ConceptNet in Context

Robyn Speer
February 8, 2020

LUMINOSO

Origins

- Open Mind Common Sense
- Created by Catherine Havasi, Push Singh, Thomas Lin, others, in 1999
- Motivating example: making search more natural
  - “my cat is sick” -> “veterinarian cambridge ma”
- Goal: teach computers the basic things that people know
- Represent this knowledge in natural language, so non-experts can contribute it and interact with it
- Hugo Liu first transformed Open Mind into a knowledge graph, ConceptNet

Collecting knowledge with crowdsourcing

4	You are likely to find water in soup	by matrix29
4	Something you find at the supermarket is soup	by aww
3	You are likely to find chicken in soup.	by jbbdude
3	Something you find in a container is soup	by bedume
3	Something you find in a jar is soup	by smtango
3	soup is a type of hot liquid	by verbosity
2	You are likely to find a bean in a soup	by Fable
2	Soup is good food	by Visionsofk
2	You are likely to find a fungus in soup	by bedume
2	Soup is a light meal	by push
2	a bowl is for soup	by friscojohn
2	Billibi is a soup	by Ariadne
2	Something you find in a can is soup	by guru1

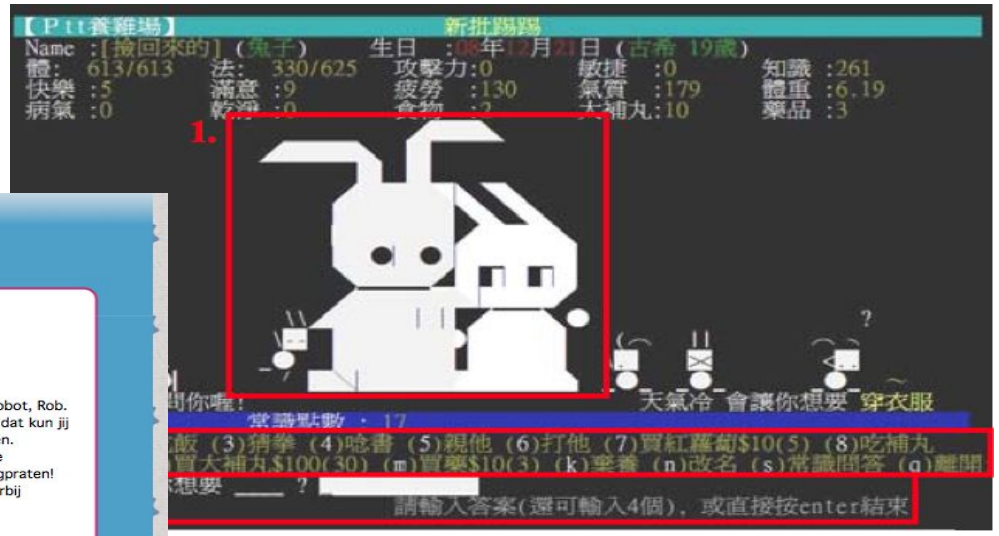
Open Mind Common Sense, around 2006

Open Mind wants to know...

Are these statements true?

- You are likely to find **soup** in **a market**.
  Yes No Sort of
- You are likely to find **food** in **soup**.
  Yes No Sort of
- You are likely to find **soup** in **the cabinet**.
  Yes No Sort of
- **soup** can be **eaten**
  Yes No Sort of
- You are likely to find **soup** in **the freezer**.
  Yes No Sort of

An international, multilingual project



OPEN MIND

common sense no Brasil

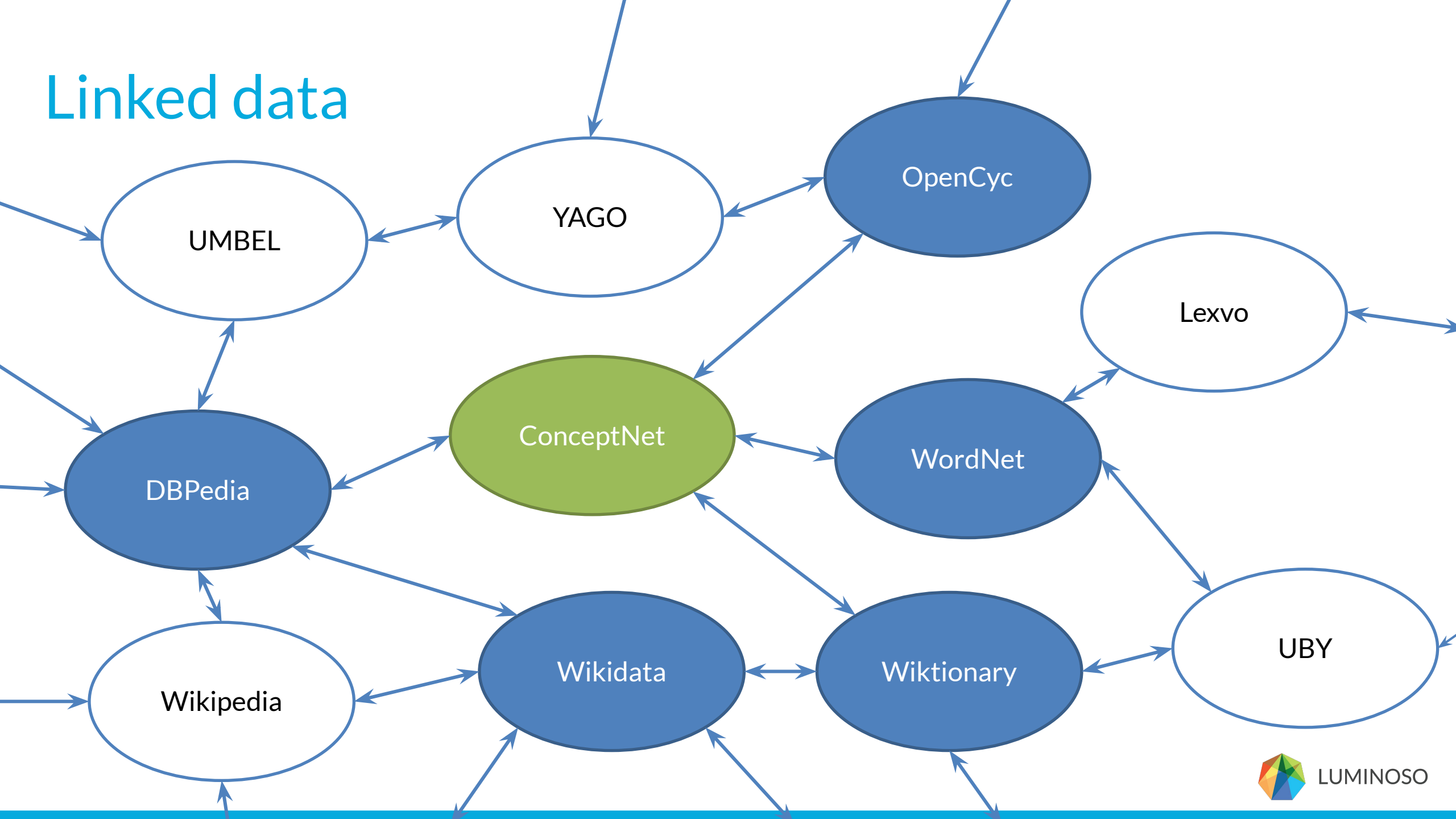
Ensinando ao computador as coisas que todos nós sabemos

