Building a knowledge graph through machine learning

NELL: Never-Ending Language Learning

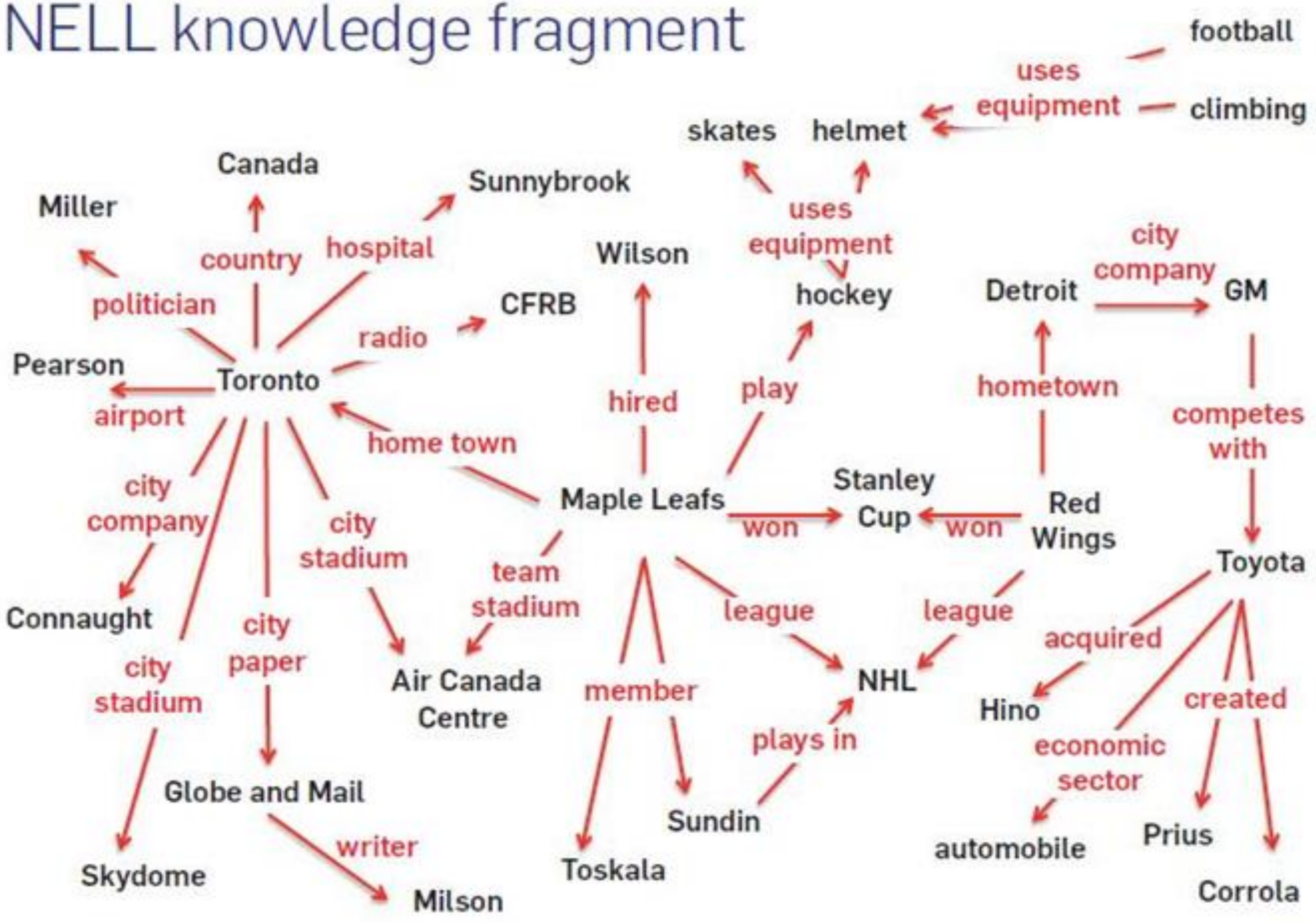
- Tom Mitchell et al. (CMU), 2010 to present.
- Learning to “read the web” 24 hours/day.

- Training data includes a collection of 1.2 billion web pages.
- Access to additional data through search engine APIs (100K calls/day).

- KB has 2.8 million instances over 1186 different categories.

- KB is freely available for download.
- You can help train NELL via Twitter.

NELL knowledge fragment



Motivation for NELL

Thesis: “we will never truly understand human or machine learning until we can build computer programs that, like people,

- Learn many different types of knowledge or functions
- From years of diverse, mostly self-supervised experience
- In a staged curricular fashion, where previously learned knowledge enables learning further types of knowledge
- Where self-reflection and the ability to formulate new representations and new learning tasks enable the learning to avoid stagnation and performance plateaus.”

