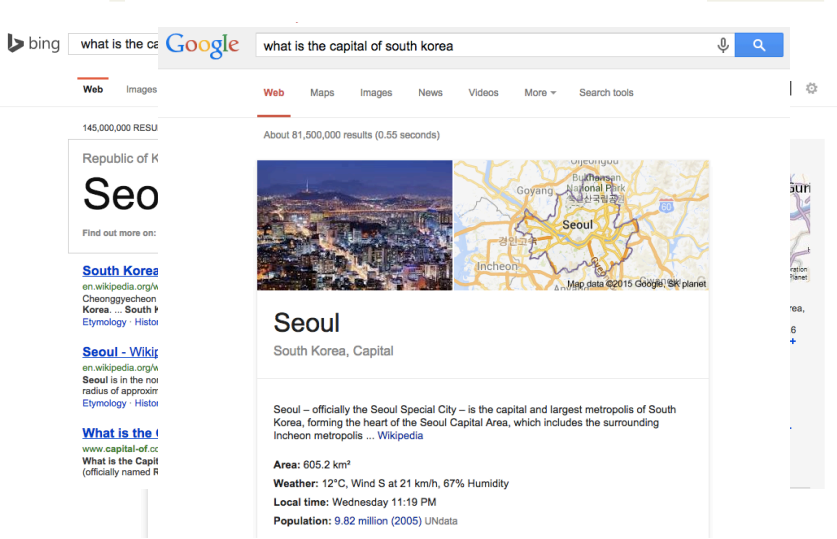
Inferencing in Information Extraction

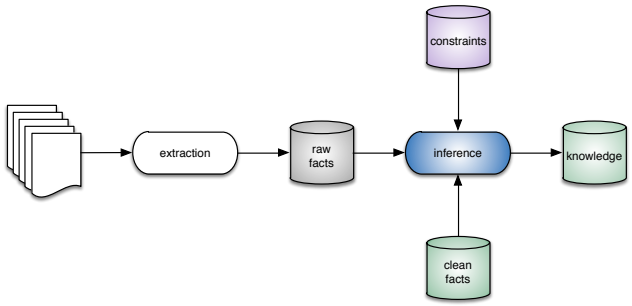
Denilson Barbosa, Univ. of Alberta, denilson@ualberta.ca
Haixun Wang, Google Research MTV, haixun@google.com
Cong Yu, Google Research NYC, congyu@google.com

Knowledge is a Core Part of Search



Overview

- Conceptual overview of knowledge base population
  - inference (or reasoning) can happen anywhere



- The inference step is needed to integrate heterogeneous facts
  - multiple sources, with different levels of coverage and credibility
  - possibly contradictory
  - extracted with different tools, on different corpora, at different points in time

Knowledge Inferencing

- Earlier Challenges
  - Creating Knowledge Base from scratch with good enough accuracy and coverage
  - Performance at Web-scale
- Current Challenges
  - Merging knowledge from different extraction sources
  - Beyond simple fact seeking
  - Long tail coverage
- Knowledge Inferencing
  - Fuse knowledge from multiple sources using probabilistic reasoning
  - Infer additional knowledge through sophisticated NLP and ML techniques
  - Using advanced, human-powered techniques to extract long tail knowledge

Outline

- Knowledge Inferencing
- Part I (Denilson): Joint Inferencing and Knowledge Fusion
- Part II (Haixun): Deep Language Inferencing with Learning
- Part III (Cong): Long Tail Knowledge Extraction
- Conclusion

Roadmap

- Knowledge Inferencing
- Part I (Denilson): Joint Inferencing and Knowledge Fusion
- Part II (Haixun): Deep Language Inferencing with Learning
- Part III (Cong): Long Tail Knowledge Extraction
- Conclusion

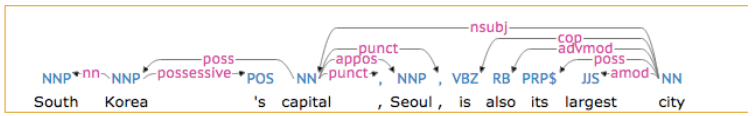
Raw facts

- SPO triples with a confidence score, possibly indicating truthfulness
  - Sources: Wikipedia Infoboxes; Open Relation Extraction; web tables; human annotations; DOM tree parsing; …
- Example: triples from the NELL SVO corpus mentioning Seoul and South Korea


```
US Treasury officials visiting South Korea  had urged   Seoul   1
South Koreaa delegation flew from   Seoul   2
South Korea s after capital Seoul   7
Yellow Sea Seoul South Korea    sits on Taedong River   12
Seoul   in  South Korea 34
Seoul daily JoongAng Ilbo   cited   South Korean government official    1
Seoul hotels South Korea    falls into  category    1
successful Seoul Olympics   raised  South Korea 's international reputation 2
Visitors    can find at 188-3 1-Ga Uljiro Jung-Gu Seoul South Korea 3
Hotel Discount Asia Hotels Websites South Korea Seoul   Get via Email   1
South Korean President Lee Myung-bak    named as    Seoul 's new pointman   1
South Korea 's Seoul Composite  declined at 1,449.48    1
South Korea Seoul Composite fell    0.72 %  12
U.S.-Japan-South Korea Legislative Exchange Program Seoul   participate in  discussions 5
South Korean workers    prepin  Seoul   1
sun sets    Seoul South Korea   1
```


Extracting raw facts

- Just one example: parsing and exploiting dependencies among words



South Korea 's capital Seoul

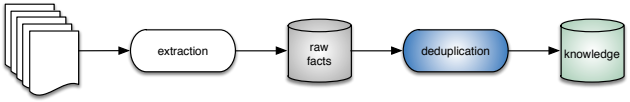
- For the trade-off between effectiveness and efficiency of typical NLP methods see [Mesquita et al. 2013]

Clustering/De-duping raw facts

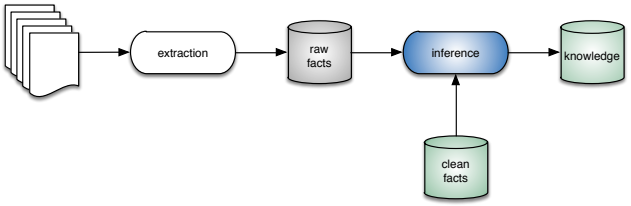
S	P	O
Bill Gates	founded	Microsoft Corp
William Gates III	founded	Microsoft
Steven Paul Jobs	co-founded	Apple Inc
Steve Jobs	founded	Next
Steve Jobs	returned to	Apple

synonyms?

