Confidential

Open Information Extraction: Approaches and Applications

- Mausam
- *Professor, Computer Science.*
  Head, School of Artificial Intelligence.
  Indian Institute of Technology, Delhi

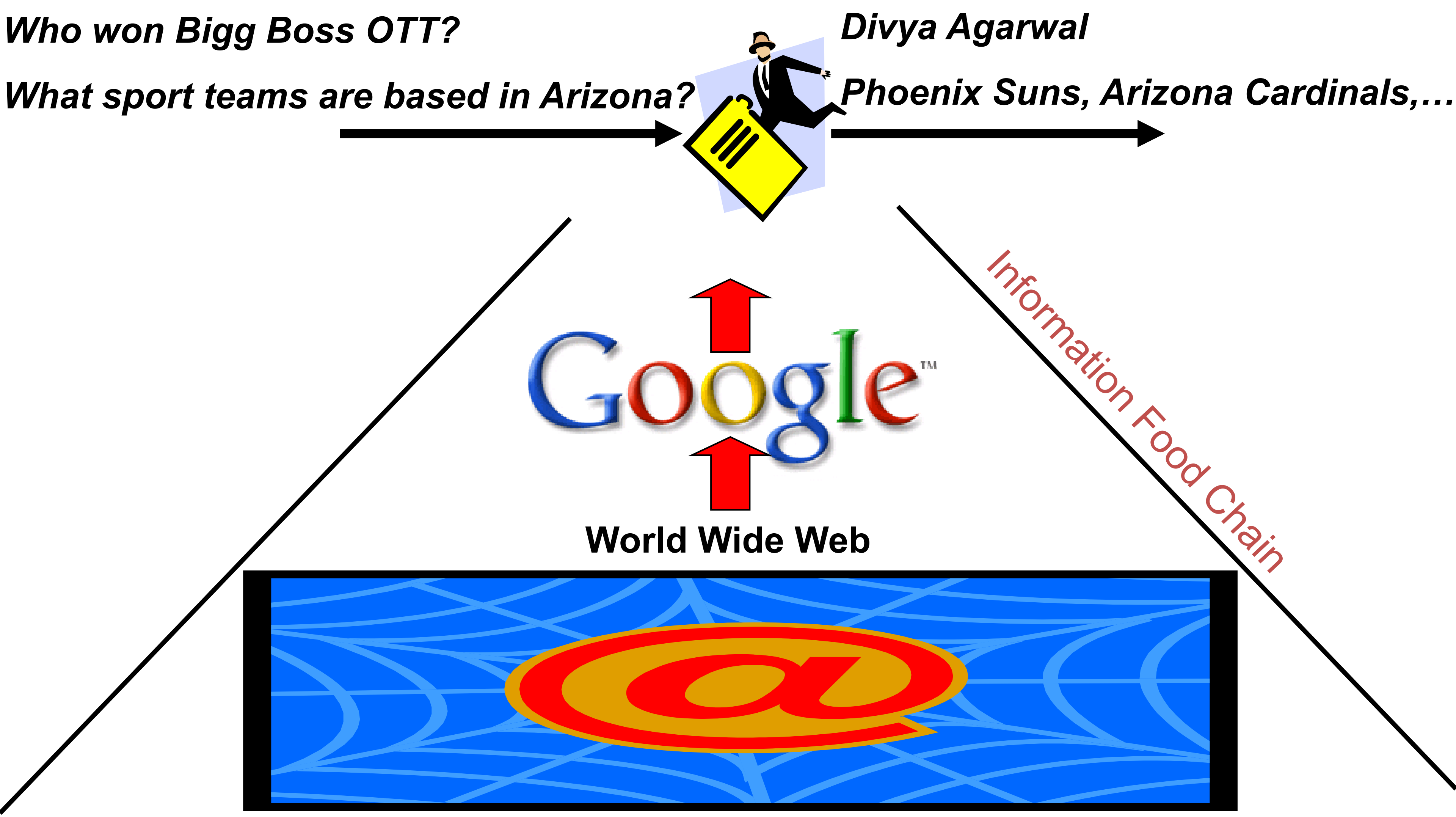
- Keshav Kolluru
- *PhD Scholar*
  Indian Institute of Technology, Delhi

"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)

Paradigm Shift: from retrieval to reading



Paradigm Shift: from retrieval to reading

Science Report

Finding: beer that doesn't give a hangover

Researcher: Ben Desbrow

Country: Australia

Organization: Griffith Health Institute

Paradigm Shift: from retrieval to reading

Compare Roku vs Fire

most apps but not iTunes	*most apps but not Vudu, iTunes*
remote	*voice-controlled remote*
good UI	*good UI*
works perfectly	*blames router*
needs laptop during travel	*connects easily during travel*

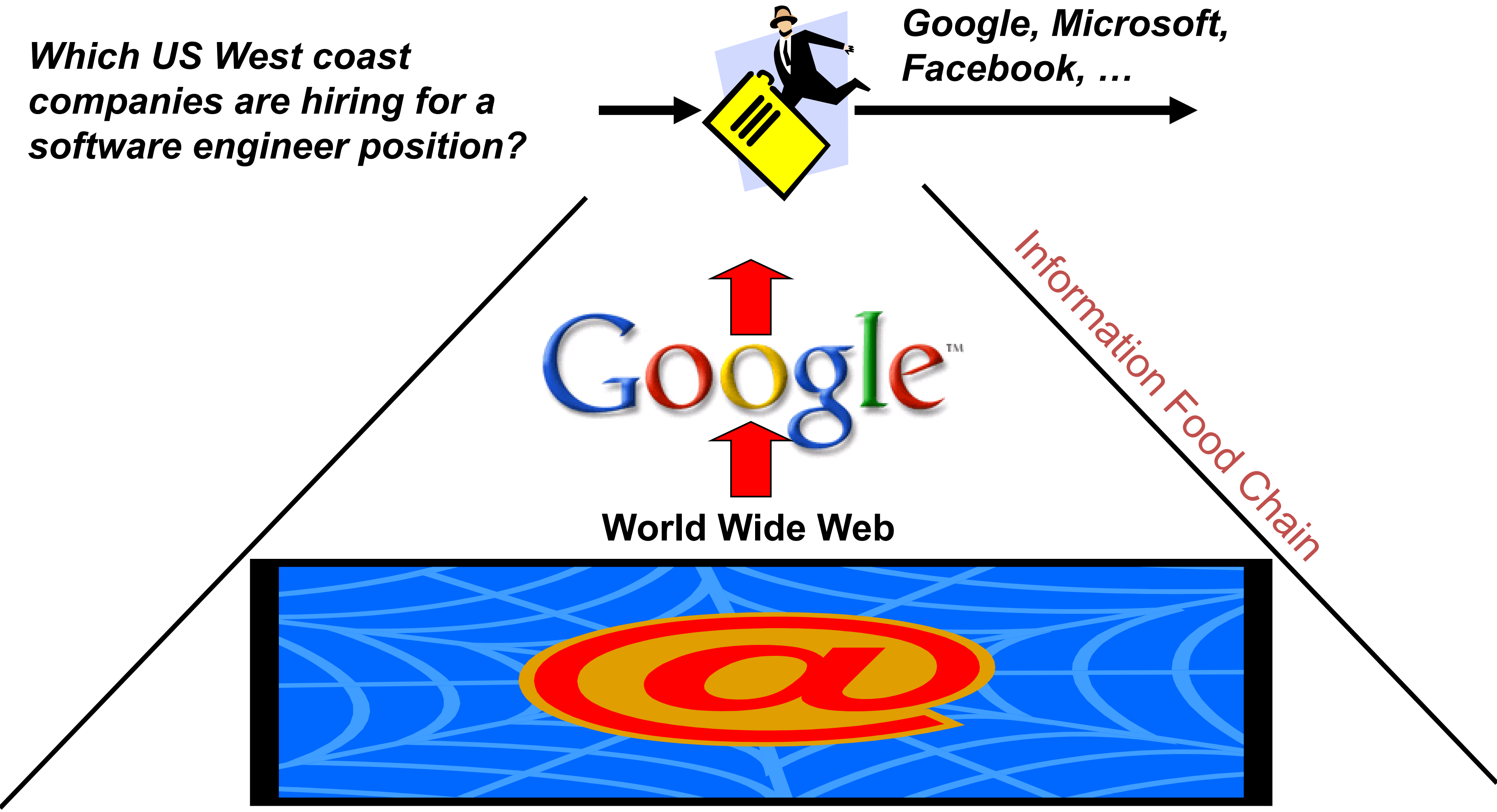
Google

Food Chain

World Wide Web

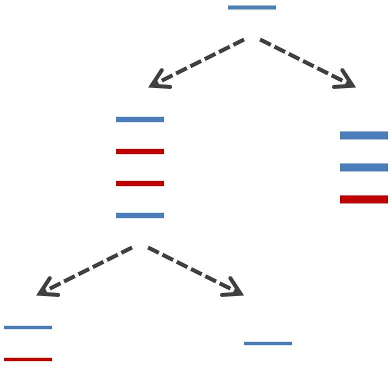
Paradigm Shift: from retrieval to reading

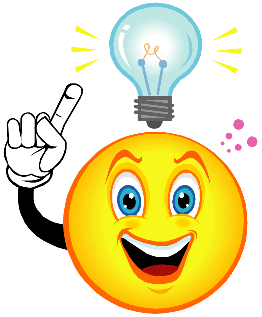
Which US West coast companies are hiring for a software engineer position?



Information Systems Pipeline

Data ⟶ Information ⟶ Knowledge ⟶ Wisdom

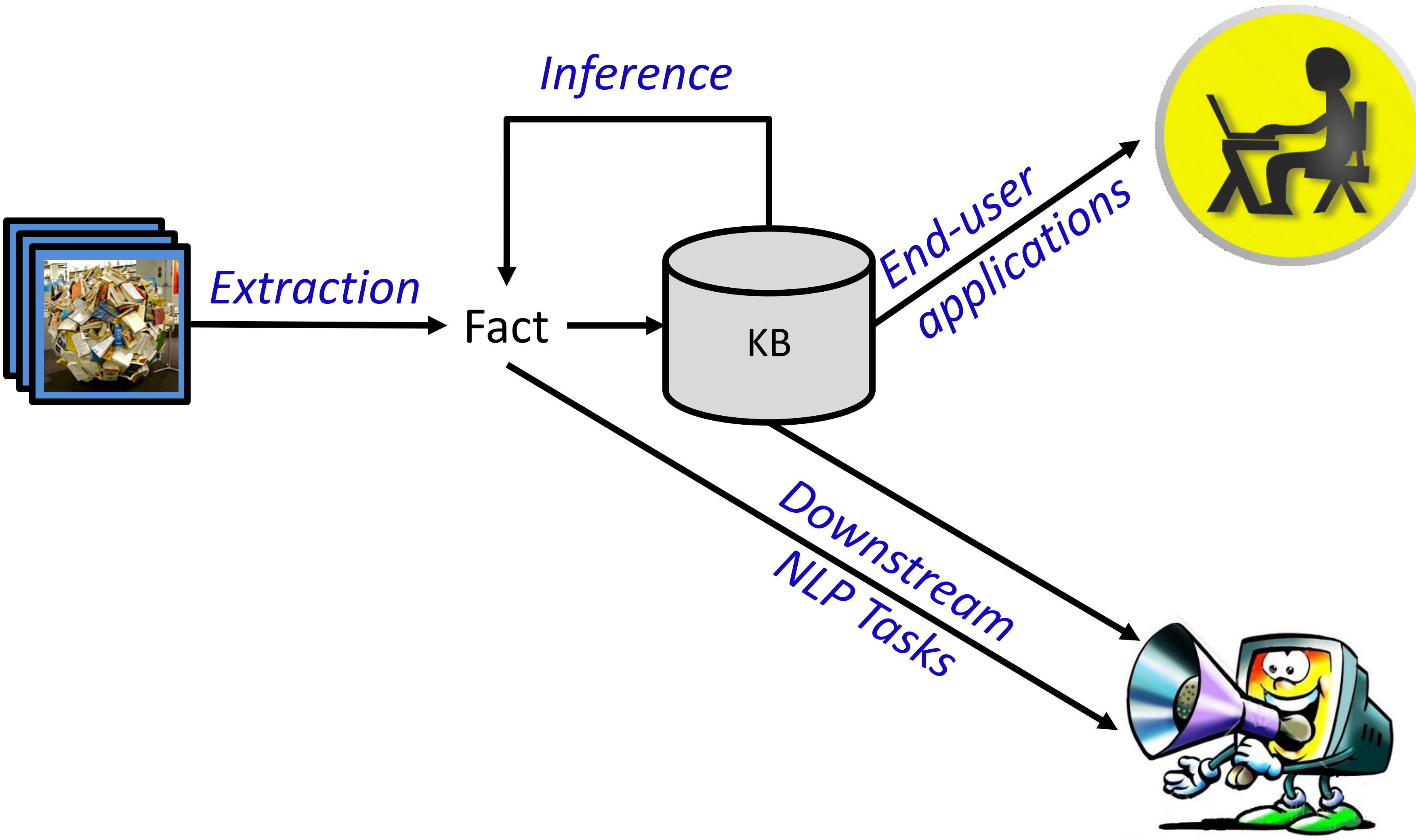




Text ⟶ Facts ⟶ Knowledge Base ⟶ Applications

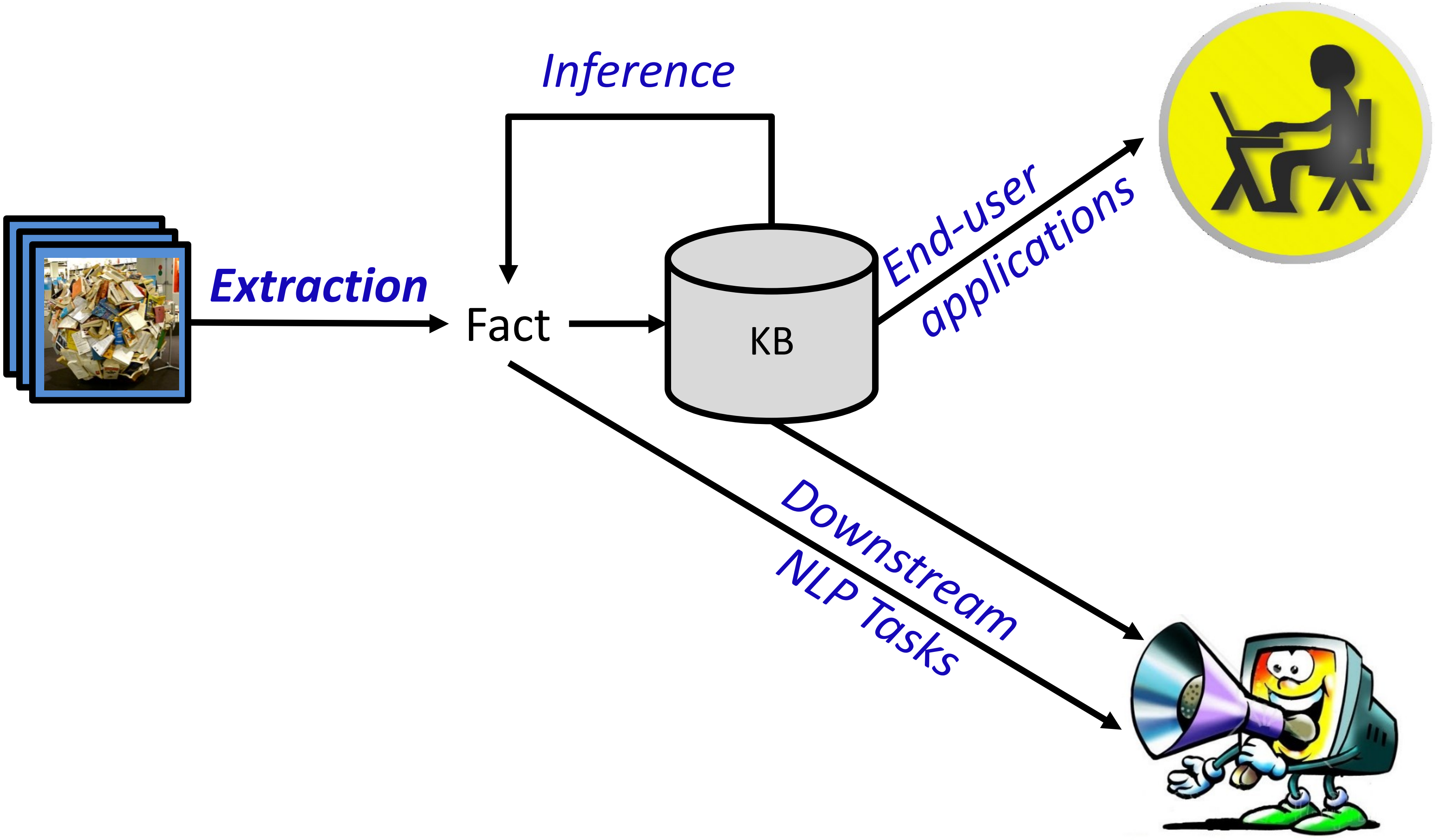
NASSCOM

Research Overview



NASSCOM

Research Overview



Closed Information Extraction

Extracting 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
(Google, Larry Page)
(Apple, Steve Jobs)
(Microsoft, Bill Gates)
…

rel:acquisition
(Google, DeepMind)
(Apple, Shazam)
(Microsoft, Maluuba)
…

Open Information Extraction

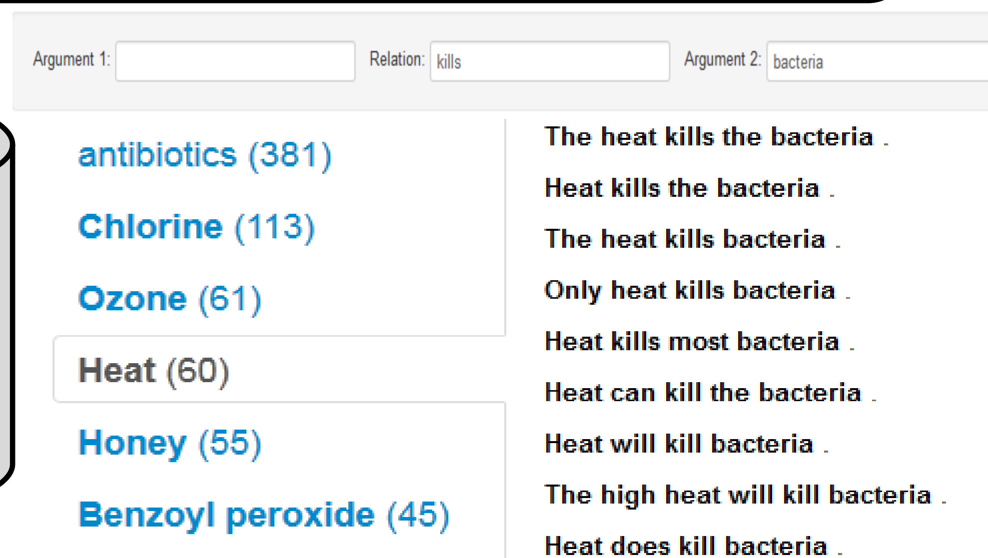
Extracting information from natural language text for *all* relations in *all* domains in a *few* passes.

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

↓ Open IE

(Steve Jobs, be the founder of, Apple), (Steve Jobs, died of, cancer)

(Google, acquired, DeepMind)
(Oranges, contain, Vitamin C)
(Edison, invented, phonograph)
…



Open Information Extraction

Extracting information from natural language text

for *a* ... ses.

"Ap ... "

(Ste

(Google, a

(Oranges, contain, V ... C)

(Edison, invented, phonograph)

...

Ontology Free!

nent 2: bacteria

Heat (60)

Honey (55)

Benzoyl peroxide (45)

the bacteria .

bacteria .

s bacteria .

at kills bacteria .

Heat kills most bacteria .

