Open Information Extraction Systems and Downstream Applications

Mausam
Computer Science and Engineering
Indian Institute of Technology
New Delhi, India

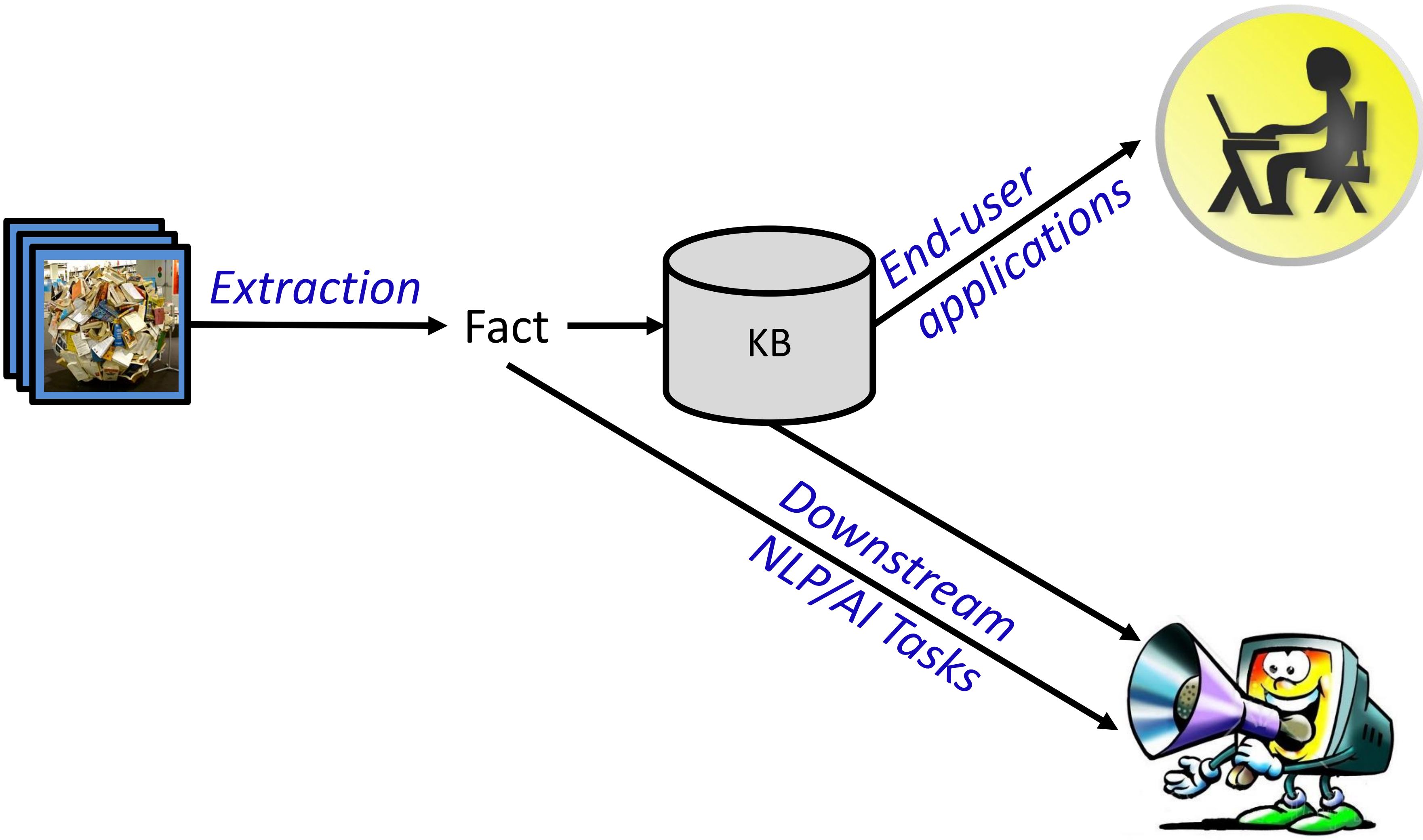
Joint work with
Oren Etzioni, Stephen Soderland, Michael Schmitz,
Ido Dagan, Ganesh Ramakrishnan, Sunita Sarawagi, Parag Singla,
Niranjan Balasubramanian, Robert Bart, Janara Christensen,
Danish Contractor, Anthony Fader, Aman Madaan, Ashish Mittal,
Harinder Pal, Abhishek Yadav, Gabriel Stanovsky

"The Internet is the world's largest library. It's just that all the books are on the floor."

- John Allen Paulos

~20 Trillion URLs (Google)

Overview


Extraction
Fact
KB
End-user
applications
Downstream
NLP/AI Tasks

Closed KBs

Populating information wrt *a given ontology* from natural language text

"Apple's founder Steve jobs died of cancer following a…"

↓ Closed IE

rel:founder_of(Apple, Steve Jobs)

rel:founder_of
(Apple, Steve Jobs)
(Google, Larry Page)
(Facebook, Mark Zuckerberg)
…

rel:acquisition
(Google, DeepMind)
(IBM, uStream)
(Microsoft, Powerset)
…

Lessons from DB/KR Research

- Large-scale Ontologies
  - expensive to create
  - very hard to maintain
  - conflict with distributed authorship

- KBs are brittle: "*can only be used for tasks whose knowledge needs have been anticipated in advance*" (Halevy IJCAI '03)

Open KBs

Broad-coverage KBs with light-weight structure

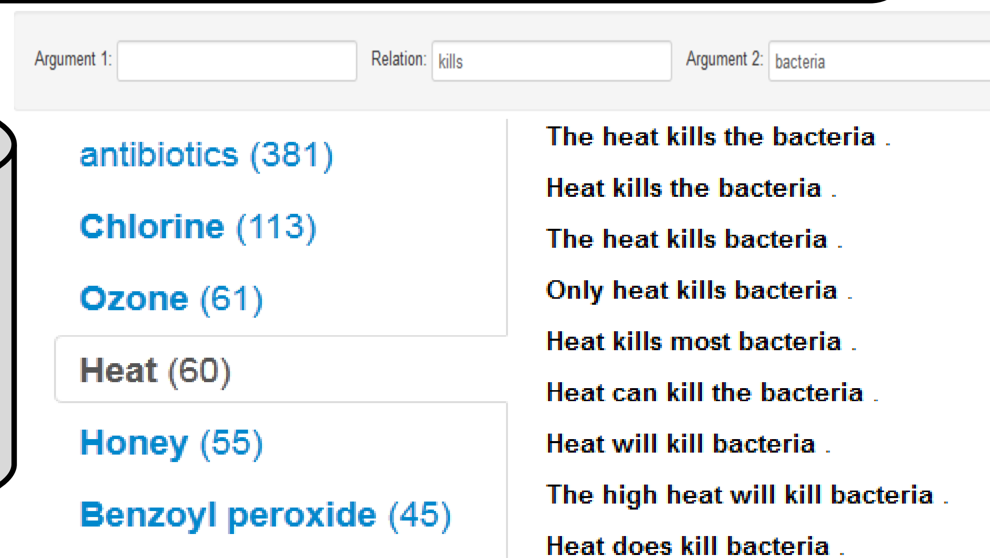
(no annotation per relation)

"When Saddam Hussain invaded Kuwait in 1990, the international.."