- Given similarity functions for S, P, and O:
  - Repeat: merge all S, P or O that are sufficiently similar
  - Until: convergence
  - Similarity functions may take semantics (e.g., types) into account
  - Ex: Resolver [Yates and Etzioni, 2009]



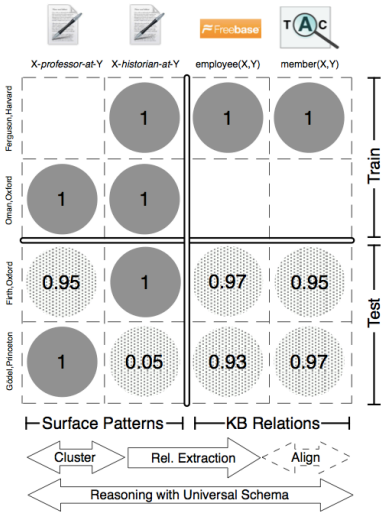
Leveraging known clean data



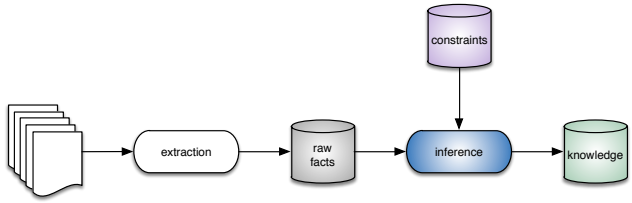
- Joint inference of relational patterns predicting relations
- To leverage clean facts from a KG we need to
  - link the entities in those facts to their mentions in SPO corpus of raw facts
  - refine the similarity function to compute the similarity between predicates in the KB and the text in the "P" column of the SPO corpus

Matrix factorization

- Ex: [Riedel et al, 2013]
- Issue: lack of good negative facts to train the model
  - probabilistic models were sensitive to the choice of "random" negative facts
- Re-cast the problem as one of recommending relations/patterns for S,O pairs
  - nicely grounded on the theory of matrix factorization



Joint inference on raw facts + constraints



- "Classical" Data Cleaning approach: filter out facts that do not conform to know set of constraints/ontological rules
- PROSPERA [Nakashole et al. 2011]
  - MapReduce-based knowledge-harvesting engine combining pattern-based gathering of relational fact candidates with weighted MaxSat-based consistency reasoning to identify the most likely correct facts

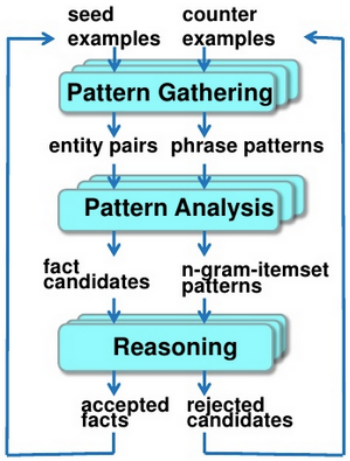
PROSPERA

- Manual set of logical constraints:

$occurs(p, e1, e2) \wedge type(e1, dom(R)) \wedge type(e2, range(R))$
$\quad \wedge expresses(p, R) \Rightarrow R(e1, e2)$ //pattern-fact duality
$occurs(p, e1, e2) \wedge type(e1, dom(R)) \wedge type(e2, range(R))$
$\quad \wedge R(e1, e2) \Rightarrow expresses(p, R)$ //pattern-fact duality
$R(e1, e2) \wedge type(R, function) \wedge different(e2, e3)$
$\quad \Rightarrow \neg R(e1, e3)$ //functional dependency
$R(e1, e2) \wedge sub(R, S) \Rightarrow S(e1, e2)$ //inclusion dependency
$R(e1, e2) \wedge inv(R, T) \Rightarrow T(e2, e1)$ //inverse relations
$T(e1, e2) \wedge inv(R, T) \Rightarrow R(e2, e1)$ //inverse relations

- Grounding based on approximate string similarity

$occurs$(and PRP alma mater, Barbara_Liskov, Stanford_University)
$\quad \wedge expresses$(and PRP alma mater, graduatedFrom)
$\quad \Rightarrow graduatedFrom$(Barbara_Liskov, Stanford_University)
$graduatedFrom$(Barbara_Liskov, Stanford_University)
$\quad \Rightarrow \neg graduatedFrom$(Barbara_Liskov, UC_Berkeley)
$graduatedFrom$(Barbara_Liskov, UC_Berkeley)
$\quad \Rightarrow \neg graduatedFrom$(Barbara_Liskov, Stanford_University)



- Can have domain-specific constraints

Evaluation—PROSPERA

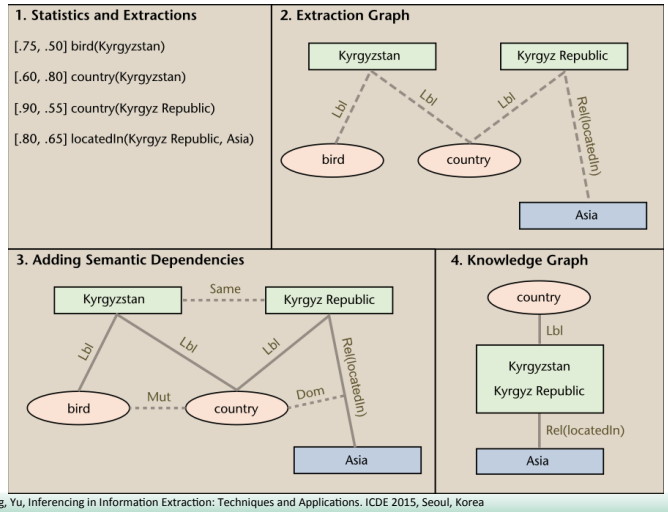
- PROSPERA vs NELL
  - Never Ending Language Learning, CMU project [Mitchell et al, 2015]

Relation	# Extractions			Precision			Precision@1000
	PROSPERA-6	NELL-6	NELL-66	PROSPERA-6	NELL-6	NELL-66	PROSPERA-6
AthletePlaysForTeam	14,685	29	456	82%	100%	100%	100%
CoachCoachesTeam	1,013	57	329	88%	100%	100%	n/a
TeamPlaysAgainstTeam	15,170	83	1,068	89%	96%	99%	100%
TeamWonTrophy	98	29	397	94%	88%	68%	n/a
AthletePlaysInLeague	3,920	2	641	94%	n/a	n/a	n/a
TeamPlaysInLeague	1,920	62	288	89%	n/a	n/a	n/a
AthleteWonTrophy	10	n/a	n/a	90%	n/a	n/a	n/a
CoachCoachesInLeague	676	n/a	n/a	99%	n/a	n/a	n/a
TeamMate	19,666	n/a	n/a	86%	n/a	n/a	100%