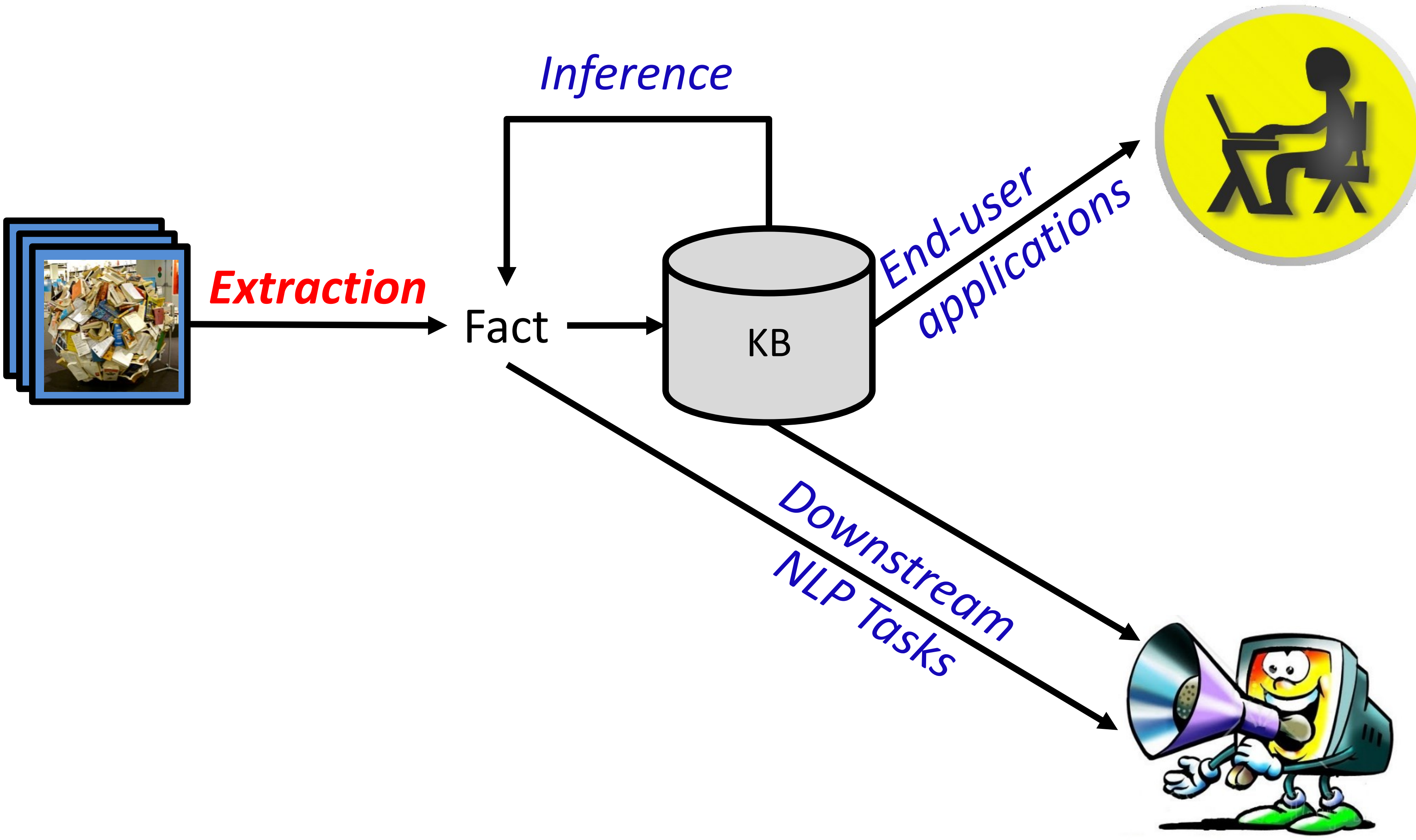
Heat can kill the bacteria .

Heat will kill bacteria .

The high heat will kill bacteria .

Heat does kill bacteria .

Overview


Inference
Extraction
Fact
KB
End-user
applications
Downstream
NLP Tasks

Demo

- http://openie.allenai.org

NASSCOM

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…
- 2017 [@IITD]: Open IE 5.0
  - compound noun phrases, numbers, lists
- 2020 [@IITD]: Open IE 6.0
  - deep neural models

training data automatically generated

taking a stronger ML leap

increasing precision, recall, expressiveness

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…
- 2017 [@IITD]: Open IE 5.0
  - compound noun phrases, numbers, lists
- 2020 [@IITD]: Open IE 6.0
  - deep neural models

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:

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

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

Lexical Constraint

Problem: "overspecified" relation phrases

Obama is offering only modest greenhouse gas reduction targets at the conference.

Solution: must have many distinct args in a large corpus

is offering only modest …

Obama the conference ≈ 1

is the patron saint of

100s ≈

Anne	mothers
George	England
Hubbins	quality footwear
….	

Number of Relations (circa 2011)

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 Extraction Algorithm

1. Identify longest relation phrases satisfying constraints

Hudson was born in Hampstead, which is a suburb of London.

arg1 arg2 arg1 arg2

2. Heuristically identify arguments for reach relation phrase

(Hudson, was born in, Hampstead)

(Hampstead, is a suburb of, London)

ReVerb: Error Analysis

- Steve Squeri, the CEO of American Express, said that a majority of employees will work from home
- After winning the Superbowl, the Giants are now the top dogs of the NFL.
- Ahmadinejad was ***elected*** as the new President of Iran.

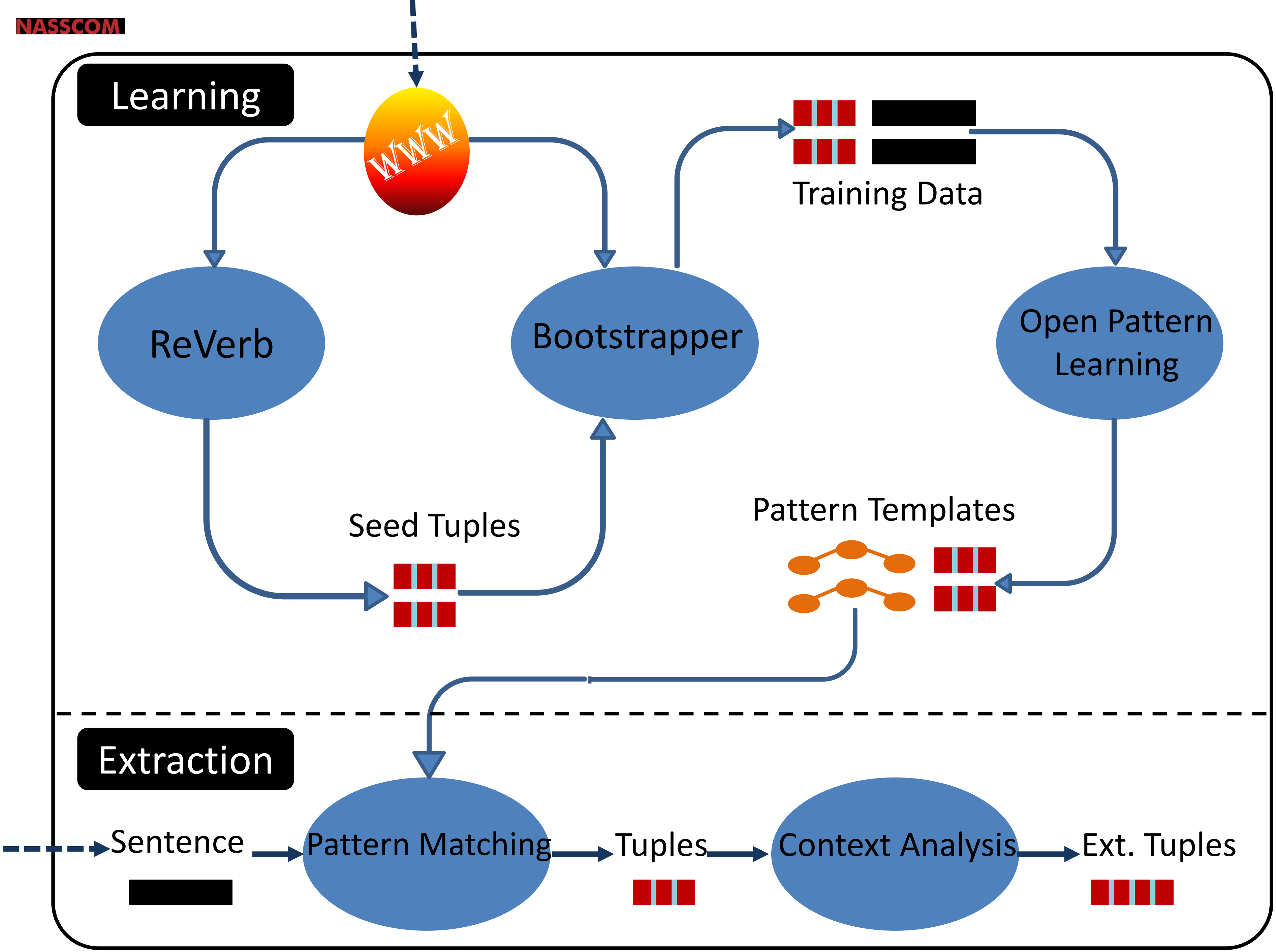
OLLIE: **O**pen **L**anguage **L**earning for **I**nformation **E**xtraction

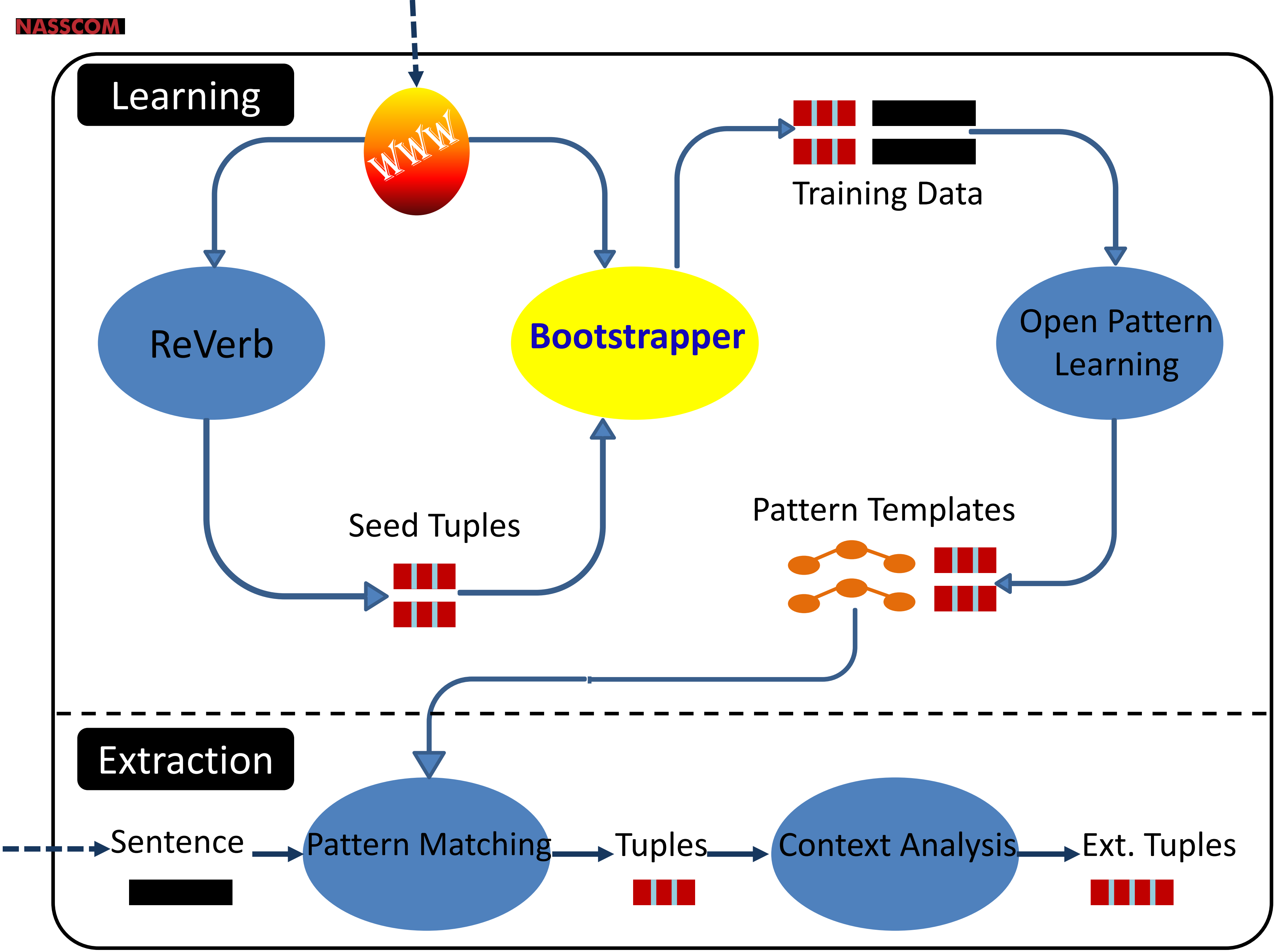
Open Information Extraction

- 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…
- 2017 [@IITD]: Open IE 5.0
  - compound noun phrases, numbers, lists
- 2020 [@IITD]: Open IE 6.0
  - deep neural models

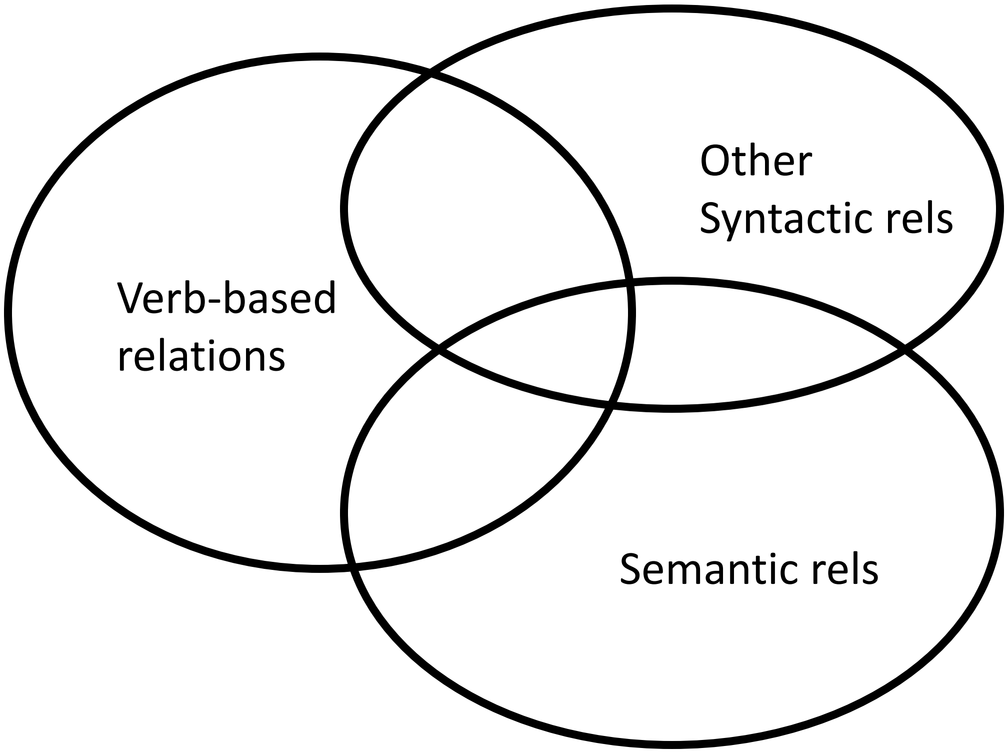
training data automatically generated

increasing precision, recall, expressiveness


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

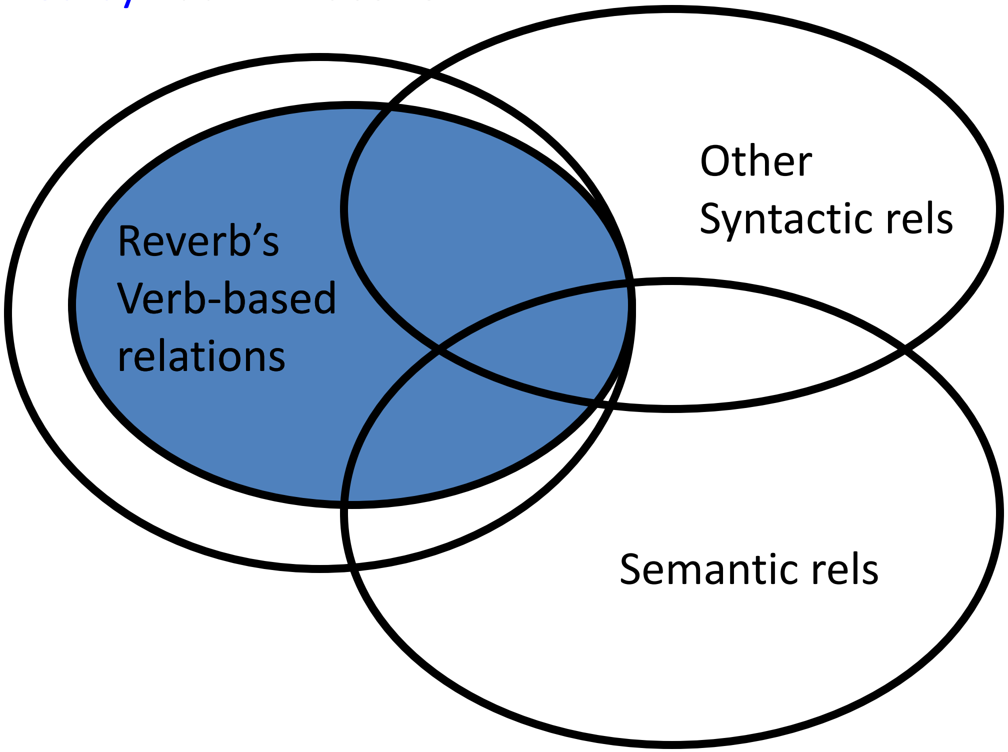

Learning
WWW
Training Data
ReVerb
Bootstrapper
Open Pattern Learning
Seed Tuples
Pattern Templates
Extraction
Sentence
Pattern Matching
Tuples
Context Analysis
Ext. Tuples

Bootstrapping Approach



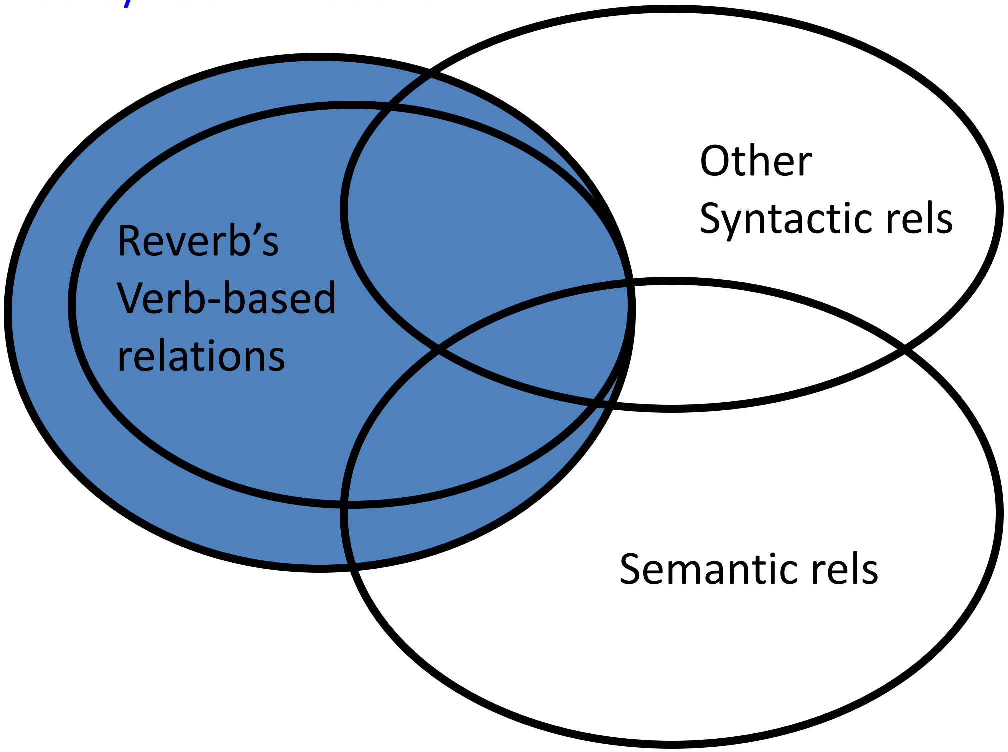
Bootstrapping Approach

Federer is coached by Paul Annacone.



Bootstrapping Approach

Federer is coached by Paul Annacone.