Open IE

(Saddam Hussain, invaded, Kuwait)

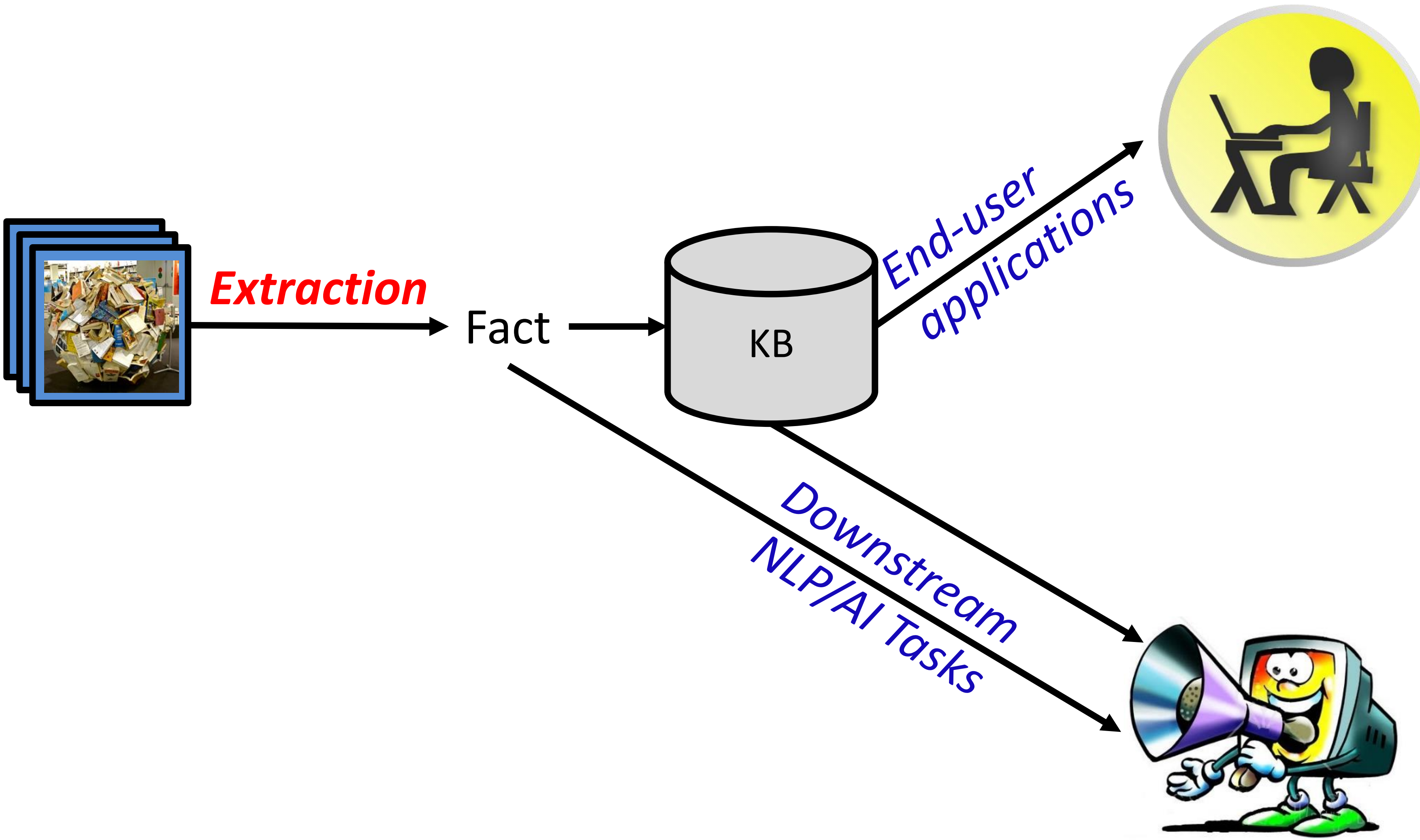
(Google, acquired, Youtube)
(Oranges, contain, Vitamin C)
(Edison, invented, phonograph)
...



Demo

- http://openie.cs.washington.edu

Overview


Extraction
Fact
KB
End-user
applications
Downstream
NLP/AI Tasks

Open Information Extraction

- 2007: Textrunner (~Open IE 1.0)
  - CRF and self-training
- 2010: ReVerb (~Open IE 2.0)
  - POS-based relation pattern
- 2012: OLLIE (~Open IE 3.0)
  - Dep-parse based extraction; nouns; attribution
- 2014: Open IE 4.0
  - SRL-based extraction; temporal, spatial…
- 2016 [@IITD]: Open IE 5.0
  - compound noun phrases, numbers, lists

increasing precision, recall, expressiveness

Fundamental Hypothesis

∃ *semantically tractable* subset of English

- Characterized relations & arguments via POS
- Characterization is compact, domain independent
- Covers 85% of binary relations in sample

ReVerb

Identify **Re**lations from **Verbs**.

1. Find longest phrase matching a simple syntactic constraint:

$$\boxed{\begin{array}{c} V \mid V P \mid V W^{*} P \\ \begin{array}{lcl} V & = & \text{verb particle? adv?} \\ W & = & (\text{noun} \mid \text{adj} \mid \text{adv} \mid \text{pron} \mid \text{det}) \\ P & = & (\text{prep} \mid \text{particle} \mid \text{inf. marker}) \end{array} \end{array}}$$

Sample of ReVerb Relations

invented	acquired by	has a PhD in
inhibits tumor growth in	voted in favor of	won an Oscar for
has a maximum speed of	died from complications of	mastered the art of
gained fame as	granted political asylum to	is the patron saint of
was the first person to	identified the cause of	wrote the book on

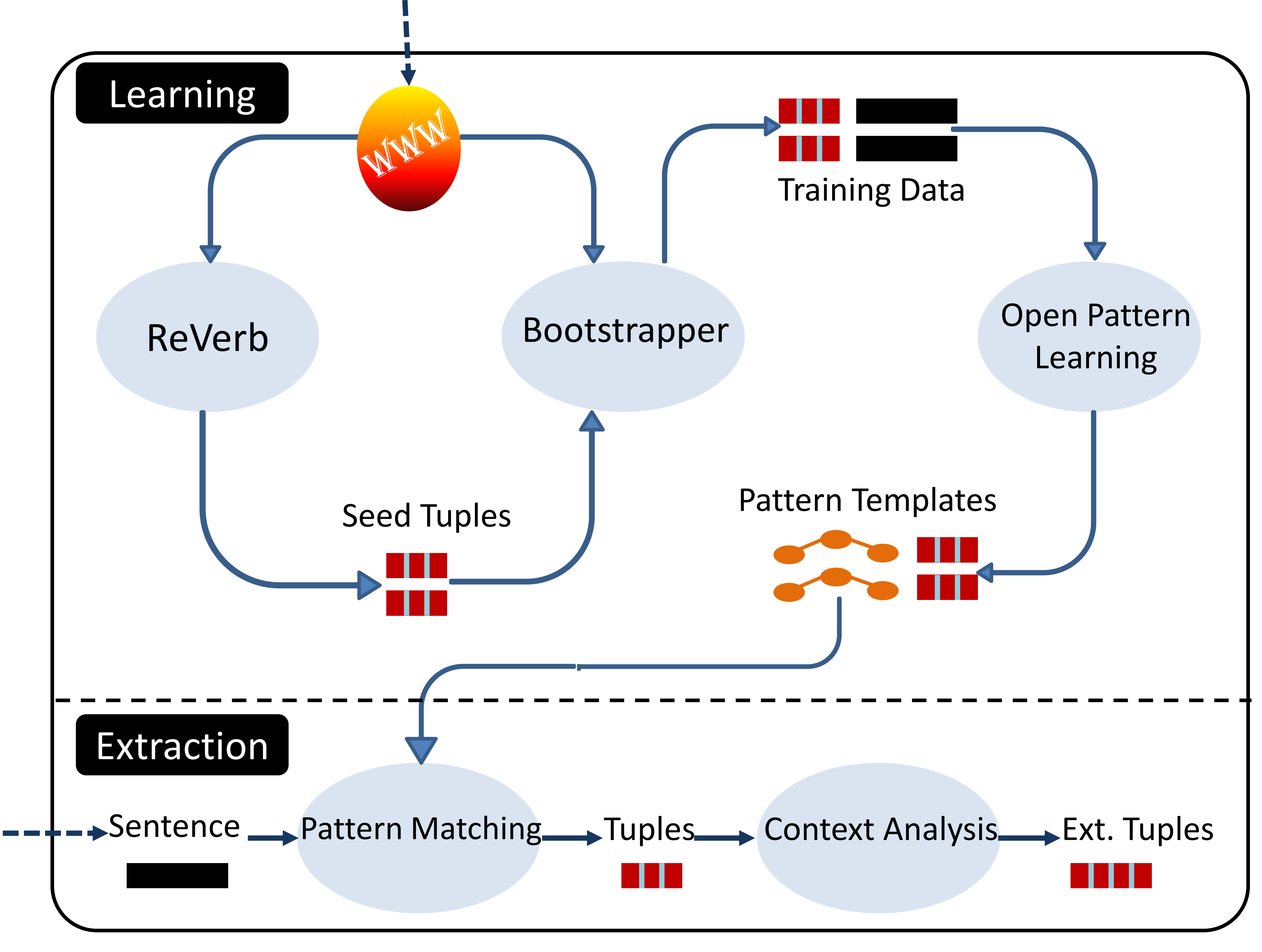
Number of Relations

DARPA MR Domains	<50
NYU, Yago	<100
NELL	~500
DBpedia 3.2	940
PropBank	3,600
VerbNet	5,000
WikiPedia InfoBoxes, f > 10	~5,000
TextRunner (phrases)	**100,000+**
ReVerb (phrases)	**1,500,000+**