Conheça os projetos do nosso grupo de pesquisa

O que é Open Mind Common Sense?

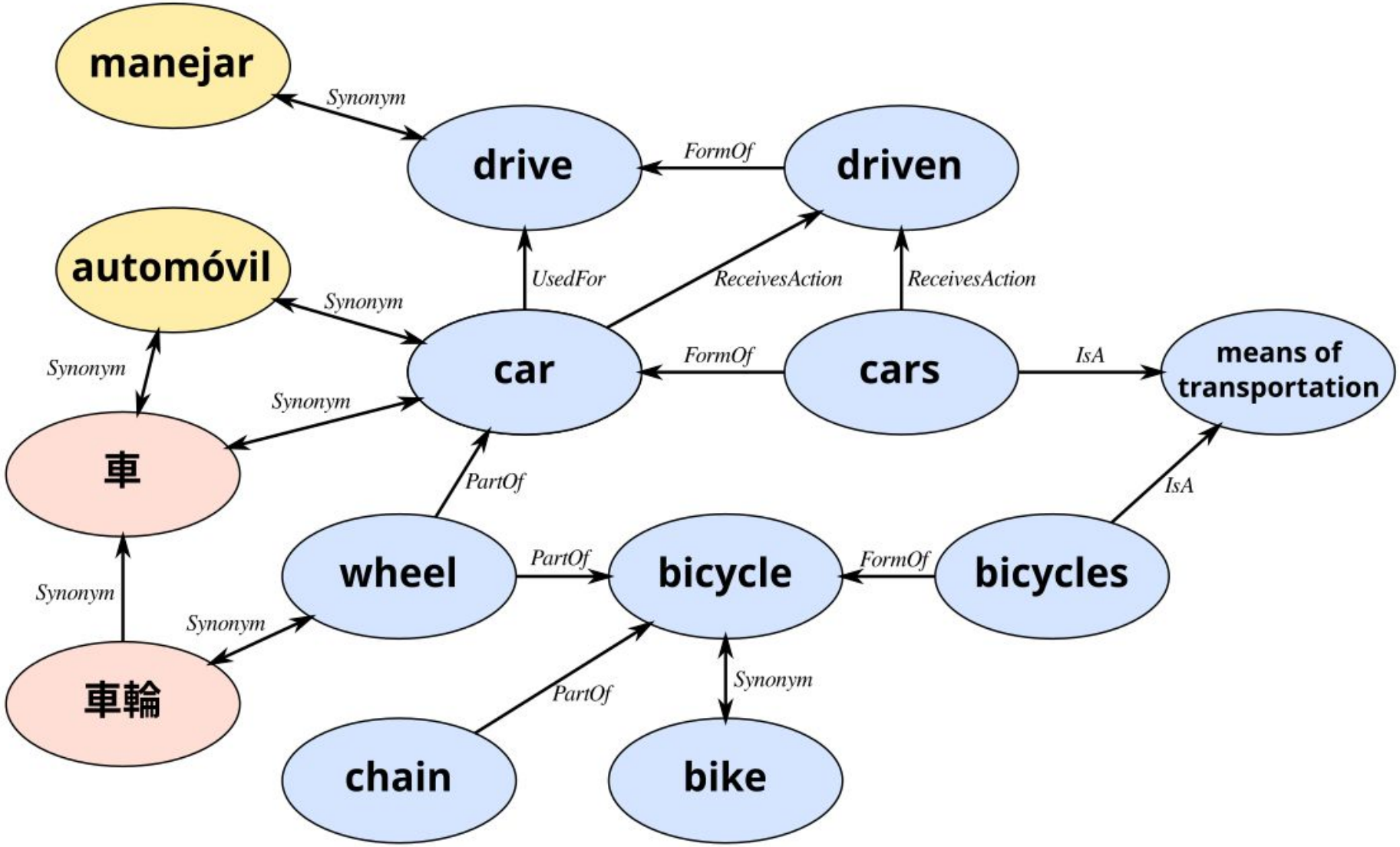
Junte-se a nossa comunidade no Orkut

Linked data



LUMINOSO

A small fragment of ConceptNet 5



LUMINOSO

ConceptNet's data sources

- **Crowdsourced knowledge**
  - Open Mind Common Sense, Wiktionary, DBPedia, Yahoo Japan / Kyoto University project
- **Games with a purpose**
  - Verbosity, nadya.jp
- **Expert resources**
  - Open Multilingual WordNet, JMDict, CEDict, OpenCyc, CLDR emoji definitions

How do we represent this in machine learning?

