

Lexical Knowledge Structures

By

Ashutosh Kumar Nirala (10305906)
MTech-II, CSE

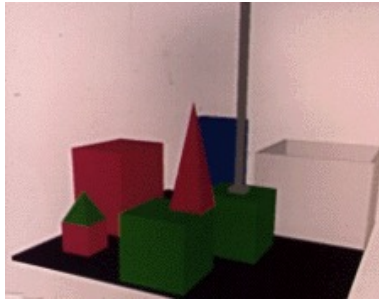
Guide - Dr. Pushpak Bhattacharyya

IIT Bombay

July 10, 2012

- Need of Lexical resources.
 - Making computers smarter.
 - From AI-NLP perspective.
 - Providing information.
- Lexical Knowledge Structures.

SHRDLU (1971)



```
READY  
How many blocks and pyramids are there.  
IDONT UNDERSTAND.  
READY  
What does the box contain?  
THE BLUE PYRAMID .  
  
READY
```

[Send] [Clear] [Exit]

SHRDLU, Demo by Terry Winograd at the MIT AI Lab (1971)

The dialog that was used as a SHRDLU demo (in 1971):¹

Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
Computer: FOUR OF THEM.
Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
Computer: YES, THE RED CUBE.
Person: IS IT SUPPORTED?
Computer: YES, BY THE TABLE.

¹taken from : <http://hci.stanford.edu/winograd/shrdlu/index.html>

SHRDLU : a success story.

- Considered a significant step forward in NLP, as it combines
 - models of human linguistic
 - reasoning methods in the language understanding process.
- But so far has not been extended further.
 - Works in simple, logical, and closed domain.
 - Can-not handle hypothesis.
 - Things are totally abstracted.

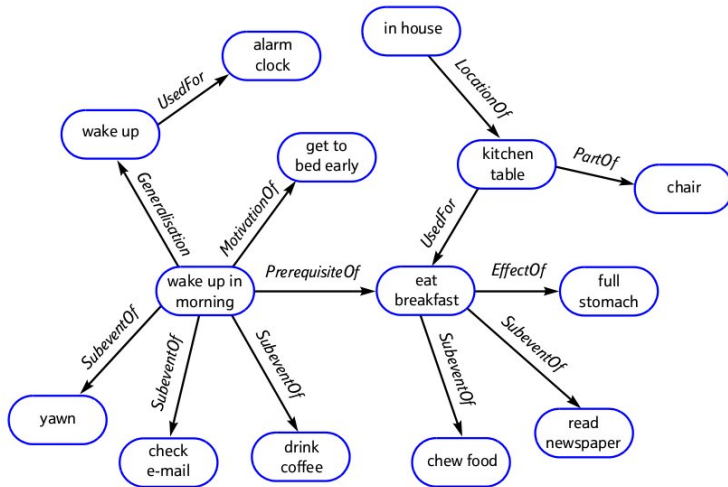
Lexical Knowledge Networks

- Cyc project, started 1984 by Doug Lenat
 - Goal is to capture all facts that the average person knows.
 - 350 man-years of effort estimated
- ConceptNet, started 1999, by MIT Media Lab
 - In 2000 become a World Wide Web collaborative project.
 - By 2004 had 300 000 concepts and 1.6 million relations.
- English WordNet, started 1985, by direction of George A. Miller
 - Lexical database that could be searched conceptually.
- YAGO ontologies 2007
 - Combines WordNet and Wikipedia.
 - Made by crawling Wikipedia in January 2007.
- VerbOcean
 - Contains relations between verbs.
 - Relations captured semi-automatically.

ConceptNet

- A common sense knowledge base from MIT Media Lab.
- Aims to capture facts,
-which enables humans in day to day activity.
 - by capturing relations between **concepts**
- Started in 1999,
- Contributed by 1000s of people.
 - via OMCS web interface. (*Till ConceptNet 4.0rc4*)
 - in ConceptNet 5, *English Wikipedia, WordNet and many other.*

Typical relations in concept net



Relations in ConceptNet, K-Lines

- There are 20 different relations (as in ConceptNet2.1)

K lines² (1.25 million assertion)

ConceptuallyRelatedTo	'bad breath''mint''f=4;i=0;'
ThematicKLine	'wedding dress''veil''f=9;i=0;'
SuperThematicKLine	'western civilisation''civilisation' 'f=0;i=12;'

²[5] : Marvin Minsky : A Theory of Memory

Relations in ConceptNet Agents, Things

AGENTS (104 000 assertions)

CapableOf	'dentist' 'pull tooth' 'f=4;i=0;'
-----------	-----------------------------------

THINGS (52 000 assertions)

IsA (Hyponym)	'horse' 'mammal' 'f=17;i=3;'
---------------	------------------------------

PartOf (Meronym)	'butterfly' 'wing' 'f=5;i=1;'
------------------	-------------------------------

DefinedAs (Gloss)	'meat' 'flesh of animal' 'f=2;i=1;'
-------------------	-------------------------------------

MadeOf	'bacon' 'pig' 'f=3;i=0;'
--------	--------------------------

PropertyOf	'fire' 'dangerous' 'f=17;i=1;'
------------	--------------------------------

Relations in ConceptNet Events, Spatial, Causal

EVENTS (38 000 assertions)

PrerequisiteEventOf	'eat breakfast''wake up in morning' 'f=2;i=0;'
FirstSubeventOf	'start fire''light match''f=2;i=3;'
SubeventOf	'eat breakfast''chew food''f=2;i=0;'
LastSubeventOf	'attend classical concert''applaud''f=2;i=1;'

SPATIAL (36 000 assertions)

LocationOf	'army''in war''f=3;i=0;'
------------	--------------------------

CAUSAL (17 000 assertions)

EffectOf	'view video''entertainment''f=2;i=0;'
DesirousEffectOf	'sweat''take shower''f=3;i=1;'

FUNCTIONAL (115 000 assertions)

UsedFor	'alarm clock''wake up''f=1;i =2;'
CapableOfReceivingAction	'drink''serve''f =0;i =14;'

AFFECTIVE (34 000 assertions)

MotivationOf	'go to bed early''wake up in morning' 'f =3;i=0;'
DesireOf	'person''not be depressed' 'f=2;i=0;'

Development Process of ConceptNet

Development Process of ConceptNet via OMCS

- Knowledge acquisition from the general public[7].
- Extraction & Normalisation phase.
- Relaxation phase.