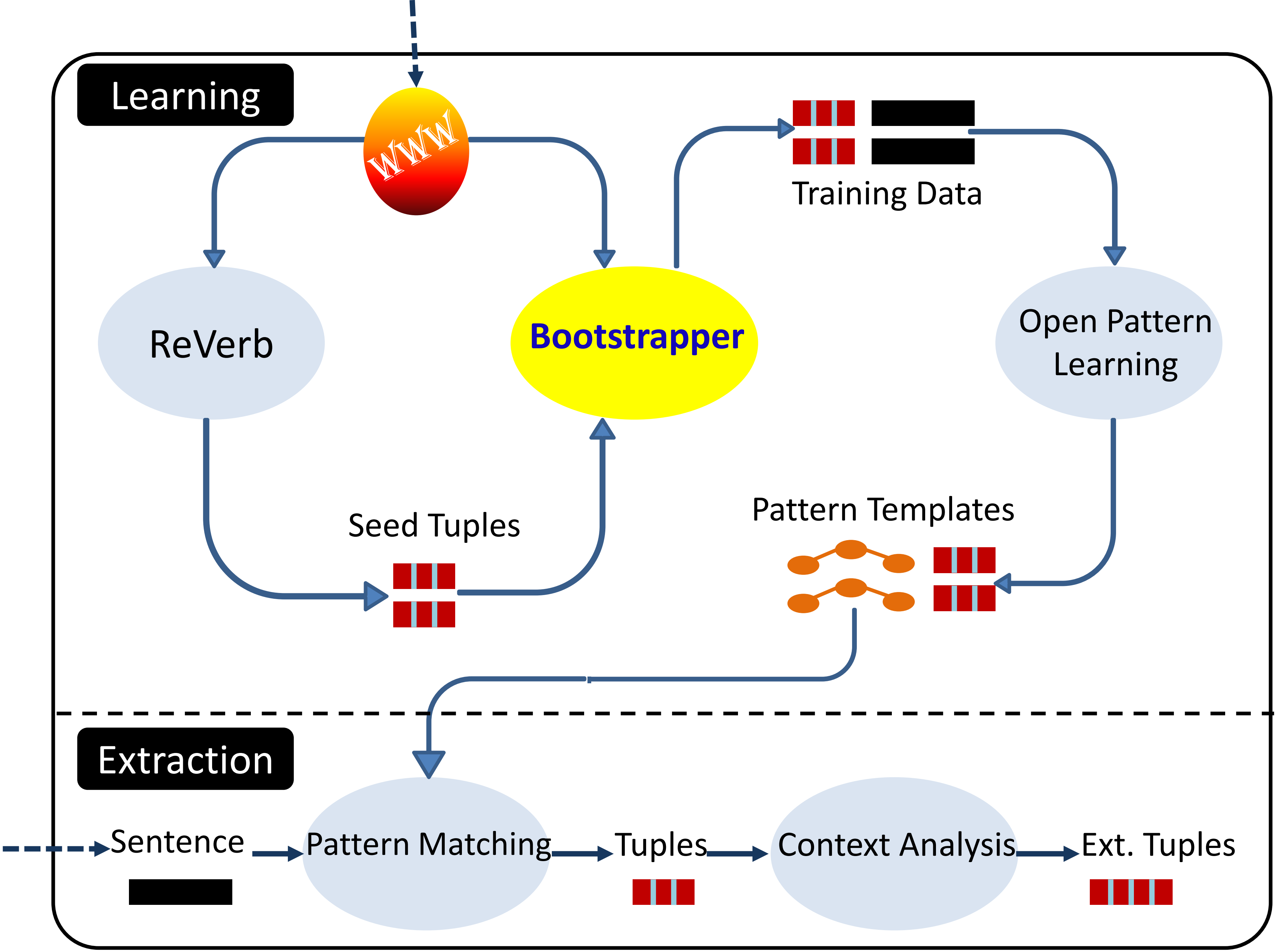
ReVerb: Error Analysis

- Larry Page, the CEO of Google, talks about multi-screen opportunities offered by Google.
- After winning the Superbowl, the Giants are now the top dogs of the NFL.
- Ahmadinejad was ***elected*** as the new President of Iran.

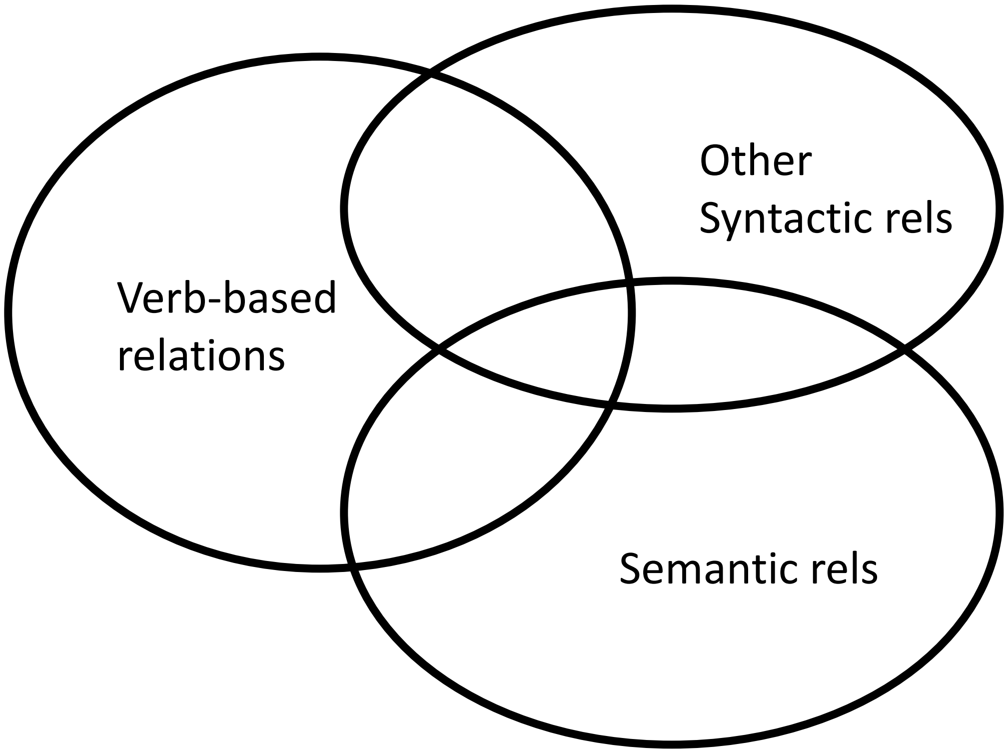
OLLIE: **O**pen **L**anguage **L**earning for **I**nformation **E**xtraction



Learning
WWW
Training Data
ReVerb
Bootstrapper
Open Pattern Learning
Seed Tuples
Pattern Templates
Extraction
Sentence
Pattern Matching
Tuples
Context Analysis
Ext. Tuples

