from text extraction (24k event phrases)
• Statement creation
  • 3 workers per subject
  • Free-form interface
  • ~12$/hour
  • 300k statements
  • 3*5 minute/subject (?) → ~$100k cost

Event

PersonX pays PersonY a compliment

Before
1. Does PersonX typically need to do anything before this event?

After
2. What does PersonX likely want to do next after this event?

3. Does this event affect people other than PersonX?
   (e.g., PersonY, people included but not mentioned in the event)
   ☐ Yes ☐ No
   a). What do they likely want to do next after this event?
Quasimodo evaluation data

• Not a KB construction effort!
• Only tiny slice of humans data for evaluation


2400 statements

<table>
<thead>
<tr>
<th>Surgeons are the ones who operate</th>
<th>Surgeons are medical specialists</th>
<th>Surgeons work in the operating rooms</th>
<th>Surgeons have a good salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgeons are precise.</td>
<td>Surgeons have studied.</td>
<td>Surgeons work in hospitals.</td>
<td>Surgeons work under sterile conditions.</td>
</tr>
<tr>
<td>Surgeons are a physicians who perform surgeries work in hospitals.</td>
<td>Surgeons work also in podiatry.</td>
<td>Surgeons work also in dentistry and the</td>
<td></td>
</tr>
<tr>
<td>a surgeon is a specialist doctor</td>
<td>a surgeon works with his hands</td>
<td>a surgeon is able to prevent and cure diseases.</td>
<td></td>
</tr>
<tr>
<td>Surgeons perform complex operations</td>
<td>Surgeons treat people</td>
<td>Surgeons work in clinics</td>
<td>Brave surgeons</td>
</tr>
<tr>
<td>surgeons save people's lives</td>
<td>the surgeons are real heroes</td>
<td>surgeons work many hours a day</td>
<td>surgeons must always be available</td>
</tr>
</tbody>
</table>

[Romero et al., CIKM 2019]
We are collecting common knowledge, where you are given a subject, and you should tell us 4 simple English **general sentences** about that subject that quickly come to your mind.

**Examples**

**Subject: Elephants**

Some possible general sentences:

- Elephants are grey.
- Elephants live in Africa.
- Elephants have big ears.
- ...

**Subject: Nurses**

Some possible general sentences:

- Nurses work in hospitals.
- Nurses are compassionate.
- Nurses wear white.
- ...

**Rules**

1. Write full grammatical sentences that start with the subject ("Elephants are grey.", not "They are grey.").
2. Sentences should contain verbs ("Elephants have trunks."), not just be word clouds (Not "Trunk" nor "Elephant trunk").
3. Sentences should be general ("Elephants are big"), not refer to your personal situation ("I saw an elephant last week").
4. Submissions not adhering to these rules will be rejected!

Still lots of misunderstandings

“*I know a surgeon*”

“*Un cirujano es un gran hombre*”

“*The heart surgeon*”

→ Second round of peer filtering might help
Wikidata

• Collaborative knowledge base construction effort
• Under umbrella of Wikimedia foundation
• Best public source on encyclopedic knowledge today
• Commonsense:
  • Comparably lower coverage
  • Roughly comparable to ConceptNet
  • Growing...to be monitored

[Ilievski et al, Arxiv 2020]
## Crowdsourcing - Summary

<table>
<thead>
<tr>
<th>Project</th>
<th>Focus</th>
<th>#statements</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet</td>
<td>Taxonomical relations</td>
<td>175k synsets</td>
<td>Expert-built</td>
</tr>
<tr>
<td>Cyc</td>
<td>General statements and rules</td>
<td>OpenCyc: ~2 M statements</td>
<td>Expert-built, closed source</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>Object properties</td>
<td>1.6 M statements</td>
<td></td>
</tr>
<tr>
<td>Atomic</td>
<td>Events</td>
<td>877k statements</td>
<td></td>
</tr>
<tr>
<td>Wikidata</td>
<td>Object properties</td>
<td>100k CSK statements</td>
<td>Editable</td>
</tr>
</tbody>
</table>

- Limited by volunteer effort/money
- Targeted domains in reach for industrial efforts
- Quality assurance important
Outline

1. Introduction to CSK
2. Crowdsourced CSKB construction
3. Text-extraction for ACSKB construction
   1. Overview
   2. Recipe
   3. Example projects
4. CSKBs: Summary and Outlook
Overview

• **Earliest projects** on CSKB construction were **manually** authored (Cyc, ConceptNet)

• Challenges in scale
  • Atomic: ~100k$ annotator expenses

• **Automated information extraction** and KB construction field with **long history**
  • Focus traditionally on crisp "encyclopedia" knowledge (cf. DBpedia, YAGO, NELL, DeepDive, ...)