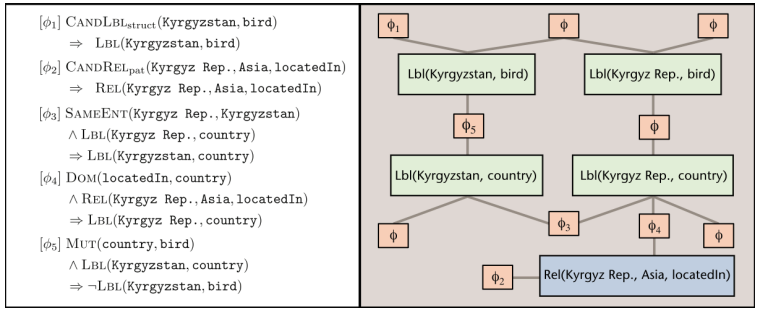
Table 1: Performance comparison between PROSPERA and NELL on sports relations

Knowledge extraction with graphical models

- [Pujara et al 2015]



Probabilistic Soft Logic



- The construction of the Knowledge Graph = finding the most likely graph given the extracted facts and ontological constraints
  - Probabilistic Soft Logic [Broecheler et al, 2010]: relaxing binary truth values to the continuous domain turns inference into a convex optimization problem that can be computed fast and scales very well

Evaluation—PSL

Method	AUC	F1
Baseline	0.873	0.828
NELL	0.765	0.673
MLN	0.899	0.836
PSL-KGI	**0.904**	**0.853**

Table 2. Comparing against Previous Work on the NELL Data set.

Knowledge graph identification using PSL demonstrates a substantive improvement. The best-performing method is shown in boldface.

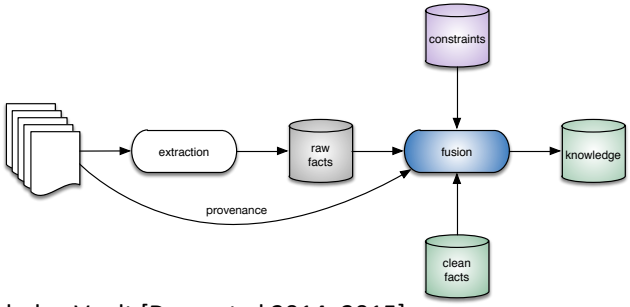
Method	AUC	F1
PSL-NoSrcs	0.900	0.852
PSL-NoER	0.899	**0.853**
PSL-NoOnto	0.887	0.826
PSL-KGI	**0.904**	**0.853**

Table 3. Comparing Variants of PSL Graph Identification.

This comparison shows the importance of ontological information, but the best performance is achieved when all of the components of knowledge graph identification are combined.

- Compared against 165th iteration of NELL
- Used YAGO to link entities to mentions in the raw facts

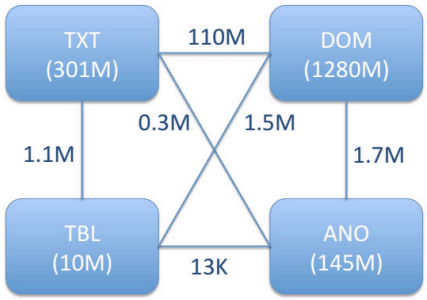
From data fusion to knowledge fusion



- Knowledge Vault [Dong et al 2014, 2015]
- Key ingredients:
  - Multiple Web Extractors operating at Web-scale
  - Graph based priors for the likelihood of each candidate triple
  - Knowledge fusion based on agreement among extractors, sources, and priors
- At scale: map-reduce fusion algorithm

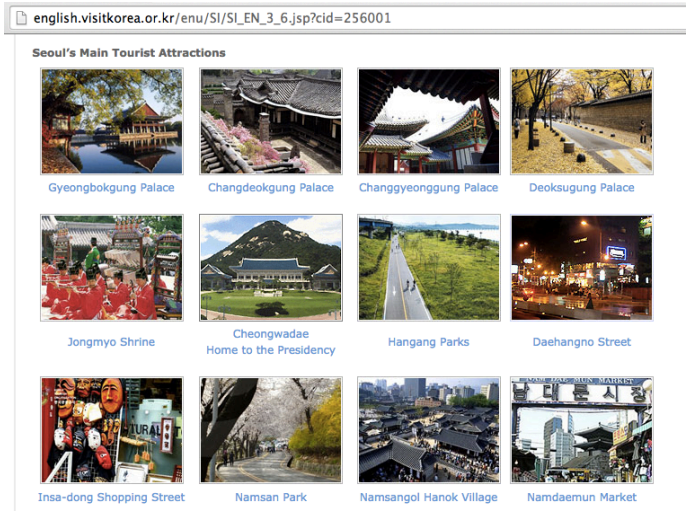
Sources

- TXT: text extractions
- DOM tree extractions
  - 75% of all extractions
- ANO: schema.org annotations
- TBL: Web tables
  - Schema matching/table understanding approach



Thanks to Xin Luna Dong (Google)

DOM Tree extractions

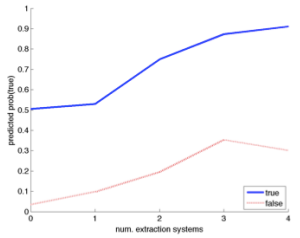


Fusion in the KV

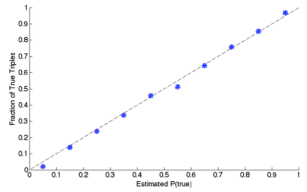
- Extract 1.6B facts from the Web, over ~4.5K relations and 1.1K entity types
- Estimate thee likelihood of each fact based on different kinds of priors
  - 271M high confidence (>0.9) candidate facts
- Train classifiers for each relation and extractor labeled according to a *local* closed world assumption
  - a candidate is deemed false only if contradicts a known fact
  - features: an indicator of the number of sources, and the mean score of the extractor for that candidate

Fusion in the KV

- Fusion on the prediction of the different classifiers and also on the number of systems that extract the candidate



- The ultimate goal is to accurately predict the likelihood of each candidate—need to map extractor scores into probabilities



Knowledge Vault—some remarks

- Impressive undertaking at massive scale
- Nice combination of machine learning and data cleaning approaches
- Where do the best facts come from?