Basic idea

- NELL learns several things:
  - Categories
  - Triples: noun phrase 1 - relation - noun phrase 2
  - New relations

- Multiple inference algorithms propose triples and gather evidence for them.
  - Linguistic information
  - Word co-occurrence
  - Image labeling
  - Etc.

- Categories and triples supported by multiple sources of evidence grow in confidence.

Ask NELL on-demand results:

3 possible entities found

Click to change visible entity:

- beef (meat)
- beef (grain)
- "beef"

---

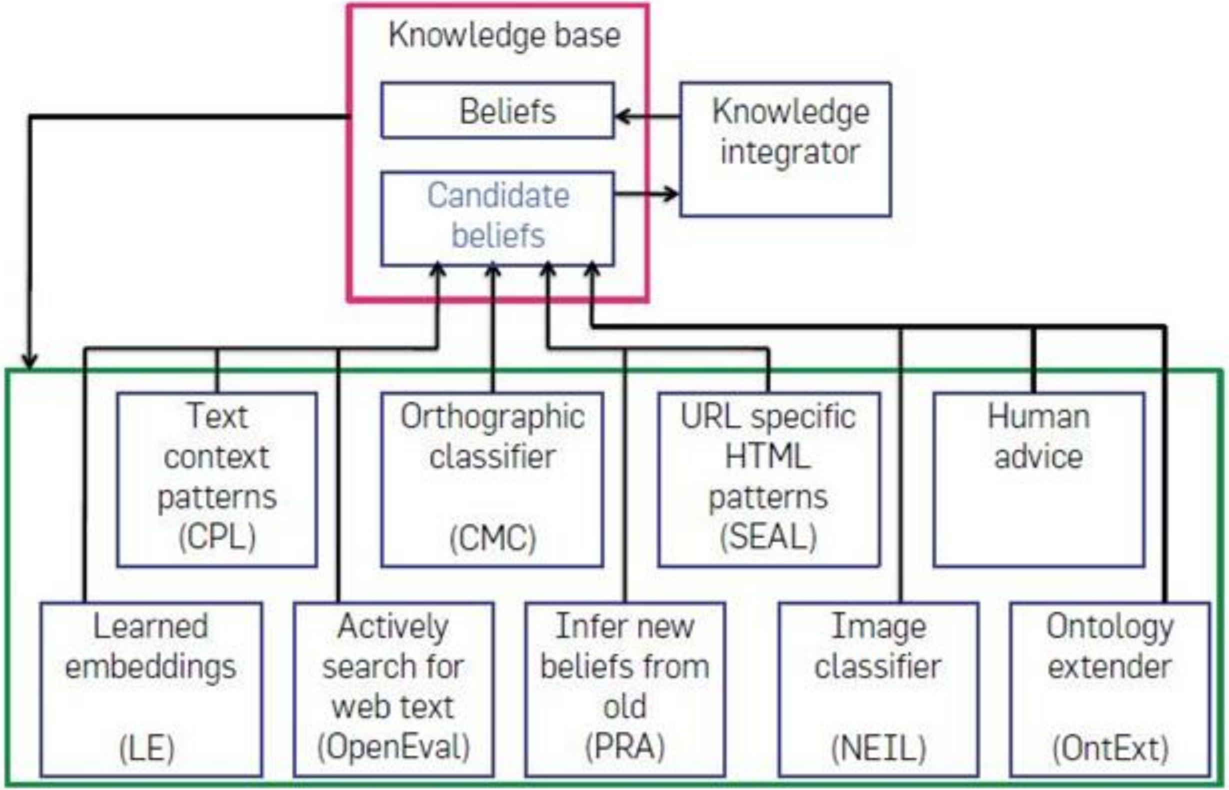
beef (meat)

literal strings: beef, Beef, BEEF

categories

- meat(100.0%)
  - Seed
  - Human feedback from bkisiel @217 on 07-mar-2011 why? using beef
  - SEAL @189 (100.0%) on 15-jan-2011 why? using beef
  - CMC @532 (97.6%) on 15-mar-2012 why? using beef
  - CPL @802 (91.1%) on 08-jan-2014 why? using beef
- agriculturalproduct(100.0%)
  - Seed
  - SEAL @230 (100.0%) on 08-apr-2011 why? using beef
  - MBL @796 (100.0%) on 15-dec-2013 why? using concept:agriculturalproduct:beef

NELL architecture



Never-ending learning

- Set of learning tasks $L = \{ L_i \}$
- Task $L_i = <T_i, P_i, E_i>$
  - T_i is a task $<X_i, Y_i>$ specifying the domain of a function $f_i^* : X_i \rightarrow Y_i$
  - P_i is a performance metric $P_i : f \rightarrow \mathbb{R}$
  - E_i is an experience
- Coupling contraints $C = \{<\phi_k, V_k>\}$
  - ϕ_k specifies degree of satisfaction of the coupling constraint among tasks
  - V_k is a vector of indices over learning tasks specifying the arguments to ϕ_k
- $f_i^* = \arg\max_{f \in F_i} P_i(f)$