• **Can we use automated IE and KBC for CSK?**
Graham Neubig
@gneubig

One commonly cited argument about the difficulty of learning common-sense reasoning is that "no-one writes down common sense". A counter-argument is "well, the web is big": instructables.com/id/How-To-Open…

How to Open a Door

Step 1: Locate Desired Door

Step 2: Locate Door Handle or Knob

Step 3: Turn Knob or Handle and Pull or Push
Challenges of automated CSKB construction

• **Underspecified text semantics**
  • “Lions attack humans” – all/some/all the time/once/..?

• **Reporting bias**
  • “woman kills” vs. “woman breathes” – 1.5M vs. 0.1M web search results
  • “pink elephant” vs. “grey elephant” – 6.9M vs. 1.9M web search results

• **Sparse observations of quadratic+ space of possible statements**
  • Do computer programmers drink water?

• **Noise and polysemy**
  • Pigs can fly - idiom
  • Lynx: Constellation, web browser, animal
Outline

1. Introduction to CSK
2. Crowdsourced CSKB construction
3. Text-extraction for ACSKB construction
   1. Overview
   2. Recipe
   3. Example projects
4. CSKBs: Summary and Outlook
Recipe – Generic design points

1. Sources
2. Extraction method
3. Type of contextualization
4. Consolidation method
Design point 1 – Source choices

• “Where to extract from?”
  • Wikipedia
  • Books and other dedicated sources
    • ARC science corpus
    • Project Gutenberg
  • Web search
  • Forums
    • Reddit
    • Quora
    • Yahoo Answers
  • Search engine query logs
  • Web crawls
    • ClueWeb
    • CommonCrawl
  • …
Extraction source - considerations

• (CS)KB projects stand and fall with source selection
• Precision: *Topic-specific sources >> random web*
  • Event knowledge – Wikihow [HowToKB, WWW 2017]
  • Cultural knowledge – Movie scripts [Knowlywood, CIKM 2015]
  • Science knowledge – Science textbooks [GenericsKB, Arxiv 2020]

• Frequency signals may be stronger from general web dumps, but considerable noise

• Intermediate setting: *Targeted web search* [TupleKB, Ascent]
Design point 2 – Extraction method options

• “How to extract”

   • Hearst patterns etc.
2. Co-occurrence [DoQ, ACL 2019]
   • Window, same sentence, ...
3. Open information extraction [TupleKB, Quasimodo, Ascent]
   • Any verb phrase
4. Relation-specific supervised learning
Extraction method - considerations

• Preferred method depends on desired knowledge representation
  • E.g.,
    • Few non-overlapping relation → Co-occurrence
    • Moderate relations → Supervised extractors
    • Many relations → OpenIE

• Has implications downstream
  • Extraction confidences (supervised extractors) for quantitative contextualization
  • Text context for qualitative contextualization
  • OpenIE with many unspecific extractions
Design point 3 – Contextualization

“What do we annotate statements with?”

1. Observation frequency [WebChild 2.0, DoQ]
   • *Elephant, has, tusks, 155*
   • *Elephant, has, tail, 84*

2. Quantitative [0,1] truth labels [TupleKB, Quasimodo]
   • *Elephant, lives in, group, 0.87*

3. Qualitative truth labels [Ascent]
   • *Elephant, lives in, group, temp: during wet season*
   • *Subgroup: Female elephant, lives in, group*
Contextualization - considerations

• **Frequencies** trivial to interpret, but do not qualify degree of truth

• **Quantitative truth labels** nontrivial semantics

• **Qualitative labels** easier to interpret, but harder to compare

• **Expressive proposals** from KR exist (e.g., modal logics)
  • Actual implementation not easy
    • Sparse realization in natural language
    • Correct extraction nontrivial
Design point 4 – Consolidation

“What do we do with redundant and competing extractions?”