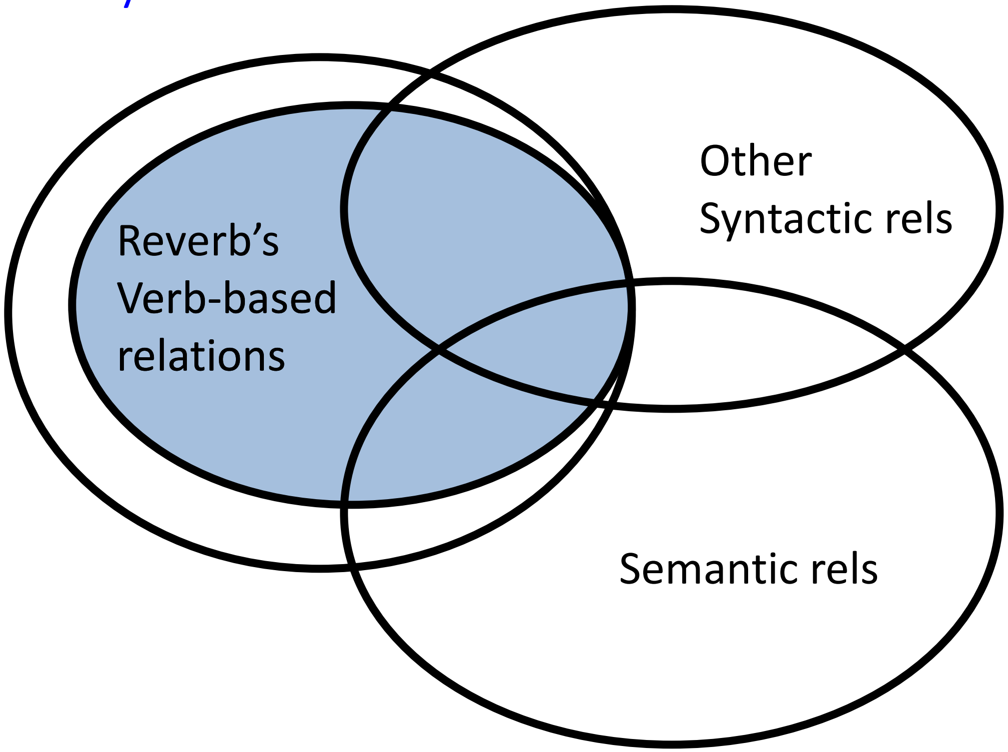

Learning
WWW
ReVerb
Bootstrapper
Training Data
Open Pattern Learning
Seed Tuples
Pattern Templates
Extraction
Sentence
Pattern Matching
Tuples
Context Analysis
Ext. Tuples

Bootstrapping Approach



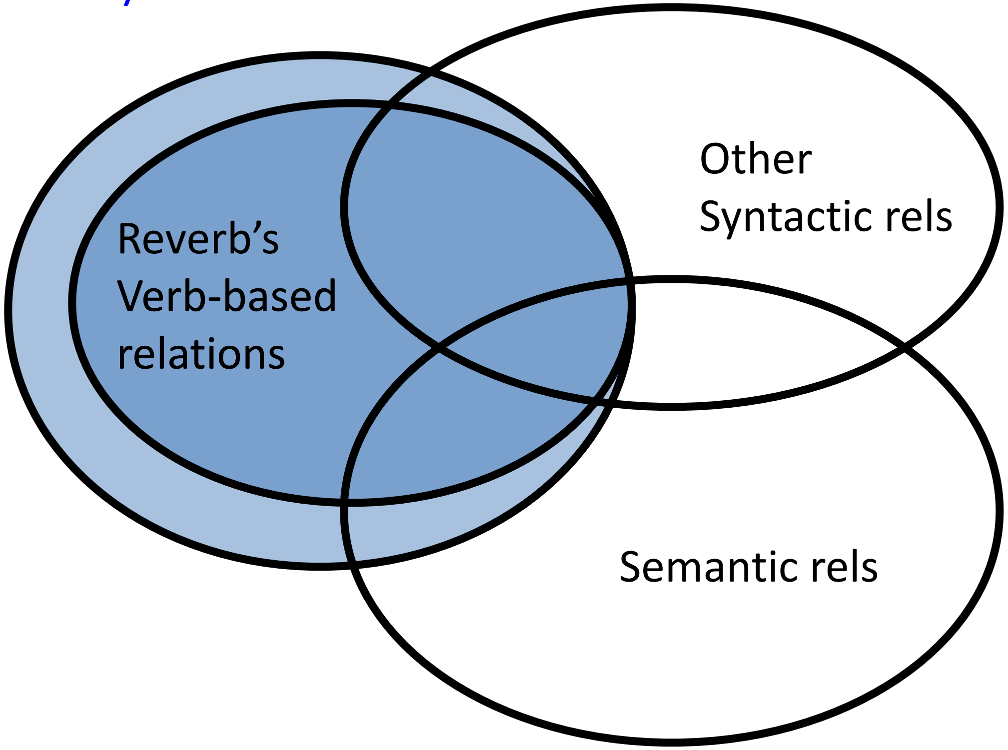
Bootstrapping Approach

Federer is coached by Paul Annacone.



Bootstrapping Approach

Federer is coached by Paul Annacone.

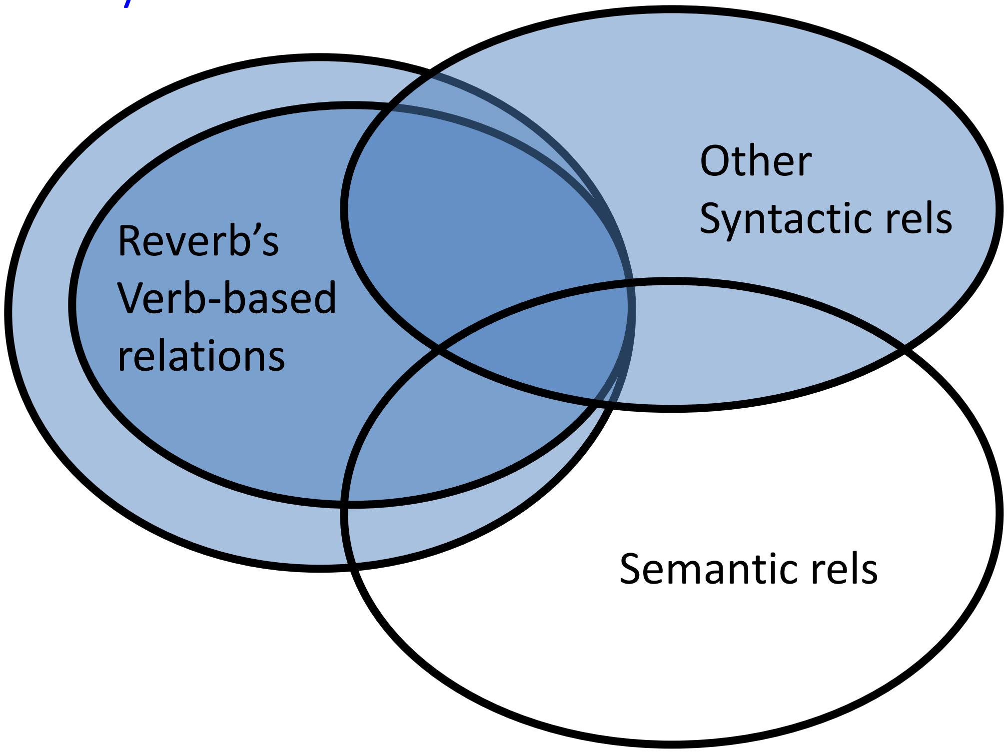


Now coached by Paul Annacone, Federer has …

Bootstrapping Approach

Federer is coached by Paul Annacone.