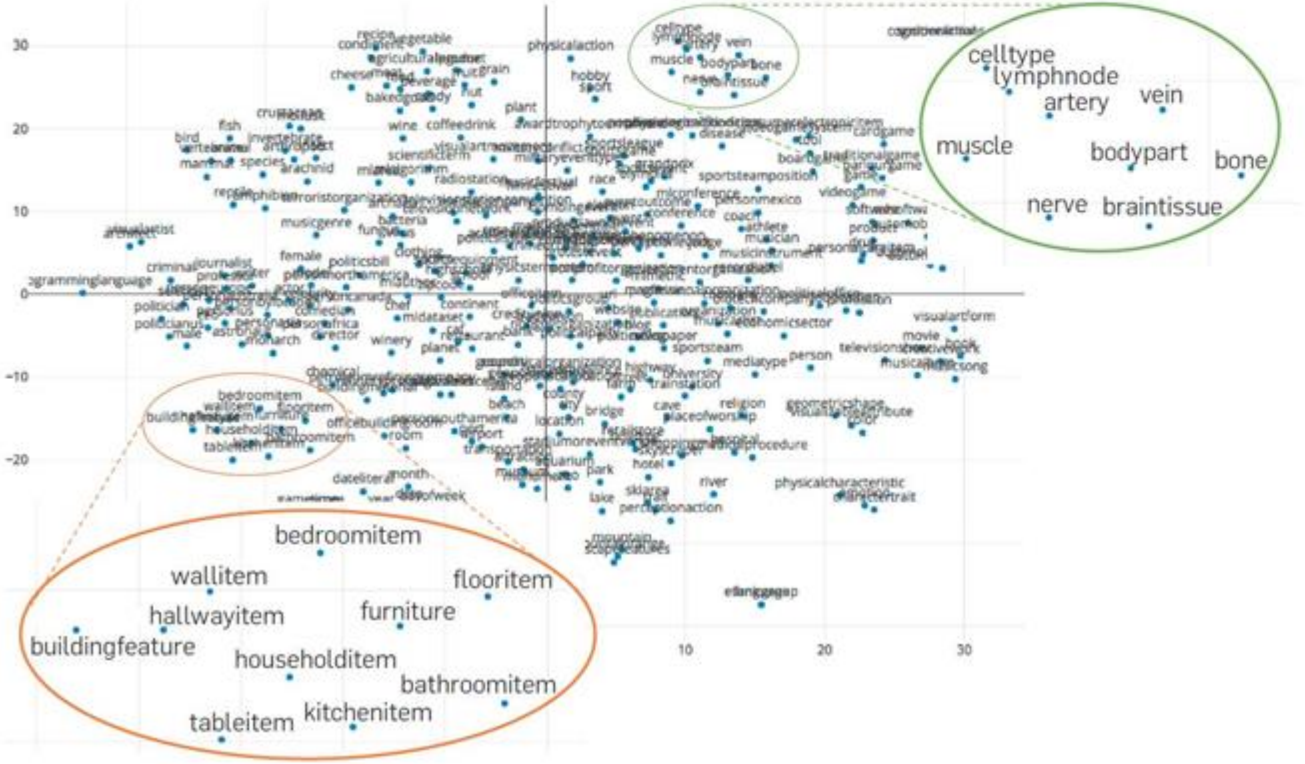
Goal: improve the quality of the task functions f_i as measured by the P_i.

NELL faces over 4100 distinct learning tasks.