- People not having special training in NLP or AI.
 - CycL like Cyc can not be used
- So a context is given, like:-
Bob had a cold. Bob went to a doctor
knowledge helpful to understand it was collected.
 - *Bob was feeling sick.*
 - *The doctor made Bob feel better.*
 - *The doctor might have worn a white coat.*

Relations are extracted by matching patterns like

- *[a | an | the] N1 (is | are) [a | an | the] [A1] N2*

→ *Dogs are mammals*

→ *Hurricanes are powerful storms*

gives

*Dog **IsA** mammal*

*Hurricane **IsA** powerful storm*

- *N1 requires [a | an] [A1] N2*

→ *Writing requires a pen*

→ *Bathing requires water*

gives:-

*pen **UsedFor** writing*

*Water **UsedFor** bathing*

- Duplicate assertions are merged and count is maintained.
- *IsA* relation is used to lift the concepts

(IsA 'apple' 'fruit')

(IsA 'banana' 'fruit')

(IsA 'peach' 'fruit')

AND

(PropertyOf 'apple' 'sweet')

(PropertyOf 'banana' 'sweet')

(PropertyOf 'peach' 'sweet')

IMPLIES

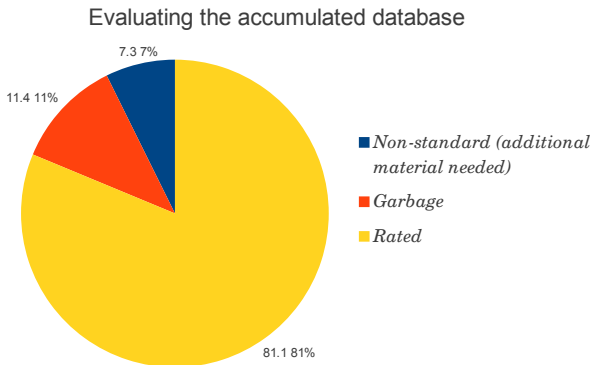
(PropertyOf 'fruit' 'sweet')

Relaxation phase (contd.)

- SuperThematicKLine relations capturing generalization are produced.
 - WordNet and FrameNet's verb synonym sets and class-hierarchies are used.
 - (*SuperThematicKLine* 'buy food' 'buy')
 - (*SuperThematicKLine* 'purchase food' 'buy')
- If noun phrase have adjectival modifier and is repeated then *PropertyOf* relation is inferred.
 - [(*IsA* 'apple' 'red round object');
(*IsA* 'apple' 'red fruit');]
It implies (*PropertyOf* 'apple' 'red');

Evaluation of accumulated data

1% of the OMCS-1 corpus was manually evaluated.



Evaluation of accumulated data

- 8 judges rated items on 4 attributes
- Scored on 1 to 5
where score 5 means total agreement with the attribute.
 - **Generality** : is item too specific?
 - score 5 : *Dew is wet*
 - score 1 : *Eritrea is part of Africa*
 - **Truth**
 - Score 1 : *Someone can be at infinity*
 - **Neutrality** : is it personal opinion?
 - Score 1 : *Idiots are obsessed with star trek.*
 - **Sense** : does the item makes sens?
 - Score 1 : *Cows can low quietly!!*

Manual Rating

- Rating, with increasing relevance [7].

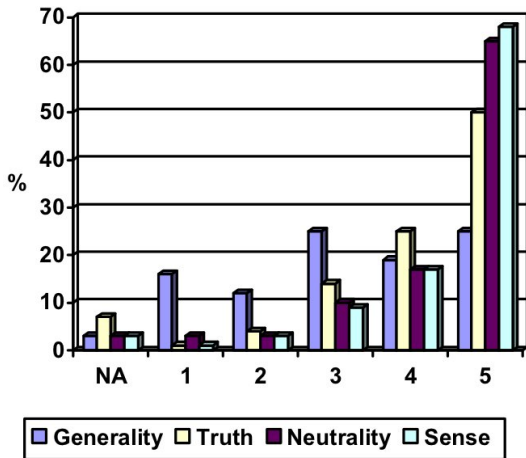
Avg Score

Generality : 3.26

Truth : 4.28

Neutrality : 4.42

Sense : 4.55



- Following observations were made
 - Templates are efficient.
 - Participants want to enter what is in their mind.
 - Participants wished interaction, access and modification to data.

Workflow model for acquisition

- User browse database
 - Finds item, assoc with a template, of interest.
- On click on template a form is presented to user.
 - Examples are also shown
 - User fills the form and submit.
- System display the **inferred relations**.
- User can accept or reject them.

- A sample web interface ³.

The screenshot displays the Open Mind Common Sense web interface. At the top, it says "Open Mind Common Sense Explain your world." and is logged in as "AstroNauts". Below the header is a navigation bar with links for "Home", "Add new knowledge", "Highest rated", "My contributions", and "Ad-hoc categories".

Example statements

- An activity a dog can do is bark
- humans can die only once
- Fish can not live out of the water
- A living being can die
- Cars can go fast

Teach OpenMind another statement of this type.

CapableOf: What can it do?

Google	can	search the web	Teach OpenMind
Pictures	can	display events	Teach OpenMind
Books	can	contain stories	Teach OpenMind
An activity	can do is		Teach OpenMind
children	sometimes	cry	Teach OpenMind

Places to start

Concepts

alberta, stay fit, a pound, a mountain, use a telephone, Cambridge, an alcoholic, Candy bars, gain more land, downtown

Vote on these statements...

- Chess is a board game
- You are likely to find a cake in a bakery
- a highway is used for travel
- Beer is an alcoholic beverage
- religion is the opiate of the masses
- water is essential to all life
- cats are cute
- You are likely to find a creek in a forest
- Cats are a curious animal
- a butterfly is an insect

Feedback

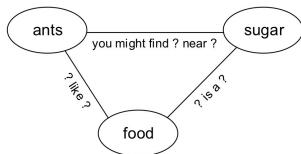
(Send it in!)

³Picture taken from A kid's Open Mind Common Sense [6]

- Method 1 : Analogies over concept
 - Slots filled in the **template** are searched for other templates.
 - *A mother can have a baby* gives
 - *A mother can hold her baby*
 - Then other relations matching this newly found template are searched
 - *A small girl can hold her small dog*
 - For each match, **slots values** are replaced with the found one.
 - *A small girl can have a small dog*
 - If user finds it correct he may confirm this.

- Method 2 : Analogies over Relations
 - **Template** are searched for other concepts.
 - *A mother can have a baby* gives
 - *A child can have a goldfish*
 - Then for new **slots values** other **Template** are searched.
 - *A child can take care of goldfish*
 - For each match, **slots values** are replaced with entered one.
 - *A mother can take care of a child*
 - If user finds it correct he may confirm this.

- Method 3 : Analogies as Inference Rules
 - It first generates a list of inference rules.
 - For this programs first tries to find a cycle.