• Similar statements may be seen several times
• Redundancy and contradictions may require additional inference

• Common consolidation methods
  1. Keep all [DoQ]
  2. Frequency cutoff [Ascent]
     • E.g., at least seen 5 times
  3. Per-statement consolidation [TupleKB, Quasimodo]
     • Feature-based classification/ranking
     • E.g., BERT-based clustering, MaxSAT, …
Consolidation - considerations

• Redundancy challenge and blessing
• Exploiting redundancy requires strong text similarity/entailment modules
• Previous projects often stuck to per-statement consolidation due to lack of strong similarity/entailment modules
• Recent advances on pretrained LMs give hope for joint consolidation (see e.g., Dice, Ascent)
Outline

1. Introduction to CSK
2. Crowdsourced CSKB construction
3. Text-extraction for ACSKB construction
   1. Overview
   2. Recipe
   3. Example projects
4. CSKBs: Summary and Outlook
Example projects

1. **Webchild** [Tandon et al., WSDM 2014]
   • Disambiguated noun-adjective pairs

2. **Quasimodo** [Romero et al., CIKM 2019]
   • Salient general triples

3. **DoQ** [Elazar et al., ACL 2019]
   • Quantitative knowledge

4. **Dice** [Chalier et al., AKBC 2020]
   • Multifaceted quantitative contextualization and joint consolidation
Quasimodo

= Query Logs and QA Forums for Salient Commonsense Definitions

• Focus on salient knowledge
  • Human associations, curiosity

• Source: Query logs and QA forum questions
• Extraction method: OpenIE
• Contextualization: Supervised precision + IDF
• Consolidation: Largely per-statement regression

[Romero et al., CIKM 2019], builds on [TupleKB - Mishra et al., TACL 2017]
Starting point: Humans vs. automated IE

Manual constructions:
- Salient but few

[ConceptNet]

**elephant is capable of...**
- carry a trunk
- remember water sources

(6 more)

Automated construction:
- Many but boring

[TupleKB]

**Elephant:**
- require, ground
- inhabit, region
- (95 more)

How to reconcile the two?
Salient knowledge: Utterance context

Key idea: Questions convey salient knowledge

• Why do cats purr?
• Why do Americans love guns?
• Why are airplanes white?

  a) So someone knows these!
  b) That someone cares enough to ask!
Salient knowledge: Premier sources

• QA forums:
  • Reddit
  • Quora
  • Yahoo answers
  • Ask.com

• Search engine query logs
  • Bing
  • Google
Tapping search engine query logs

- Autocomplete gives only 10 suggestions/query
  → Exhaustive suffix probing
    - Why do cats a
    - Why do cats b
    - Why do cats ...
    - Why do cats aa
    - Why do cats ab
    - ...

why do cats
why do cats purr
why do cats like boxes
why do cats meow
why do cats knead
why do cats sleep so much
why do cats hate water
why do cats like catnip
why do cats lick you
why do cats have whiskers
Question templates

<table>
<thead>
<tr>
<th>Pattern</th>
<th>In Query Logs</th>
<th>In QA Forums</th>
</tr>
</thead>
<tbody>
<tr>
<td>how does</td>
<td>19.4%</td>
<td>7.5%</td>
</tr>
<tr>
<td>why is</td>
<td>15.8%</td>
<td>10.4%</td>
</tr>
<tr>
<td>how do</td>
<td>14.9%</td>
<td>38.07%</td>
</tr>
<tr>
<td>why do</td>
<td>10.6%</td>
<td>9.21%</td>
</tr>
<tr>
<td>how is</td>
<td>10.1%</td>
<td>4.31%</td>
</tr>
<tr>
<td>why does</td>
<td>8.97%</td>
<td>5.46%</td>
</tr>
<tr>
<td>why are</td>
<td>8.68%</td>
<td>5.12%</td>
</tr>
<tr>
<td>how are</td>
<td>5.51%</td>
<td>1.8%</td>
</tr>
<tr>
<td>how can</td>
<td>3.53%</td>
<td>10.95%</td>
</tr>
<tr>
<td>why can’t</td>
<td>1.77%</td>
<td>1.40%</td>
</tr>
<tr>
<td>why can</td>
<td>0.81%</td>
<td>0.36%</td>
</tr>
</tbody>
</table>
Statement extraction

• Questions → statements → tuples using OpenIE

Why are lions hunting zebras?