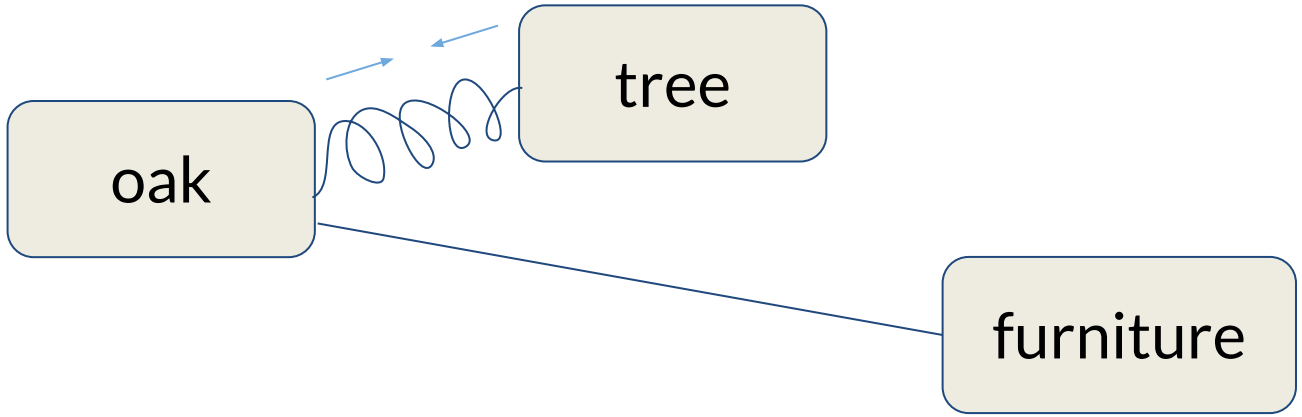
Knowledge graphs as word embeddings

- We started representing ConceptNet as embeddings in 2007
  - Enabled new capabilities that were difficult to evaluate
- When word embeddings became popular, they were instead based on distributional semantics (CBOW, skipgrams, etc.)
- Retrofitting (Manaal Faruqui, 2015) revealed the power of distributional semantics **plus** a knowledge graph
  - Apply knowledge-based constraints after training
  - For some reason this works better than during training

Retrofitting with a knowledge graph

- Terms that are connected in the knowledge graph should have vectors that are closer together

- Many extensions now:
  - "Counter-fitting" moves antonyms farther apart (Mrkšić et al., 2016)
  - "Morph-fitting" accounts for morphology (Vulić et al., 2017)
  - Applied to the union of vocabularies instead of the intersection (our work)