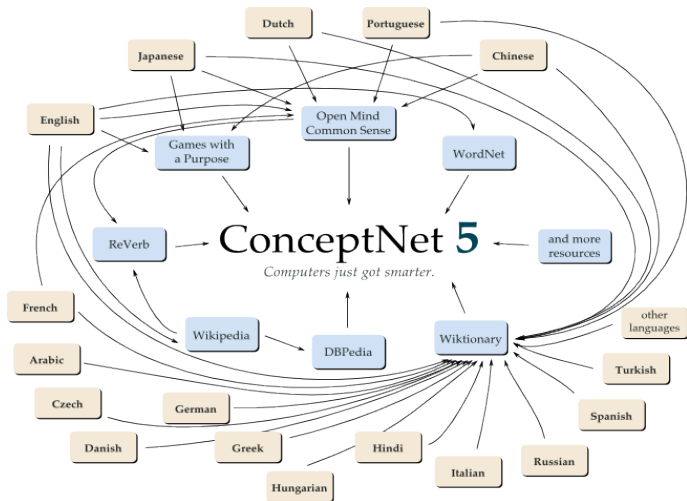


- Rules are automatically extracted using OMCS-1 database.
- More matches \Rightarrow better rules.
- If two links for rules are discovered program can infer third
 - User enters : *Bats like darkness*
 - If db has : *You might find bats near cave interiors*
 - and the corresponding rule, then it will infer
 - *Cave interior is a darkness*

- Clarification by suggesting common words as replacement.
 - common words extracted as frequency from OMCS-1 corpus.
 - Replacement using synonym dictionaries.
- Users are prompted for WSD.
 - Automated methods suggest sense tags.
 - User only need to provide one or two senses.
- Concepts are linked to topic.
 - Linking maintained as topic vectors.
 - Facilitates wide knowledge retrieval.

ConceptNet5

ConceptNet5 contains concepts from a no of sources.⁴



⁴taken from : <http://conceptnet5.media.mit.edu/>

- ConceptNet5 released on 2011 October 28
- ConceptNet5.1 released on 2012 April 30
- Multiple sources.
- Concepts in other languages.

本 — *MadeOf* → 紙

本は紙でできている。(A book is made of paper.)

ईसाई धर्म — *TranslationOf* → christianity

ईसाई धर्म is Hindi for *Christianity*

- Available as full download and
Core download without relations from other resources.

Graphical structure of ConceptNet5

- Available in multiple formats.
- Hypergraph, edges about relations.
 - justified by other assertions, knowledge sources or processes.
 - each justification have positive or negative weight.
 - Negative means not true.
- Relations could be interlingual
or automatically extracted relations, specific to a language.

- Uniform Resource Identifier.

eg : *<http://conceptnet5.media.mit.edu/web/c/en/gandhi>*

- every object has URI.
- standard place to look it up.
- meaningful
- for edges it is hash - for uniqueness.

URI hierarchy (contd)

Different kinds distinguished from first element.

- [/a/](#) assertions.
- [/c/](#) concepts (words, phrases from a language).
- [/ctx/](#) context in which assertion is true.
- [/d/](#) datasets.
- [/e/](#) unique id for edges.
- [/l/](#) license for redistributing information in an edge.
 - [/l/CC/By](#) Creative Commons.
 - [/l/CC/By-SA](#) Attribution-ShareAlike.
- [/r/](#) language independent relation like [/r/IsA](#)
- [/s/](#) knowledge sources
 - human contributors, Web sites or automated processes.

- Each concept has minimum three components

/c/hi/बेटा/

- */c/* to indicate it is a concept.
- language part, ISO abbreviated.
- concept text.

- Optional fourth component for POS

/c/en/read/v

- Optional fifth component for a particular sense.

/c/en/read/v/interpret_something_that_is_written_or_printed

Fields in ConceptNet5.1

```
{
  "endLemmas": "fruit",
  "rel": "/r/IsA",
  "end": "/c/en/fruit",
  "features": [
    "/c/en/apple /r/IsA -",
    "/c/en/apple - /c/en/fruit",
    "- /r/IsA /c/en/fruit"
  ],
  "license": "/l/CC/By",
  "sources": [
    "/s/rule/sum_edges"
  ],
  "startLemmas": "apple",
  "text": [
    "fruit",
    "apple"
  ],
  "uri": "/a[/r/IsA/,/c/en/apple/,/c/en/fruit/]",
  "weight": 244.66679999999999,
  "dataset": "/d/conceptnet/5/combined-core",
  "start": "/c/en/apple",
  "score": 1049.3064999999999,
  "context": "/ctx/all",
  "timestamp": "2012-05-25T03:41:00.346Z",
  "nodes": [
    "/c/en/fruit",
    "/c/en/apple",
    "/r/IsA"
  ],
  "id": "/e/3221407ec935683f2b7079b0495f164e1e321cd4"
}
```

- **Lookup** : When URI is known.

Example

<http://conceptnet5.media.mit.edu/data/5.1/c/en/apple>

- **Search** : when URI is not known

- Performed with base URL + criteria (in GET)

BASE URL :

<http://conceptnet5.media.mit.edu/data/5.1/search>

WITH criteria :

<http://conceptnet5.media.mit.edu/data/5.1/search?text=apple>

- **Association** : for finding similar concepts.

Arguments for Search

Passed as GET parameter

- **{id, uri, rel, start, end, context, dataset, license}** : matches start of the field.
- **nodes** : if start of any node matches.
- **text, {startLemmas, endLemmas, relLemmas}** : matches anywhere.
- **surfaceText** matches surface text but is case sensitive
- **minWeight, limit, offset**
- **features** : needs exact match.
- **filter** :
 - **core** : no ShareAlike resources included
 - **core-assertions** : one result per assertion

- BASE URL :
<http://conceptnet5.media.mit.edu/data/5.1/assoc>
- SOURCE CONCEPT : *[/list/<language><term list>](#)*
 - multiple terms are ', ' separated.
 - @ specifies a weight (relative to other elements)
- GET PARAMETERS
 - *[limit=n](#)*
 - *[filter=URI](#)*

<http://conceptnet5.media.mit.edu/data/5.1/assoc/list/en/cat,food@0.5?limit=1&filter=/c/en/dog>

Applications Developed using ConceptNet

Goal-Oriented Search Engine With Commonsense ⁵

Address

© PEN  MIND
commonsense

"I want "

<<< Use Commonsense Concepts? <<< Use Commonsense Investigator?
 <<< I typed in Natural Language. <<< Use Commonsense Generalizations?
 <<< Print context vector.