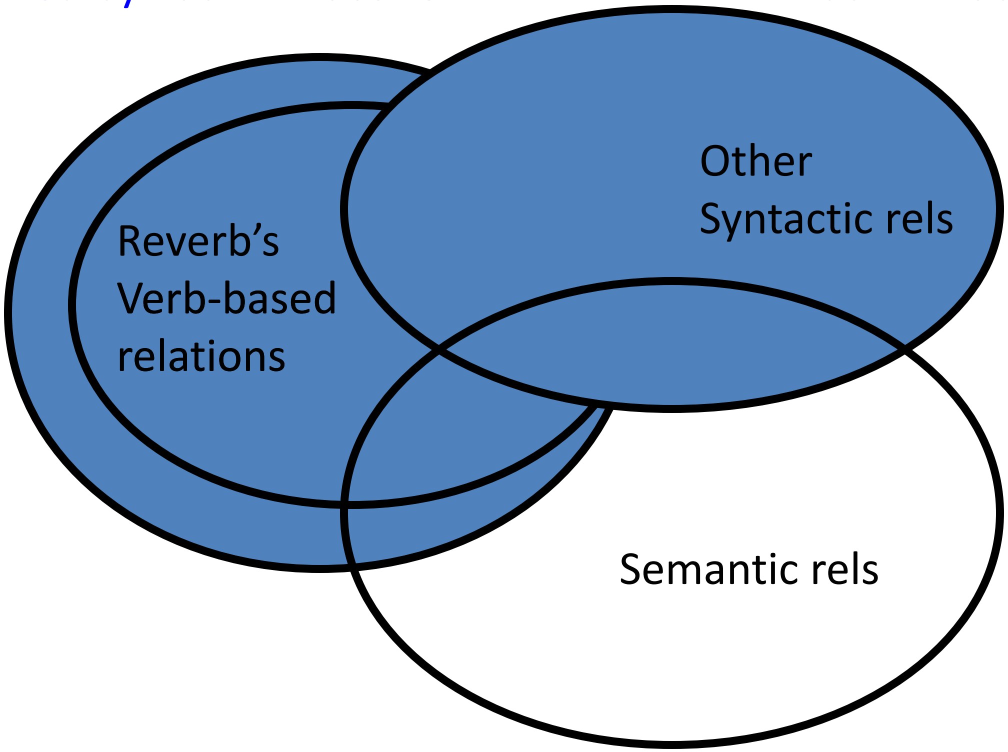


Now coached by Paul Annacone, Federer has …

Bootstrapping Approach

Federer is coached by Paul Annacone.

Paul Annacone, the coach of Federer,

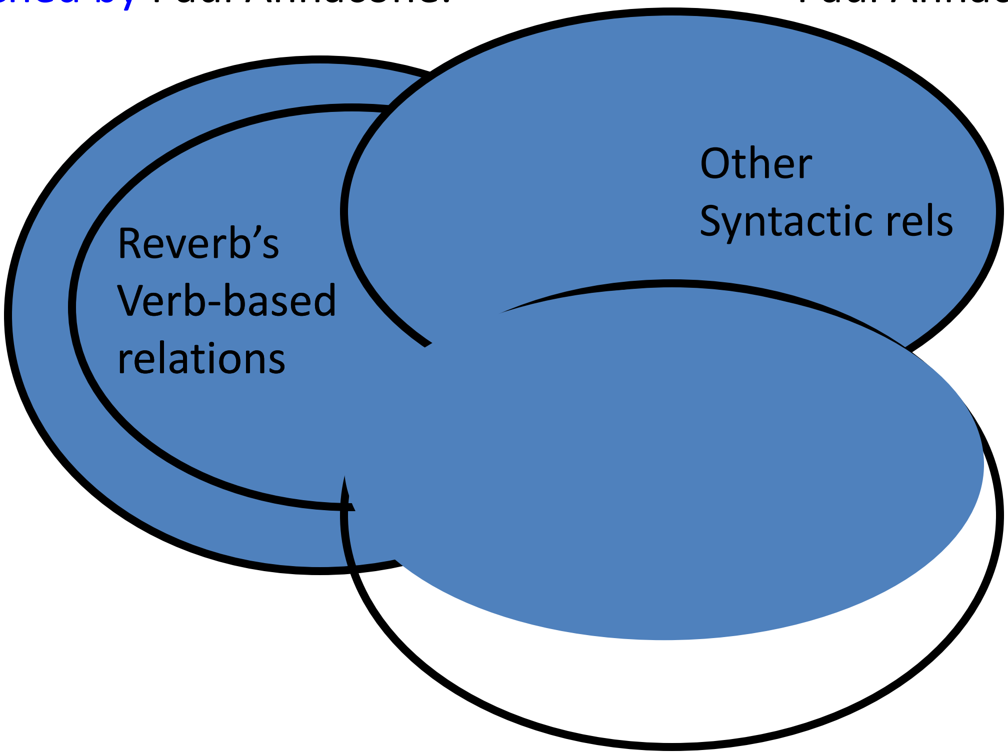


Now coached by Paul Annacone, Federer has …

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.

Bootstrapping

High Quality ReVerb Extractions

Extraction Lemmas (seeds)

Web Sentences

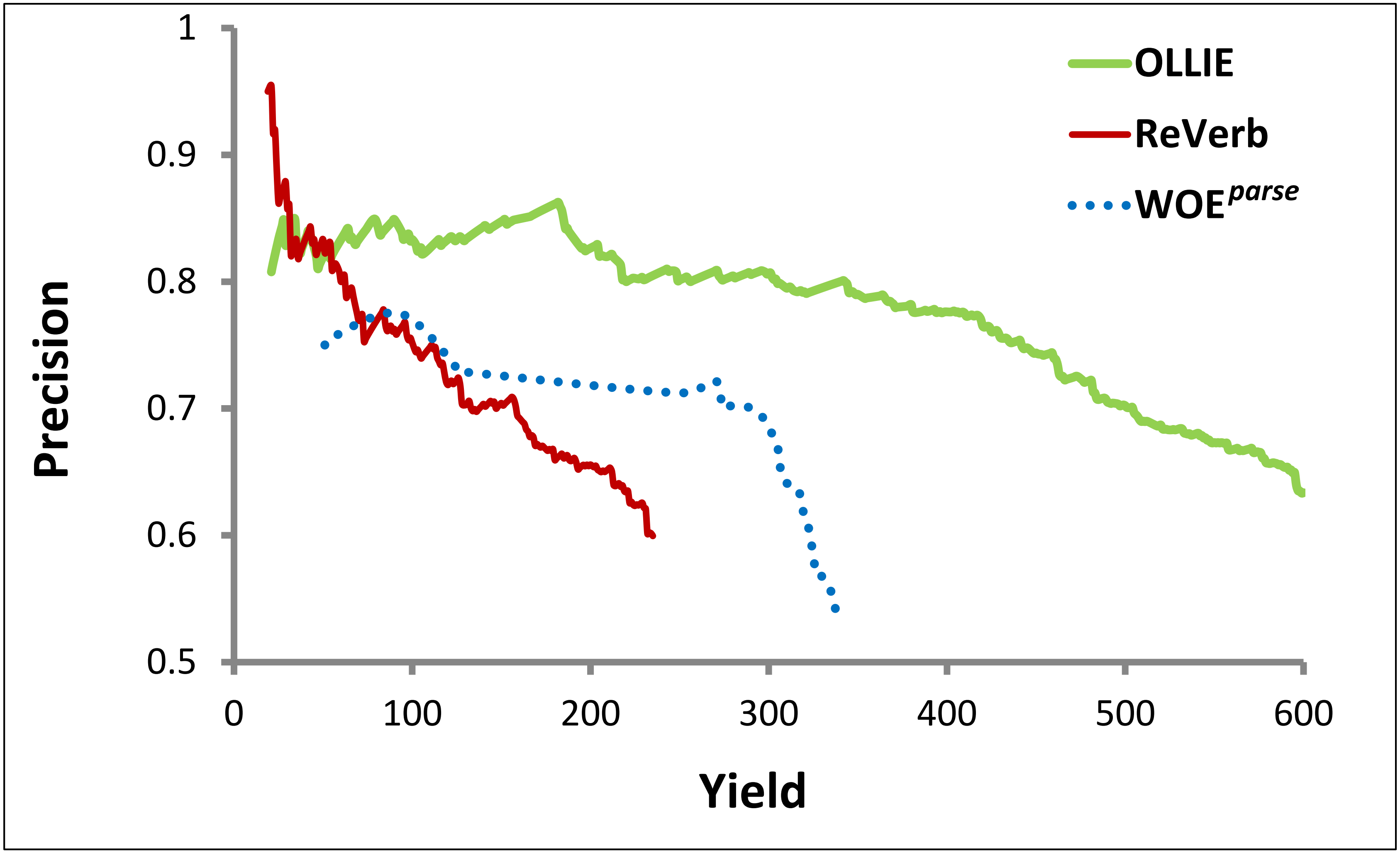
(Ahmadinejad, is the current president of, Iran)

ahmadinejad, president, iran

Ahmadinejad, who is the president of Iran, is a puppet for the Ayatollahs.

Evaluation

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



NASSCOM

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…
- 2017 [@IITD]: Open IE 5.0
  - compound noun phrases, numbers, lists
- 2020 [@IITD]: Open IE 6.0
  - deep neural models

increasing precision, recall, expressiveness

RelNoun: Nominal Open IE

Constructions	Phrase	Extraction
Verb1	Francis Collins is the director of NIH	(Francis Collins; is the director of; NIH)
Verb2	the director of NIH is Francis Collins	(Francis Collins; is the director of; NIH)
Appositive1	Francis Collins, the director of NIH	(Francis Collins; [is] the director of; NIH)
Appositive2	the director of NIH, Francis Collins,	(Francis Collins; [is] the director of; NIH)
Appositive3	Francis Collins, the NIH director	(Francis Collins; [is] the director [of]; NIH)
AppositiveTitle	Francis Collins, the director,	(Francis Collins; [is]; the director)
CompoundNoun	*NIH director Francis Collins*	*(Francis Collins; [is] director [of]; NIH)*
Possessive	NIH's director Francis Collins	(Francis Collins; [is] director [of]; NIH)
PossessiveAppositive	NIH's director, Francis Collins	(Francis Collins; [is] director [of]; NIH)
AppositivePossessive	Francis Collins, NIH's director	(Francis Collins; [is] director [of]; NIH)
PossessiveVerb	NIH's director is Francis Collins	(Francis Collins; is director [of]; NIH)
VerbPossessive	Francis Collins is NIH's director	(Francis Collins; is director [of]; NIH)

Compound Noun Extraction Baseline

- 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

Continuing with Fundamental Hypothesis

- Rule-based system to characterize relational noun phrases
  - Classifies and filters orgs
  - List of demonyms for location conversion
  - Bootstrap a list of relational noun *prefixes*
    - vice, ex, health, …

Experiments

[Pal & Mausam AKBC'16]

System	Precision	Yield
OLLIE-NOUN	0.29	136
RELNOUN 1.1	0.53	60
+ Compound Noun Baseline	0.37	100
+ ORG filtering	0.39	100
+ demonyms	0.52	158
RelNoun 2.0 → + compound relational nouns	**0.69**	**209**

NASSCOM

Numerical Open IE

[Saha, Pal, Mausam ACL'17]

"Hong Kong's labour force is 3.5 million."

Open IE 4: (Hong Kong's labour force, is, 3.5 million)

Open IE 5: (Hong Kong, has labour force of, 3.5 million)

"James Valley is nearly 600 metres long."

Open IE 4: (James Valley, is, nearly 600 metres long)

Open IE 5: (James Valley, has length of, nearly 600 metres)

"James Valley has 5 sq kms of fruit orchards."

Open IE 4: (James Valley, has, 5 sq kms of fruit orchards)

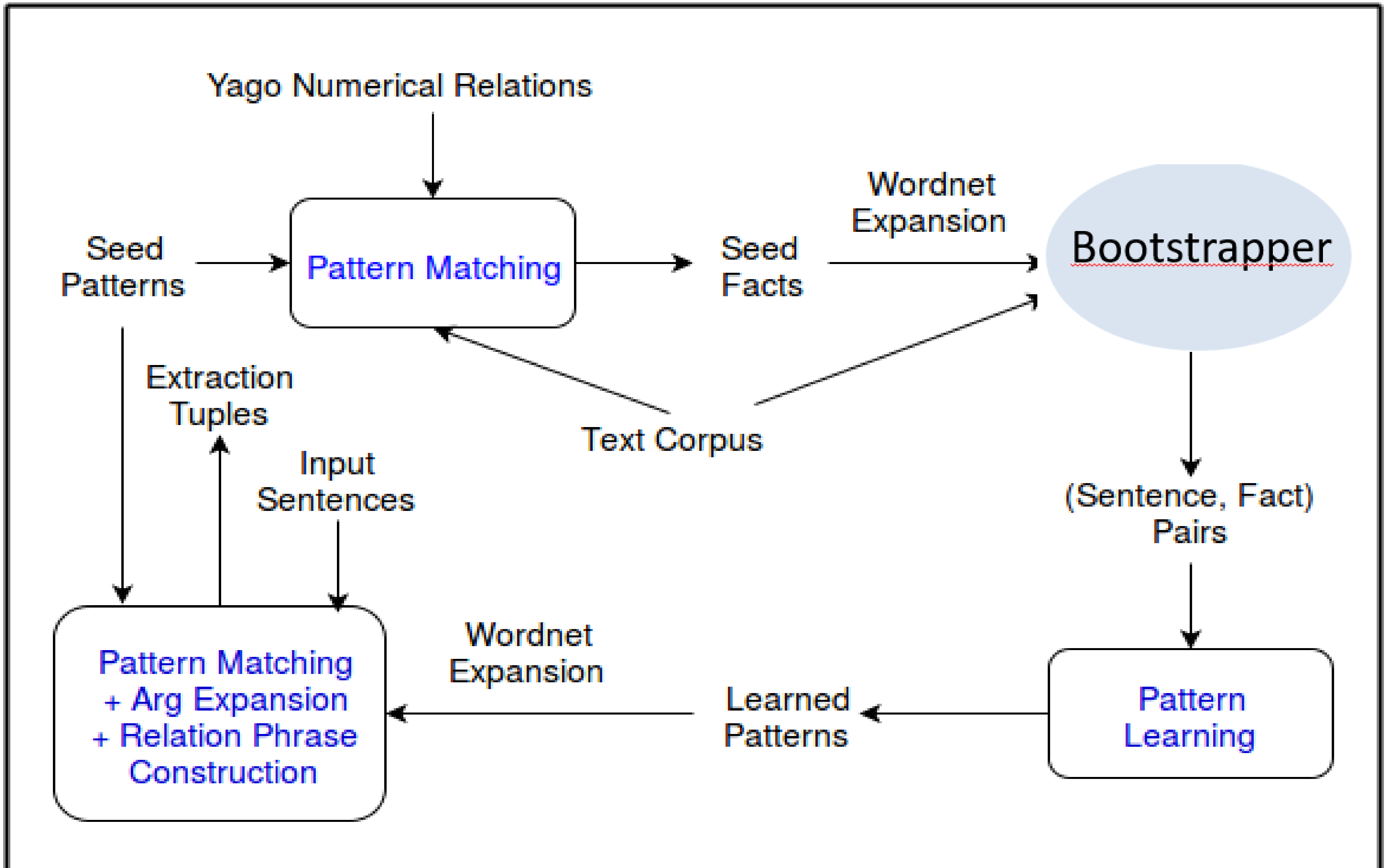
Open IE 5: (James Valley, has area of fruit orchards, 5 sq kms)

Peculiarities of Numerical IE

- Numbers are weak entities
- Units
  - Multiple units for same relation
  - Implicit relations may be expressed via units
- Sentence may express change in quantity
- Relation/argument scoping
  - literacy rate of India
  - *rural* literacy rate of India
  - literacy rate of *South* India

Bootstrapping for Numerical Open IE

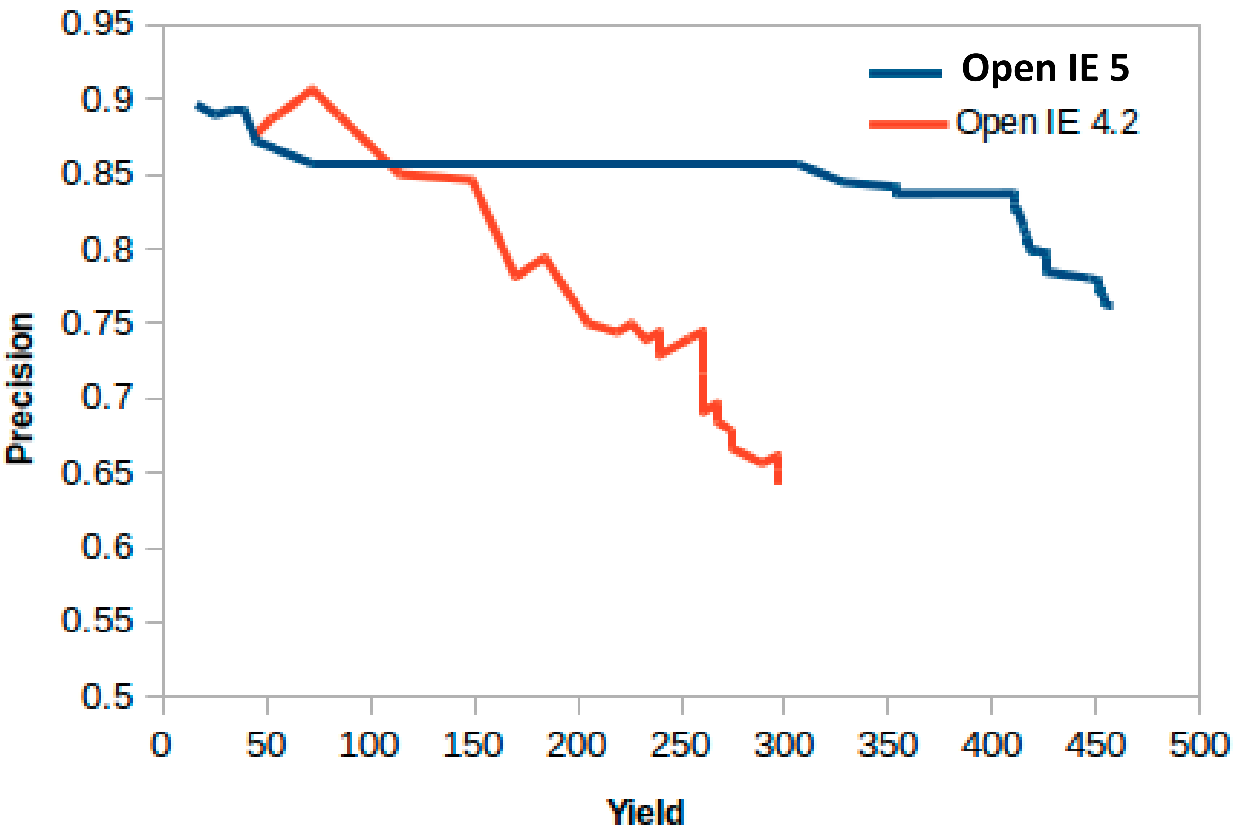
[Saha, Pal, Mausam ACL'17]



Experiments

[Saha, Pal, Mausam ACL'17]

Open IE 5 achieves 1.5x yield and 15 point precision gain on *numerical* facts over Open IE 4.2.



Nested Lists in Open IE

[Saha, Mausam COLING'18]

"President Biden met the leaders of India and China."

Open IE 4: (President Biden, met, the leaders of India and China)

Open IE 5: (President Biden, met, the leaders of India)

(President Biden, met, the leaders of China)

Language Model for Disambiguation

"President Biden met (the leaders of India) and (China)."

- President Biden met the leaders of India
- President Biden met China

"President Biden met the leaders of (India) and (China)."