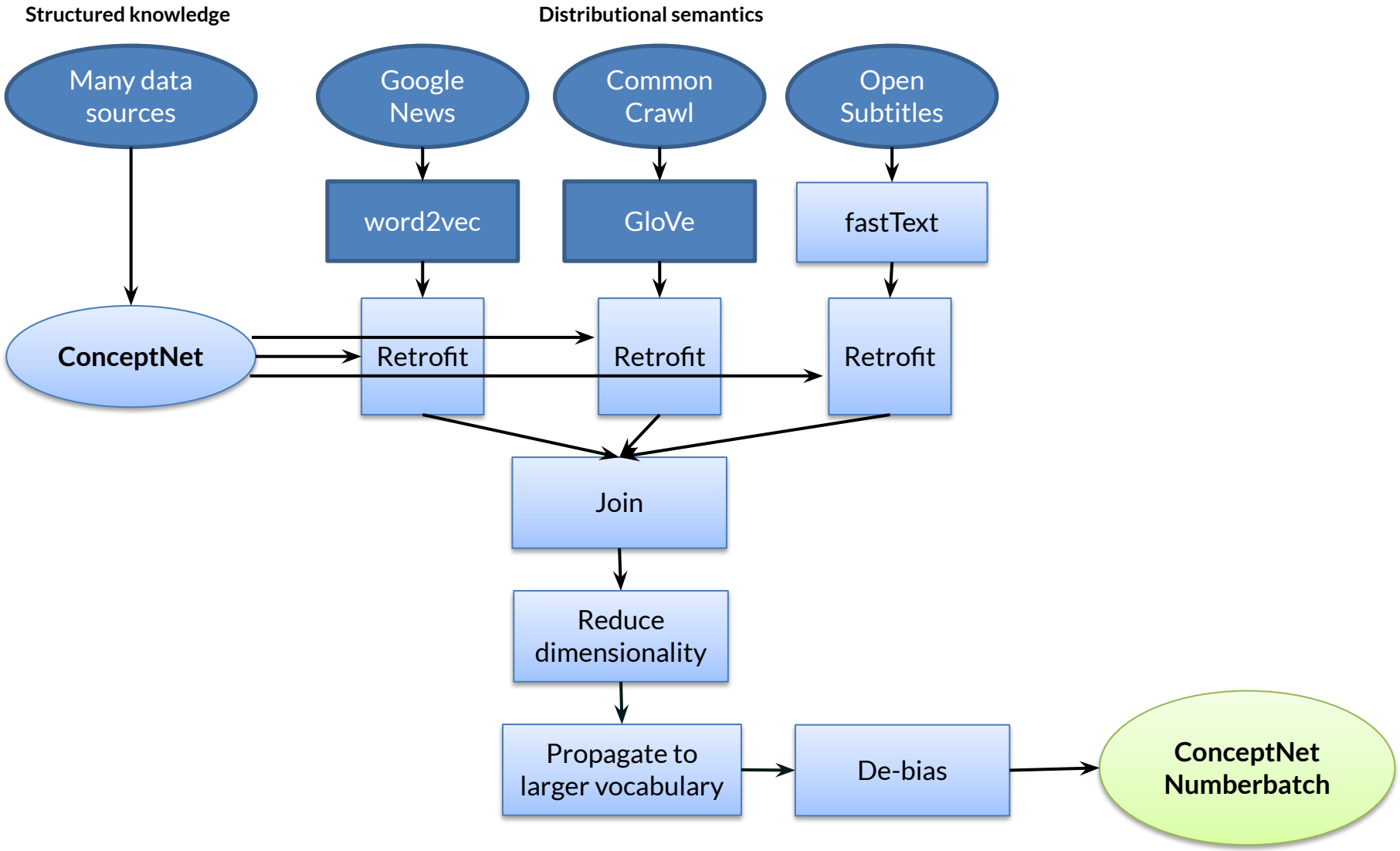


LUMINOSO

ConceptNet Numberbatch

- Word embeddings with common sense built in
- Hybrid of ConceptNet and distributional semantics, via our variant of retrofitting
- Multilingual by design
- Open source, open data