Category classification tasks

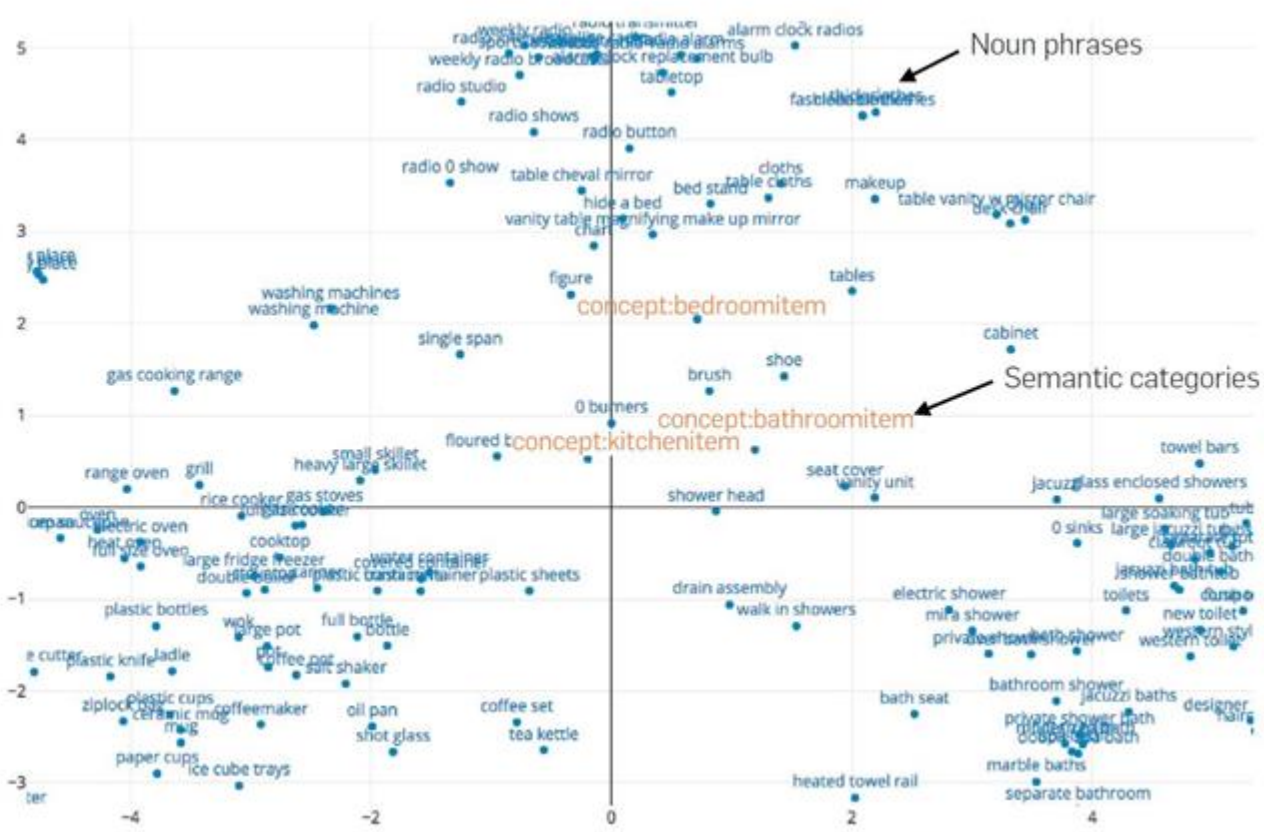
1. Character string features of the noun phrase: Coupled Morphological Classifier system (CMC)
2. Distribution of text contexts found around this noun phrase in the 1.2 billion page database: Coupled Pattern Learner system (CPL)
3. Distribution of text contexts found through active web search (OpenEval).
4. HTML structure of web pages that mention the noun phrase: Set Expander for Any Language system (SEAL)
5. Visual images associated with the noun phrase: Never Ending Image Learner (NEIL)
6. Learned vector embeddings (feature vectors) of the noun phrase: LE (Learned Embeddings)

Learned embeddings of 280 categories



(a) Embeddings of the semantic categories.

Bedroom, bathroom, and kitchen room items



(b) Embeddings of the noun phrases and semantic categories.

Relation classification

Does “Pittsburgh” + “US” satisfy the relation CityLocatedInCountry(x,y) ?

There are 461 relations in the ontology.

Four methods are used for relation classification:

1. Distribution of text contexts from CPL
2. Distribution of text context from OpenEval
3. HTML structure from SEAL
4. Learned vector embeddings from LE

Entity resolution

- Functions to classify whether pairs of noun phrases are synonyms.

- Noun phrases are kept distinct from the entities to which they refer.

- Necessary to deal with polysemy.
  - “Coach” can be either a person or a vehicle.

- Two methods are used:
  - String similarity
  - Similarities in beliefs about the entities

- NELL learns for each category what are the good types of knowledge to take as evidence for synonymy.

Inference rules among belief triples

- Functions that propose new beliefs to be added to the KB.

- For each relation, the corresponding function is represented by a collection of restricted Horn Clause rules learned by the Path Ranking Algorithm (PRA)

Sample of self-discovered NELL relations

- athleteWonAward
- animalEatsFood
- languageTaughtInCity
- clothingMadeFromPlant
- beverageServedWithFood
- fishServedWithFood
- athleteBeatAthlete
- athleteInjuredBodyPart
- arthropodFeedsOnInsect
- animalEatsVegetable
- plantRepresentsEmotion
- foodDecreasesRiskOfDisease
- clothingGoesWithClothing
- bacteriaCausesPhysCondition
- buildingFeatureMadeFromMaterial
- emotionAssociatedWithDisease
- foodCanCauseDisease
- agriculturalProductAttractsInsect
- arteryArisesFromArtery
- countryHasSportsFans
- bakedGoodServedWithBeverage
- beverageContainsProtein
- animalCanDevelopDisease
- beverageMadeFromBeverage