Paul Annacone, the coach of Federer,

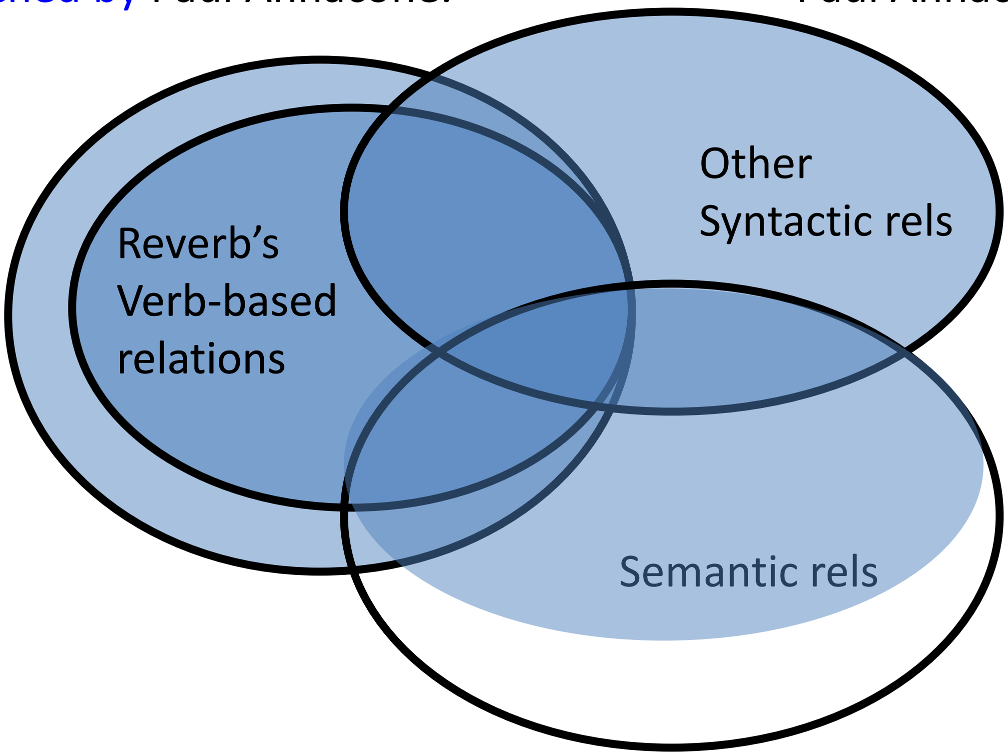


Now coached by Paul Annacone, Federer has …

Bootstrapping Approach

Federer is coached by Paul Annacone.

Paul Annacone, the coach of Federer,

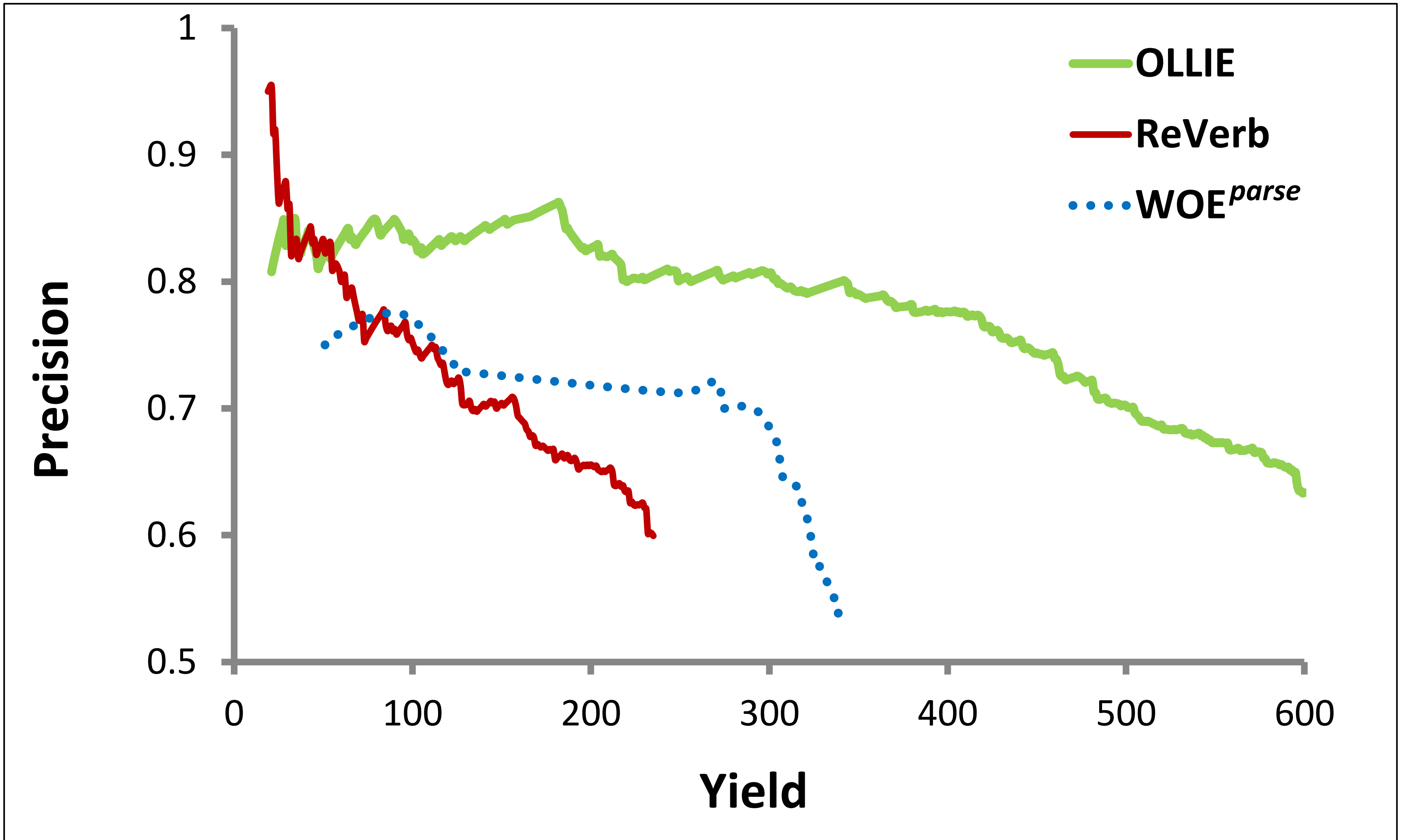


Now coached by Paul Annacone, Federer has ...

Federer *hired* Annacone as his new coach.

Evaluation

[Mausam, Schmitz, Bart, Soderland, Etzioni - EMNLP'12]



Context Analysis

"John refused to visit Vegas."

↓

(John, visit, Vegas)

"Early astronomers believed that the earth is the center of the universe."

↓

(earth, is the center of, universe)

"If she wins California, Hillary will be the nominated presidential candidate."

↓

(Hillary, will be nominated, presidential candidate)

Context Analysis

"John refused to visit Vegas."

↓

(John, refused to visit, Vegas)

"Early astronomers believed that the earth is the center of the universe."

↓

[(earth, is the center of, universe) Attribution: early astronomers]

"If she wins California, Hillary will be the nominated presidential candidate."

↓

[(Hillary, will be nominated, presidential candidate) Modifier: if she wins California]

Open Information Extraction

- 2007: Textrunner (~Open IE 1.0)
  - CRF and self-training
- 2010: ReVerb (~Open IE 2.0)
  - POS-based relation pattern
- 2012: OLLIE (~Open IE 3.0)
  - Dep-parse based extraction; nouns; attribution
- 2014: Open IE 4.0
  - SRL-based extraction; temporal, spatial…
- 2016 [@IITD]: Open IE 5.0
  - compound noun phrases, numbers, lists

increasing
precision,
recall,
expressiveness

Compound Noun Extraction

[Pal & Mausam - AKBC'16]

- NIH Director Francis Collins

  (Francis Collins, is the Director of, NIH)

- Challenges
  - New York Banker Association — ORG NAMES
  - German Chancellor Angela Merkel — DEMONYMS
  - Prime Minister Modi
  - GM Vice Chairman Bob Lutz