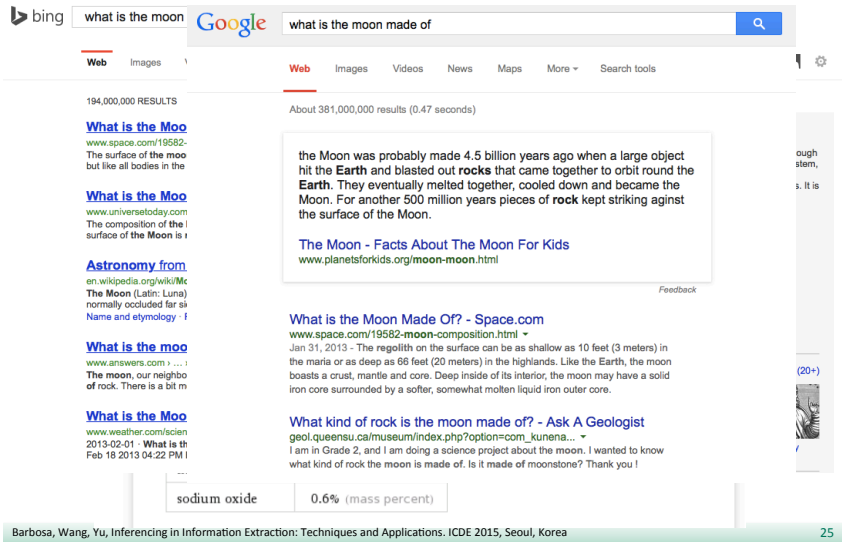
System	#	# > 0.7	# > 0.9	Frac. >0.9	AUC
TBL	9.4M	3.8M	0.59M	0.06	0.856
ANO	140M	2.4M	0.25M	0.002	0.920
TXT	330M	20M	7.1M	0.02	0.867
DOM	1200M	150M	94M	0.08	0.928
FUSED-EX.	1600M	160M	100M	0.06	0.927

Table 2: Performance of different extraction systems.

Quiz

- What is the moon made of?
  - A) cheese
  - B) rocks and metals
  - C) valuable metals
  - D) all of the above
- By moon here we mean Earth's (the planet you're on right now) only natural satellite

What is the moon made of?



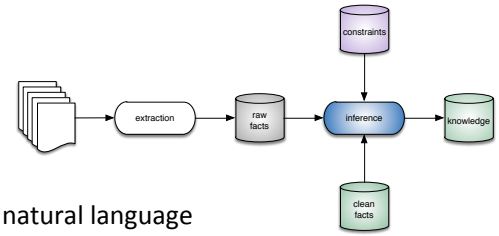
What are we (humans) teaching the machines?

- From the NELL corpus:

```
denilson@noronha:/data1/stup/nell$ \grep -e "\bMoon\b" -e "\bmoon\b" nell_hazy_svo_604m | grep "\bmade of\b"
41.95 Beautiful Moon Goddess pendant    made of fine lead free pewter   1
moon    made of green cheese    1
Moon    made of cheese  7
Moon    made of fragmented rock 1
Silver Moon A large reflective disk     made of water   1
Moon    made of Cheese  4
popular being Snow-Skin Durian Moon-cake        made of selected durians        1
moon    made of valuable metal  6
real moon       made of nothing 5
being Soshie Moon       made of different stripes       3
Ascolta Sun Kil Moon    made of ashes   1
Galileo made of moon    1
Antique brooch  made of beautiful moon opalescent glass 1
Shoelace        made of moon bittorrent shoes   2
Moon    considering being made of       green cheese    1
moon Ascolta Sun Kil Moon       made of ashes   4
Moon    made of green cheese    1
Moon    made of Swiss cheese    7
Moon Goddess pendant    made of fine lead-free pewter   1
intricate Celestial Moon molds  made of sturdy polystyrene great        4
Agave Denimsmith 's Nectar Blue Moon La Sirena jeans    made of Japanese pinstripe denim        3
al-Muqanna moon made of iron sheet      2
Moon    made of cheddar 1
TOR     made of 2 Free Jamario Moon     4
Mom E 's Clapotis       made of Blue Moon Fiber Arts Geisha     2
Moon    made of earth   7
moon    made of cheese  3
one Fry had made of on  moon    2
moon    made of Before  advent  1
```


Challenges

- Validation!
- The many kinds of knowledge
  - factual, common-sense, ontological, ephemeral, …

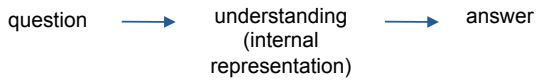


- Dealing with intricacies of natural language
  - context of the extractions; author bias;
- The need for accurate *truthfulness* values
- The need for logical constraints
  - can we learn useful constraints from data?
- Fusion with generalizations and *temporal facts*
  - "historian at" → "faculty member at" → "professor at" → "works at"
- **Validation!**

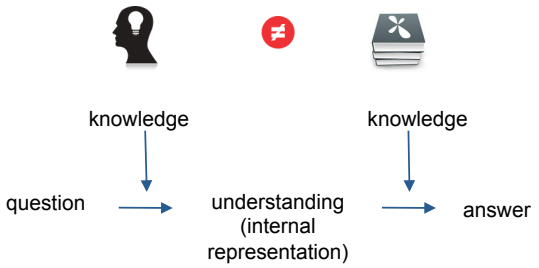
Roadmap

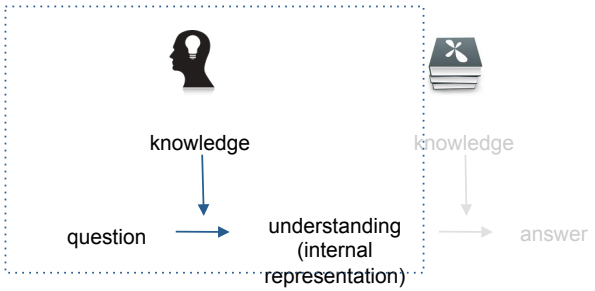
- Knowledge Inferencing
- Part I (Denilson): Joint Inferencing and Knowledge Fusion
- Part II (Haixun): Deep Language Inferencing with Learning
- Part III (Cong): Long Tail Knowledge Extraction
- Conclusion

Text understanding is the key



There are two types of knowledge





"What people know when they know a language."