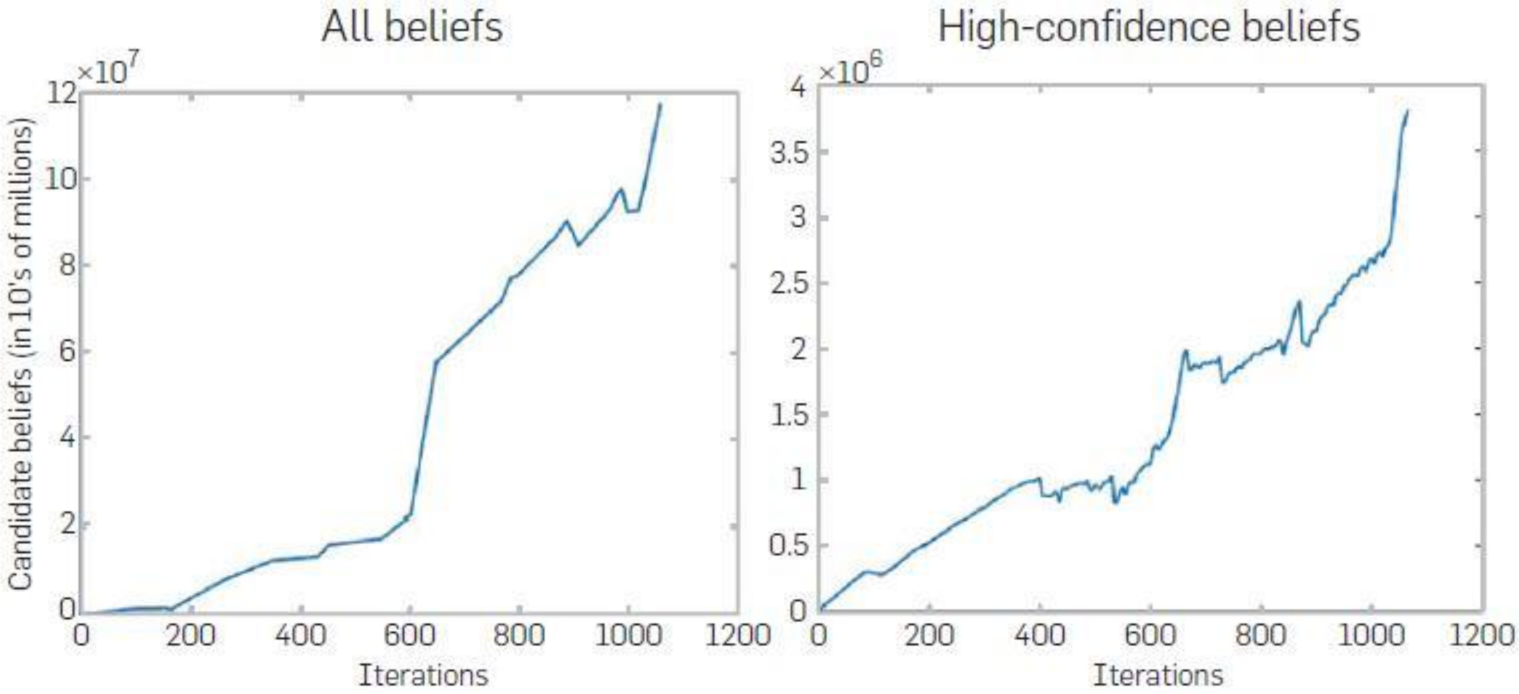
Coupling constraints

- **Multi-view co-training coupling**: do alternative methods for (1) classifying noun phrases into categories or (2) classifying noun phrase pairs into relations, yield the same conclusions?
- **Subset/superset coupling**: when a new category is added, find its parents. Make sure that $(\forall x)\ C_1(x) \Rightarrow C_2(x)$
- **Multi-label mutual-exclusion coupling**: when a new category is added, find categories disjoint from it. Makes sure that $(\forall x)\ C_1(x) \Rightarrow \neg C_2(x)$
- **Coupling relations to the argument types**: a relation x-R-y requires arguments of the appropriate category for x and y.
- **Horn clause coupling**: when NELL learns a rule of form