  COMPOUND RELATIONAL NOUNS

Numerical IE

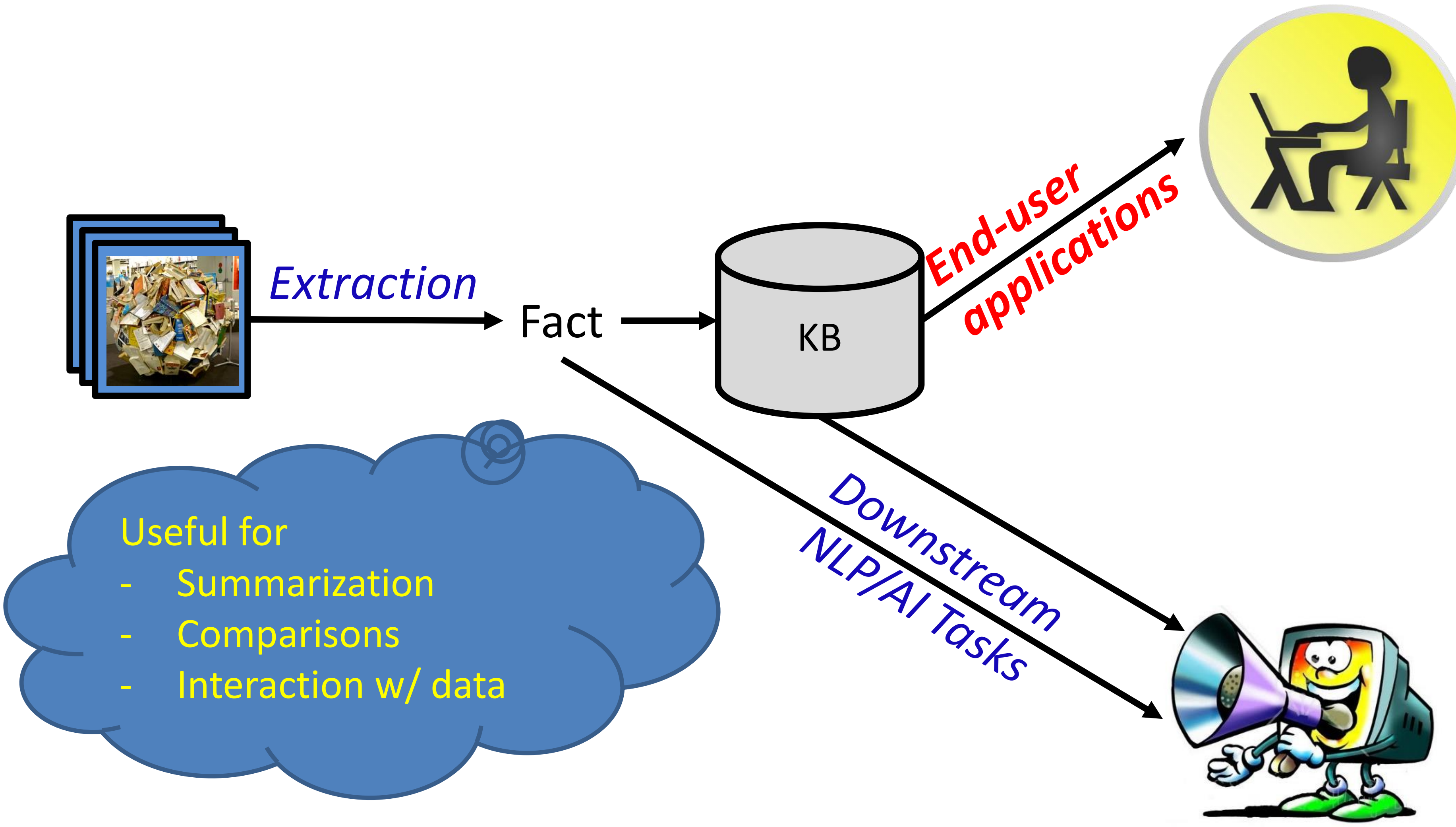
[Madaan, Mittal, Mausam, Ramakrishnan, Sarawagi AAAI'16]

"Venezuela with its inflation rate 96% is suffering from a major…"

↓ Numerical Open IE

(Venezuela, inflation rate, 96 %)

Overview



Extraction
Fact
KB
End-user applications
Downstream NLP/AI Tasks
Useful for
- Summarization
- Comparisons
- Interaction w/ data

Extractions: a great way to summarize

[redacted] (17)

[redacted] (is) operative (of) al-Qa'ida (3)
[redacted] (is) military chief (of) Al-Qaeda (1)
[redacted] (is) strategic planner (of) Al-Qaeda (2)
[redacted] (is) the military commander of Al-Qaeda (1)
[redacted] had married Abu'l-Walid 's eldest daughter (3)
[redacted] is also still listed on the FBI 's Most Wanted Terrorists list (1)
[redacted] is in Iran (1)
[redacted] remained in Pakistan (2)
[redacted] remains listed on the FBI 's list of Most Wanted Terrorists (1)
[redacted] represents the rebirth of Al-Qaeda (2)
[redacted] (is) coordinator (of) operations (1)
[redacted] gave his blessing to that attack (1)
[redacted] somehow gave the blessing for that (1)
[redacted] was apprehended by Iranian authorities (1)
[redacted] would be a major coup (1)
[redacted] has served as its security chief (1)
[redacted] illustrates his interests (3)

[redacted] (16)

Extractions: a great way to compare

[Contractor, Mausam, Singla - NAACL'16]

Cluster Labels	Granada (Spain)	New York City (U.S.)
art, arch.	moorish architecture religious art fine art beautiful architecture	contemporary art modern american art medieval art egyptian art
palace, courtyard	brick-walled courtyard lovely courtyard area nasrid royal palace alhambra palace	
museum, finest	alhambra museum archaeological museum world heritage site splendid arabic shops	fine art museums guggenheim museum islamic art collection metropolitan museum
gardens, park	partal gardens palace gardens pleasant gardens moorish style gardens	flushing meadows park central park renowned gardens natl. recreational area

Extractions: a great way to compare

[Contractor, Mausam, Singla - NAACL'16]

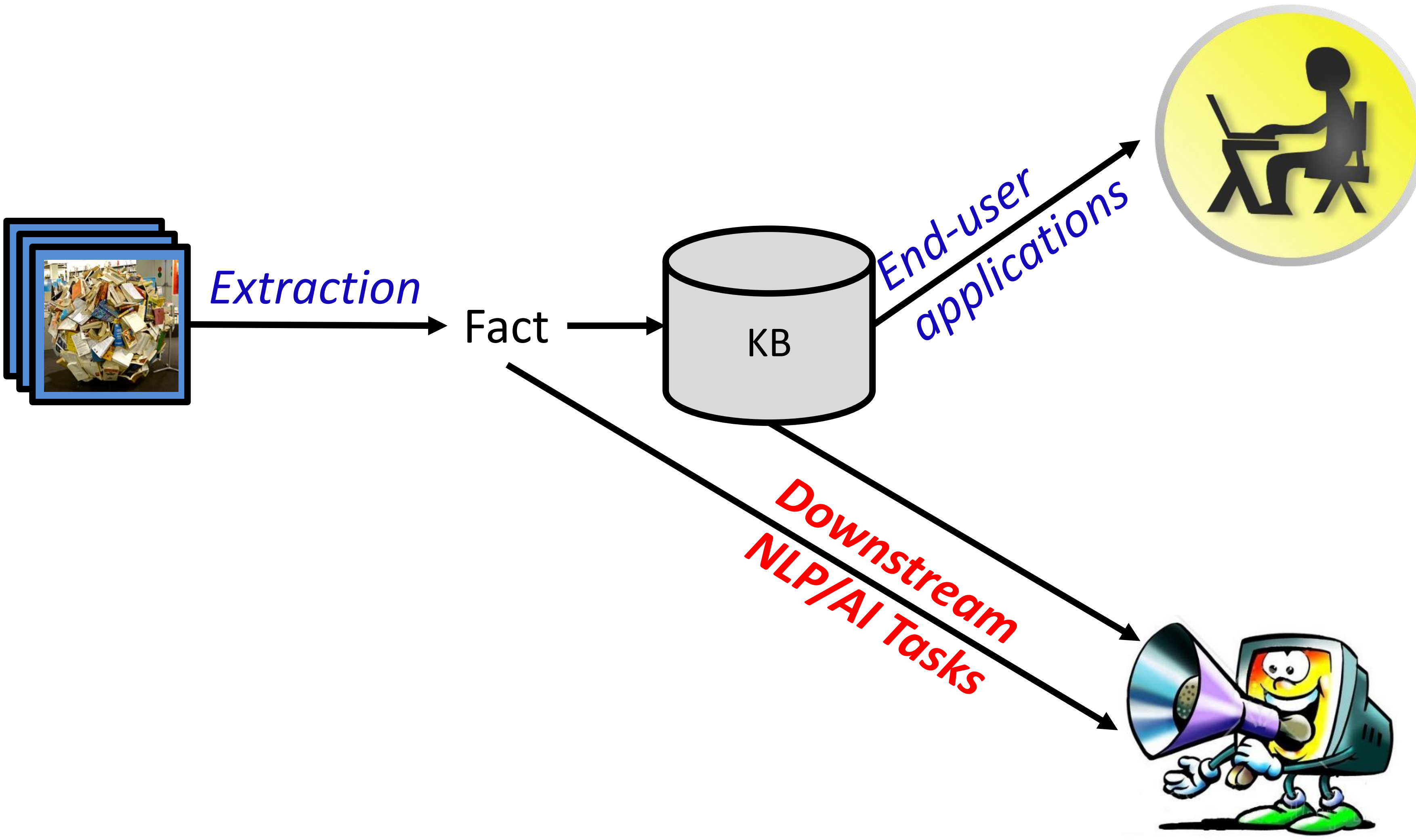
Cluster Labels	Granada (Spain)	New York City (U.S.)
art, arch.	moorish architecture religious art fine art beautiful architecture	contemporary art modern american art medieval art egyptian art
palace, courtyard	brick-walled courtyard lovely courtyard area nasrid royal palace alhambra palace	
museum, finest	alhambra museum archaeological museum world heritage site splendid arabic shops	fine art museums guggenheim museum islamic art collection metropolitan museum
gardens, park	partal gardens palace gardens pleasant gardens moorish style gardens	flushing meadows park central park renowned gardens natl. recreational area