Alyssa searched the web for "help me get rid of the mice in my kitchen" --> Distilled into "rid mice kitchen"

Alyssa investigated your query - **help me get rid of the mice in my kitchen** - and concluded that the solution is to look for: **"pest control" in Cambridge, MA**

[Cambridge Pest Control Services](#)
 Description: 43% - Articles & General info: The Smart Cambridge Yellow Pages(cambridge... Cambridge Pest Control Services. Best Pest Control Services, Inc. 63 Elm St. Somerville, MA 02144. 617-625... Date Not Available)
http://cambridge.zami.com/Pest_Control_Services

[PEST CONTROL - PEST MANAGEMENT](#)
 Description: 42% - Directories & Lists: Northeast Document Conservation Center 100 Brickstone Square Andover, MA 01810-1494 Tel:(978) 470-1010 Fax:(978) 475-6021. TECHNICAL LEAFLET. PRESERVATION SUPPLIERS AND

⁵taken from : <http://agents.media.mit.edu/projects/goose/>

- Parses the query into semantic frame.
- Classify into common sense sub-domain.
- Reformulation
 - Apply reasoning using inference chain.
 - Heuristically guided.
 - Termination on application-level rule.
 - extract the reformulated search term.
 - Search on commercial search engine.
- Re-ranking
 - Based on weighted concepts.

GOOSE : a scenario [3]

- Goal : *I want help solving this problem*
and query, *my golden retriever has a cough*
- Parsing gives

Problem Attribute	[cough]
Problem Object	[golden retriever]
- commonsense sub-domain classified : **animals** with the chain
 - *A golden retriever is a kind of dog.*
 - *A dog may be a kind of pet.*
 - *Something that coughs indicates it is sick.*
 - *Veterinarians can solve problems with pets that are sick.*
 - *Veterinarians are locally located.*
- The reformulated search is
Veterinarians, Cambridge MA
Location obtained from user profile. Page containing concepts closer to veterinarians is ranked high

GOOSE Results [3]

Search Task	no of successful inferences	Avg. score GOOSE	Avg. score Google
Solve household problem	7/8	6.1	3.5
Find someone online	4/8	4.0	3.6
Research a product	1/8	5.9	6.1
Learn more about	5/8	5.3	5.0

Other applications [4]

- Commonsense ARIA
 - Suggests photos while writing email or Web pages.
 - Uses manually marked tags.
 - Add tags when photo is used.
 - Use common sense for better search [7]
 - Given : *Susan is Jane's sister*
 - Commonsense : *in a wedding, the bridesmaid is often the sister of the bride*
 - Jain's photo can be retrieved if tag is *Susan and her bridesmaids*
- MAKEBELIEVE : interactively invents a story.
 - Uses causal projection chains to create storyline.
- GloBuddy : dynamic foreign language phrasebook.
 - Translates related concepts.
 - eg : *I am at a restaurant* generates *people, waiter, chair, eat* with translations.
- Suggesting words in mobile text-messages by inferring context

YAGO : A Large Ontology from Wikipedia and WordNet⁶

⁶[9] : Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum

Google searches web pages.

The screenshot shows a Google search interface. At the top, there are navigation links: +You, Search, Images, Maps, Play, YouTube, News, Gmail, Documents, Calendar, and More. The Google logo is on the left, and a search bar contains the text "which politician was born in same year as Obama". A blue search button with a magnifying glass icon is on the right. Below the search bar, the word "Search" is displayed in red, followed by the text "About 411,000,000 results (0.35 seconds)".

On the left side, there is a vertical navigation menu with the following items: Web, Images, Maps, Videos, News, Shopping, and More. Below this menu is a link for "Show search tools".

The main content area displays search results. The first result is under the "Web" category and is a link to the Wikipedia page for "Barack Obama, Sr. - Wikipedia, the free encyclopedia". The snippet for this result reads: "Obama Sr. was **born** in Rachuonyo District on the shores of Lake Victoria just outside Kendu That **same year**, **Obama** Sr. published a paper entitled "Problems Facing Our 1A. <http://www.boston.com/news/politics/2008/articles/2008/09/21/> ...".

The second result is from "snopes.com" and is titled "Barack Obama Birth Certificate". The snippet reads: "17 Mar 2012 – At the time of **Obama's** birth, it also shows that his father is aged 25 **years** old, and that **Obama's** father was **born** in "Kenya, East Africa".

The third result is from "www.washingtonpost.com" and is titled "For Obama, gay marriage stance borne of a long evolution - The ...". The snippet reads: "10 May 2012 – It's been something **Obama** has struggled with since 1996. ... to demilitarize the language of **politics** after Gabby Giffords was shot last **year**?"

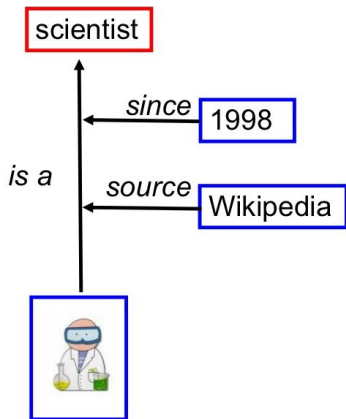
- Combines high coverage with high quality.
 - Uses infoboxes and category of Wikipedia.
 - Overall precision of 95%
 - decidable.
- YAGO model uses extension to RDFS.
- Expresses entities, facts, relation between facts and properties of relation.

YAGO data model, few examples

- *Elvis won a Grammy Award*
 - (Elvis Presley, HASWONPRIZE, Grammy Award)
- words are entities as well.
 - Quotes to distinguish from other entities.
 - (“Elvis”, MEANS, Elvis Presley)
 - Allows to deal with synonyms and ambiguity
 - (“Elvis”, MEANS, Elvis Costello)
- Similar entities are grouped into classes.
 - (Elvis Presley, TYPE, singer)
- Classes & relations are entities as well.
 - (singer, SUBCLASSOF, person)
 - (subclassOf, TYPE, atr)

n-ary relations

- Expressing multiple relations⁷
 - Every edge is given an edge identifier.