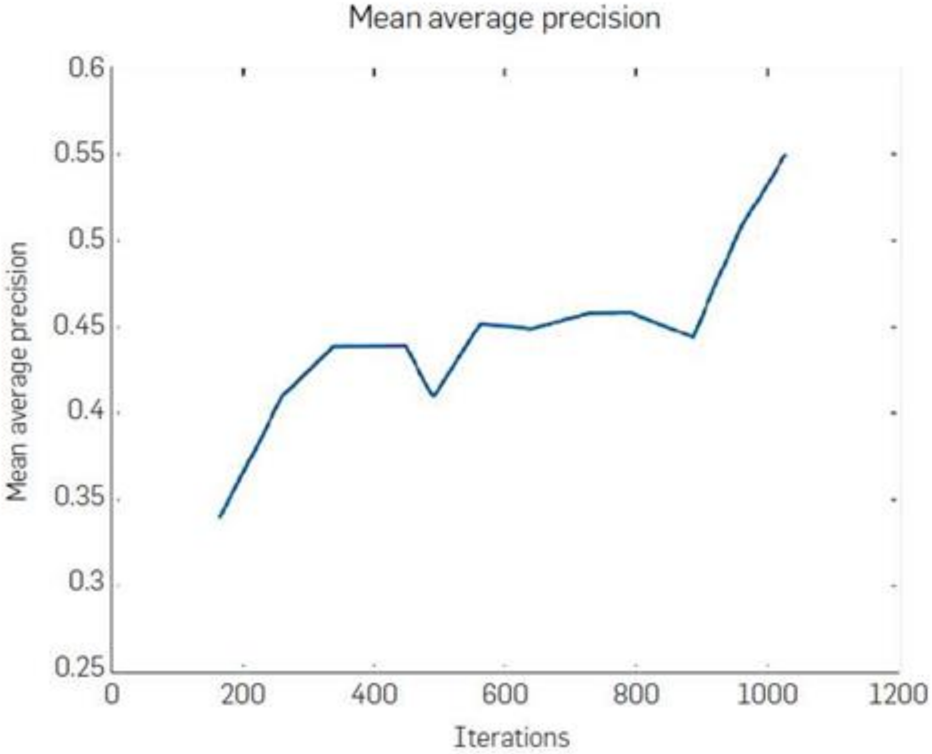
  $$(\forall x,y,z)\ R_1(x,y) \wedge R_2(y,z) \Rightarrow R_e(x,z)$$

  this serves as a coupling constraint between the R_i and category labels.

Growth of the KB over time

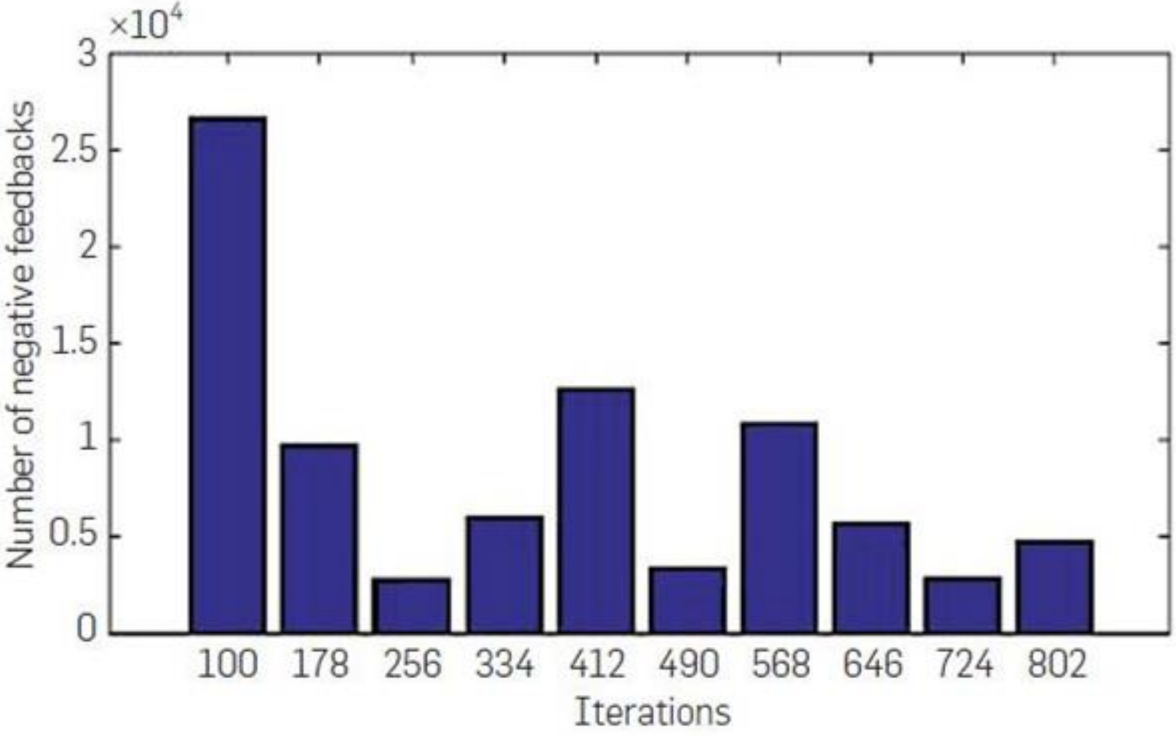


Performance improvement over time



Mean average precision over the 1000 most confident predictions for a sample of 18 categories and 13 relations in NELL's ontology.