Building ConceptNet Numberbatch

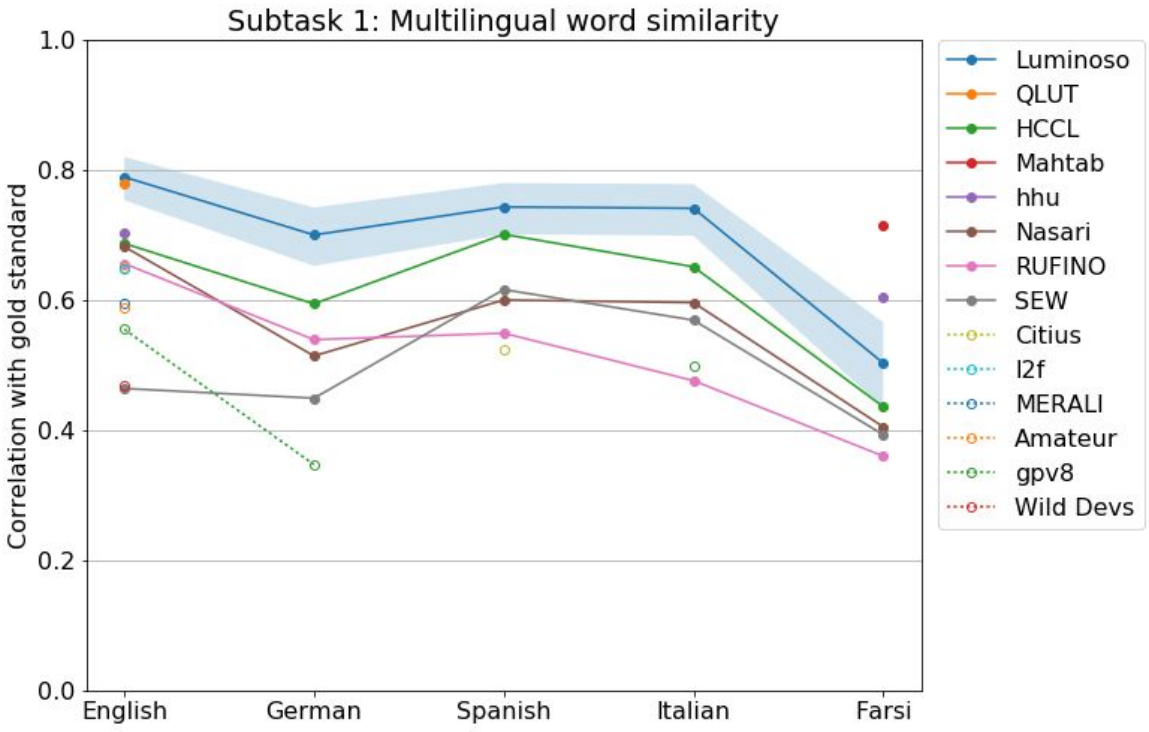


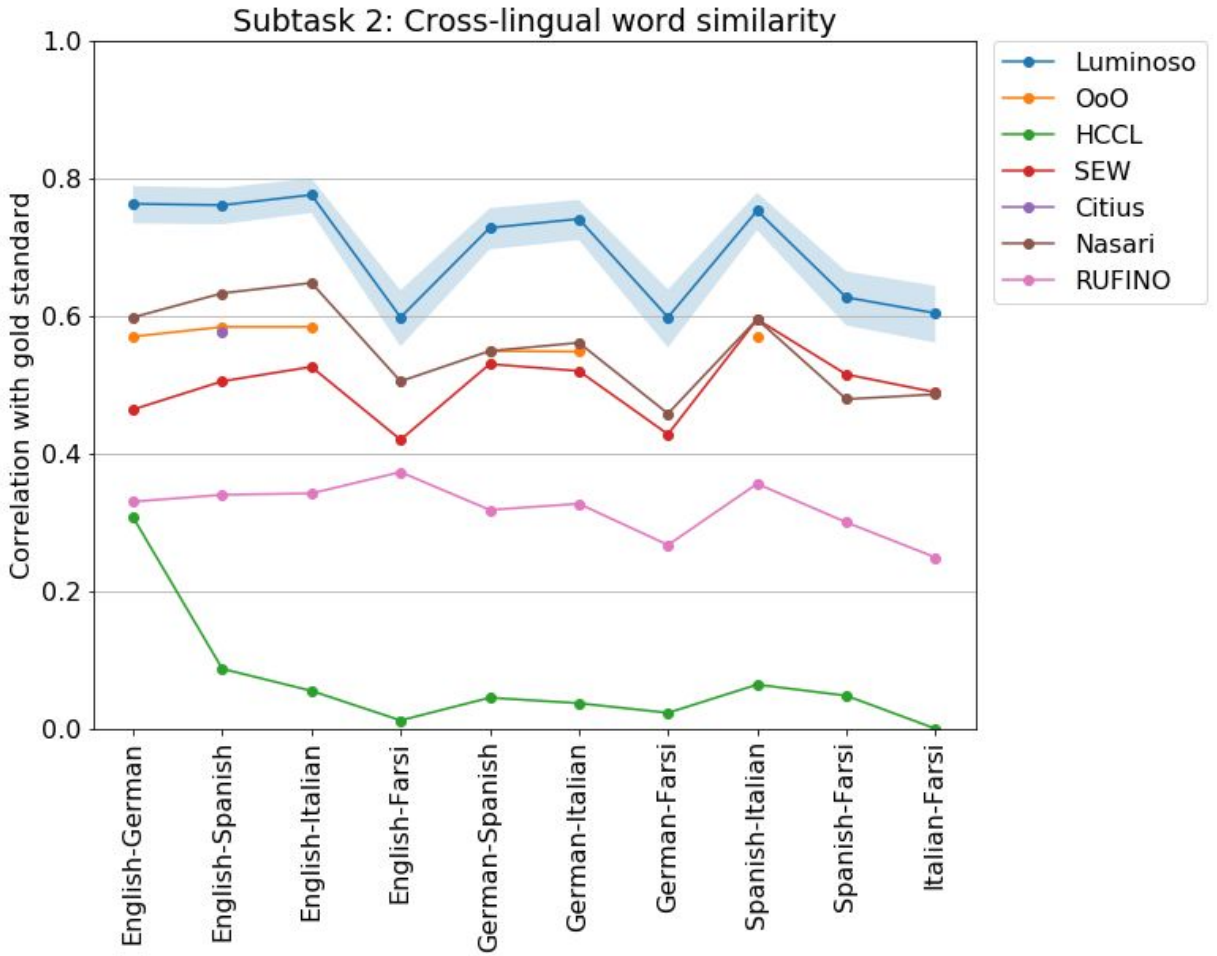
LUMINOSO

Benchmarks

Hey wow, this actually works

Intrinsic evaluation: Word relatedness (SemEval 2017)





LUMINOSO

Intrinsic evaluation: Distinguishing attributes (SemEval 2018)

Term 1	Term 2	Attribute	Discriminative?
lambs	cattle	wool	✔
		***Lambs** produce **wool**, while **cattle** do not.*	
shoulder	leg	arm	✔
		*A **shoulder** is attached to an **arm**, while a **leg** is not.*	
train	subway	rails	✖
		*Both a **train** and a **subway** involve **rails**, so rails are not a discriminative attribute here.*	
finger	soup	water	✖
		***Soup** may be related to **water**, but this is the wrong direction. In this task, a discriminative attribute must be related to the first term and not the second.*	

- We got **74%** accuracy (2nd place) by directly querying ConceptNet Numberbatch
- Additional features trained on the provided training data didn't help on the test set
- All top systems used knowledge graphs

LUMINOSO

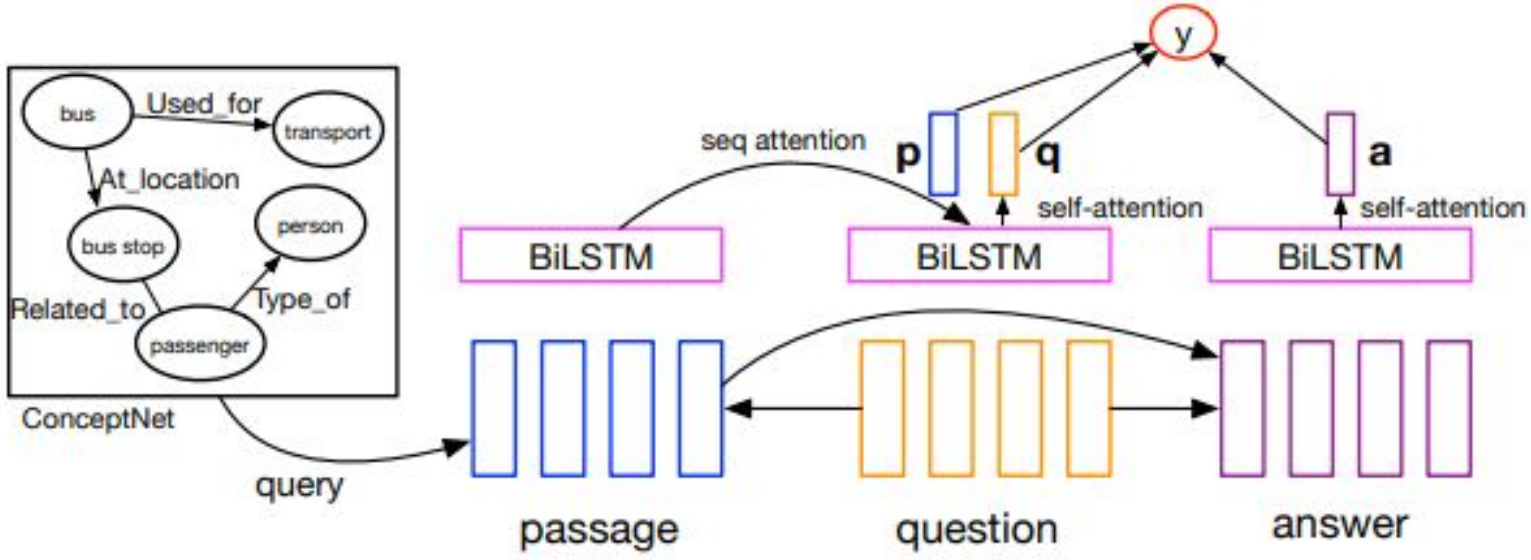
Extrinsic evaluation: Story understanding