- #1 (Sam, IS_A, scientist)
- #2 (#1, SINCE, 1998)
- #3 (#1, SOURCE, Wikipedia)

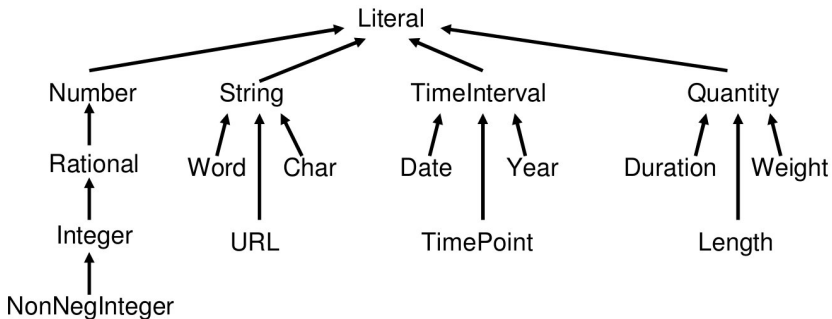
⁷picture taken from presentation by Fabian M. Suchanek

- **common entities** : which are neither facts nor relations.
E.g.# : singer, person, Elvis Presley
- **individuals** : common entities which are not classes. E.g.# :
Elvis Presley
- Its a **reification graph.** defined over
 - set of common entities nodes C ,
 - set of edge identifiers I
 - set of relation names R
 - reification graph is an injective total function
 $G_{C,I,R} : I \rightarrow (C \cup I) \times R \times (C \cup I)$

- Any YAGO ontologies must have following relations (R)
 - type : (Elvis Presley, TYPE, singer)
 - subClassOf : (singer, SUBCLASSOF, person)
 - domain : (subClassOf, DOMAIN, class)
 - range : (subRelationOf, RANGE, relation)
 - subRelationOf : (fatherOf, SUBRELATIONOF, parentOf)
- Common entities (C) must contain the classes
 - entity
 - class
 - relation
 - atr : acyclic transitive relation

Classes for all literals⁸.

- Classes for all literals⁸.



⁸Graph from [10] : YAGO report 2007

Semantics : Rewrite rule

$$\{f_1, \dots, f_n\} \hookrightarrow f$$

i.e., given facts f_1 to f_n , fact f is inferred.

- $\Phi \hookrightarrow (\text{domain}, \text{RANGE}, \text{class})$
 $\Phi \hookrightarrow (\text{domain}, \text{DOMAIN}, \text{relation})$
 - i.e., **range** for **domain** (which is a *relation*) will be a class.
 - But, “**domain**” **relation** can only be applied to a **relation**.
 - So, any relation’s domain will always be *some* class.
 - E.g.# (`isCitizenOf`, `domain`, `person`)
- $\Phi \hookrightarrow (\text{range}, \text{RANGE}, \text{class})$
 $\Phi \hookrightarrow (\text{range}, \text{DOMAIN}, \text{relation})$
 - E.g.# (`isCitizenOf`, `range`, `country`)

Semantics : Rewrite rule (contd.)

- $\Phi \leftrightarrow (\text{subClassOf}, \text{DOMAIN}, \text{class})$
- $\Phi \leftrightarrow (\text{subClassOf}, \text{RANGE}, \text{class})$
- $\Phi \leftrightarrow (\text{subClassOf}, \text{TYPE}, \text{atr})$
 - E.g1. #
(NonNegInteger, SUBCLASSOF, Integer) &
(Integer, SUBCLASSOF, Number)
So : (NonNegInteger, SUBCLASSOF, Number)
 - E.g2. #
(wordnet_carnival_100511555, SUBCLASSOF,
wordnet_festival_100517728) &
(wordnet_festival_100517728, SUBCLASSOF,
wordnet_celebration_100428000)
So : (wordnet_carnival_100511555, SUBCLASSOF,
wordnet_celebration_100428000)

Semantics : Rewrite rule (contd.)

- $\Phi \leftrightarrow (\text{type}, \text{RANGE}, \text{class})$
- $\Phi \leftrightarrow (\text{subRelationOf}, \text{DOMAIN}, \text{relation})$
- $\Phi \leftrightarrow (\text{subRelationOf}, \text{RANGE}, \text{relation})$
- $\Phi \leftrightarrow (\text{subRelationOf}, \text{TYPE}, \text{atr})$
- E.g. #
(happenedOnDate, SUBRELATIONOF, startedOnDate) &
(startedOnDate, SUBRELATIONOF, startsExistingOnDate)
So :
(happenedOnDate, SUBRELATIONOF, startsExistingOnDate)
- For literal class for each edge $X \rightarrow Y$
 $\Phi \leftrightarrow (X, \text{subClassOf}, Y)$

Semantics : Rewrite rule (contd)

Given

- $r, r_1, r_2 \in R$, where
 - $r, r_1 \neq \text{type}$, and
 - $r, r_2 \neq \text{subRelationOf}$
- $x, y, c, c_1, c_2 \in I \cup C \cup R$, where
 - $c, c_2 \neq \text{atr}$

Then,

- $\{(r_1, \text{subRelationOf}, r_2), (x, r_1, y)\} \leftrightarrow (x, r_2, y)$
 - E.g.# : $\{(\text{motherOf}, \text{SUBRELATIONOF}, \text{parentOf}), (\text{Kunti}, \text{MOTHEROF}, \text{Arjun})\} \leftrightarrow (\text{Kunti}, \text{PARENTOF}, \text{Arjun})$
- $\{(r, \text{type}, \text{atr}), (x, r, y), (y, r, z)\} \leftrightarrow (x, r, z)$
 - E.g1. #
 $\{(\text{NonNegInteger}, \text{SUBCLASSOF}, \text{Integer}), (\text{Integer}, \text{SUBCLASSOF}, \text{Number})\} \leftrightarrow$
So : $(\text{NonNegInteger}, \text{SUBCLASSOF}, \text{Number})$

Semantics : Rewrite rule (contd)

- $\{(r, \text{domain}, c), (x, r, y)\} \leftrightarrow (x, \text{type}, c)$
 - E.g.#
 $\{(\text{Sonia_Gandhi}, \text{ISCITIZENOF}, \text{India}),$
 $(\text{isCitizenOf}, \text{DOMAIN}, \text{person})\} \leftrightarrow$
 $(\text{Sonia_Gandhi}, \text{TYPE}, \text{person})$
- $\{(r, \text{range}, c), (x, r, y)\} \leftrightarrow (y, \text{type}, c)$
 - E.g.#
 $\{(\text{Sonia_Gandhi}, \text{ISCITIZENOF}, \text{India}),$
 $(\text{isCitizenOf}, \text{RANGE}, \text{country})\} \leftrightarrow$
 $(\text{India}, \text{TYPE}, \text{country})$
- $\{(x, \text{type}, c_1), (c_1, \text{subClassOf}, c_2)\} \leftrightarrow (x, \text{type}, c_2)$
 - E.g.#
 $\{(\text{Elvis Presley}, \text{TYPE}, \text{singer}),$
 $(\text{singer}, \text{SUBCLASSOF}, \text{person})\} \leftrightarrow$
 $(\text{Elvis Presley}, \text{TYPE}, \text{person})$

Given $\mathbf{F} = (I \cup C \cup R) \times R \times (I \cup C \cup R)$

- **Theorem 1**

[Convergence of \longrightarrow]

Given a set of facts $F \subset \mathbf{F}$, the largest set S with $F \longrightarrow S$ is finite and unique.