- President Biden met the leaders of India
- President Biden met the leaders of China

Complex Example

"Gates, an American investor and co-founder of Microsoft, stepped down as CEO of Microsoft in January 2000, but remained as chairman and created the position of chief software architect for himself and transferred his duties to Ray Ozzie and Craig Mundie."	
Extraction	**Systems**
1. (Gates; stepped down as; CEO of Microsoft)	[OC, O4, C]
2. (Gates; stepped down as CEO of Microsoft; in January 2000)	[OC, O4]
3. (Gates; is; an American investor)	[OC]
4. (Gates; is an investor from; United States)	[OC, O4]
5. (Gates; is co-founder of; Microsoft)	[OC]
6. (Gates; is; an American investor and co-founder of Microsoft)	[C]
7. (Gates; remained as; chairman)	[OC, O4, C]
8. (Gates; created; the position of chief software architect for himself)	[OC, O4, C]
9. (Gates; transferred; his duties)	[OC]
10. (Gates; transferred his duties to; Ray Ozzie)	[OC]
11. (Gates; transferred his duties to; Craig Mundie)	[OC]
12. (His; has; duties)	[C]
13. (Gates; transferred his duties to Ray Ozzie; the position of chief software architect for himself)	[C]
14. (Gates; transferred his duties to Craig Mundie; the position of chief software architect for himself)	[C]

Experiments

[Saha, Mausam COLING'18]

	Precision	Yield
Open IE 4.2	79.1	172
ClausIE	67.2	204
Open IE 5	**81.2**	**315**

Code for Open IE 5 available at
https://github.com/dair-iitd/OpenIE-standalone
(downloaded over 9000 times)

(Intermediate) Take Home

- Find a high precision subset
  - even regular expressions are good for low data
  - significant subset of a language is semantically tractable
- Bootstrap training data
  - increase recall while maintaining high precision
  - going down the long tail of syntactic expressions
- Focus on specific constructions
  - nested lists, compound nouns, numerical expressions

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…
- 2017 [@IITD]: Open IE 5.0
  – compound noun phrases, numbers, lists
- 2020 [@IITD]: Open IE 6.0
  – deep neural models

taking a stronger ML leap

increasing precision, recall, expressiveness

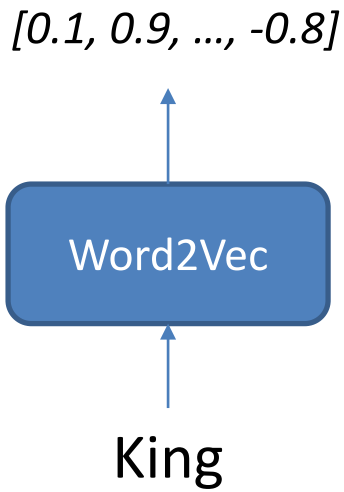
Primer on Deep Learning for NLP

- Word2Vec: Vector representation of words
- Transformers: Attention-based models
- BERT: Pretrained Representations
- Seq2Seq: Encoder-Decoder models

Word2Vec

[Mikolov, et. al., Neurips’13]

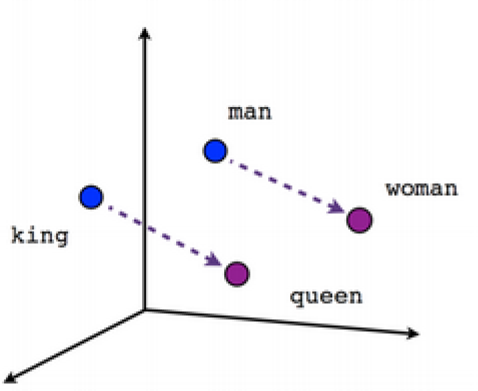
Vector representation of words



Word2Vec

[Mikolov, et. al., Neurips'13]

- vec(*King*) - vec(*Man*) + vec(*Woman*) = vec(***Queen***)
- *A person is known by the _________ he keeps*
- *A person is known by the <u>company</u> he keeps*
- *A **word** is known by the <u>company</u> it keeps*



Male-Female

Transformer

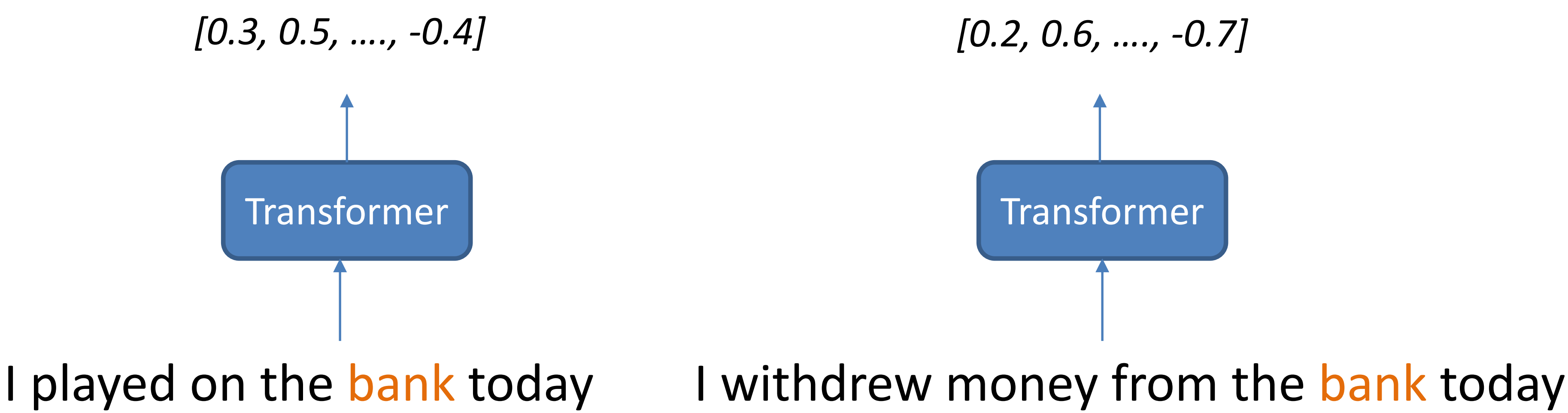
[Vaswani, et. al., Neurips'17]

- One static vector per word is very limiting!
- What about words that have multiple meanings?
- Bank – financial institution or river bank
- Transformers:
  Generate context-based word embeddings

NASSCOM

Transformer

[Vaswani, et. al., Neurips'17]



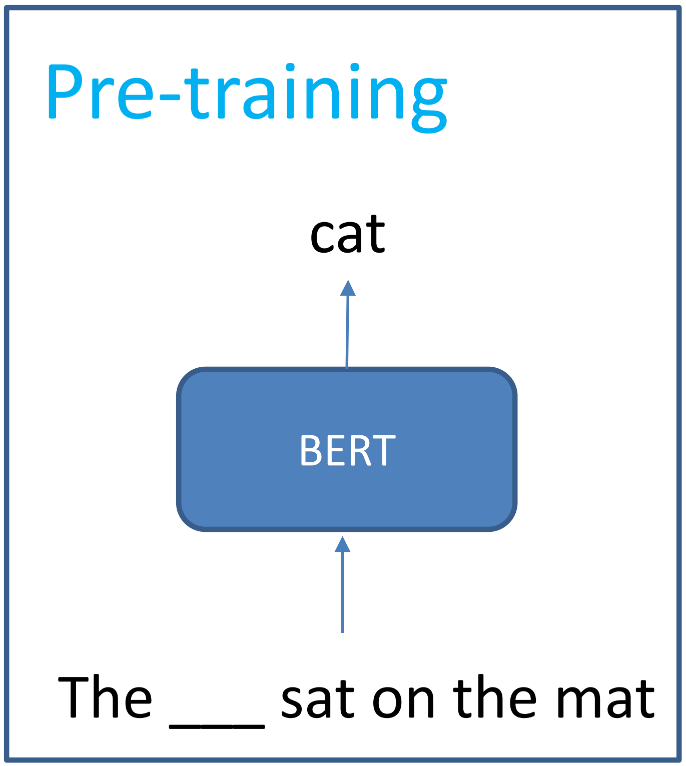
BERT

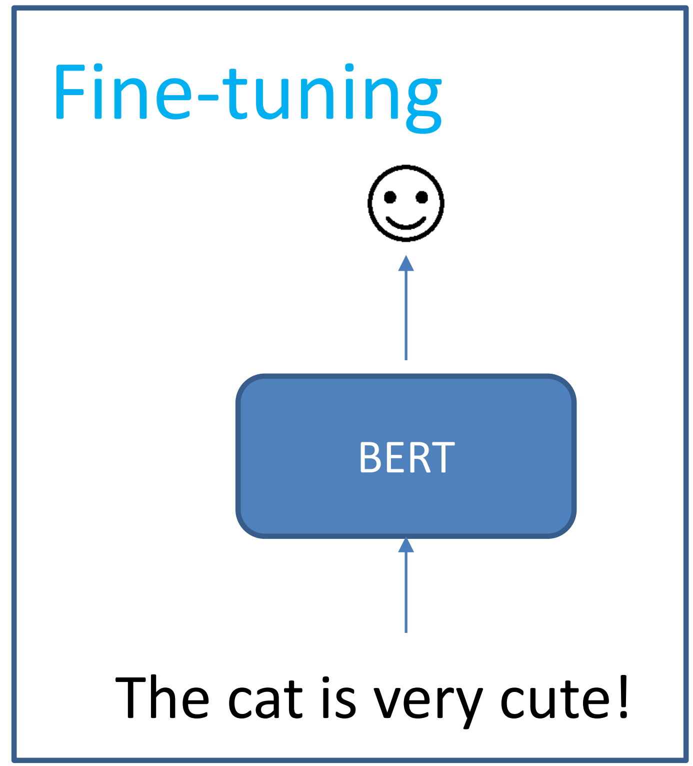
[Devlin, et. al., NAACL'18]

- Training model on each task independently
- Requires learn language from scratch
- Tedious approach!
- BERT pre-training learns language separately
- Frees the model to learn task-specific details

BERT

[Devlin, et. al., NAACL’18]





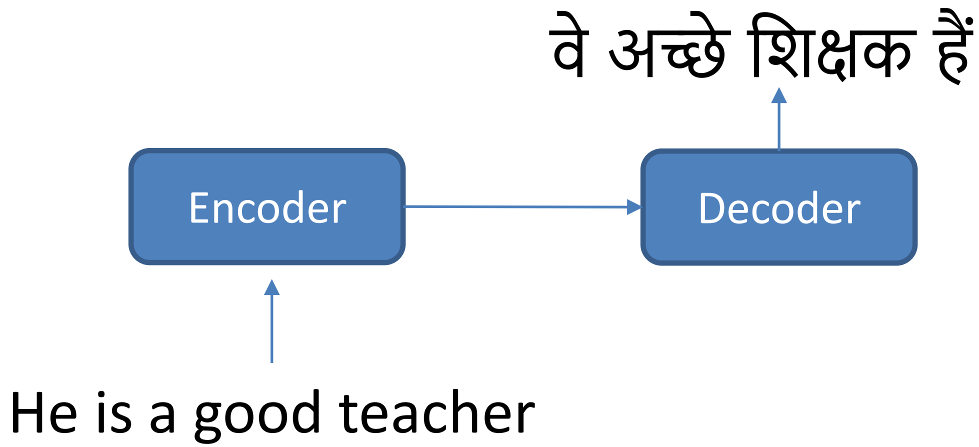
Seq2Seq

- NLP tasks often require generating sequences
- *Machine Translation, Summarization, Chatbots*

- Seq2Seq use an Encoder-Decoder architecture

- Encoder embeds the input
- Decoder generates the sequence

Seq2Seq



Neural OpenIE Extraction

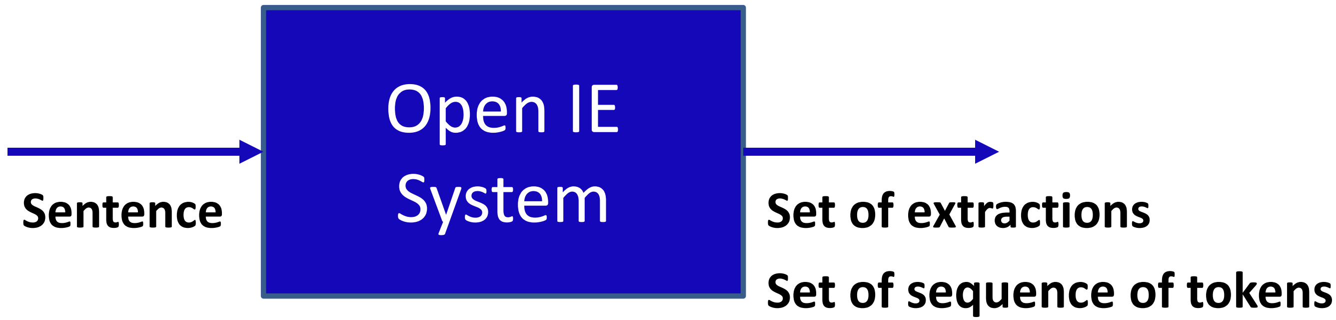
From text:

1. Generative models (IMoJIE, *ACL'20*)
2. Labeling models (OpenIE6, *EMNLP'20*)
3. Multilingual models (AACTrans, *Submitted*)

From Knowledge Bases:

1. Open Knowledge Bases (CEAR, *Submitted*)

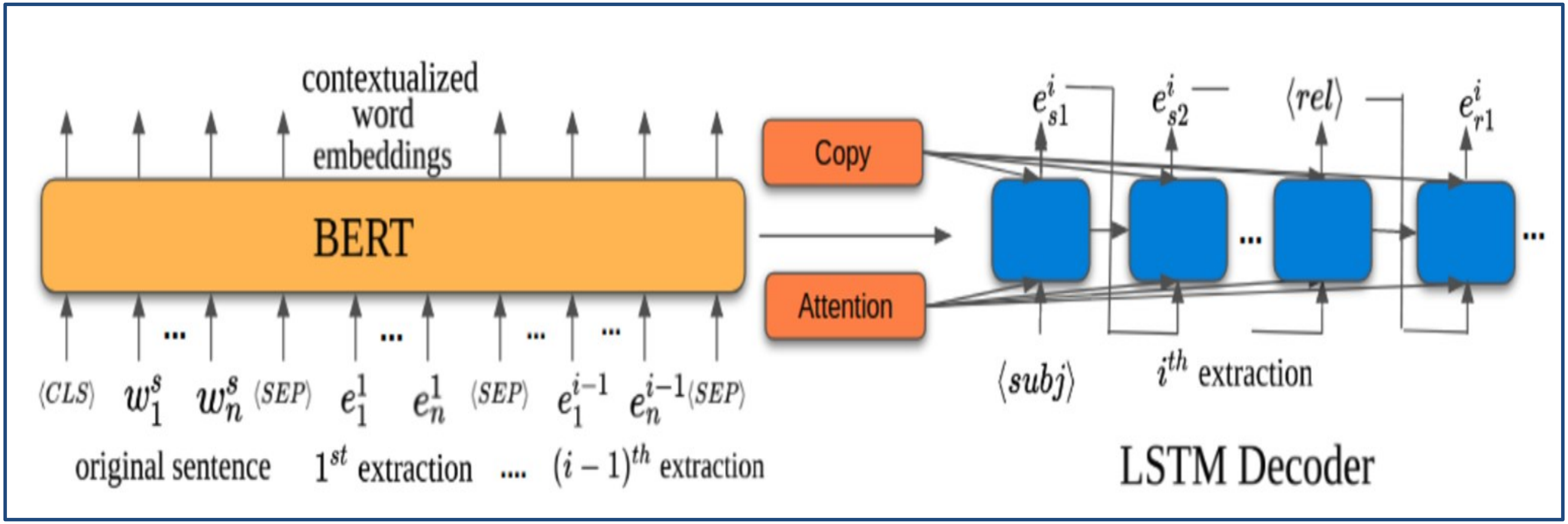
Neural Models



- How to output a set?
  - one at a time: like a sequence
- How to handle large output lengths?
  - output one extraction at a time
- How to ensure model does not repeat same tuple?
  - give all previous extractions as input

IMoJIE: Iterative Memory Based Joint Open IE

[Kolluru, Aggarwal, Rathore, Mausam, Chakrabarti ACL'20]

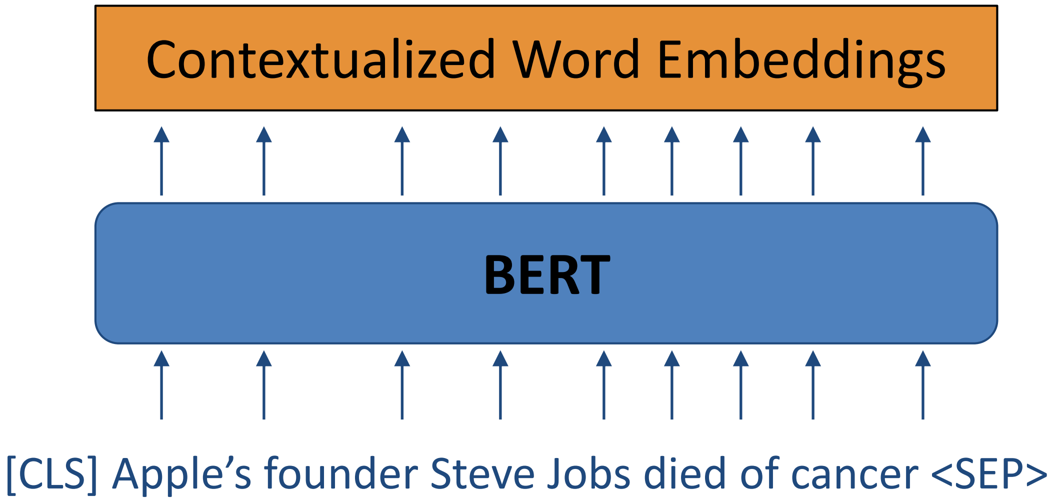


Terminology

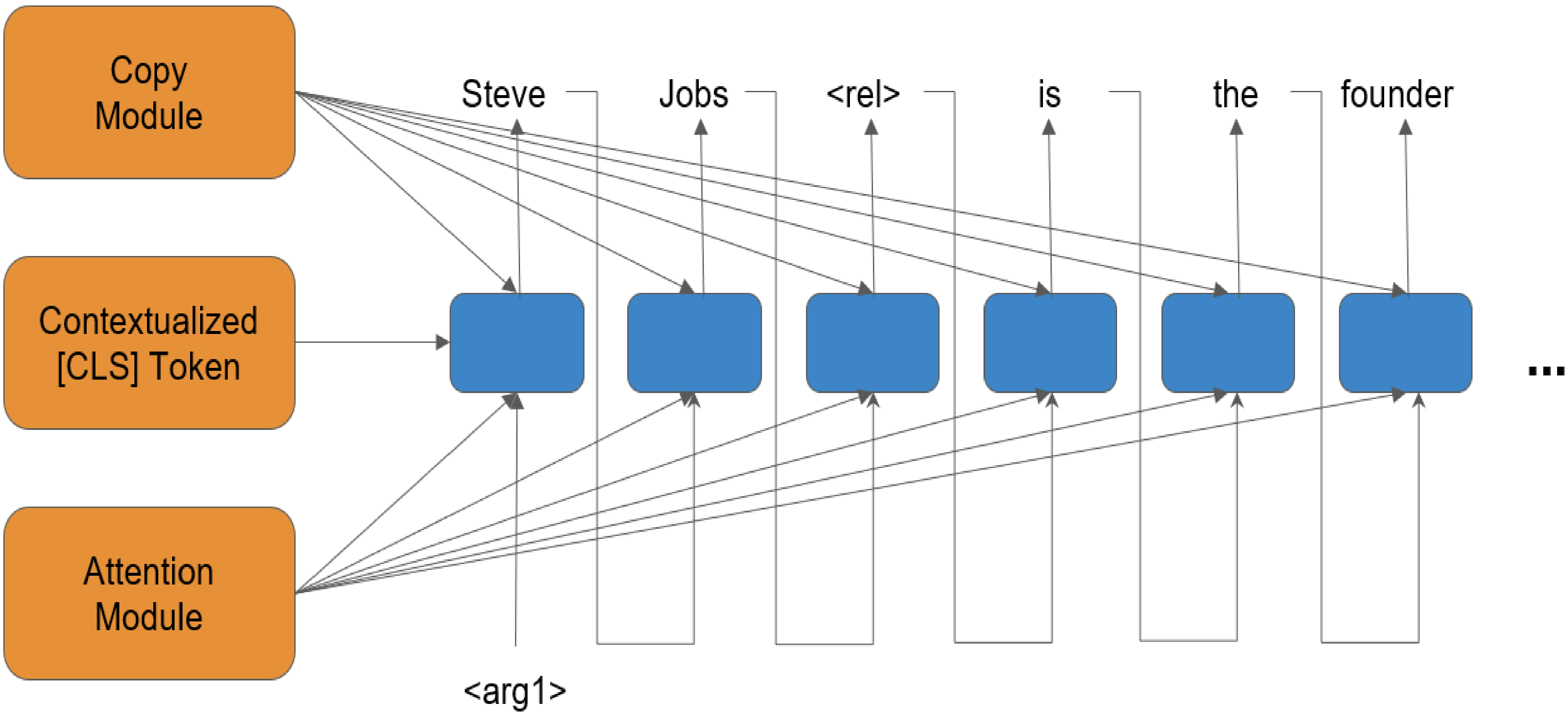
<arg1>, <rel>,
<arg2>
<subj>, <rel>, <obj>