- SemEval-2018 task: answer simple multiple-choice questions about a passage

> **Text**: It was a long day at work and I decided to stop at the gym before going home. I ran on the treadmill and lifted some weights. I decided I would also swim a few laps in the pool. Once I was done working out, I went in the locker room and stripped down and wrapped myself in a towel. I went into the sauna and turned on the heat. I let it get nice and steamy. I sat down and relaxed. I let my mind think about nothing but peaceful, happy thoughts. I stayed in there for only about ten minutes because it was so hot and steamy. When I got out, I turned the sauna off to save energy and took a cool shower. I got out of the shower and dried off. After that, I put on my extra set of clean clothes I brought with me, and got in my car and drove home.
>
> **Q1**: Where did they sit inside the sauna?
>
> (a) on the floor
> (b) on a bench
>
> **Q2**: How long did they stay in the sauna?
>
> (a) about ten minutes
> (b) over thirty minutes

Story understanding at SemEval-2018

- Winning system: TriAN (Three-way Attention and Relational Knowledge for Commonsense Machine Comprehension)
  - Liang Wang et al., Yuanfudao Research
  - Concatenated each input embedding with a *relation embedding*, trained to represent what ConceptNet relations exist between the word and the passage



LUMINOSO

Other benchmarks

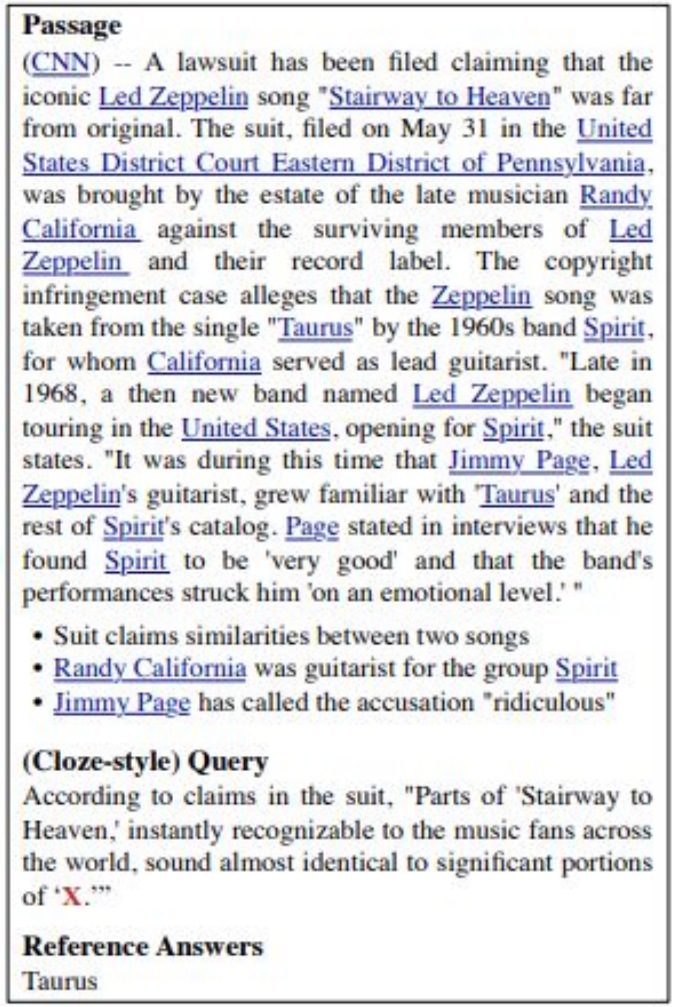
- Story Cloze Test
  - GPT-1 was a breakthrough, but Jiaao Chen et al. (2018) improved on it slightly with ConceptNet
- OpenBookQA
  - ConceptNet didn't help, but Ai2's own science knowledge graph Aristo did (Todor Mihaylov et al., 2018)
- CommonsenseQA
  - Generating synthetic training data using ConceptNet helps (Zhi-Xiu Ye et al., 2019)

Has the situation changed?

- Transformer models were big news in 2019
- Language models such as BERT, XLNet, and GPT-2 indicate some level of implicit common sense understanding

ReCoRD / COIN shared task (2019)

- Run by Simon Ostermann, Sheng Zhang, Michael Roth, and Peter Clark for EMNLP
- Answer questions based on news stories, some of which are intended to require common sense reasoning
- Winning system: XLNet plus rule-based answer verification (Xiepeng Li et al.)
- None of the top 3 systems used external knowledge



Passage

(CNN) -- A lawsuit has been filed claiming that the iconic Led Zeppelin song "Stairway to Heaven" was far from original. The suit, filed on May 31 in the United States District Court Eastern District of Pennsylvania, was brought by the estate of the late musician Randy California against the surviving members of Led Zeppelin and their record label. The copyright infringement case alleges that the Zeppelin song was taken from the single "Taurus" by the 1960s band Spirit, for whom California served as lead guitarist. "Late in 1968, a then new band named Led Zeppelin began touring in the United States, opening for Spirit," the suit states. "It was during this time that Jimmy Page, Led Zeppelin's guitarist, grew familiar with 'Taurus' and the rest of Spirit's catalog. Page stated in interviews that he found Spirit to be 'very good' and that the band's performances struck him 'on an emotional level.' "

- Suit claims similarities between two songs
- Randy California was guitarist for the group Spirit
- Jimmy Page has called the accusation "ridiculous"

(Cloze-style) Query

According to claims in the suit, "Parts of 'Stairway to Heaven,' instantly recognizable to the music fans across the world, sound almost identical to significant portions of '**X**.'"

Reference Answers

Taurus

LUMINOSO

Why Do Masked Neural Language Models Still Need Common Sense Knowledge?