- **Corollary 1**

[Decidability]

The consistency of a YAGO ontology is decidable.

- **Theorem 2**

[Uniqueness of the Canonical Base]

The canonical base of a consistent YAGO ontology is unique.

- Can be computed by greedily removing derivable facts.

- Can't state : f is FALSE
- Primary relation of n-ary relation is always true.
 - E.g *Elvis was a singer from 1950 to 1977*
 - #1 : (Elvis, TYPE, singer)
 - #2 : (#1, DURING, 1950-1977)
- Intentional predicates (like BELIEVES^{THAT}) NOT POSSIBLE

Sources and Information Extraction

- WordNet
 - Uses hypernyms/hyponyms relation
 - Conceptually it is DAG in WordNet
- Wikipedia
 - XML dump of Wikipedia
 - categories.
 - infobox.
 - 2,000,000 articles in english wikipedia (Nov 2007) YAGO.
 - 3,867,050 articles in english wikipedia (Feb. 2012) YAGO2.
- YAGO2⁹: geo-location information from Geonames¹⁰

⁹YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages

¹⁰from <http://www.geonames.org/>

Two steps (YAGO 1)

- Extraction from Wikipedia
- Quality Control.

Extraction from Wikipedia


- Page title is a candidate for **individual**.
- **Infoboxes**

- Each row has attribute value.
- manual rules designed for 170 (200 for YAGO2) frequent attributes

E.g:

relation : BIRTHDATE
domain : person
range : timeInterval

Albert Einstein



Albert Einstein in 1921

Born	14 March 1879 Ulm, Kingdom of Württemberg, German Empire
Died	18 April 1955 (aged 76) Princeton, New Jersey, United States

Navigation icons: back, forward, search, etc.

- Infobox type establishes the article entity class.
E.g. # city infobox or person infobox.
 - however, for *Economy of a country*, type is country.
- Each row can generate fact. (Arg_1 , RELATION, Arg_2)

Usually

- Arg_1 is article entity.
 - RELATION determined by attribute.
 - Arg_2 value of the attribute.
- *Inverse attribute* : entity becomes Arg_2

E.g. #

- if attribute is *official namee*
(entity HASOFFICIALNAME officialname) is not generated
(officialname MEANS entity) is generated instead

- Infobox type *may* disambiguate **meaning** of attribute
E.g.#
 - length of car is in space
 - length of song is in duration
- Value is parsed¹¹ as an instance of the range of target relation.
 - Regular expression is used to parse *numbers, dates* and *quantities*
 - Units of measurement normalized to ISO units.
- If range is not a literal class
 - Wikipedia link is searched for entity.
 - If search fails corresponding attribute is ignored.

¹¹[8] LEILA, A link type parser is used

- Category system of Wikipedia is exploited
- Broadly categories could be
 - conceptual categories, like
Naturalized citizens of a country
 - category for administrative purposes, like
Articles with unsourced statements
 - categories giving relational information like
1879 births
 - categories indicating thematic vicinity like
Physics
- Only conceptual category can be class for individual.

Identifying Conceptual Category

- Administrative and relation categories are very low.
 - less than a dozen
 - manually excluded
- Shallow linguistic parsing splits category name
Naturalized citizens of Japan is split as
 - pre-modifier Naturalized
 - head citizens
 - post-modifier of Japan
- Plural head usually means *conceptual category*

Defining hierarchy of classes using WordNet

- Wikipedia categories are organized as DAG
 - reflects only thematic structure of Wikipedia
Elvis is in the category Grammy Awards
 - So WordNet is used to define hierarchy over leaf category of Wikipedia.
- Each synset of WordNet becomes a class.
 - Proper nouns are removed.
 - Identified If
WordNet synset has a common noun with Wikipedia page.
 - Some information is lost only common nouns become class.
- subClassOf relation taken from hyponyms relation of WordNet
 - A is subClassOf of B in YAGO, if
synset A is hyponyms of synset B in WordNet

Defining hierarchy of classes using WordNet

- Lower classes of Wikipedia are connected to higher class of WordNet
 - E.g. # **American people in Japan** is a *subclass of* **person**
 - First category name is split in *pre*, *head* and *post*.
pre American head people post in Japan
 - head is stemmed to its singular form
people → person
 - If pre + head is in WordNet, desired class is achieved
American person
 - *else*, only head compound is searched
 - The match with highest frequency synset is used.
 - Exception like **capital** whose predominant sense in WordNet (**financial asset**) and Wikipedia (**capital city**) differed were manually corrected

- A **means** relation is established *between* each word of **WordNet synset**
E.g.# (metropolis, means, city)
- Wikipedia redirects are used to give **means** relation
E.g.# (Einstein, Albert, means, Albert Einstein)
- givenNameOf and familyNameOf relations are used using person names
E.g.# (Albert, givenNameOf, Albert Einstein)
E.g.# (Einstein, familyNameOf, Albert Einstein)

Category heuristics

- Relational category pages gives info about article
 - E.g. # category **Rivers in Germany** ensures article entity has **locatedIn** relation with Germany.
 - Regular expressions heuristics are used to get category names like **Mountains | Rivers in (.*)**
- Exploiting Language Category
 - Categories like *fr:Londers*, and articles in them like *the city of London* gives relation **London isCalled “Londres” inLanguage French**

- **Canonicalization**

- **Redirect Resolution:**

- facts are obtained from infobox.
 - Some links might be to the Wikipedia redirect pages.
 - Such incorrect arguments are corrected.

- Duplicate facts are removed.

- more precise facts are kept

E.g.# out of birthDate 1935-01-08 and 1935 only 1935-01-08 is kept.

- **Type Checking**

- **Reductive** : facts are dropped if

- class for an entity can not be detected.
 - first argument is not in the domain of the relation.

- **Inductive** : class for an entity is inferred

- Works well with person - E.g.# if entity has birthDate then person is inferred.

- Meta relations are stored like normal relation.
 - URL for each individual is stored with **describes**
 - **foundIn** relation are stored as *witness*.
 - **using** relation stores technique of extraction.
 - **during** relation stores the time of extraction.
- **File format** : model is independent of storage.
 - simple text files are used as internal format
 - Estimated accuracy between 1 and 0 is stored as well.
 - XML version of text file and RDFS version are available.
 - database schema is simply
FACTS(faactId, arg1, relation, arg2, accuracy)
 - Software to load in Oracle, Postgres or MySQL is provided.