Transform → Lions are hunting zebras

OpenIE → (lions, are hunting, zebras)

Normalize → (lion, hunt, zebras)

Score → (lion, hunt, zebras), 0.73
# Anecdotal Examples

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practical human knowledge</td>
<td>(car, slip on, ice)</td>
</tr>
<tr>
<td>Problems linked to a subject</td>
<td>(pen, can, leak)</td>
</tr>
<tr>
<td>Emotions linked to events</td>
<td>(divorce, can, hurt)</td>
</tr>
<tr>
<td>Human behaviors</td>
<td>(ghost, scare, people)</td>
</tr>
<tr>
<td>Visual facts</td>
<td>(road, has_color, black)</td>
</tr>
<tr>
<td>Cultural knowledge (USA)</td>
<td>(school, have, locker)</td>
</tr>
<tr>
<td>Comparative knowledge</td>
<td>(light, faster than, sound)</td>
</tr>
</tbody>
</table>
Results – Precision

Sample from a list of common subjects (popular animals and occupations)

5 = best, 1 = worst
Results – Recall

Crowd task:

Tell us 3 things that come to your mind when thinking of lions.

1. Lions ...
2. Lions ...
3. Lions ...

![Graph](image)
Extrinsic evaluation

Where would I not want a fox?

thumb up: hen house, thumb down: england, mountains, english hunt, thumb down: california

<table>
<thead>
<tr>
<th>KB</th>
<th>Elementary NDMC</th>
<th>Middle NDMC</th>
<th>CommonsenseQA2</th>
<th>Trivia</th>
<th>Examveda</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Questions (Train/Test)</td>
<td>623/541</td>
<td>604/679</td>
<td>9741/1221</td>
<td>1228/452</td>
<td>1228/765</td>
<td>10974/3659</td>
</tr>
<tr>
<td>Random</td>
<td>25.5</td>
<td>23.7</td>
<td>21.0</td>
<td>25.9</td>
<td>25.4</td>
<td>22.0</td>
</tr>
<tr>
<td>word2vec</td>
<td>26.2</td>
<td>28.3</td>
<td>27.8</td>
<td>27.4</td>
<td>25.6</td>
<td>27.2</td>
</tr>
<tr>
<td>Quasimodo</td>
<td><strong>38.4</strong></td>
<td><strong>34.8</strong></td>
<td>26.1</td>
<td><strong>28.1</strong></td>
<td>32.6</td>
<td><strong>31.3</strong></td>
</tr>
<tr>
<td>ConceptNet</td>
<td>28.5</td>
<td>26.4</td>
<td><strong>29.9</strong></td>
<td>24.4</td>
<td>27.3</td>
<td>27.5</td>
</tr>
<tr>
<td>TupleKB</td>
<td>34.8</td>
<td>25.5</td>
<td>25.3</td>
<td>22.2</td>
<td>27.4</td>
<td>27.5</td>
</tr>
<tr>
<td>WebChild</td>
<td>26.2</td>
<td>25.1</td>
<td>25.2</td>
<td>25.9</td>
<td>27.1</td>
<td>24.1</td>
</tr>
</tbody>
</table>

Accuracy in multiple-choice question answering.
(Simple question-answer connectedness scheme)
Example projects

1. Webchild 1.0 [Tandon et al., WSDM 2014]
   • Disambiguated noun-adjective pairs
2. Quasimodo [Romero et al., CIKM 2019]
   • Salient general triples
3. DoQ [Elazar et al., ACL 2019]
   • Quantitative knowledge
4. Dice [Chalier et al., AKBC 2020]
   • Multifaceted quantitative contextualization and joint consolidation
Distribution over quantities (DoQ)

• Understanding numerical properties and the way they relate to words.