NASSCOM

IMoJIE Encoder – Step 1

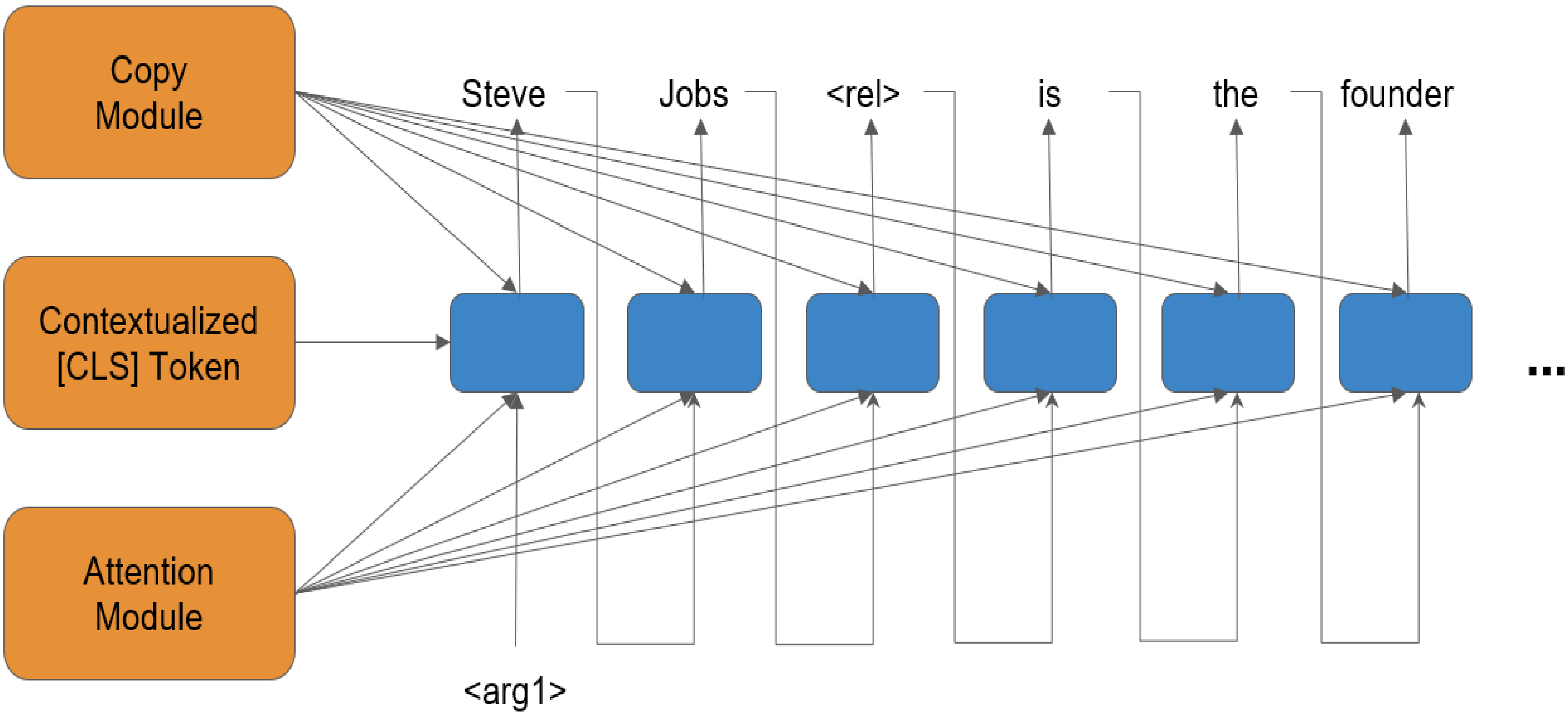


IMoJIE Decoder – Step 1



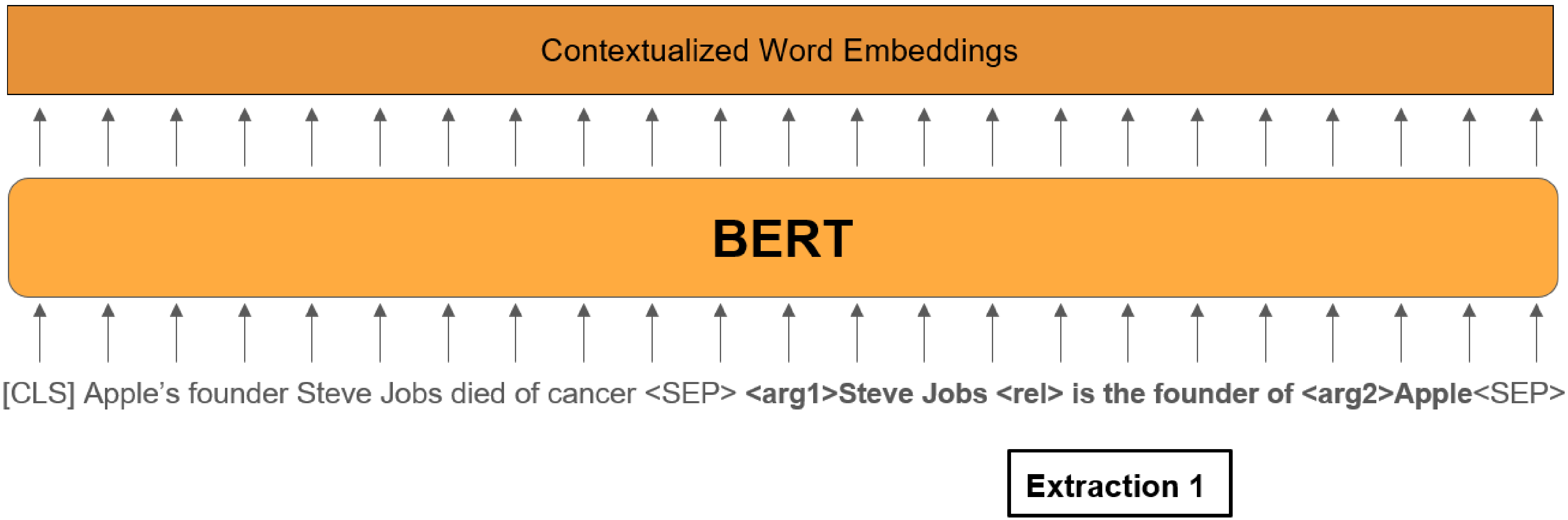
NASSCOM

IMoJIE Decoder – Step 1



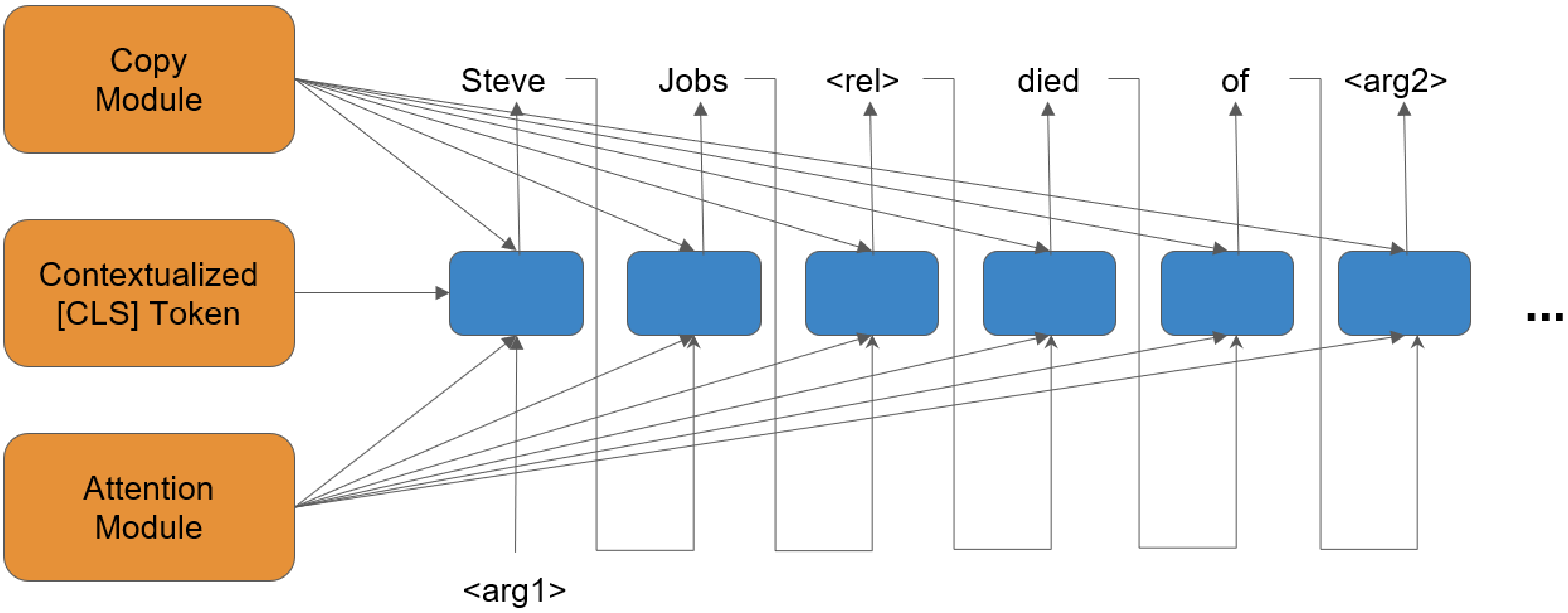
Extraction 1 : <arg1> Steve Jobs <rel> is the founder of <arg2> Apple

IMoJIE Encoder – Step 2

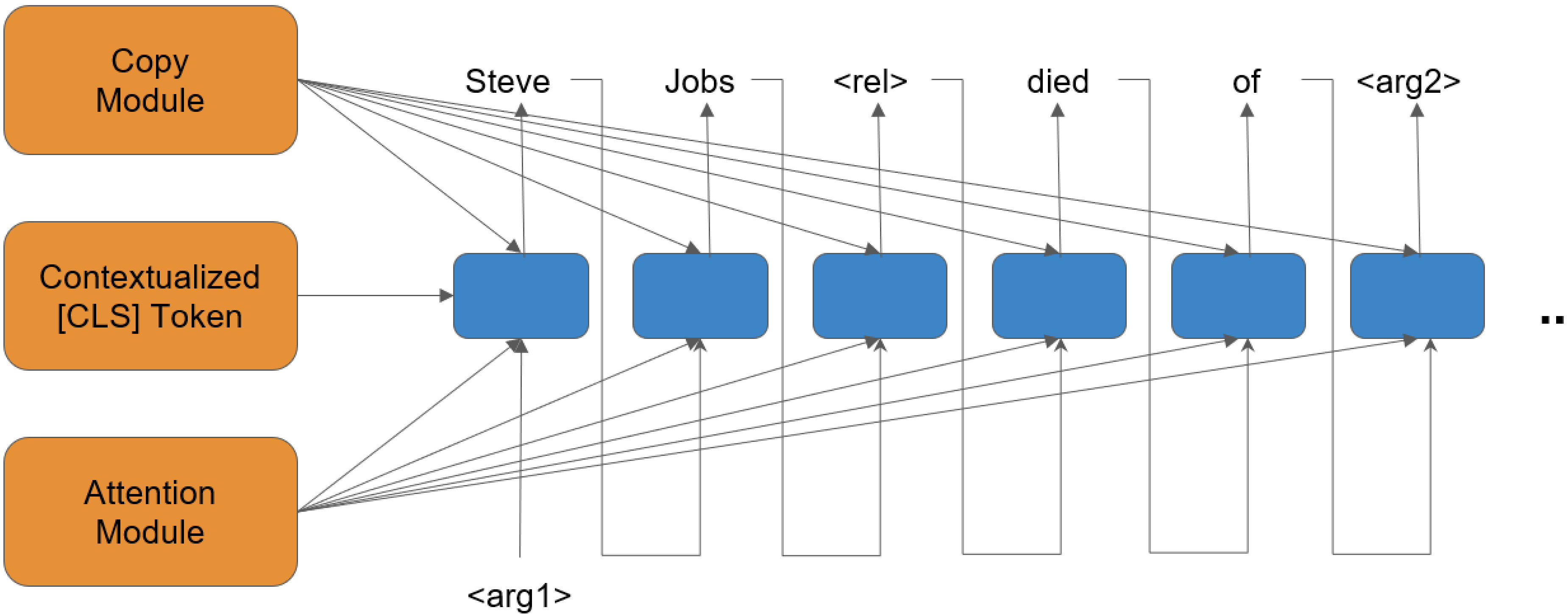


NASSCOM

IMoJIE Decoder – Step 2



IMoJIE Decoder – Step 2

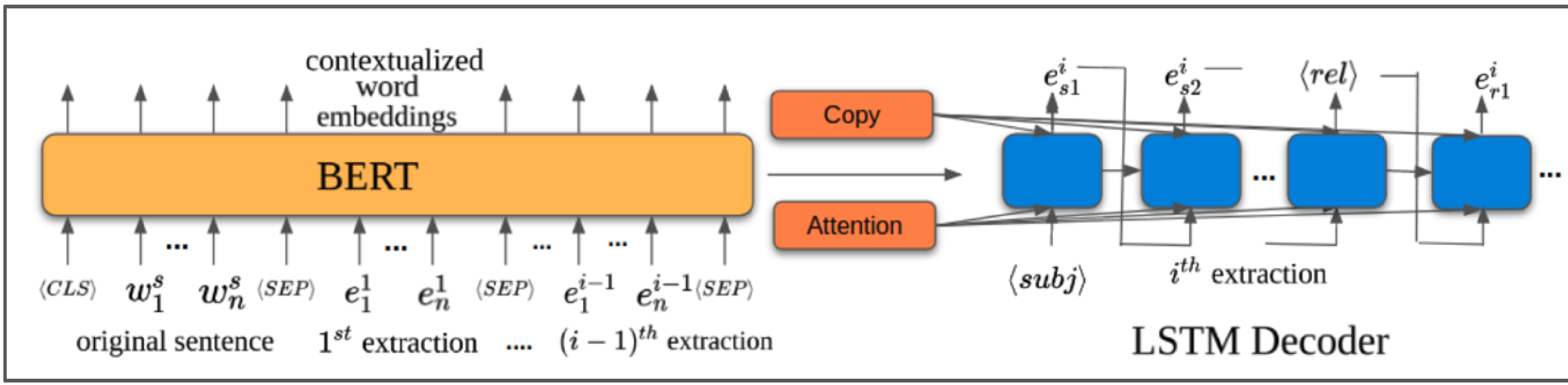


Extraction 2 : <arg1> Steve Jobs <rel> died of <arg2> cancer

IMoJIE

Extraction 1 : <arg1> Steve Jobs <rel> is the founder of <arg2> Apple

Extraction 2 : <arg1> Steve Jobs <rel> died of <arg2> cancer



Terminology

<arg1>, <rel>, <arg2>

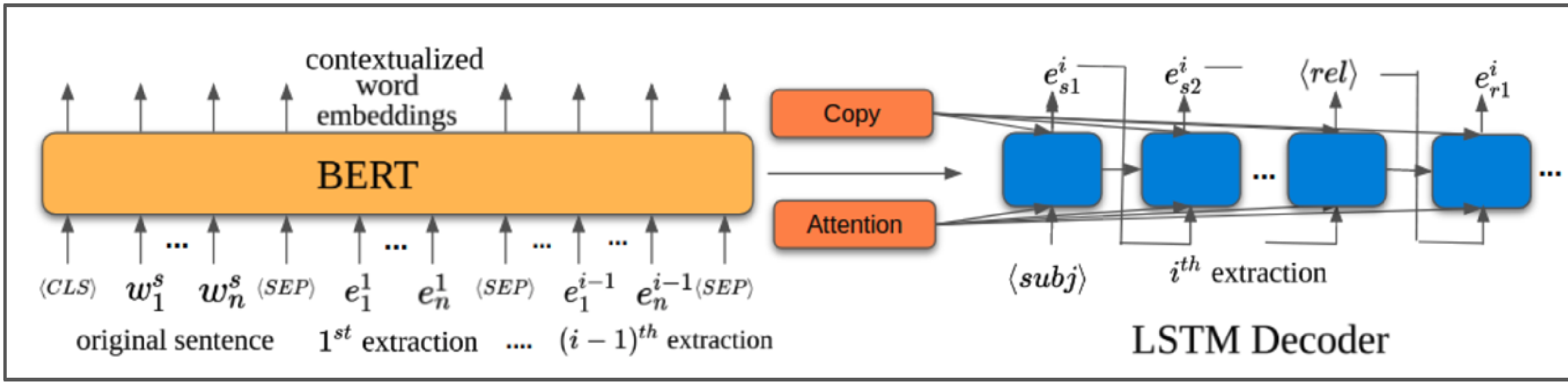
<subj>, <rel>, <obj>

IMoJIE

Slow!

Extraction 1 : <arg1> Steve Jobs <rel> is the founder of <arg2> Apple

Extraction 2 : <arg1> Steve Jobs <rel> died of <arg2> cancer



Terminology

<arg1>, <rel>, <arg2>

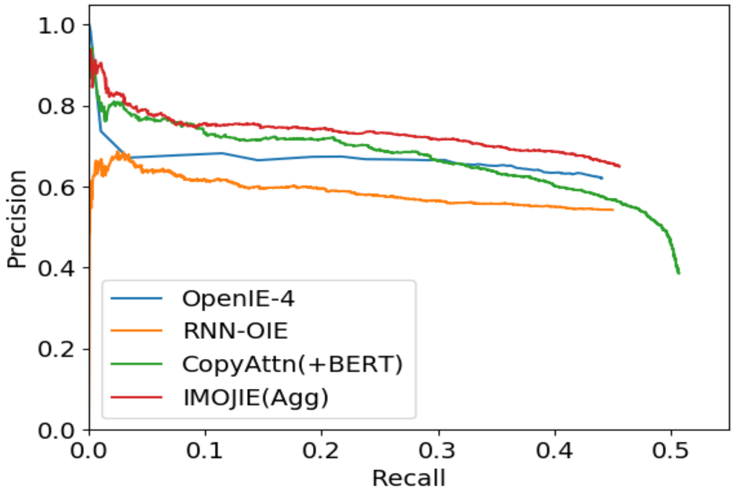
<subj>, <rel>, <obj>

Evaluation using CaRB

[Bharadwaj, Aggarwal, Mausam EMNLP'19]

- CaRB uses a matching strategy to compare system extractions with reference extractions and produces a precision, recall value

- We compute 3 metrics:
  - *Optimal F1*: Maximum F1 value
  - *AUC*: Area under the curve
  - *Last F1*: F1 at last point in curve



Results

System	CaRB		Speed
	F1	AUC	Sentences/sec.
Open IE 4	51.6	29.5	20.1
RnnOIE	49.0	26.0	**149.2**
IMoJIE	**53.5**	**33.3**	2.6

- Trade-off between speed and accuracy
- IMoJIE is 4.5 F1 better than RnnOIE ☺
- IMoJIE is 60x slower than RnnOIE! ☹

- Code, training data, pretrained models at https://github.com/dair-iitd/imojie
  downloaded 3500+ times

Labeling for OpenIE

Apple's founder Steve Jobs died of cancer [is] [of] [from]

ARG2 REL ARG1 ARG1 NONE NONE NONI
REL REL NONE

NONE NONE ARG1 ARG1 REL REL ARG2
NONE NONE NONE

NASSCOM

Labeling for OpenIE

Apple's founder Steve Jobs died of cancer [be] [of] [from]