Knowledge about Language: An Example

- Mary puts on a coat every time she leaves the house.
  she = Mary (possible)
- She puts on a coat every time Mary leaves the house.
  she = Mary (impossible)
- Every time Mary leaves the house she puts on a coat.
  she = Mary (possible)
- Every time she leaves the house Mary puts on a coat.
  she = Mary (possible)

from Prof. Maria Polinsky, Harvard

Generative vs. Information Theoretic Linguistics

The most famous Chomsky quote:

- *Colorless green ideas sleep furiously.*
- *Furiously sleep ideas green colorless.*

. . . It is fair to assume that neither sentence (1) nor (2) ... has ever occurred in an English discourse. *Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English.* Yet (1), though nonsensical, is grammatical, while (2) is not.

Could Chomsky be wrong?

- Colorless green ideas sleep furiously.
- Furiously sleep ideas green colorless.

- Assumption: probabilistic models assign zero probability to unseen events.

Modeling a sequence of words

Pereira, Fernando (2000) "Formal grammar and information theory: together again?"

Two words:

$$p(x,y) = p(x)p(y|x) = p(x)\sum_c p(y|c)p(c|x)$$

hidden variable

A sequence of words:

$$p(w_1, w_2, \cdots, w_n) = p(w_1)\prod_{i=1}^{n-1} p(w_{i+1}|w_i)$$

Statistical inferencing

- Assume # hidden classes = 15
- Using EM to learn the model from a newspaper corpus, we get:

$$\frac{p(\text{Colorless green ideas sleep furiously})}{p(\text{Furiously sleep ideas green colorless})} \approx 2 \times 10^5$$

How can machines understand short Text (e.g., queries)

Hua et al, ICDE 2015

- watch harry potter 
- harry potter watch 

Short Text (e.g., search queries)

watch for kids

i) 

ii) 

iii) 

Document captions

ASUS Intel Core i5 8GB DDR3 1TB HDD Capacity
ASUS Intel Core i5 8GB DDR3 1TB HDD Capacity Desktop PC Windows 7 Professional P8H61E (BP6320-I53470163B)

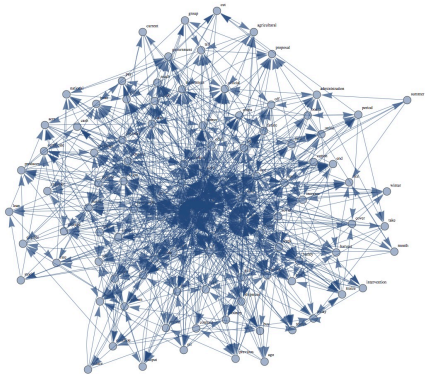
(2) | Write a Review

In stock. Limit 5 per customer.

- Intel Core i5 3470(3.20GHz)
- 8GB DDR3 1TB HDD Capacity
- Windows 7 Professional
- Energy star certification

A semantic network for text understanding

Probase [Wu et al, 2012]

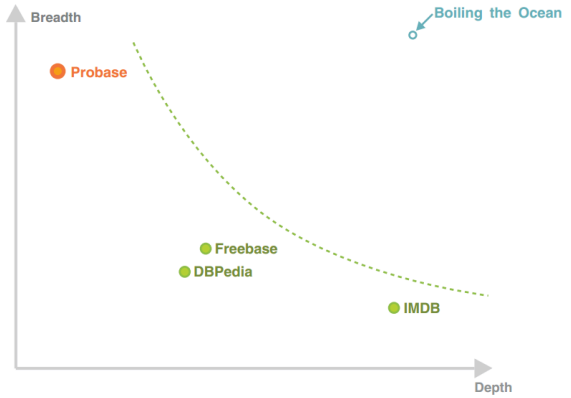


Nodes: Concepts (noun/verb/preposition phrases, etc)
Edges: isA, head-modifier, sibling, co-occurrence

Knowledge in Probase

- It is about *what people know when they know a language*.
- It is not about *right or wrong*. It's *usage*.
- Words/phrases/constructs evoke a network of concepts which lead to understanding.

What can Probase do?



isA Extraction

- Hearst pattern

  NP such as NP, NP, …, and|or NP
  such NP as NP,* or|and NP
  NP, NP*, or other NP
  NP, NP*, and other NP
  NP, including NP,* or | and NP
  NP, especially NP,* or|and NP

- … is a … pattern

  NP is a/an/the NP

- *domestic animals* such as *cats* and *dogs* …
- animals other than *cats* such as *dogs* …
- *China* is a *developing country*.
- *Life* is a box of *chocolate*.

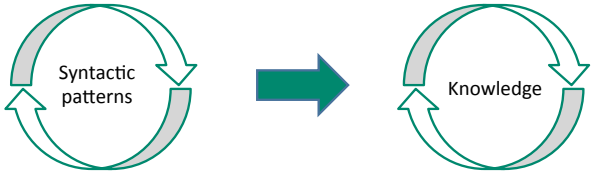
How hard is it?

- … *cities in Asian countries* such as *Japan* and *China* …
- … *cities in Asian countries* such as *Beijing* and *Tokyo* …

… **animals** other than **cats** such as **dogs** …

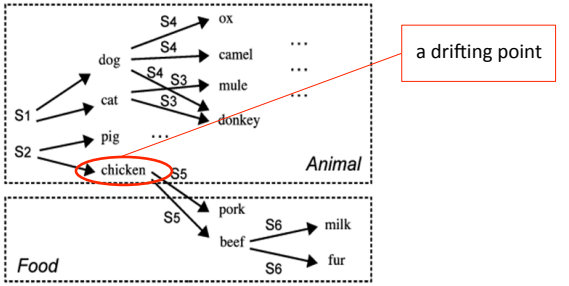
… **household pets** other than **animals** such as **reptiles**, aquarium fish …

Iterative Information Extraction



Challenge: Semantic Drifts

Overcoming Semantic Drift in Information Extraction, EDBT 2014



S1="Animals **such as** dogs and cats, grow fast."
S2="Land animals **such as** chicken and pigs – all of which live on land"

S3="Postures are often named after animals, **such as** mule, donkey and cat."
S4="... innkeeper, angels, and animals **such as** ox, camels, donkeys and dog"
S5="Common food from animals such as pork, beef and chicken"

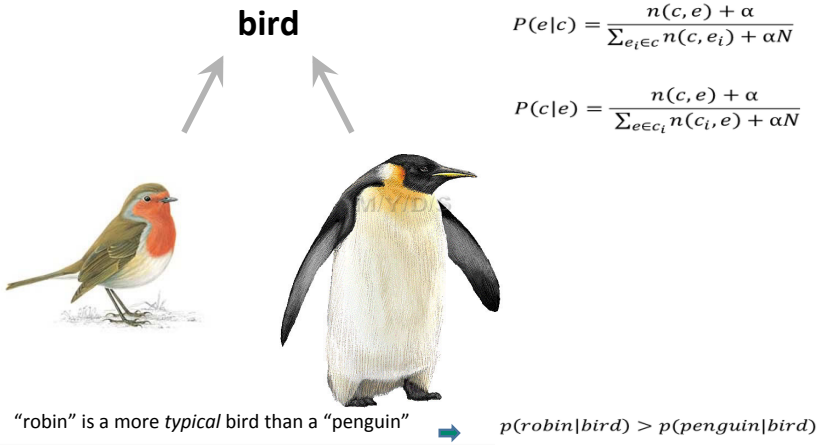
S6="Products from animals **such as** fur, milk and beef are given to families..."