- Presumably you just saw this talk by Sunjae Kwon
- MNLMs seem to understand a lot but they still struggle with things that actually require common-sense
- So try augmenting your system with an attention model of edges in a knowledge graph

A simplistic answer to why we need knowledge

- Language models describe text that is *likely*
- Statements that are too obvious are unlikely

The first thing you do when you start driving is get a new airbag. If you're familiar with the issue the airbag must always be deployed and the airbag cannot collapse without first being re-installed to minimize injury. In short, if

You would use a pen to verni-tate and the ink would not be completely dry in the center, but it wouldn't stay dry, either.

You create a book by reading a book, which means that you create a model of the universe that is a representation of a book (a table, a board, an image, a shape) that you can modify as needed. In this way your model of the universe can be very elaborate or very small. The important thing to realize is that you

Headphones are used for the vast majority of games. A few titles that have headphones are the following: - Halo 4 - Mario Kart 8 - Assassin's Creed Syndicate Halo 3: ODST - The Longest Journey - Resident Evil 6 Uncharted:

Jogging causes erythropoietin accumulation and the increased production of the hormones GH and IGF 1 . This may be particularly important to those with a prediabetes or insulin dependent diabetes . Research into the effects of

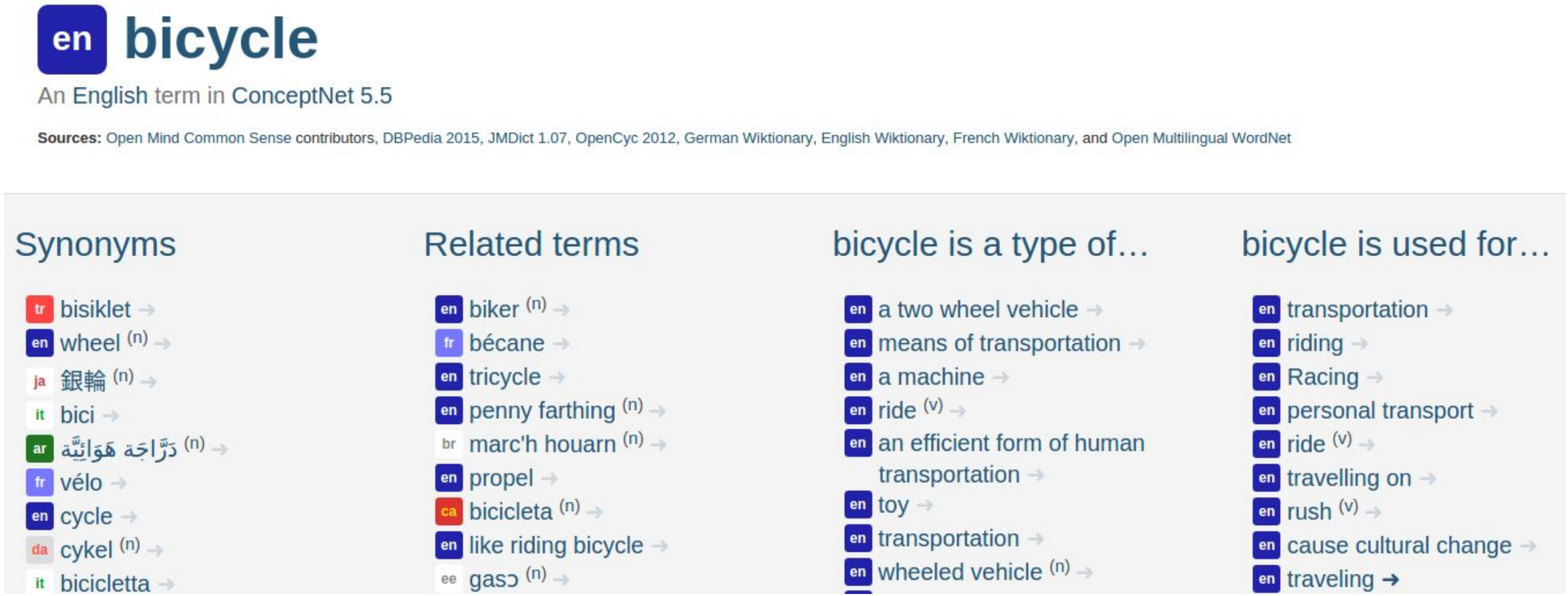
(nonsensical “knowledge” produced by the GPT-2 model at talktotransformer.com)

Other languages exist

- Most neural language models only learn English, unless they're specifically designed for translation
- The corpora in other languages aren't big enough or representative enough
- ConceptNet's representation connects many languages (100 languages have over 10k terms each)

Using ConceptNet

conceptnet.io – a browsable interface



- Links to other resources such as the documentation wiki and the Gitter chat

LUMINOSO

api.conceptnet.io – a Linked Data API


```
{
  "@context": [
    "http://api.conceptnet.io/ld/conceptnet5.5/context.ld.json",
    "http://api.conceptnet.io/ld/conceptnet5.5/pagination.ld.json"
  ],
  "@id": "/c/en/bicycle",
  "edges": [
    {
      "@id": "/a/[/r/AtLocation/,/c/en/bicycle/,/c/en/garage/]",
      "dataset": "/d/conceptnet/4/en",
      "end": {
        "@id": "/c/en/garage",
        "label": "the garage",
        "language": "en",
        "term": "/c/en/garage"
      },
      "license": "cc:by/4.0",
      "rel": {
        "@id": "/r/AtLocation",
        "label": "AtLocation"
      },
```


LUMINOSO

How should we represent ConceptNet in question answering?