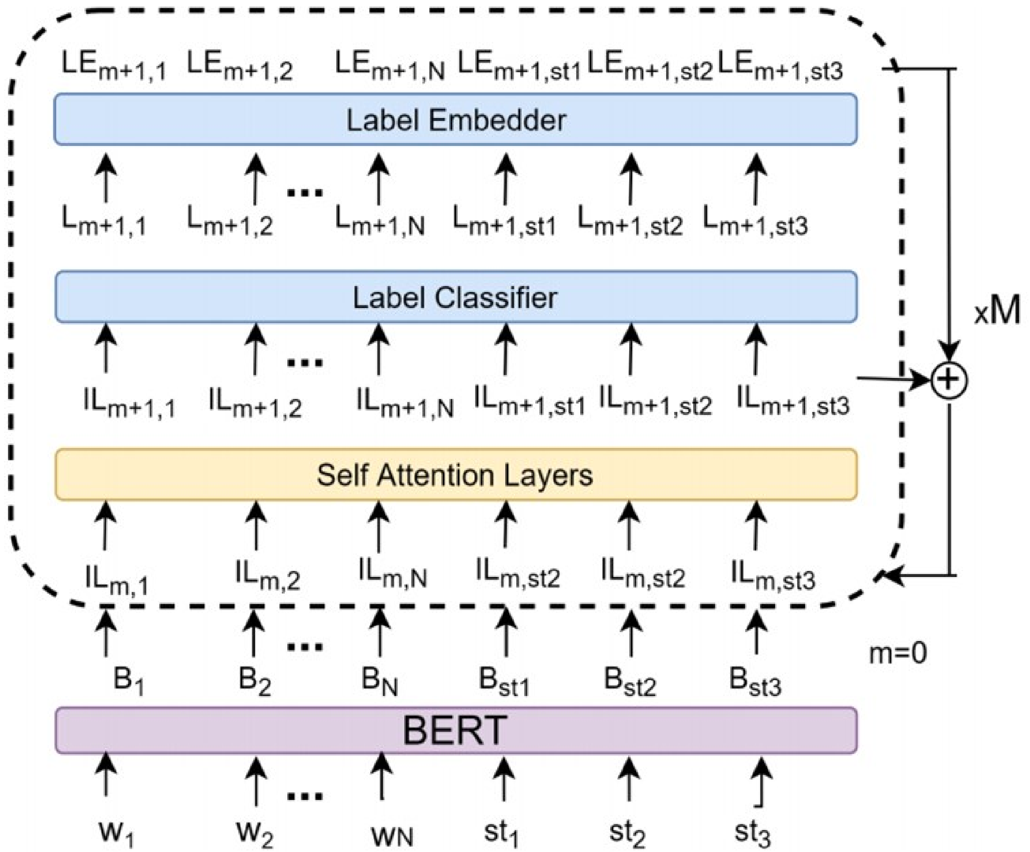
ARG2 REL ARG1 ARG1 NONE NONE NONE REL REL NONE

NONE NONE ARG1 ARG1 REL REL ARG2 NONE NONE NONE

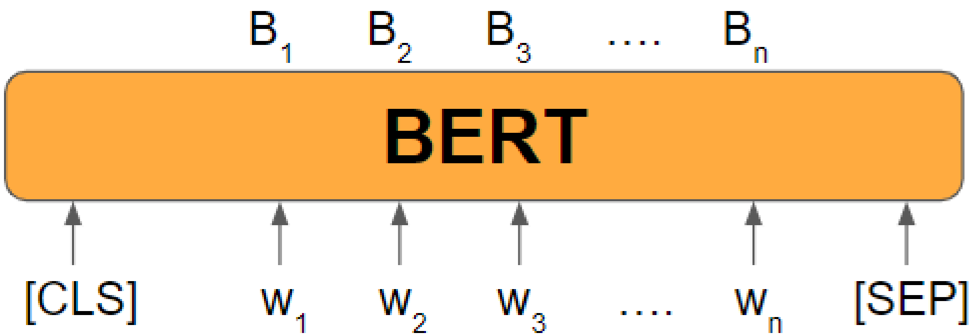
(Steve Jobs, [be] the founder [of], Apple)
(Steve Jobs, died of, cancer)

IGL – *Iterative* Grid Labeling

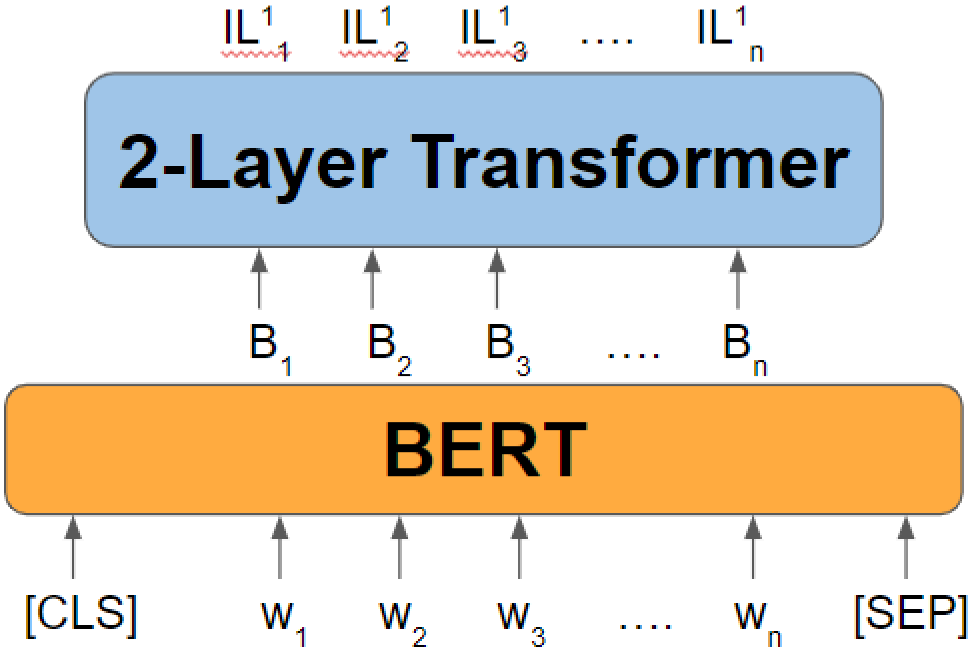
[Kolluru, Adlakha, Aggarwal, Mausam, Chakrabarti EMNLP'20]



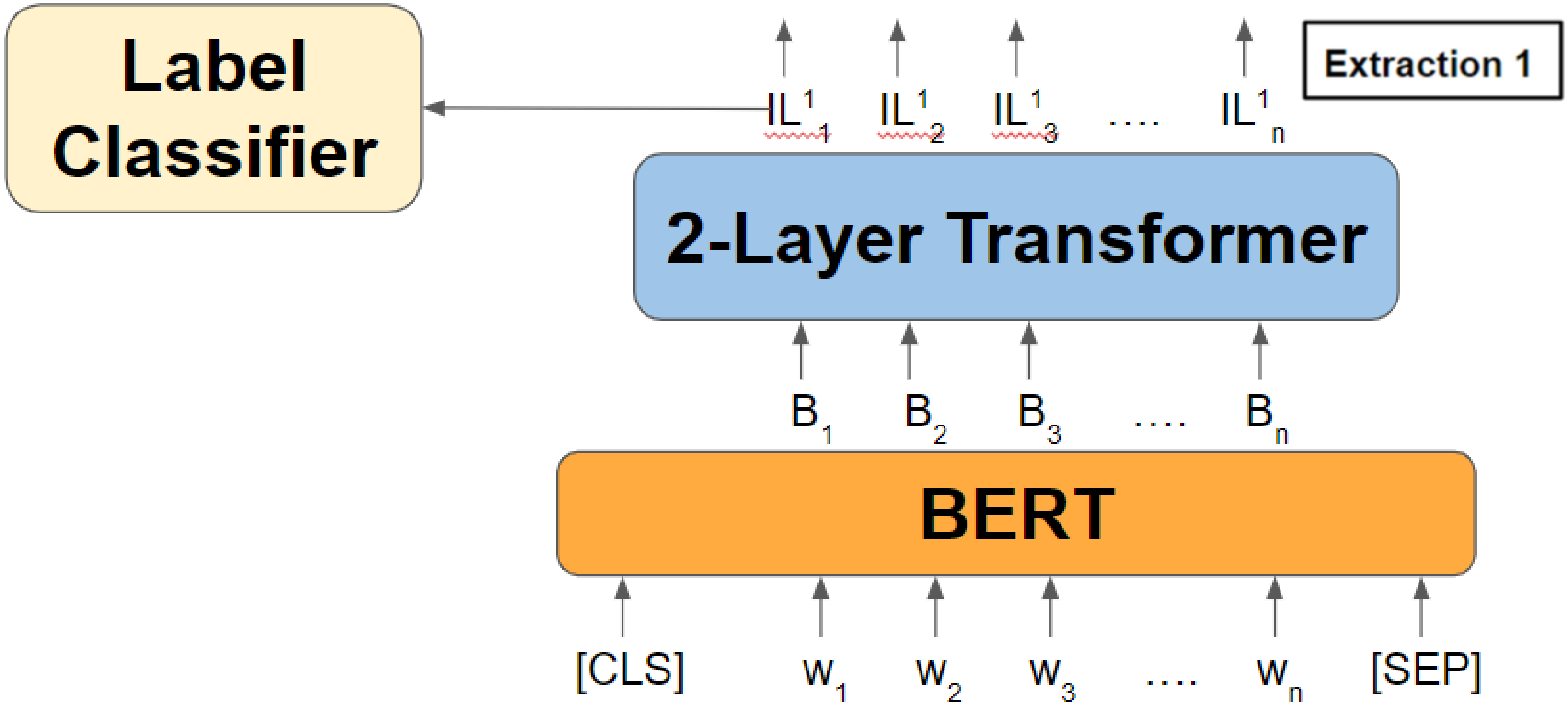
IGL – *Iterative* Grid Labeling



IGL – *Iterative* Grid Labeling

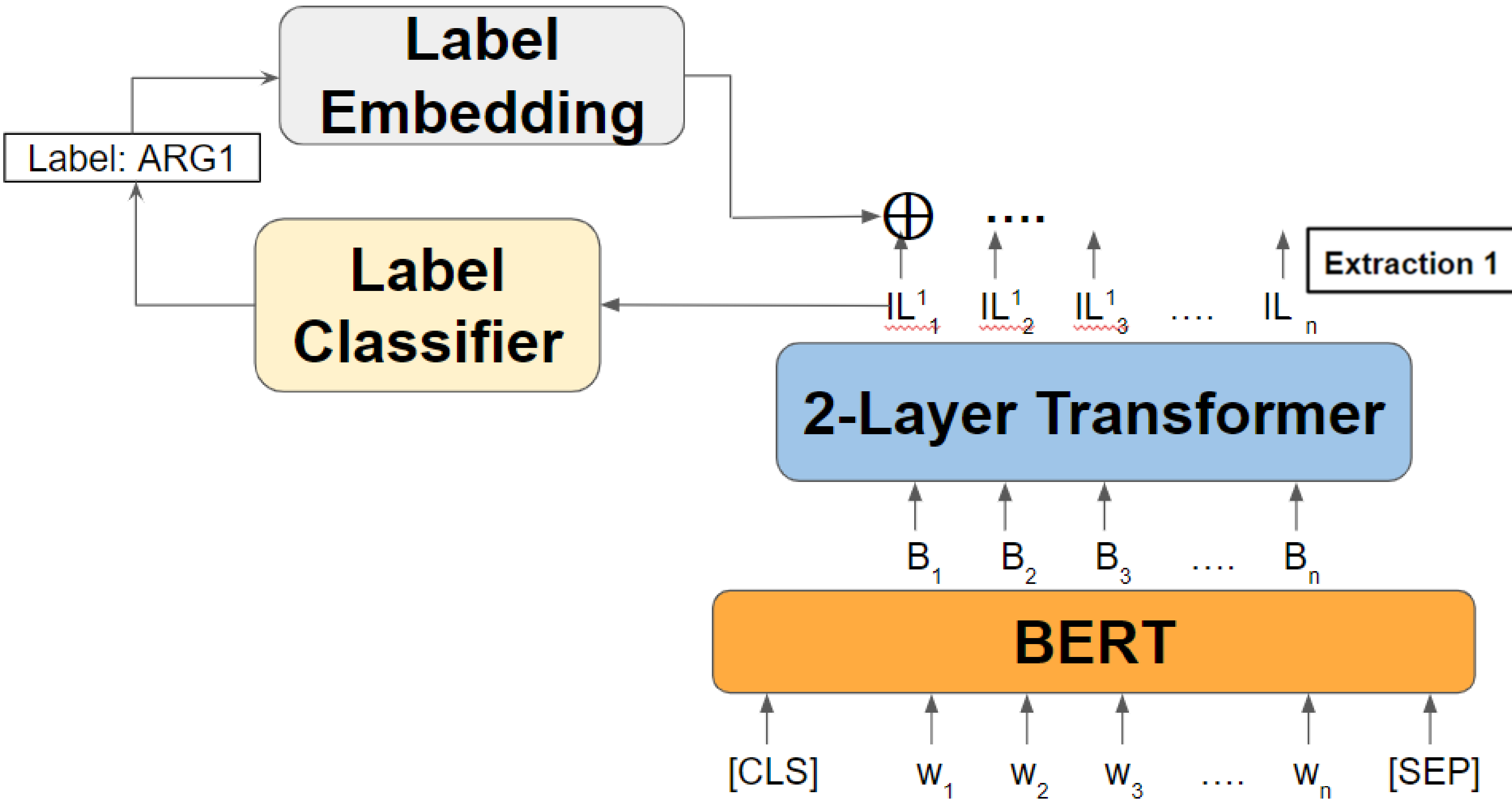


IGL – *Iterative* Grid Labeling

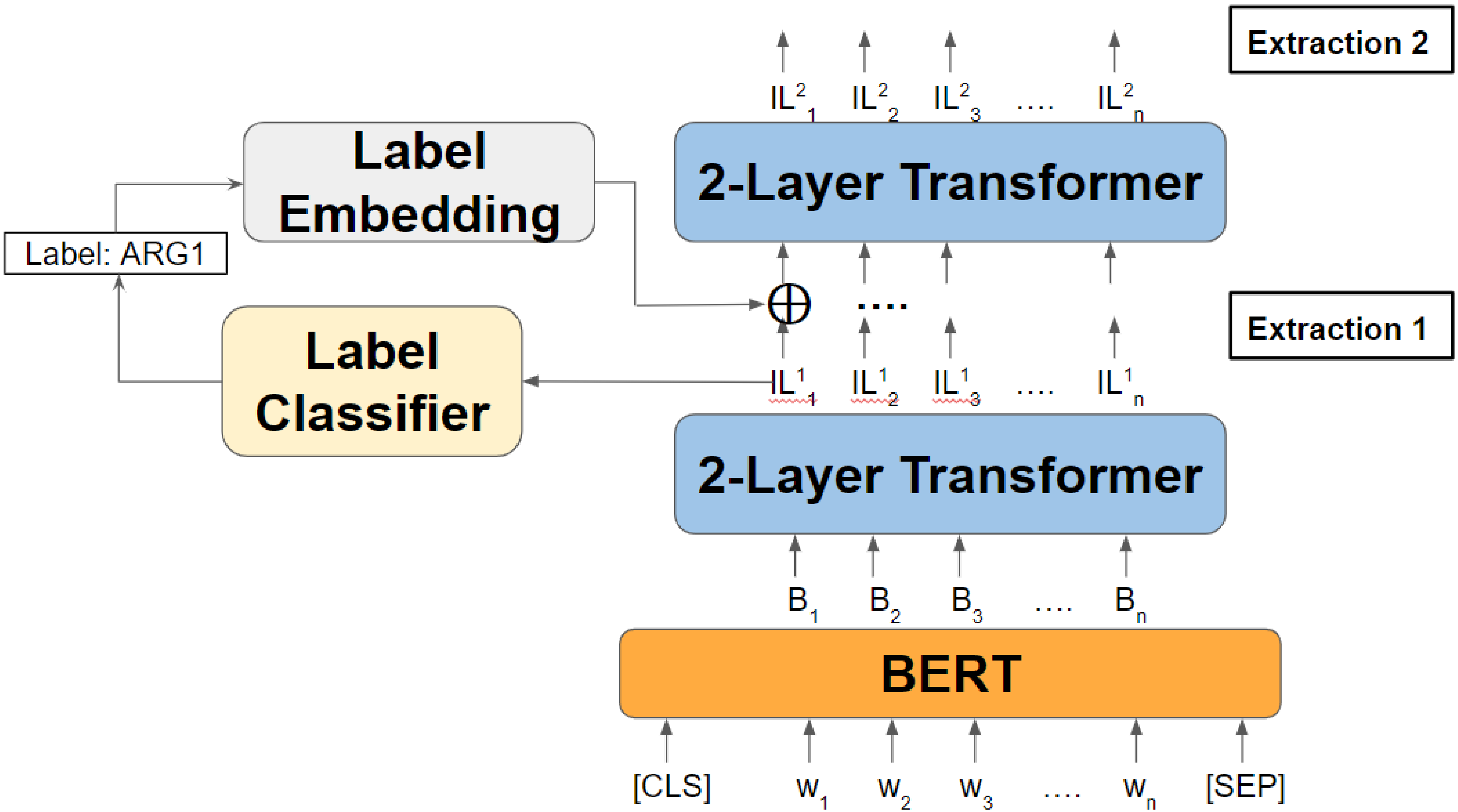


NASSCOM

IGL – *Iterative* Grid Labeling



IGL – *Iterative* Grid Labeling



IGL – Iterative *Grid* Labeling

E4	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
E3	**ARG1**	NONE	**REL**	**REL**	**REL**	**ARG2**	**ARG2**	NONE
E2	**ARG1**	NONE	**REL**	**REL**	NONE	**ARG2**	**ARG2**	NONE
E1	**ARG1**	**ARG1**	NONE	NONE	**REL**	NONE	**ARG2**	NONE
	w1	w2	w3	w4	w5	w6	w7	w8

Results

System	CaRB		Speed
	F1	AUC	Sentences/sec.
RnnOIE	49.0	26.0	**149.2**
IMoJIE	**53.5**	33.3	2.6
IGL-OIE	52.4	**33.7**	142.0

- IGL-IE 60x faster than IMoJIE
- IGL-IE 1.1 F1 lower than IMoJIE

IGL for OpenIE

Known-tradeoff between Speed & Accuracy

- Full generation is more powerful than labeling
- Full generation is much slower than labeling

Solution: Constraints

[Nandwani, Pathak, Mausam, Singla NeurIPS'19]

What makes a good set of extractions?

"Obama gained popularity after Oprah endorsed him for the presidency"

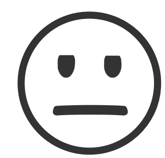
(Obama, gained, popularity)

What makes a good set of extractions?

"Obama gained popularity after Oprah endorsed him for the presidency"

(Obama, gained, popularity)

(Oprah, endorsed, him)



What makes a good set of extractions?

"Obama gained popularity after Oprah endorsed him for the presidency"

(Obama, gained, popularity) ☺

(Oprah, endorsed him for, the presidency)

What makes a good set of extractions?

"Obama gained popularity after Oprah endorsed him for the presidency"

(Obama, gained, popularity)

↓

(Obama, gained, popularity)
(Oprah, endorsed, him)

↓

(Obama, gained, popularity)
(Oprah, endorsed him for, the presidency)

What changed?

What makes a good set of extractions?

"Oprah", "endorsed", "presidency" should have been in the set of extractions

Because they convey ***information!***

POSC Constraints:

All words with POS tags as *nouns (N), verbs (V), adjectives (JJ), and adverbs (RB)* should be part of at least one extraction.