Evaluating YAGO

- Randomly selected facts were presented to judges along with Wiki pages.
- pages were rated *correct*, *incorrect* or *don't know*
- Only facts that stem from heuristics were evaluated
 - Portion stems from WordNet is not evaluated.
 - Non-heuristics relations like **describes**, **foundIn** are not evaluated.
- 13 judges evaluated 5200 facts.

Precision of heuristics

	Heuristic	#Eval	Precision
1	hasExpenses	46	100.0% \pm 0.0%
2	hasInflation	25	100.0% \pm 0.0%
3	hasLaborForce	43	97.67441% \pm 0.0%
4	during	232	97.48950% \pm 1.838%
5	ConceptualCategory	59	96.94342% \pm 3.056%
6	participatedIn	59	96.94342% \pm 3.056%
7	plays	59	96.94342% \pm 3.056%
8	establishedInYear	57	96.84294% \pm 3.157%
9	createdOn	57	96.84294% \pm 3.157%
10	originatesFrom	57	96.84294% \pm 3.157%
	...		
72	WordNetLinker	56	95.11911% \pm 4.564%
	...		
74	InfoboxType	76	95.08927% \pm 4.186%
75	hasSuccessor	53	94.86150% \pm 4.804%
	...		
88	hasGDPPPP	75	91.22189% \pm 5.897%
89	hasGini	62	91.00750% \pm 6.455%
90	discovered	84	90.98286 \pm 5.702%

- Rules are interpreted - no longer hard coded.
- Becomes Addition YAGO2 facts.
- **Factual rules**
 - Declarative translations of
 - all the manually defined *exceptions* and *facts* (total 60) in the code of YAGO1
 - “capital” **hasPreferredMeaning** [wordnet_capital_108518505](#)
 - Literal types come with regular expression to match them.

- **Implication rules** stored as
 - “\$1 \$2 \$3; \$2 subpropertyOf \$4;” **implies** “\$1 \$4 \$3”
- **Replacement rules** for cleaning HTML tags, normalizing units etc
 - “\{\{USA\}\}” **replace** “[[United States]]”
- **Extraction rules** stores regular expression rules¹². for deriving fact.

¹²the *regex* is as defined for : regular expression syntax of `java.util.regex`

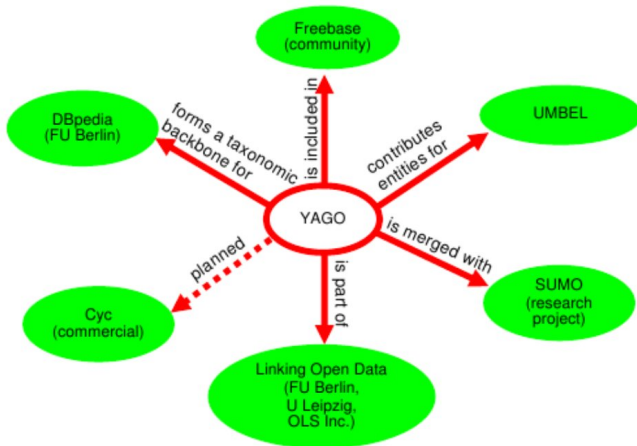
- **Temporal Dimension:** Assign begin and/or end of time spans to all entries, facts, events, etc.
- **Geo-Spatial Dimension:** assign location in space to all entities having a permanent location.
 - GeoNames¹³ is tapped.
- **Textual Dimension:**
 - relation like `hasWikipediaAnchorText`, `hasCitationTitle`, etc, are extracted from Wikipedia
 - multi-lingual data from Universal Wordnet is added.

¹³from <http://www.geonames.org/>

YAGO : Application

YAGO in development of ontologies

YAGO in development of ontologies ¹⁴



¹⁴picture taken from presentation of Besnik fetahu

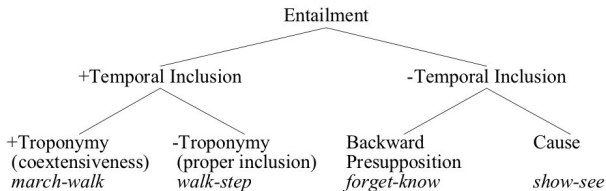
- **Querying**
- **Semantic Search** : Basis for search engines like NAGA and ESTER
 - NAGA uses YAGO KB for graph-based information retrieval.
 - ESTER combines ontological search with text search.

- Freely available at
<http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html>

VerbOcean

- Developed at University of Southern California.
- Captures semantic relation between 29,165 verb pairs [1].
 - by mining the Web for Fine-Grained Semantic Verb Relation

- WordNet provide relations between verbs
 - but at a coarser level.



- No entailment of *buy* by *sell*.
- VerbOcean relates verbs
 - doesn't group them in classes.

- **Similarity**

- produce :: create
- reduce :: restrict

- **Strength** : Subclass of **Similarity**

intensity or completeness of change produced.

- taint :: poison
- permit :: authorize
- surprise :: startle
- startle :: shock

● **Antonymy**

- Switching thematic roles of the verb
 - buy :: sell
 - lend :: borrow
- Between stative verbs
 - live :: die
 - differ :: equal
- Between siblings sharing a parent
 - walk :: run
- Entailed by common verb
 - fail :: succeed both entailed by try
- In happens-before relation
 - damage :: repair
 - wrap :: unwrap

- **Enablement** between V_1 and V_2 if V_1 is *accomplished* by V_2 .
 - assess :: review
 - accomplish :: complete
- **Happens-before** : Related verbs refer to *temporally disjoint intervals*.
 - detain :: prosecute
 - enroll :: graduate
 - schedule :: reschedule

- Associated verb pairs are extracted.
- Scored on Lexico-syntactic patterns.
- Semantic relation extracted on score of the patterns.
- Pruning.

Extracting Associated verb pairs

- 1.5GB¹⁵ newspaper corpus is considered.
- Verbs are associated if they link same sets of words.
 - Corpus is searched¹⁶ for verbs, relating same words.
 - The path considered is : *subject-verb-object*.
 - E.g.# Verbs associated with X *solves* Y (top 20)

Y <i>is solved by</i> X	X <i>resolves</i> Y
X <i>finds a solution to</i> Y	X <i>tries to solve</i> Y
X <i>deals with</i> Y	Y <i>is resolved by</i> X
X <i>addresses</i> Y	X <i>seeks a solution to</i> Y
X <i>does something about</i> Y	X <i>solution to</i> Y
Y <i>is resolved in</i> X	Y <i>is solved through</i> X
X <i>rectifies</i> Y	X <i>cope with</i> Y
X <i>overcomes</i> Y	X <i>eases</i> Y
X <i>tackles</i> Y	X <i>alleviates</i> Y
X <i>corrects</i> Y	X <i>is a solution to</i> Y
X <i>makes</i> Y <i>worse</i>	X <i>irons out</i> Y

¹⁵corpus consists of San Jose Mercury, Wall Street Journal and AP Newswire articles from the TREC-9 collection.