Human correction of NELL

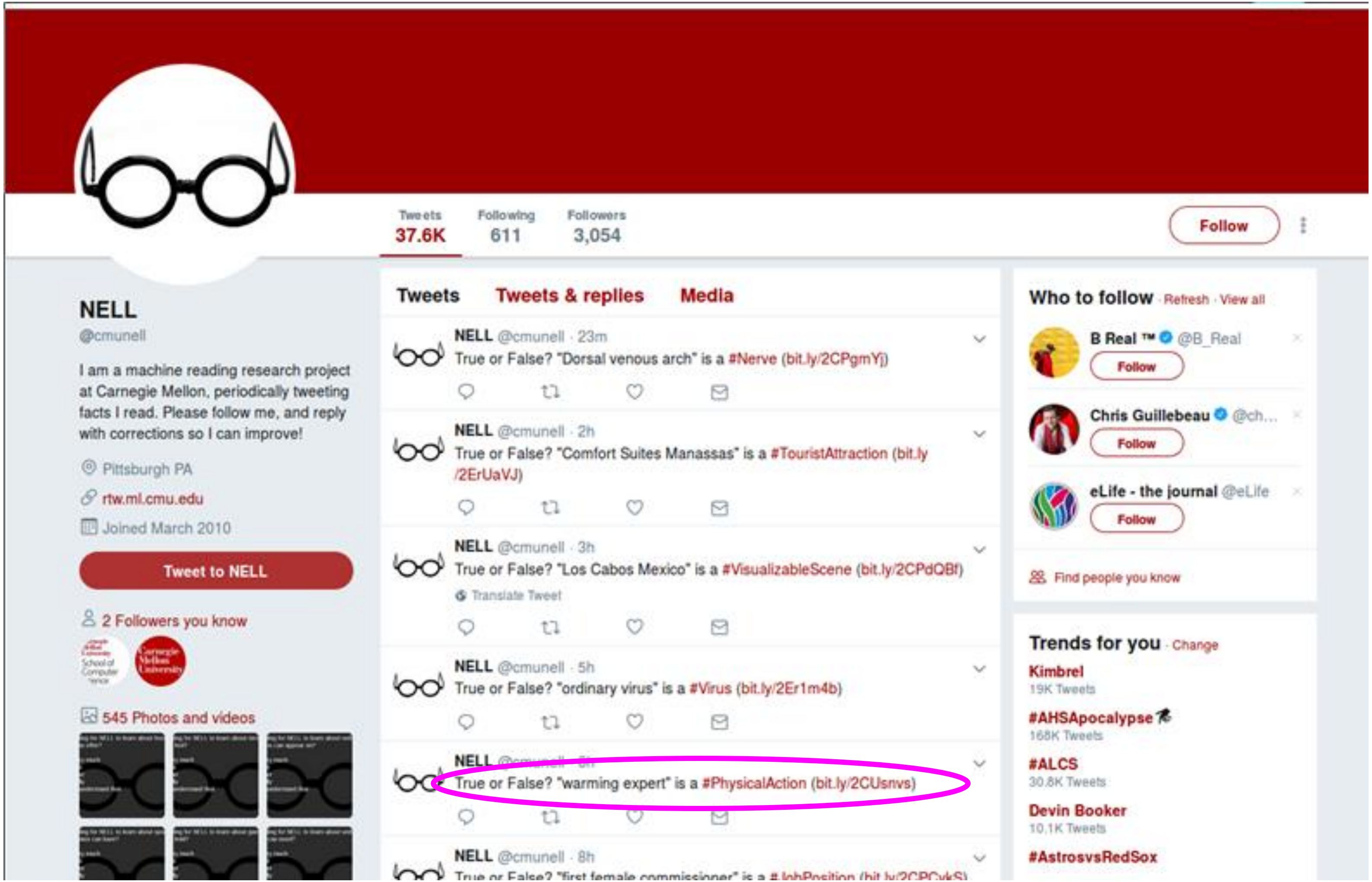


Average 2.4 negative feedback items per month per predicate.

Recently-Learned Facts

Refresh

Instance	Iteration	date learned	confidence
union_of_the_forces_for_progress is a political party	1111	06-jul-2018	100.0
tom_shannahan is an African person	1111	06-jul-2018	92.2
battle_of_heavenfield is a military conflict	1111	06-jul-2018	100.0
humberto_fuentes is a professor	1111	06-jul-2018	91.3
flamingo_gardens is an aquarium	1111	06-jul-2018	100.0
board_certified_anesthesiologist is a profession that is a kind of anesthetist	1114	25-aug-2018	93.8
steve001 is an athlete who injured his/her arm	1112	24-jul-2018	99.2
samsung is a company also known as sony	1114	25-aug-2018	93.8
the companies herald_tribune and new_york compete with eachother	1111	06-jul-2018	100.0
william_anderson died in the state or province va	1116	12-sep-2018	100.0