Extractions: a great way to compare

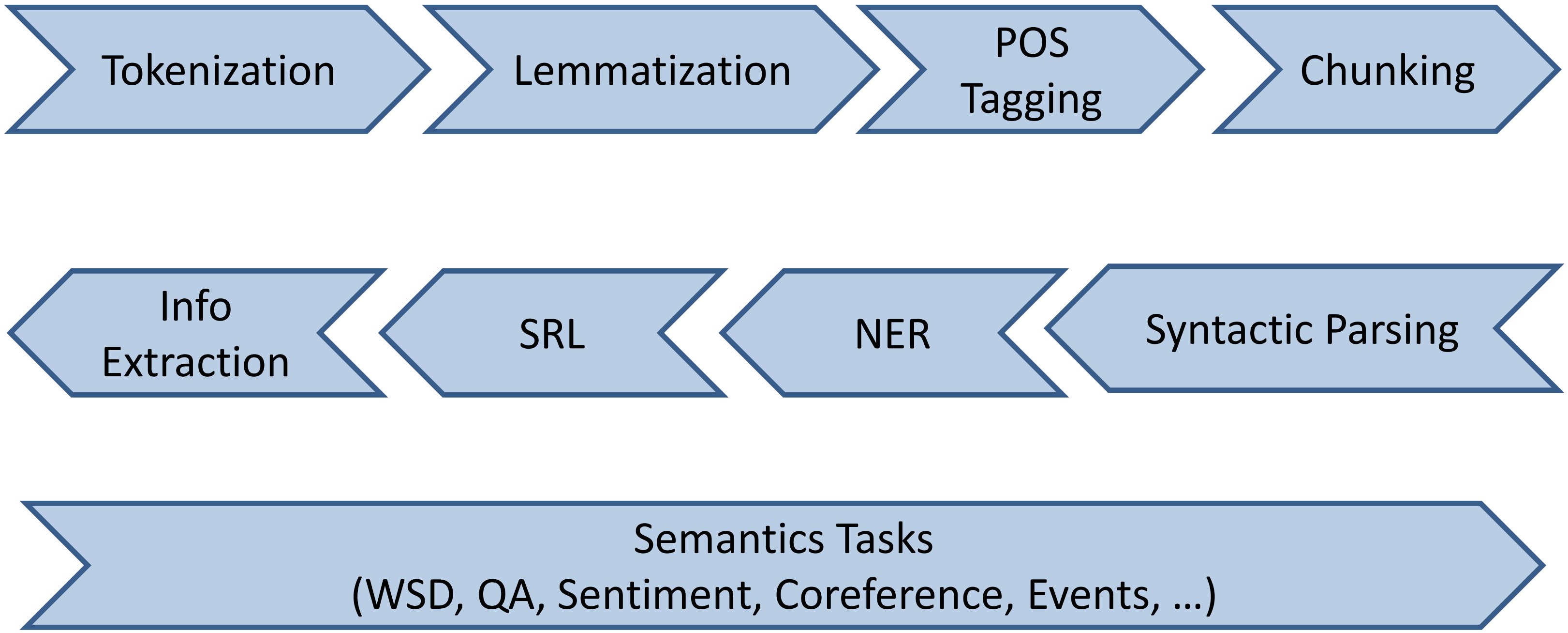
[Contractor, Mausam, Singla - NAACL'16]

Cluster Labels	Granada (Spain)	New York City (U.S.)
art, arch.	moorish architecture religious art fine art beautiful architecture	contemporary art modern american art medieval art egyptian art
palace, courtyard	brick-walled courtyard lovely courtyard area nasrid royal palace alhambra palace	
museum, finest	alhambra museum archaeological museum world heritage site splendid arabic shops	fine art museums guggenheim museum islamic art collection metropolitan museum
gardens, park	partal gardens palace gardens pleasant gardens moorish style gardens	flushing meadows park central park renowned gardens natl. recreational area