¹⁶using DIRT (Discovery of Inference Rules from Text) algorithm Lin and Pantel (2001)[2]

Lexico-syntactic patterns

- 35 Lexico-syntactic patterns are used.
- Different Lexico-syntactic patterns indicate different relations.
 - Manually selected,
 - by examining, known semantic relations, verb pairs.
 - Tense variations are accounted.
 - **Xed** *instantiates* on **sing** and **dance** as **sung** and **danced**.
- Web is googled for each associated verb pair with these patterns.
- Patterns indicating *narrow similarity*
 - X **ie** Y
 - X**ed** **ie** Y**ed**
 - *Kile, the software, has **produced** ie **created** this presentation.*

Lexico-syntactic patterns (contd.)

- Patterns indicating *broad similarity*
 - Xed and Yed
 - to X and Y
 - *The enemy camp was **bombarded and destroyed***
- Patterns indicating **strength**
 - X even Y
 - Xed even Yed
 - X and even Y
 - Xed and even Yed
 - Y or at least X
 - Yed or at least Xed
 - not only Xed but Yed
 - not just Xed but Yed
 - *Better **purchase** or at least **borrow** this book*

- Patterns indicating **enablement**

- Xed * by Ying the
- Xed * by Ying or
- to X * by Ying the
- to X * by Ying or
 - *You have an option to **choose** by selecting the values from a drop down.*

- Patterns indicating **antonymy**

- either X or Y
- either Xs or Ys
- either Xed or Yed
- either Xing or Ying
- whether to X or Y
- Xed * but Yed
- to X * but Y
 - *People either **hate** or **adore** movies like Prometheus*

- Patterns indicating **happens-before**
 - to X and then Y
 - to X * and then Y
 - Xed and then Yed
 - Xed * and then Yed
 - to X and later Y
 - Xed and later Yed
 - to X and subsequently Y
 - Xed and subsequently Yed
 - to X and eventually Y
 - Xed and eventually Yed
 - *The enemy forces were **crushed** immediately and later **annihilated** completely*

Scoring the verb pair on the pattern

- Strength of association is computed between
 - verb pair V_1 and V_2 and
 - A lexico-syntactic pattern p
- An approach inspired by mutual information

$$S_p(V_1, V_2) = \frac{P(V_1, p, V_2)}{P(p) \times P(V_1) \times P(V_2)}$$

Scoring the verb pair on the pattern

- Expanding & approximating the formula
 - For symmetric relations (**similarity**, **antonymy**)

$$S_p(V_1, V_2) \approx \frac{\frac{\text{hits}(V_1, p, V_2)}{N} + \frac{\text{hits}(V_2, p, V_1)}{N}}{\frac{2 * \text{hits}_{est}(p)}{N} \times \frac{\text{hits}("to V_1") \times C_v}{N} \times \frac{\text{hits}("to V_2") \times C_v}{N}}$$

- For asymmetric relations (**strength**, **enablement**, **happens-before**)

$$S_p(V_1, V_2) \approx \frac{\frac{\text{hits}(V_1, p, V_2)}{N}}{\frac{\text{hits}_{est}(p)}{N} \times \frac{\text{hits}("to V_1") \times C_v}{N} \times \frac{\text{hits}("to V_2") \times C_v}{N}}$$

- Where,
 - N : No of words indexed by the search engine $\approx 7.2 \times 10^{11}$)
 - $\text{hits}(S)$: of documents containing S, as returned by Google
 - C_v : Correction factor
to account for count of all tenses of verb from "to V"
 - $\text{hits}_{est}(p)$: pattern counted as estimated from a 500M POS tagged corpus.

Extracting semantic relation

- if $S_p(V_1, V_2) > C_1 (= 8.5)$
 - then semantic relation, S_p , as indicated by the pattern p is inferred between (V_1, V_2)
- Also for asymmetric relations
 - $S_p(V_1, V_2)/S_p(V_2, V_1) > C_2$ (taken as 5)

- If the pattern matching was low (< 10)
 - mark unrelated.
- *happens-before*
 - If **not**-detected
 - Un-mark *enablement*, if it is detected.
- *strength*
 - if detected
 - Un-mark *similarity*, if it is detected.
- Out of *strength*, *similarity*, *opposition* and *enablement*
 - Output the one with highest score.
 - and still marked.
- If no relation detected so far.
 - mark unrelated.

- Overall accuracy : 65.5%
- Human also agree on only 73% cases.
- Overall accuracy

<i>SEMANTIC RELATION</i>	<i>SYSTEM TAGS</i>	<i>Tags Correct</i>	<i>Preferred Tags Correct</i>
similarity	41	63.4%	40.2%
strength	14	75.0%	75.0%
antonymy	8	50.0%	43.8%
enablement	2	100%	100%
no relation	35	72.9%	72.9%
happens before	17	67.6%	55.9%



Timothy Chklovski and Patrick Pantel.

Verbocean: Mining the web for fine-grained semantic verb relations.

In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain, July 2004. Association for Computational Linguistics.



D. Lin and P Pantel.

Discovery of inference rules for question answering.

WWW '07, page 343–360. *Natural Language Engineering* 7(4), 2001.



Hugo Liu, Henry Lieberman, and Ted Selker.

Goose: A goal-oriented search engine with commonsense. pages 253–263. Springer-Verlag, 2002.



Hugo Liu and Push Singh.

Conceptnet: A practical commonsense reasoning toolkit.
BT Technology Journal, 22:211–226, 2004.



Marvin Minsky.

K-lines: A theory of memory.
Massachusetts Institute of Technology, (AIM-516), 1979.



Pim Nauts.

A kidsâ open mind common sense : Solving problems in
commonsense computing with a little help from children.
2009.





Mueller E T Lim G Perkins T Singh P, Lin T and Zhu W L.
Open mind commonsense: knowledge acquisition from the
general public.

*Proceedings of the First International Conference on
Ontologies, Databases, and Applications of Semantics for
Large Scale Information Systems, Lecture Notes in Computer
Science No 2519 Heidelberg, Springer, 2002.*



Fabian M. Suchanek.

Leila: Learning to extract information by linguistic analysis.
In *In Workshop on Ontology Population at ACL/COLING*,
pages 18–25, 2006.

-  Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum.
Yago: a core of semantic knowledge.
In Proceedings of the 16th international conference on World Wide Web, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.
-  Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum.
Yago: a core of semantic knowledge, long report.
New York, NY, USA, 2007.