Constrained Iterative Grid Labeling (CIGL)

System	CaRB		Speed
	F1	AUC	Sentences/sec.
RnnOIE	49.0	26.0	**149.2**
IMoJIE	53.5	33.3	2.6
IGL-OIE	52.4	33.7	142.0
CIGL-OIE	**54.0**	**35.7**	142.0

- CIGL 0.5 F1 improvement over IMoJIE
- CIGL 60x faster than IMoJIE

Nested Lists in Open IE

[Saha, Mausam COLING'18, Kolluru etal EMNLP'20]

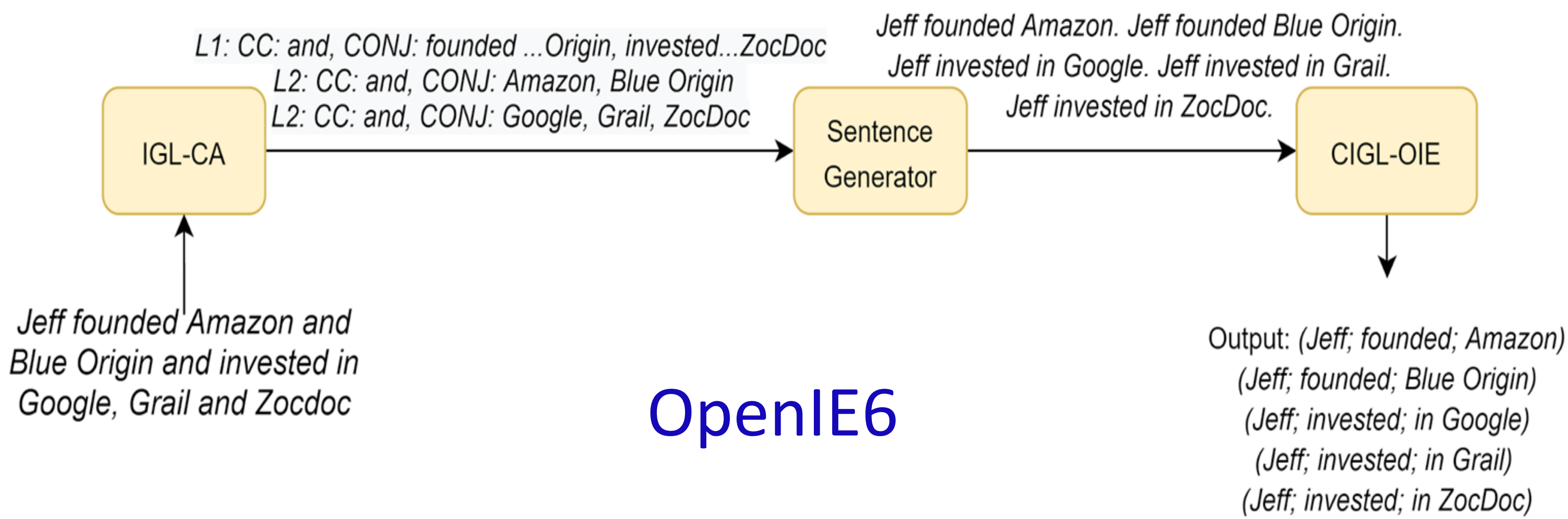
"President Biden met the leaders of India and China."

Open IE 4: (President Biden, met, the leaders of India and China)

Open IE 6: (President Biden, met, the leaders of India)

(President Biden, met, the leaders of China)

Augmenting OpenIE with Coordination Analysis



OpenIE6

Code, training data, pretrained models at
https://github.com/dair-iitd/openie6
downloaded 1500+ times

Take Home

- Find a high precision subset
  - even regular expressions are good for low data
  - significant subset of a language is semantically tractable
- Bootstrap training data
  - increase recall while maintaining high precision
  - going down the long tail of syntactic expressions
- Focus on specific constructions
  - nested lists, compound nouns, numerical expressions
- Constraints in neural models
  - allow AI experts to correct neural models and enable train-test analyze cycles

Multilingual OpenIE

- OpenIE has primarily focused on English
- Extending OpenIE to other languages
- Challenge: Creating/Curating training data
  - manual annotation is expensive
- Solution: Translate English data

Issues with normal Translation

- Need to translate sentence and extractions
- Independent translation leads to *inconsistencies*
- Lexical Inconsistencies: *Usage of synonyms*
- Semantic Inconsistencies: *Changes meaning*

Examples of Inconsistencies

Lexical Inconsistency	
English Sentence	*The shield of Athena Parthenos, sculpted by Phideas, depicts a fallen Amazon*
English Extraction	<s> *The shield of Athena Parthenos* </s> <r> *depicts* </r> <o> *a fallen Amazon* </o>
Spanish Sentence	El escudo de Atena Parthenos, sculptado por Phideas, representa un Amazonas fallecido
Spanish Extraction (Indp)	<s> El escudo de Atena Parthenos </s> <r> representa </r> <o> un Amazonas caído </o>
Spanish Extraction (Const)	<s> El escudo de Atena Parthenos </s> <r> representa </r> <o> un Amazonas fallecido </o>
Semantic Inconsistency	
English Sentence	*The discovery was remarkable as the skeleton was almost identical to a modern Kuvasz*
English Extraction	<s> *skeleton* </s> <r> *was* </r> <o> *almost identical to a modern Kuvasz* </o>
Spanish Sentence	Un descubrimiento notable porque fósil era casi idéntica a un Kuvasz moderno
Spanish Extraction (Indp)	<s> skeleto </s> <r> era </r> <o> casi idéntica a una Kuvasz moderna </o>
Spanish Extraction (Const)	<s> fósil </s> <r> era </r> <o> casi idéntica a un Kuvasz moderno </o>

Other Desiderata

Sentence Extractions	George Bluth Sr., patriarch of the Bluth family, is the founder and former CEO of the Bluth Company. <s> George Bluth Sr. </s> <r> is patriarch of </r> <o> the Bluth family </o> <s> George Bluth Sr. </s> <r> is </r> <o> the founder and former CEO of the Bluth Company </o> <s> George Bluth Sr. </s> <r> is </r> <o> patriarch of the Bluth family </o>
Telugu English Extraction	షరోన్ యొక్క దీర్ఘకాల ప్రత్యర్థి బెంజమిన్ నెతన్యాహును లికుడ్ నాయకుడిగా ఎన్నుకున్నారు *Sharon's longtime rival Benjamin Netanyahu was elected as leader of Likud* <s> షరోన్ యొక్క దీర్ఘకాల ప్రత్యర్థిని </s> <o> లికుడ్ నాయకుడిగా </o> <r> ఎన్నుకున్నారు </r>
Hindi English Extraction	जॉन लैंबर्ट ने सरकार के साधन के रूप में जाना जाने वाला एक नया संविधान सामने रखा *John Lambert put forward a new constitution known as the Instrument of Government* <s> एक नया संविधान </s> <o> सरकार के साधन के रूप में </o> <r> जाना जाता है </r>

Consistent Translation

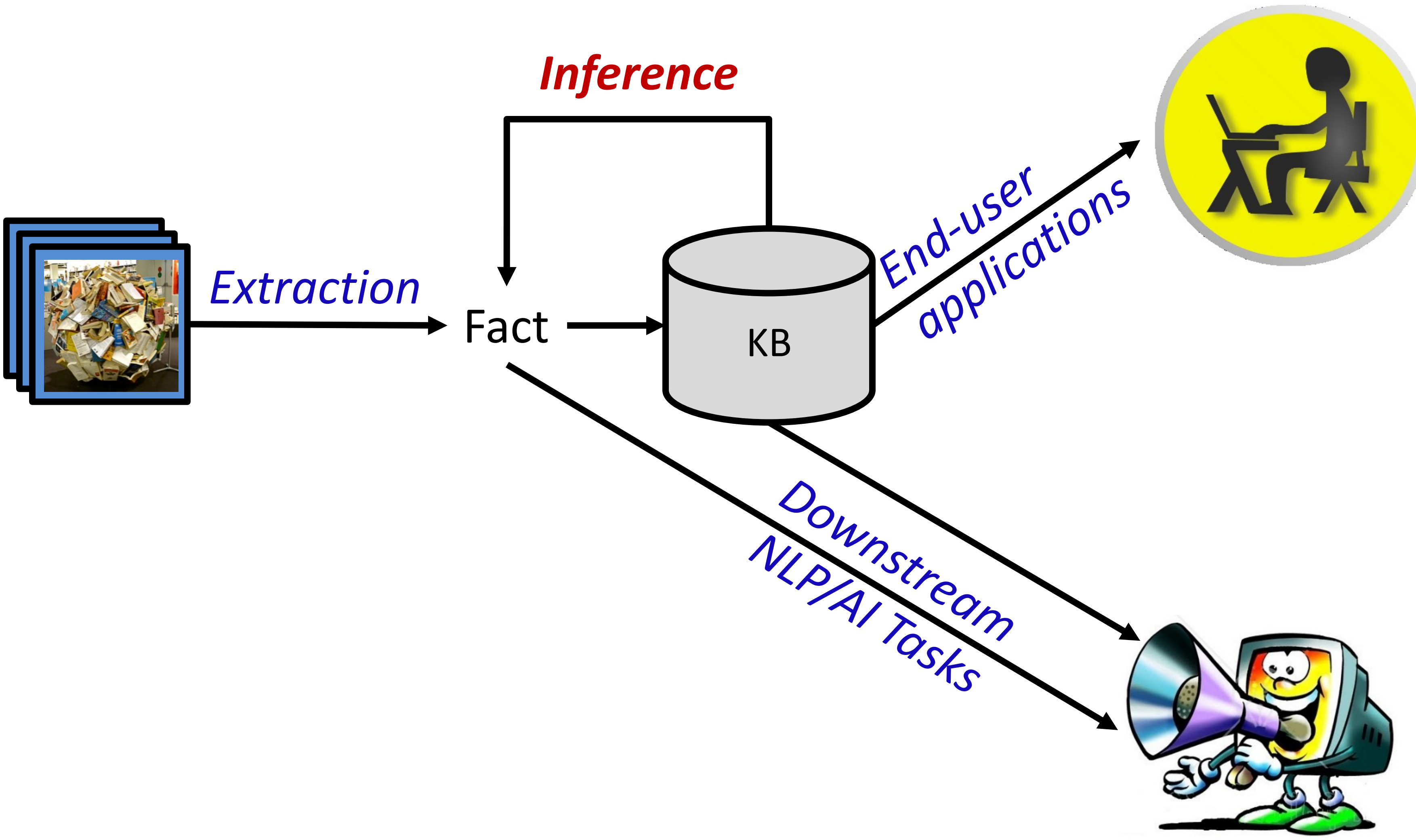
- Introduce a new type of translation: AACT
- Alignment-Augmented Consistent Translation
- Two translations are consistent to each other
  - Uses word-alignments b/w English-*F* translations

Experimental Validation

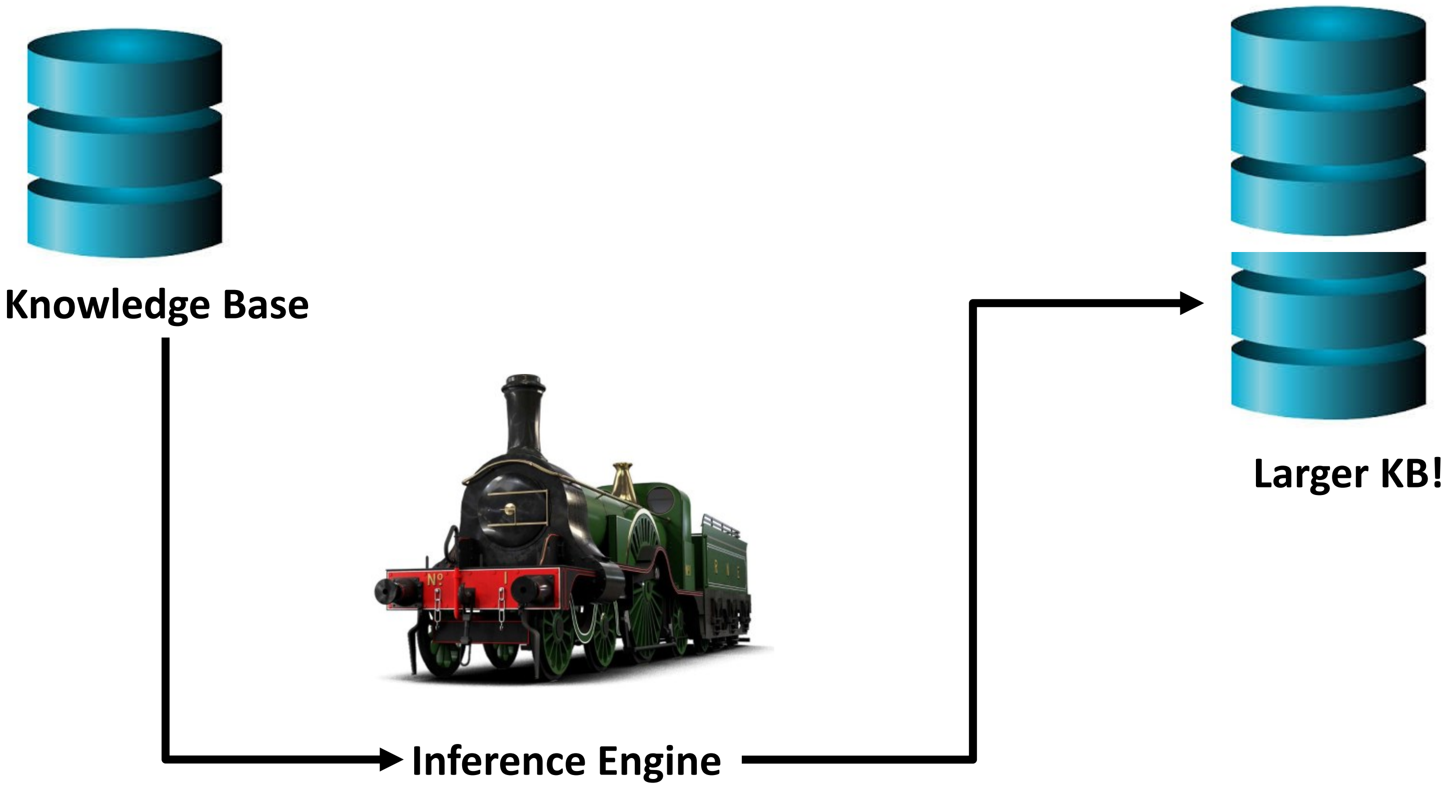
[Kolluru, Mohammed, Mittal, Chakrabarti, Mausam Unpublished'21]

- Experiments over five languages:
- *Spanish, Portuguese, Chinese, Hindi, Telugu*
- Improvement of 19.5% F1 and 10.6% AUC over prior multilingual models

Talk Outline



KB Inference



OpenIE Inference

- 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)

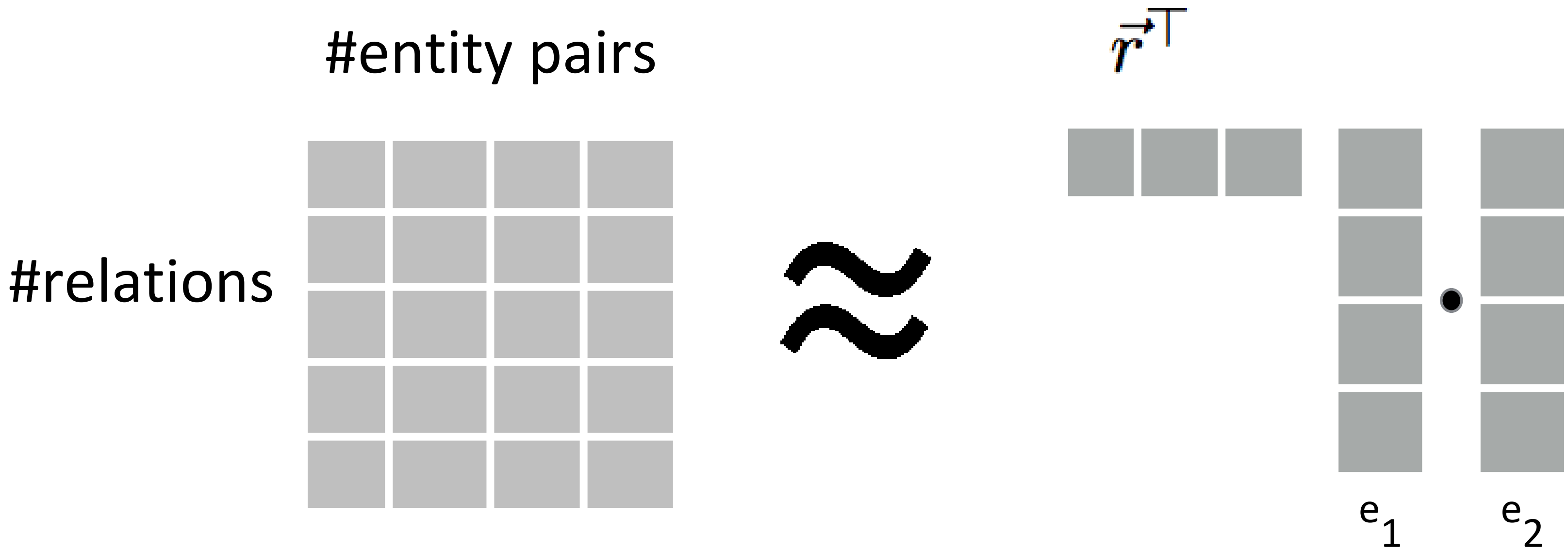
NASSCOM

Embeddings for entities/relations

iron	≈	0.2	0.5	0.6	-0.7
iron nail	≈	0.2	0.6	0.8	-0.6
conducts	∼	0.1	0.4	-0.2	-0.7
electricity	≈	0.9	-0.4	-2.5	-0.7

Represent entities (entity pairs) and relations in a continuous $\mathbf{R}^d$ / $\mathbf{C}^d$ space.

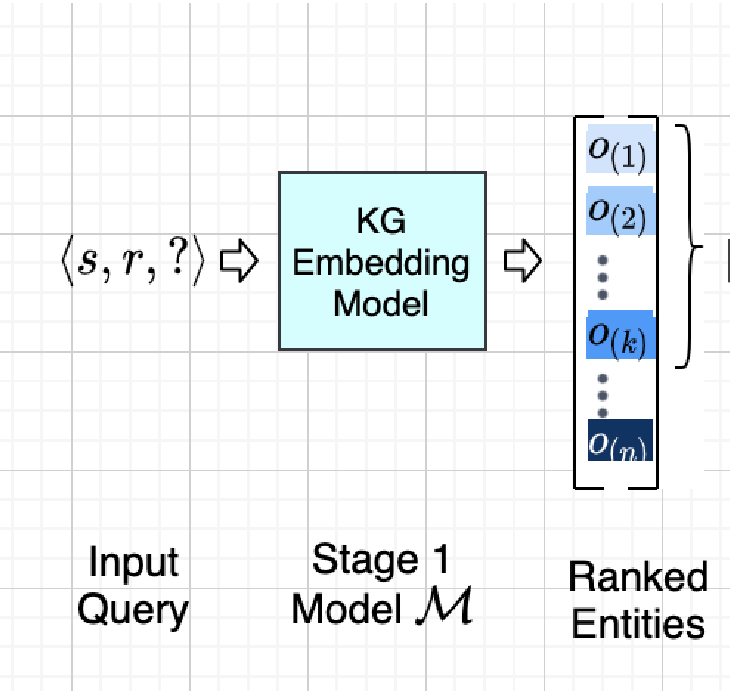
Tensor Factorization
(DistMult/ComplEx)



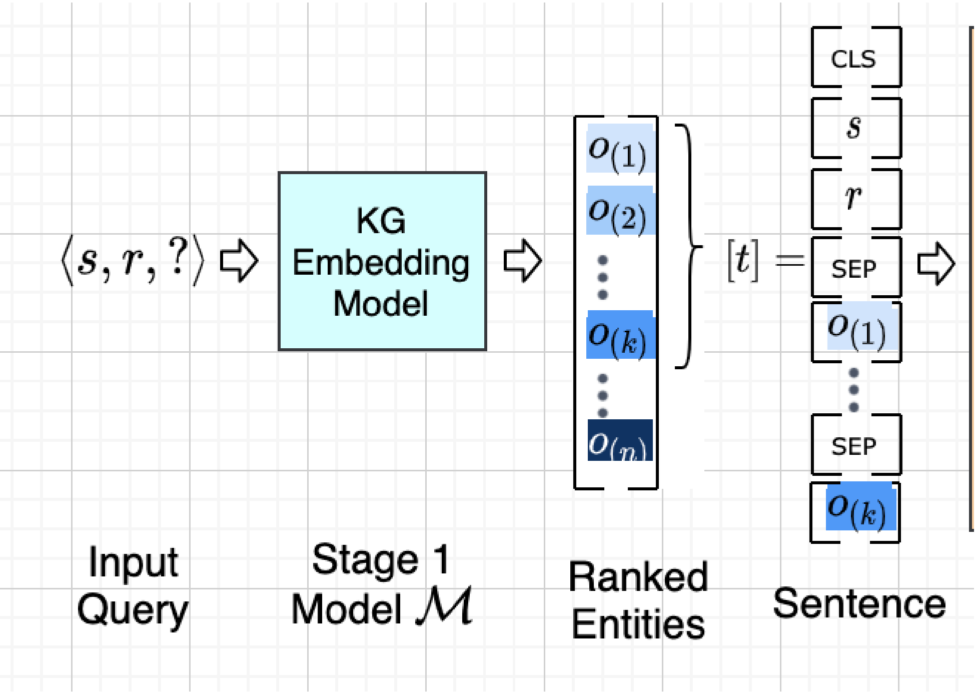
(iron nail, conducts, electricity)