• Focus on items which can be measured objectively

[Elazar et al., ACL 2019]
Distribution over quantities (DoQ)

- **Source:** Google search engine document index
- **Extraction scheme:** Text window co-occurrence of subject, quantity and dimension keyword
- **Contextualization:** Frequency
- **Consolidation:** none/distribution
Example - Measurement Detection

“These breeds can vary in weight from a 0.46 kg teacup poodle ...”

Detect numerical measurements using rules:
kg/kgs/kilogram -> Mass
Normalize (kg -> g)
Example - Co-Occurring objects

“These breeds can vary in weight from a 0.46 kg teacup poodle…”

Detect objects of interest (Nouns, Adjectives and Verbs) using a POS tagger.
Example - Aggregating Measurements
Resource Statistics - DoQ

- Distributions over Quantities (DoQ)
- A very large and diverse resource

- ~120M Unique tuples (object, measurement)
  - ~350K with >= 1000 occurrences
- Measurement types:
  - Length, mass, currency, temperature, ...
- 27 In total
Intrinsic Evaluation

- Extract the median of “popular” noun distributions
- Expand to a range
  - 20 mm → 10-100 mm
- Ask annotators if the item fits the range
  - “Is the usual length of a screw between 10-100mm?”
- 69% agreement with predictions
- Not perfect, but a reasonable start for acquiring such knowledge
Comparable Objects - Cool Results

![Speed Chart](Image)

- Speed (in km/h)
- Vehilces: ship, boat, truck, car

![Hour of Day Chart](Image)

- Hour of Day: breakfast, brunch, lunch, dinner
Comparable Objects - Some Issues

"Elevation ranges from 3,000 feet ... above sea level."

That’s a small sea!
Comparable Objects - Some Issues

“**Alfalfa** is the most cultivated legume ... reaching around **454 million tons** ...”

That’s a heavy alfalfa

https://alivebynature.com/the-right-way-to-eat-alfalfa-sprouts/
Comparable Objects - Case Study

Collected temperatures of US States

“Real” average

Predicted median
Summary

1. **Sources**
   • Domain-specific selection pays off

2. **Extraction method**
   • OpenIE vs. trained extractors

3. **Contextualization**
   • Expressivity-extractability tradeoff
   • Quantitative vs. qualitative

4. **Consolidation**
   • Advances in text similarity detection enable joint consolidation

**State of the art**

• Automatically extracted CSKBs competitive with manually-built projects
  • Usually huge gains in recall, moderate loss in precision
# Overview – major projects

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TupleKB</td>
<td>Science triples</td>
<td>Targeted web search</td>
<td>OpenIE</td>
<td>Single precision</td>
<td>Supervised per-statement</td>
<td>0.3 M</td>
</tr>
<tr>
<td>Quasimodo</td>
<td>General triples</td>
<td>User questions</td>
<td>OpenIE</td>
<td>Single precision</td>
<td>Supervised per-statement</td>
<td>4 M (v1.3)</td>
</tr>
<tr>
<td>DoQ</td>
<td>Quantity triples</td>
<td>Web crawls</td>
<td>Co-occurrence</td>
<td>Frequency</td>
<td>-</td>
<td>(120 M)</td>
</tr>
<tr>
<td>Dice</td>
<td>General triples</td>
<td>Existing structured CSKBs</td>
<td>-</td>
<td>Four quantitative facets</td>
<td>Joint MaxSAT</td>
<td>-</td>
</tr>
<tr>
<td>Ascent</td>
<td>General triples</td>
<td>Targeted web search</td>
<td>Facet-based OpenIE</td>
<td>Qualitative facets, subject constraints, frequency</td>
<td>Similarity clustering</td>
<td>8.6 M</td>
</tr>
</tbody>
</table>
Outline

1. Introduction to CSK
2. Crowdsourced CSKB construction
3. Text-extraction for ACSKB construction
   1. Overview
   2. Recipe
   3. Example projects
4. CSKBs: Summary and Outlook
Evaluation

• Intrinsic evaluation
  • Size
  • Precision
  • Salience
  • Recall
  → Based on user judgments/input

• Extrinsic evaluation
  • Wide set of academic benchmarks available
    • AllenAI science challenge perhaps most prominent
  • Often focus on reasoning, not just knowledge
AllenAI challenge