Scores

- Typicality
- Vagueness
- BLC (basic level of categorization)
- Ambiguity
- Similarity

foundation for inferencing

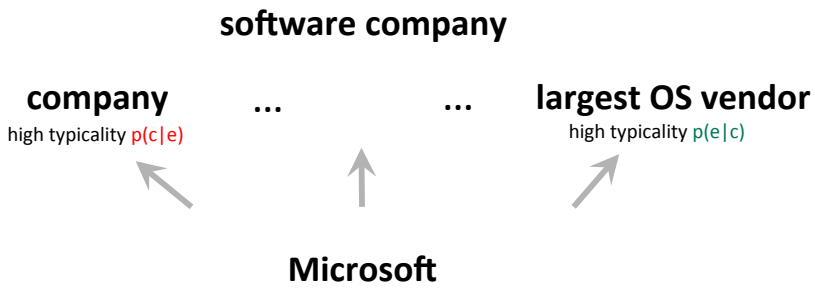
Typicality



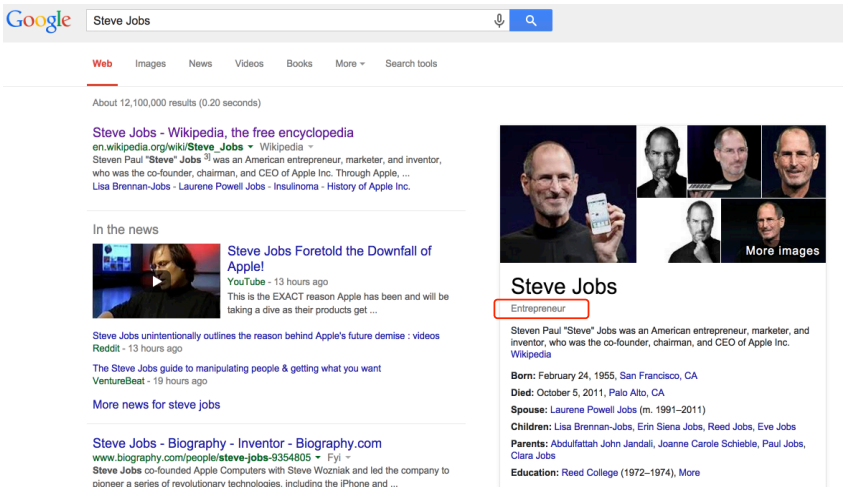
"robin" is a more *typical* bird than a "penguin"

$p(robin|bird) > p(penguin|bird)$

BLC (basic level of categorization)



BLC (basic level of categorization)



PMI

$$PMI(e,c) = \log \frac{P(e,c)}{P(e)P(c)} = \log P(e|c) - \log P(e)$$

For given e, log P(e) is a constant.

PMI degenerates into log typicality.

A better measure

$$R(e,c) = p(e|c) \cdot p(c|e)$$

$$\log R(e,c) = \log \frac{P(e,c)^2}{P(e)P(c)} = PMI(e,c) + \log P(e,c)$$

Concept Learning

China India

country

Concept Learning

China Brazil India

emerging market

body smell taste

wine

china population

country

collector of fine china

earthenware

Bayesian

$$P(c_k|E) = \frac{P(E|c_k)P(c_k)}{P(E)} \propto P(c_k)\prod_{i=1}^{M} P(e_i|c_k).$$

- For a mixture of instances and properties: Noisy-Or model

$$P(c|t_l) = 1 - (1 - P(c|t_l, z_l = 1))(1 - P(c|t_l, z_l = 0))$$

where $z_l = 1$ indicates t_l is an entity, $z_l = 0$ indicates t_l is a property

- Bayesian rule gives:

$$P(c|T) \propto P(c)\prod_l^L P(t_l|c) \propto \frac{\prod_l P(c|t_l)}{P(c)^{L-1}}$$

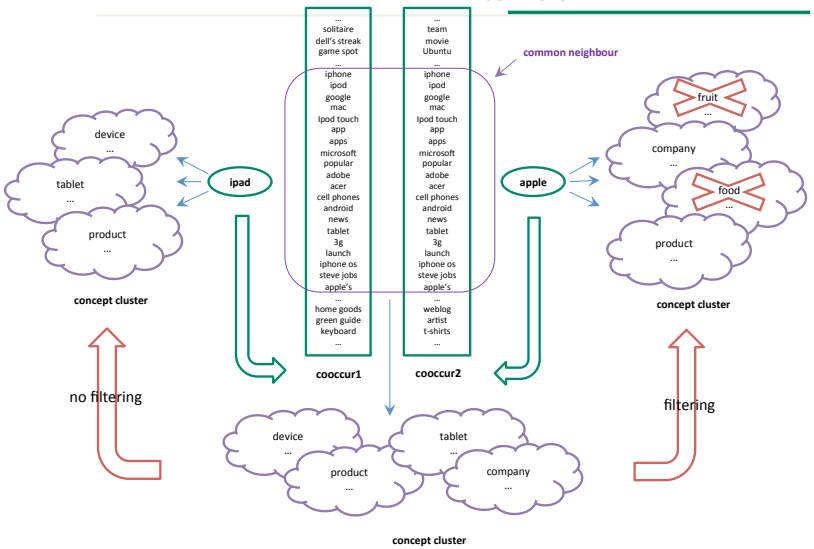
apple

iPad

company

device

Context Dependent Conceptualization
IJCAI 2013



Probabilistic model for query understanding

Given a query Q = q, our goal is to produce a structured output Y = y, which represents our understanding of q.