Overview


Extraction
Fact
KB
End-user
applications
Downstream
NLP/AI Tasks

Pipeline for Semantics Tasks

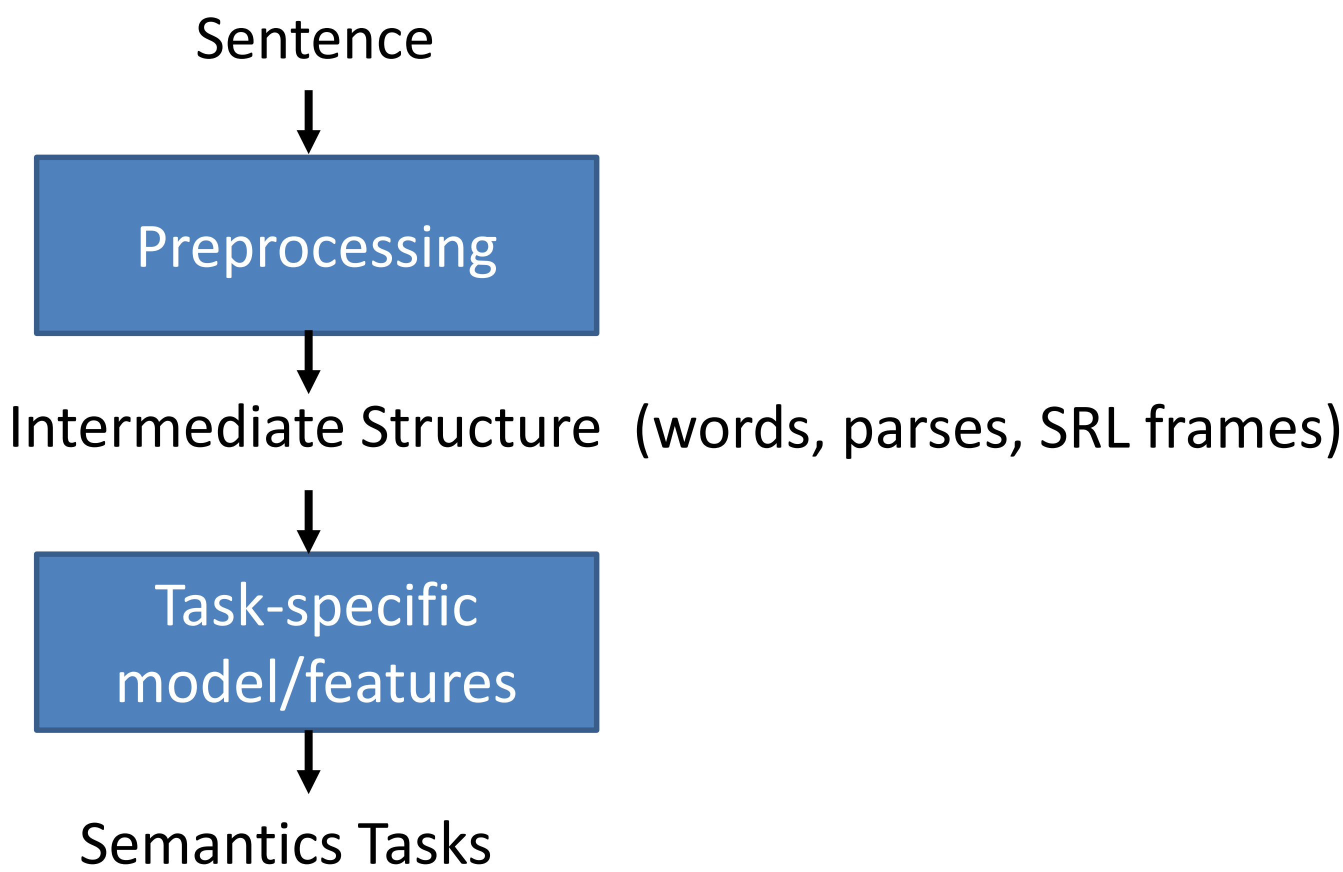


Pipeline for Semantics Tasks

Sentence Preprocessing

Semantics Tasks
(WSD, QA, Sentiment, Coreference, Events, …)

Abstracted Pipeline



Research Question

[Stanovsky, Dagan, Mausam - ACL'15]

Can Open IE be useful as an alternative intermediate structure?

Lexical Similarity/Analogies

- We experiment by switching **representations**
  - We compute Open IE based embeddings instead of lexical or syntactic context-based embeddings

"John refused to visit Vegas"

Target	Lexical	Dependency	SRL	Open IE
	John	nsubj_John	A0_John	0_John
	to	xcomp_visit	A1_to	1_to
refused	visit		A1_visit	1_visit
	Vegas		A1_Vegas	2_Vegas

Results

- Lexical similarity

Near-state-of-art for the amount of training data

	Open IE	Lexical	Deps	SRL
bruni	**.757**	.735	.618	.491
luong	**.288**	.229	.197	.171
radinsky	**.681**	.674	.592	.433
simlex	.39	.365	**.447**	.306
ws353-rel	**.647**	.64	.492	.551
ws353-sym	**.77**	.763	.759	.439
ws353-full	**.711**	.703	.629	.693

Functional similarity

Results

- Lexical analogy   a:a* :: b:b*?

	Open IE	Lexical	Deps	SRL
Google (Add)	**0.714**	0.651	0.34	0.352
Google (Mult)	**0.729**	0.656	0.367	0.362
MSR (Add)	**0.529**	0.438	0.4	0.389
MSR (Mult)	**0.55**	0.455	0.434	0.406

State of the art

$$\arg\max_{b^* \in V} \left(\cos\left(b^*, b\right) - \cos\left(b^*, a\right) + \cos\left(b^*, a^*\right)\right)$$

$$\arg\max_{b^* \in V} \frac{\cos\left(b^*, b\right) \cos\left(b^*, a^*\right)}{\cos\left(b^*, a\right) + \varepsilon}$$

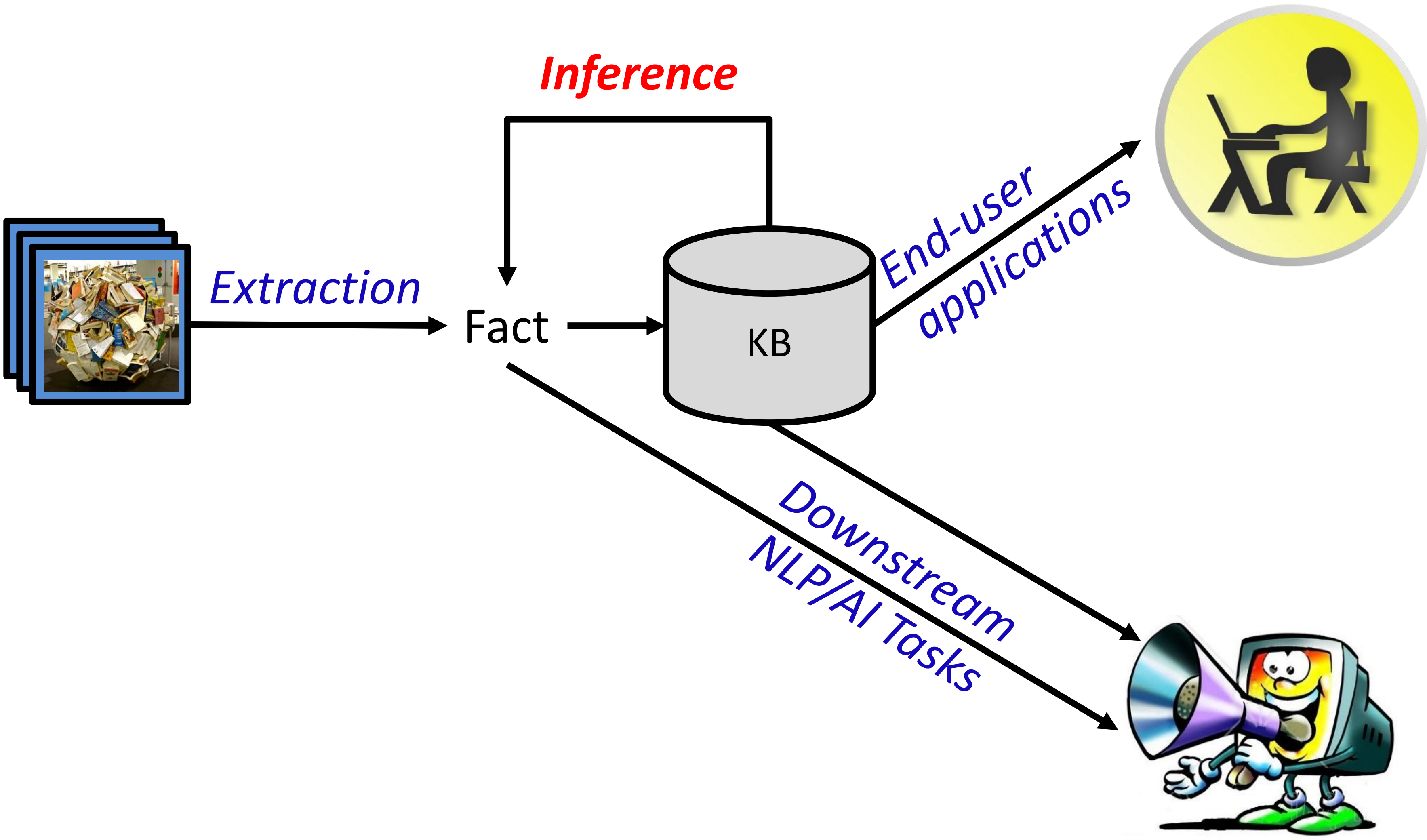
Why does Open IE do better?

- Word Analogy
  - Captures domain and functional similarity (***gentlest***: gentler), (***loudest***:?)
    - Lexical: higher-pitched X **[Domain Similar]**
    - Syntactic: thinnest X **[Functionally Similar]**
    - SRL: unbelievable X **[Functionally Similar?]**
    - Open-IE: louder ✓

Other NLP Applications

- Atomic relations → Event schemas
- Unsupervised sentence similarity
- Reading comprehension tasks
- Open IE → Closed IE
  - Software (OREO): http://homes.cs.washington.edu/~mausam/software.html

Future Work


Inference
Extraction
Fact
KB
End-user applications
Downstream NLP/AI Tasks

Key Future Direction

- Large-scale inference over Open IE

(iron, is a good conductor of, electricity)

↓

(iron nail, conducts, electricity)

(David Beckham, was born in, London)

↓

(David Beckham, was born in, England)

Thanks

Google

IBM Research

DARPA

Bloomberg

DEPARTMENT OF THE NAVY
USN
ONR
Science & Technology

Visvesvaraya PhD Scheme for Electronics and IT

KiSTi
www.kisti.re.kr

IARPA
BE THE FUTURE