Tweets
37.6K
Following
611
Followers
3,054
Follow
NELL
@cmunell
I am a machine reading research project at Carnegie Mellon, periodically tweeting facts I read. Please follow me, and reply with corrections so I can improve!
Pittsburgh PA
rtw.ml.cmu.edu
Joined March 2010
Tweet to NELL
2 Followers you know
545 Photos and videos
Tweets
Tweets & replies
Media
NELL @cmunell · 23m
True or False? "Dorsal venous arch" is a #Nerve (bit.ly/2CPgmYj)
NELL @cmunell · 2h
True or False? "Comfort Suites Manassas" is a #TouristAttraction (bit.ly/2ErUaVJ)
NELL @cmunell · 3h
True or False? "Los Cabos Mexico" is a #VisualizableScene (bit.ly/2CPdQBf)
Translate Tweet
NELL @cmunell · 5h
True or False? "ordinary virus" is a #Virus (bit.ly/2Er1m4b)
NELL
True or False? "warming expert" is a #PhysicalAction (bit.ly/2CUsnvs)
NELL @cmunell · 8h
Who to follow · Refresh · View all
B Real ™ @B_Real
Follow
Chris Guillebeau @ch...
Follow
eLife - the journal @eLife
Follow
Find people you know
Trends for you · Change
Kimbrel
19K Tweets
#AHSApocalypse
168K Tweets
#ALCS
30.8K Tweets
Devin Booker
10.1K Tweets
#AstrosvsRedSox


NELL Knowledge Base Browser
CMU Read the Web Project
categories
relations
relatedto
numberofinjuredinearthquake
generalizationof
actorsuchasactor
astronautssuchasaustronauts
weaponssuchasweapons
criminalssuchascriminals
hobbiessuchashobbies
amphibiansuchasamphibian
aquariumssuchasaquariums
athletessuchasathletes
automobileenginesuchasautomobilee
videogamessuchasvideogames
professiontypehasprofession
musicgenressuchasmusicgenres
airportsuchasairport
televisionshowssuchastelevisionshow
animaltypehasanimal
animalsuchasinvertebrate
animalsuchasinsect
animalsuchasmollusk
animalsuchasfish
inverseofarthropodandotherarthrop
chemicaltypehaschemical
agriculturalproductincludingagricultur
warming_expert (physicalaction)
literal strings: warming expert
Help NELL Learn!
NELL wants to know if this belief is correct.
If it is or ever was, click thumbs-up. Otherwise, click thumbs-down.
warming_expert is a physical action
categories
physicalaction(93.1%)
CPL @1095 (73.3%) on 17-jan-2018 ["thanks to global _" "scientists , global _"] using warming_expert
CMC @1111 (74.3%) on 25-jun-2018 [SUFFIX=ing 1.71053 SUFFIX=ng 1.39252 PREFIX=warm 1.28563 PREFIX=warmi 1.26497 FIRS
SUFFIX=rming 0.70570 SUFFIX=rt -0.30740 FULL_POS=VBG_NN -0.98892 WORDS -3.58440] using warming_expert

Ask NELL on-demand results:

2 possible entities found

Click to change visible entity:

- love_triangle (astronaut)
- "love triangle"

---

love_triangle (astronaut)

literal strings: Love triangle, love triangle, love_triangle, love-triangle

categories

- **astronaut**(96.9%)
  - SEAL @529 (96.9%) on 11-mar-2012 why? using love_triangle

rtw.ml.cmu.edu/rtw//kbbrowser/askwhy?head=love_triangle+gen

love_triangle generalizations astronaut

SEAL @529 (96.9%) on 11-mar-2012 using love_triangle

http://www.enotes.com/topic/List_of_astronauts_by_year_of_selection
http://en.wikipedia.org/wiki/List_of_astronauts_by_year_of_selection
http://ms.wikipedia.org/wiki/Senarai_pemilihan_Angkasawan
http://www.thelivingmoon.com/47john_lear/01archives/List_of_space_known_astronauts.html
http://citizendia.org/List_of_astronauts_by_selection

close

International Mission Specialists: Pedro Duque (Spain), Christer Fuglesang (Sweden), Umberto Guidoni (Italy), Steven MacLean (Canada), Mamoru Mohri (Japan), Soichi Noguchi (Japan), Julie Payette (Canada), Philippe Perrin (France), Gerhard Thiele (Germany).

Brown, Clark and McCool were crewmembers on the final *Columbia* mission. Mark and Scott Kelly are twin brothers; James Kelly is not related. Loria resigned from his shuttle mission due to injury and never flew before retiring from the astronaut corps. Nowak, who flew on STS-21, was arrested on February 5, 2007, after confronting a woman entangled in a love triangle with a fellow astronaut. She was dismissed by NASA on March 6, the first astronaut to be both grounded and dismissed (prior astronauts who were grounded due to non-medical issues usually resigned or retired).