Y consists of three parts:

1. segmentation (S),
2. term labels or concepts (Z), and
3. a tree (E) that represents the dependency structure.

Find the best understanding of q

$$\arg\max_{y} P(Y = y|Q = q)$$

Modeling P(Y|Q)

$$P(Y|Q) = P(S, Z, E|Q) = P(Z, E|S, Q)\,P(S|Q)$$

1. S = {t1, …, tN} is a sequence of terms where each term is in the vocabulary.
2. Z = {z1, …, zN} is the term labels associated with each term in the segmentation S.
3. E is a set of directed edges between the terms.

Modeling P(Z, E|S, Q)

$$P(Z, E|S, Q) = P(Z_0|S, Q) \prod_{(i,j)\in E} P(E_{ij} = 1, Z_j = z_j|Z_i = z_i, S, Q)$$

- Z0 is the root of the tree.
- This splits the probability of observing the tree E as a product of probabilities of individual edges in the tree.

Modeling $P(E_{ij} = 1, Z_j = z_j|Z_i = z_i, S, Q)$

- The edge and the term label of the node are jointly inferred.
- Edge: a relationship between two concepts. They depend on each other.

P(Z, E | S, Q) = P(Z | S, Q) P(E | Z, S, Q)

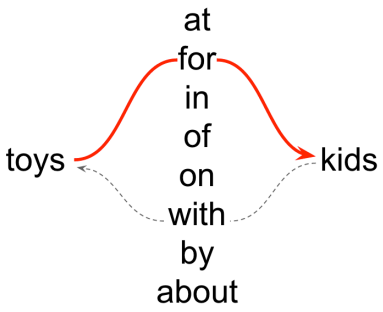
labeling relation

Head Modifier Analysis

Query	Intent
toy kids	find toys for kids
kids toy	find toys for kids

Head Modifier Analysis

- "toys for kids" appears more frequently than "kids with toys" in search
- when "toys" and "kids" appear together in a query, "toys" is usually the head



Challenge

Head-modifier signals are not from a robust model.

- Signals are **context** agnostic. We want to know, given two terms t1 and t2 in a query q, how likely t1 is the head of t2?
- It is not clear how to **generalize** to any two terms in any query.

Roadmap

- Knowledge Inferencing
- Part I (Denilson): Joint Inferencing and Knowledge Fusion
- Part II (Haixun): Deep Language Inferencing with Learning
- Part III (Cong): Long Tail Knowledge Extraction
- Conclusion

Long Tail Knowledge Extraction

Logical and linguistic inferencing are core horizontal techniques that will help us address multiple challenges in information extraction.

Long tail knowledge extraction is a specific application challenge

- As knowledge becomes the norm in Web search, long tail content has become increasingly important to users' overall search experience
  - Long tail entities in popular domains
  - Long tail domains

It's not a new problem:

Information extraction from Wikipedia: moving down the long tail

Fei Wu, Raphael Hoffmann, Daniel S. Weld, KDD 2008

But a very difficult one: if it is in Wikipedia, it is not really long tail!

Crowd Sourcing to the Rescue

Fully automated approaches rarely work well

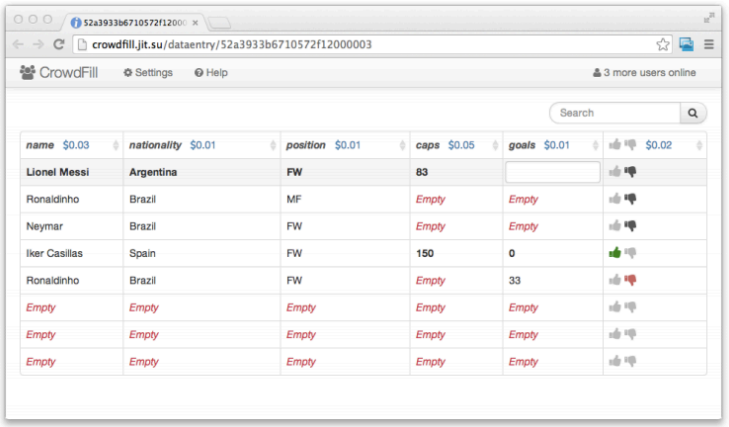
- E.g., a lot of the long tail content may not even be online!

Two recent crowd sourcing approaches

- CrowdFill, Park and Widom, SIGMOD 2014
- Quizz, Ipeirotis and Gabrilovich, WWW 2014

CrowdFill

No more micro-task, table-filling!



Advantages and Challenges

Advantages

- Workers observe and learn from each other
- Workers can prioritize the tasks themselves

name	nationality	position	caps	goals
		='FW'		≥30
	='Brazil'			≥30
	='Spain'		≥100	

- Quality goals: the system does not want to accept arbitrary entries
  - **Cardinality**: the final table must have at least *n* rows.
  - **Value**: some rows may have pre-specified values in some columns
  - **Predicate**: some rows may have conditions in some columns
    - A more general Value constraint

Concurrent Execution Mechanism

Central server keeps track of all primitive operations from worker clients

- Client: insert(r), fill(r, A, v)
- Server: insert(r), replace(r, q, V)

For each filled value from a worker, a row with *new* internal identity is created, this is crucial for concurrent edits

- If there are simultaneous edits, two new rows will be created and conflict resolution can happen later without delayed the propagation of the updated values

Compensating Workers

Task owner provides a total budget for filling a table, instead of pricing each individual cells.

Workers are compensated based on how much value they contributed to the final table:

- Filling more number of cells
- Filling more difficult cells
  - Weighted columns
  - Weighted primary key cells

Compensating Workers

Task owner provides a total budget for filling a table, instead of pricing each individual cells.