- ~8000 real school questions

Which property of a mineral can be determined just by looking at it? (A) luster [correct] (B) mass (C) weight (D) hardness

A student riding a bicycle observes that it moves faster on a smooth road than on a rough road. This happens because the smooth road has (A) less gravity (B) more gravity (C) less friction [correct] (D) more friction

Table 2: Grade-level distribution of ARC questions
Accuracy Over Time

Accuracy

Submission Date

01/01/2018
01/01/2019
01/01/2020
01/01/2021
01/01/2022

Running Best
Submissions
A Breakthrough for A.I. Technology: Passing an 8th-Grade Science Test

By Cade Metz
Sept. 4, 2019

SAN FRANCISCO — Four years ago, more than 700 computer scientists competed in a contest to build artificial intelligence that could pass an eighth-grade science test. There was $80,000 in prize money on the line.

They all flunked. Even the most sophisticated system couldn’t do better than 60 percent on the test. A.I. couldn’t match the language and logic skills that students are expected to have when they enter high school.

But on Wednesday, the Allen Institute for Artificial Intelligence, a prominent lab in Seattle, unveiled a new system that passed the test with room to spare. It correctly answered more than 90 percent of the questions on an eighth-grade science test and more than 80 percent on a 12th-grade exam.
Benchmarks on commonsense reasoning

• CommonsenseQA
  • [https://www.tau-nlp.org/commonsenseqa](https://www.tau-nlp.org/commonsenseqa)
  • Talmor et al. NAACL, 2019

• CommonGen
  • Lin et al., Arxiv, 2020

• MC-TACO
  • Zhou et al., EMNLP 2019

• Semeval 2020 Task 4
  • [https://competitions.codalab.org/competitions/21080](https://competitions.codalab.org/competitions/21080)

• Further listings:
  • [https://leaderboard.allenai.org/](https://leaderboard.allenai.org/)
CommonsenseQA

Where would I not want a fox?
👍 hen house, 👎 england, 👎 mountains, 👎 english hunt, 👎 california

Why do people read gossip magazines?
👍 entertained, 👎 get information, 👎 learn, 👎 improve know how, 👎 lawyer told to

What do all humans want to experience in their own home?
👍 feel comfortable, 👎 work hard, 👎 fall in love, 👎 lay eggs, 👎 live forever
CommonGen

Concept-Set: a collection of objects/actions.

- dog  
- frisbee  
- catch  
- throw

Generative Commonsense Reasoning

Expected Output: everyday scenarios covering all given concepts.

- A dog leaps to catch a thrown frisbee. [Humans]
- The dog catches the frisbee when the boy throws it.
- A man throws away his dog's favorite frisbee expecting him to catch it in the air.

GPT2: A dog throws a frisbee at a football player. [Machines]

Unilm: Two dogs are throwing frisbees at each other.

BART: A dog throws a frisbee and a dog catches it.

T5: dog catches a frisbee and throws it to a dog
Example 1. (event ordering)

*Paragraph:* Growing up on a farm near St. Paul, L. Mark Bailey didn't dream of becoming a judge.

*Question:* What did Mark do right after he found out that he became a judge?

- had a nice dinner
- he buy a 45-acre horse farm
- he celebrated
- mark retired
- he dream of participating himself

Example 2. (event duration)

*Paragraph:* Growing up on a farm near St. Paul, L. Mark Bailey didn't dream of becoming a judge.

*Question:* How many years did it take for Mark to become a judge?

- 63 years
- 7 weeks
- 7 years
- 7 seconds
- 7 hours
References

• Based on tutorial @ KI 2020

• Related tutorials
  • Commonsense Reasoning for Natural Language Processing, Sap et al., ACL 2020 (NLP)
  • Common Sense Knowledge Graphs (CSKGs), Ilievski et al, ISWC2020 (Semantic Web)


Further references


Take home

1. Structured CSK important interpretable and scrutable building block for trustworthy AI
2. Coherence and density require consolidation
   1. Multi-source validation
   2. Constraint-based reasoning
3. Semantics of CSK still with gaps
   1. Opportunity for deliberate KR
4. Advance of neural models suggest hybrid architectures
   1. Neural model for bridging language gaps (see also next lecture)