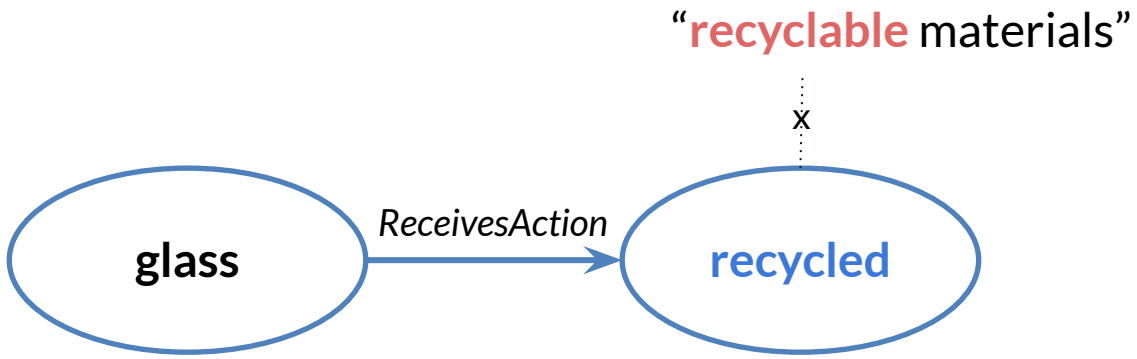
- Everything changes so fast that I can't bless one technique
- Encoding ConceptNet edges as if they were sentences, in an attention model, seems to work well in multiple systems
- Alternatively, ConceptNet can augment training data
- If the thing you need background knowledge for is straightforward enough… word embeddings and retrofitting are still an option

Recommendation: Combine ConceptNet with task-specific training data

- ConceptNet isn't going to know everything it needs to know for your task
- Knowing so many specific things is beyond its scope
- ConceptNet is noisy: it might know one thing about your topic except it's wrong
- Use it as a starting point or a constraint

Recommendation: Don’t assume completeness

- ConceptNet has ~15 million facts in English
- There are many more than 15 million facts of general knowledge
- Word forms might be slightly different
- Fuzzy matching (perhaps via embeddings) is important



LUMINOSO

Recommendation: download the data

- If you just need to iterate all the edges in ConceptNet, you don't need all the Python and PostgreSQL setup
- conceptnet.io -> Wiki -> Downloads

LUMINOSO

blog.conceptnet.io

- Tutorials built using ConceptNet
- Updates to ConceptNet and related open-source tools
- AI fairness

LUMINOSO

Extra slides

Inferring common sense with CoMET

- Bosselut et al. (2019), at Ai2
- Uses ConceptNet as a training set instead of a knowledge resource
- Fine-tune a GPT language model to generate ConceptNet statements
  - (but only in English)

Seed	Relation	Completion	Plausible
piece	PartOf	machine	✓
bread	IsA	food	✓
oldsmobile	IsA	car	✓
happiness	IsA	feel	✓
math	IsA	subject	✓
mango	IsA	fruit	✓
maine	IsA	state	✓
planet	AtLocation	space	✓
dust	AtLocation	fridge	
puzzle	AtLocation	your mind	🤔
college	AtLocation	town	✓
dental chair	AtLocation	dentist	✓
finger	AtLocation	your finger	
sing	Causes	you feel good	✓
doctor	CapableOf	save life	✓
post office	CapableOf	receive letter	✓
dove	SymbolOf	purity	✓
sun	HasProperty	big	✓
bird bone	HasProperty	fragile	✓
earth	HasA	many plant	✓
yard	UsedFor	play game	✓
get pay	HasPrerequisite	work	✓
print on printer	HasPrerequisite	get printer	✓
play game	HasPrerequisite	have game	✓
live	HasLastSubevent	die	✓
swim	HasSubevent	get wet	✓
sit down	MotivatedByGoal	you be tire	✓
all paper	ReceivesAction	recycle	✓
chair	MadeOf	wood	✓
earth	DefinedAs	planet	✓

Recommendation: make sure text normalization matches

Example text: “SETTINGS” (English)

- Wrong: `/c/en/SETTINGS, /c/en/setting, /c/en/set`
- Right: `/c/en/settings`

Example text: “aujourd’hui” (French)

- Wrong: `/c/fr/aujourd, /c/fr/hui`
- Right: `/c/fr/aujourd'hui`

Use `conceptnet5.nodes.standardized_concept_uri`, or the simple `text_to_uri.py` included with Numberbatch

Align, Mask, and Select

- Zhi-Xiu Ye et al. (2019)
- Improve performance on CommonsenseQA by generating synthetic training questions from Wikipedia and ConceptNet
- Distractors are other nodes in ConceptNet

(1) A triple from ConceptNet
(population, AtLocation, city)
(2) **Align** with the English Wikipedia dataset to obtain a sentence containing "population" and "city"
The largest **city** by **population** is Birmingham, which has long been the most industrialized city.
(3) **Mask** "city" with a special token "[QW]"
The largest **[QW]** by **population** is Birmingham, which has long been the most industrialized city?
4) **Select** distractors by searching (population, AtLocation, $*$) in ConceptNet
(population, AtLocation, Michigan) (population, AtLocation, Petrie dish) (population, AtLocation, area with people inhabiting) (population, AtLocation, country)
5) Generate a multi-choice question answering sample
question: The largest **[QW]** by **population** is Birmingham, which has long been the most industrialized city? **candidates**: *city*, Michigan, Petrie dish, area with people inhabiting, country

Knowledge graphs in Portuguese NLP

Gonçalo Oliveira, H. (2018), *Distributional and Knowledge-Based Approaches for Computing Portuguese Word Similarity*

- Knowledge graphs (including ConceptNet) improve Portuguese semantic evaluations
- Best results come from combining multiple knowledge graphs representing different variants of Portuguese

OpenBookQA (Ai2)

- “Can a Suit of Armor Conduct Electricity?” (Todor Mihaylov et al., 2018)
- QA over elementary science questions
- ConceptNet did not improve baseline results
- Ai2 built their own knowledge graph, Aristo, that focused on science knowledge and did improve the results