Workers are compensated based on how much value they contributed to the final table:

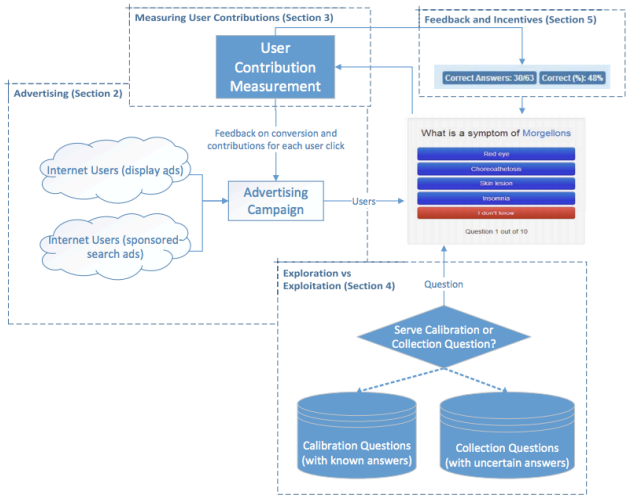
- Filling more number of cells
- Filling more difficult cells
  - Weighted columns
  - Weighted primary key cells

- Preliminary results: $10, ~11 minutes, to obtains 20 rows of SoccerPlayer information

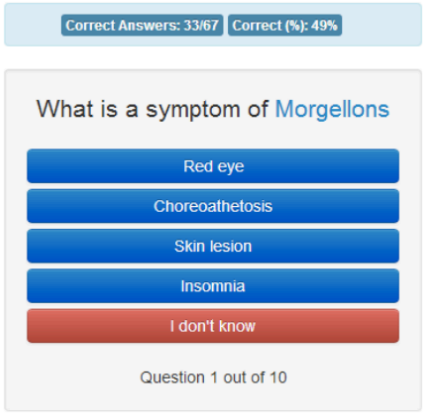
Quizz

- Key challenges
  - Finding knowledgeable users is extremely hard for long tail topics
  - Making those people work for money is even harder
    - Domain experts who are educated or otherwise busy professionals
- Key insights
  - Using advertising campaign to recruit users
  - Using games to keep users engaged
- Quizz: acquiring knowledge about long tail topics using crowdsourcing mechanisms that are difficult to obtain using automated means.

Overview of the Quizz System



The Quizz



Ad Campaign to Recruit Users

- Users are not paid, they are simply interested in the topic
- The advertising campaign does cost money!
- Using *disease symptoms* as an example topic
  - Patients
  - Doctors

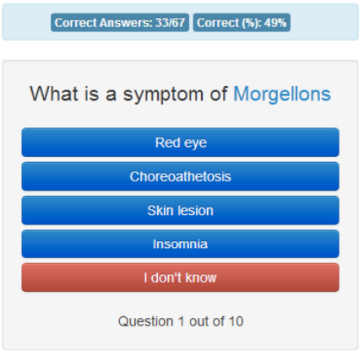
Quiz on disease symptoms
Test how well you can recognize various disease symptoms
www.quizz.us

Campaign Optimization

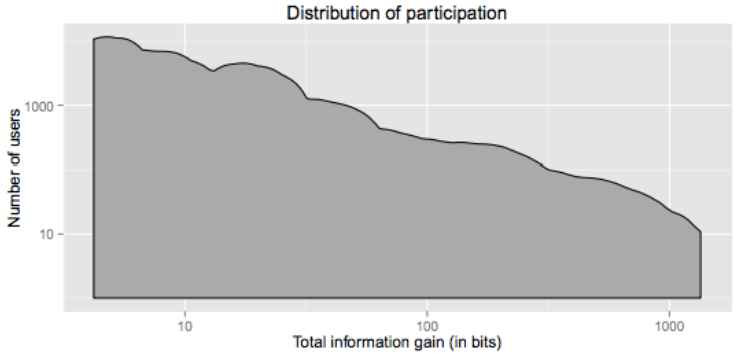
- Two important desirable traits about participating users
  - Enjoying the quizz and thus happy to complete them
  - Competence in answering the quizz questions correctly
- The Ad Campaign needs to optimize for attracting users who are both enjoying the quizz and competent.

The Quizz

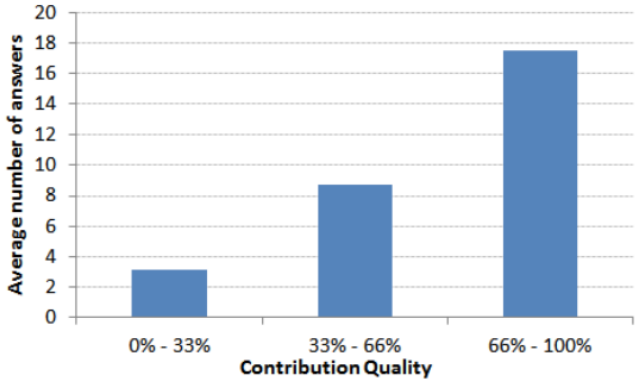
- Calibration questions
  - Estimate user quality
- Collection questions
  - Where real utilities are
- Using information gain to measure the value of a user session
  - Higher with better quality
  - Higher with more volume



User Participation



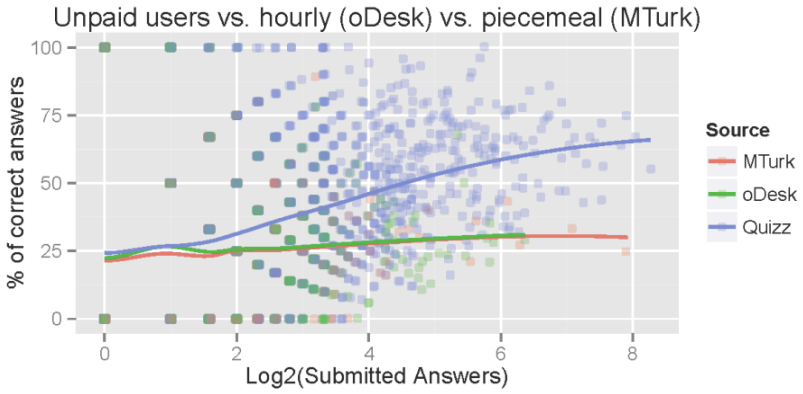
User Participation



Cost of Acquisition

Quiz	Cost/Answer @99%	@95%	@90%
Disease Causes	$0.07	$0.05	$0.04
Disease Symptoms	$0.02	$0.01	$0.01
Treatment Side Effects	$0.13	$0.10	$0.07
Artist and Albums	$0.16	$0.13	$0.09
Latest Album	$0.09	$0.07	$0.05
Artist and Song	$0.54	$0.42	$0.31
Film Directors	$0.07	$0.05	$0.04
Movie Actors	$0.22	$0.18	$0.13
Average	*$0.16*	*$0.12*	*$0.09*

Quality of Answers



Conclusion

- Knowledge extraction is becoming mature
  - A combination of human and automated techniques make the quality of Knowledge, especially for popular entities in popular domains, very high
  - But abundant advanced challenges still remain

- Heavy/Logical inference happens too late: we need to understand the source (text) so as to qualify the raw facts

- Knowledge of language and statistical inferencing hold the key for text understanding and information extraction.

- More challenges lie in the long tail