NASSCOM

CEAR: Cross-Entity Aware Reranker for Knowledge Base Completion

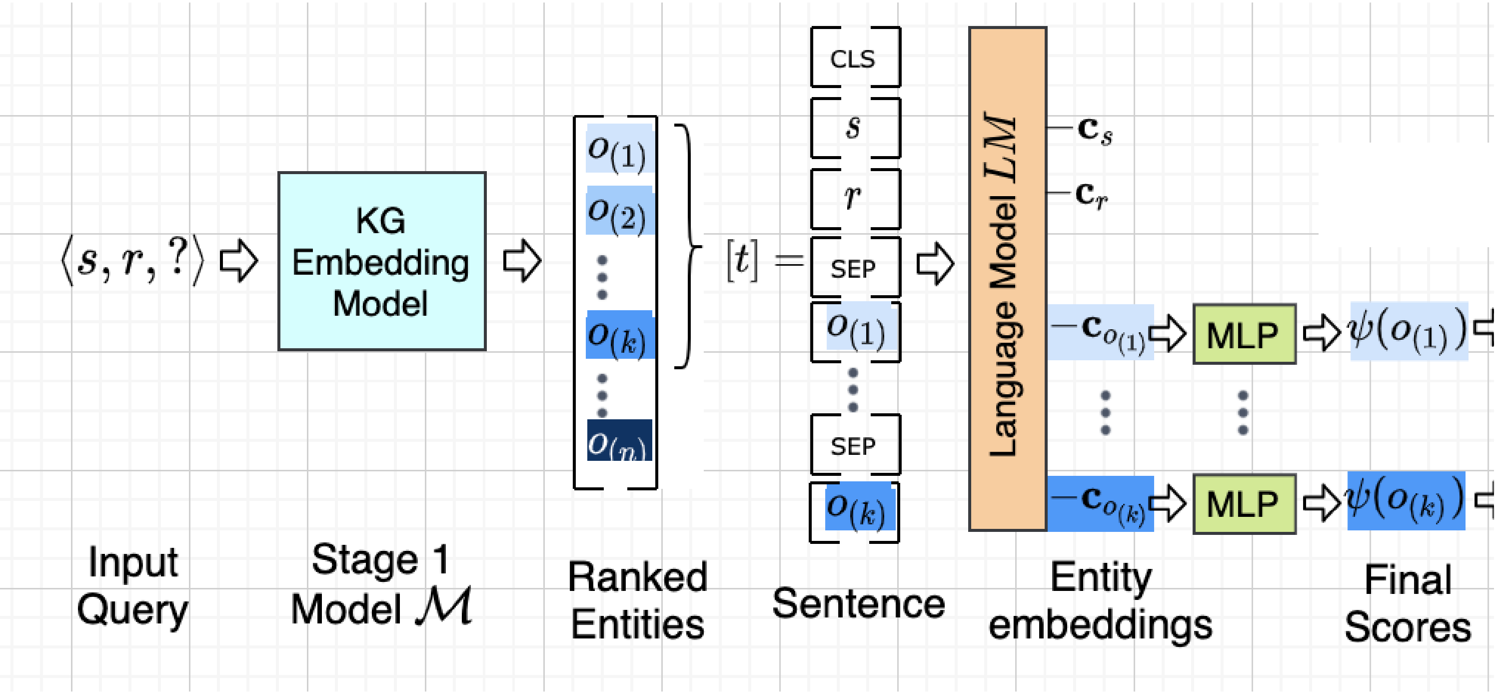


CEAR: Cross-Entity Aware Reranker for Knowledge Base Completion

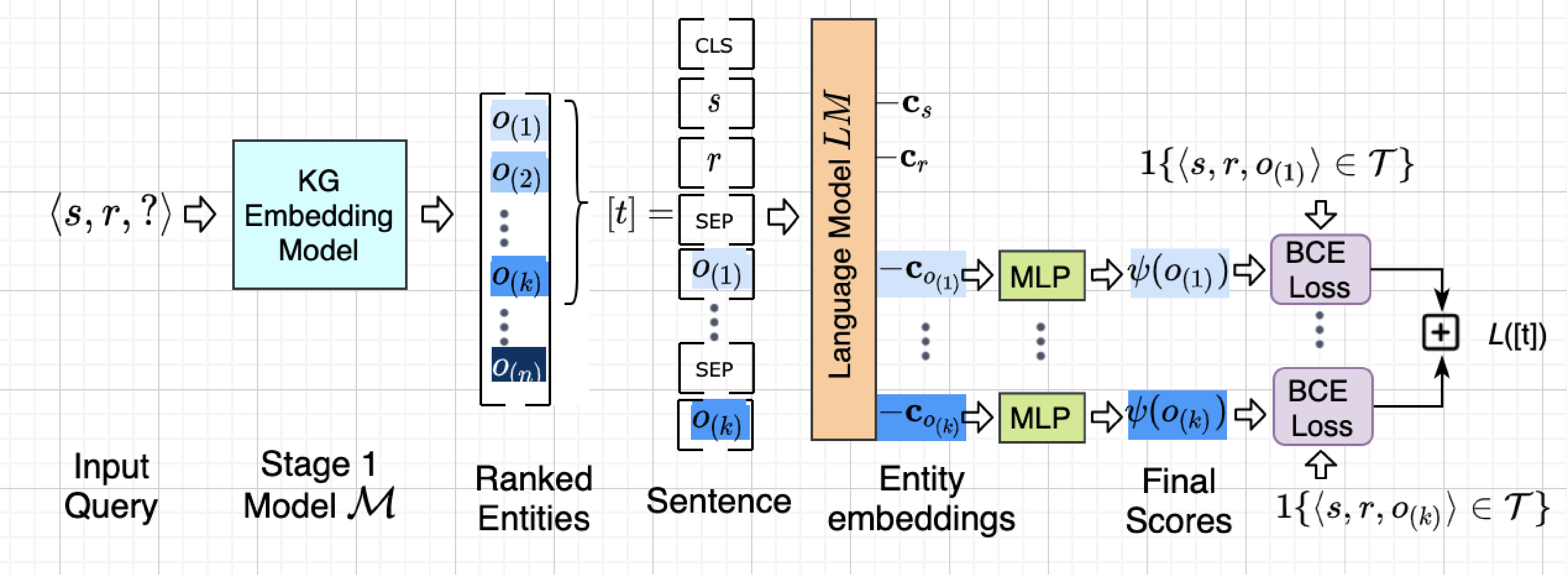


NASSCOM

CEAR: Cross-Entity Aware Reranker for Knowledge Base Completion



CEAR: Cross-Entity Aware Reranker for Knowledge Base Completion



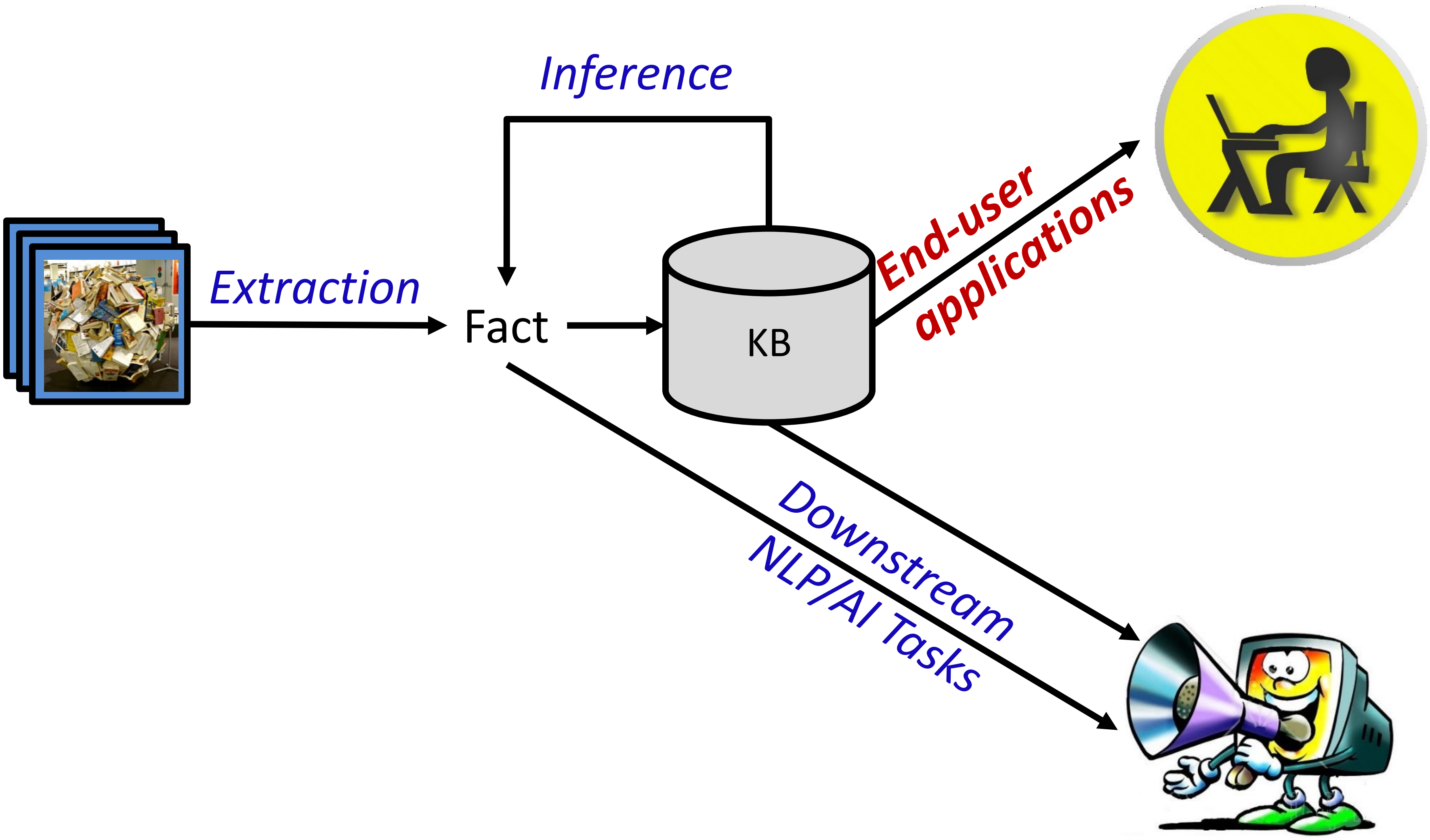
Results on OpenKB

[Kolluru, Chauhan, Nandwani, Singla, Mausam Unpublished'21]

Method	H@1	H@10	H@50
ComplEx-LSTM	2.1	7.0	14.6
ExtremeText	6.4	16.3	**26.0**
CEAR (ComplEx-LSTM)	3.8	9.1	14.6
CEAR (ExtremeText)	**7.4**	**17.9**	**26.0**

Table 3: Link Prediction performance on OLPBENCH.

Overview


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

Information Overload

Today a person is subjected to more new information in a day than a person in the middle ages in his entire life!

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)

Alzheimer's Disease Literature

[Tsutsui, Ding, Meng iConference'17]

Table 3: Two step paths from AD to HD or HIV

AD	is the most common cause of	cognitive impairment	is an early symptom of	HD
	are significantly associated with	depression	is common in	
	is characterized by	vascular dysfunction	may occur in	
	is associated with increased	neuronal death	is also a pathological hallmark in	
	is strongly correlated with	the apoe genotype	does **not** affect the course of	
AD	frequently exhibit	delirium	sometimes accompany	HIV
	is the common cause of	dementia	is a common complication of	
	affect	neurons	are **not** infected by	
	causes pro-inflammatory effects in	endothelial cells	were **not** infected with	

Health Claims in News Headlines

[Yuan, Yu COLING Workshop’18]

Information Extractor	Precision	Recall	F-measure
REVERB	.61	.31	.41
OLLIE	.62	.46	.53
OPENIE-5.0	**.67**	**.57**	**.62**
SemRep	.23	.08	.13

NASSCOM

Entity Comparisons are Ubiquitous

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]

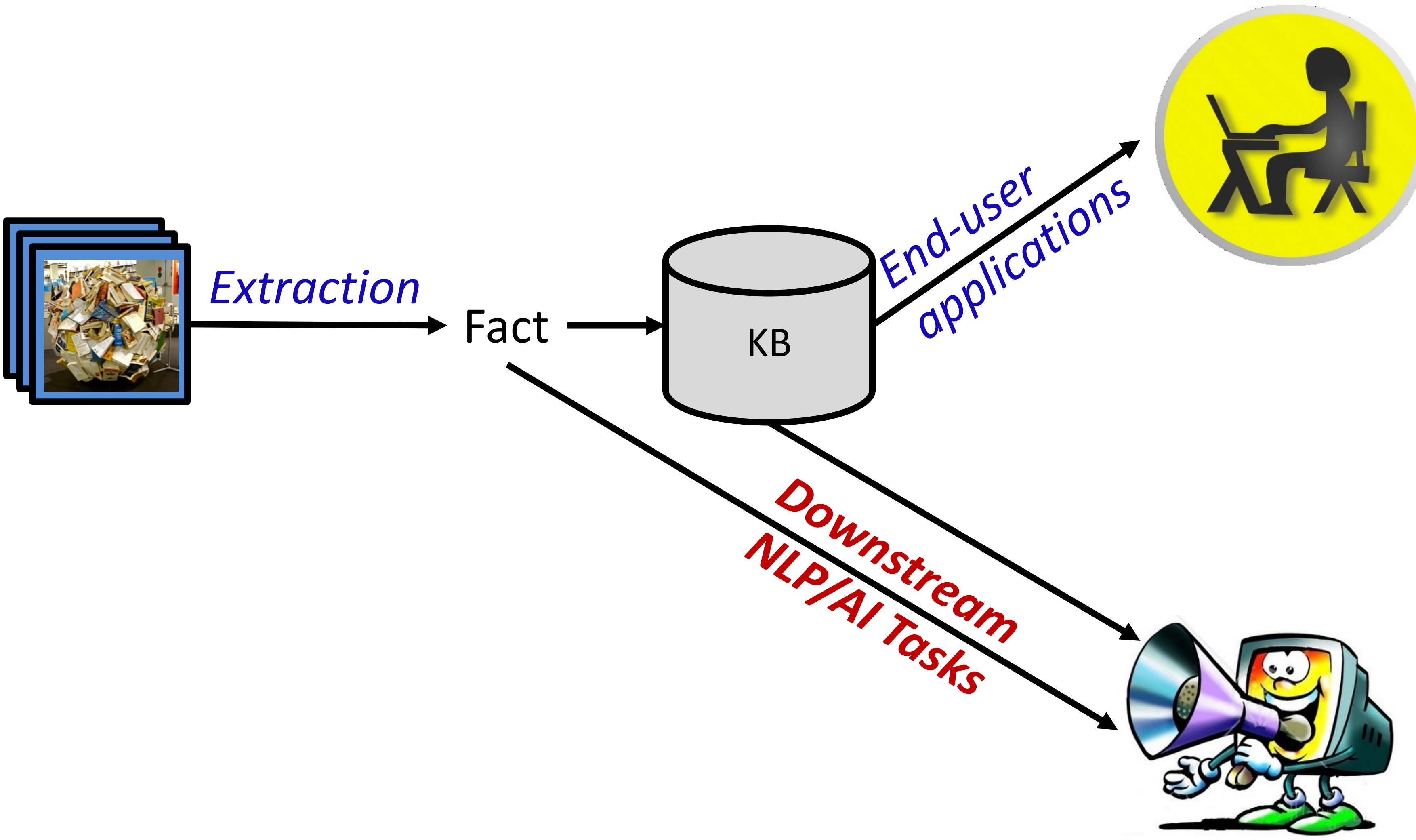
Cluster Labels	Granada (Spain)	New York City (U.S.)
art, arch.	moorish architecture religious art fine art beautiful architecture	contemporary art modern american art medieval art egyptian art
palace, courtyard	brick-walled courtyard lovely courtyard area nasrid royal palace alhambra palace	
museum, 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

Talk Outline



NLP Applications

- Improving Word Vectors
- Unsupervised KB Construction
  - Event schema induction
  - Multi-document Summarization
  - Complex Question Answering

Lexical Similarity/Analogies

[Stanovsky, Dagan, Mausam, ACL 15]

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

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

NASSCOM

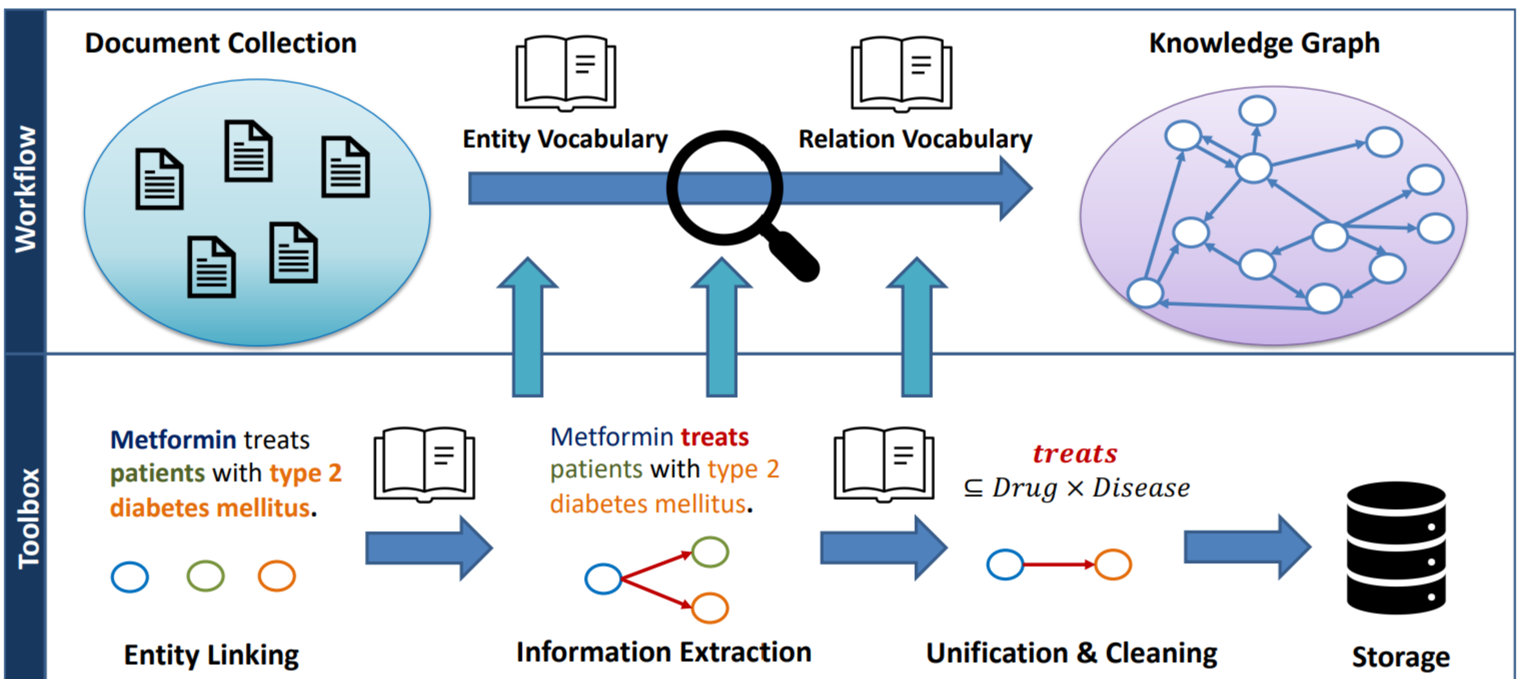
Why does Open IE do better?

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

Unsupervised KB Construction

[Kroll, Pirklbauer, Balke, JCDL'21]

- Manual domain-specific KB construction
- Expensive and Time consuming
- OpenIE can help in automation



A Probabilistic Model of Relations in Text

[Balasubramanian, Soderland, Mausam, Etzioni – AKBC-WEKEX'12]

- Rel-grams =
  a model of relation co-occurrence.
  Probability of seeing sequence of Open IE tuples.

- A resource with 27 million entries, compiled from 1.8 million news articles

Available at relgrams.cs.washington.edu

rv-n15.cs.washington.edu:10000/search?a1=&r=treat&a2=disease&v=RELARG2&m=cprobs&k=10

rel-grams Match constraints on first relation.

	Argument 1
treat	Relation
disease	Argument 2

Select view for the second relation.
RELARG2

Sort by measure
Bi-gram probability: P_k(s|f)

Co-occurrence window size (k).
10 Search

High probability tuples following (X, treat, disease):

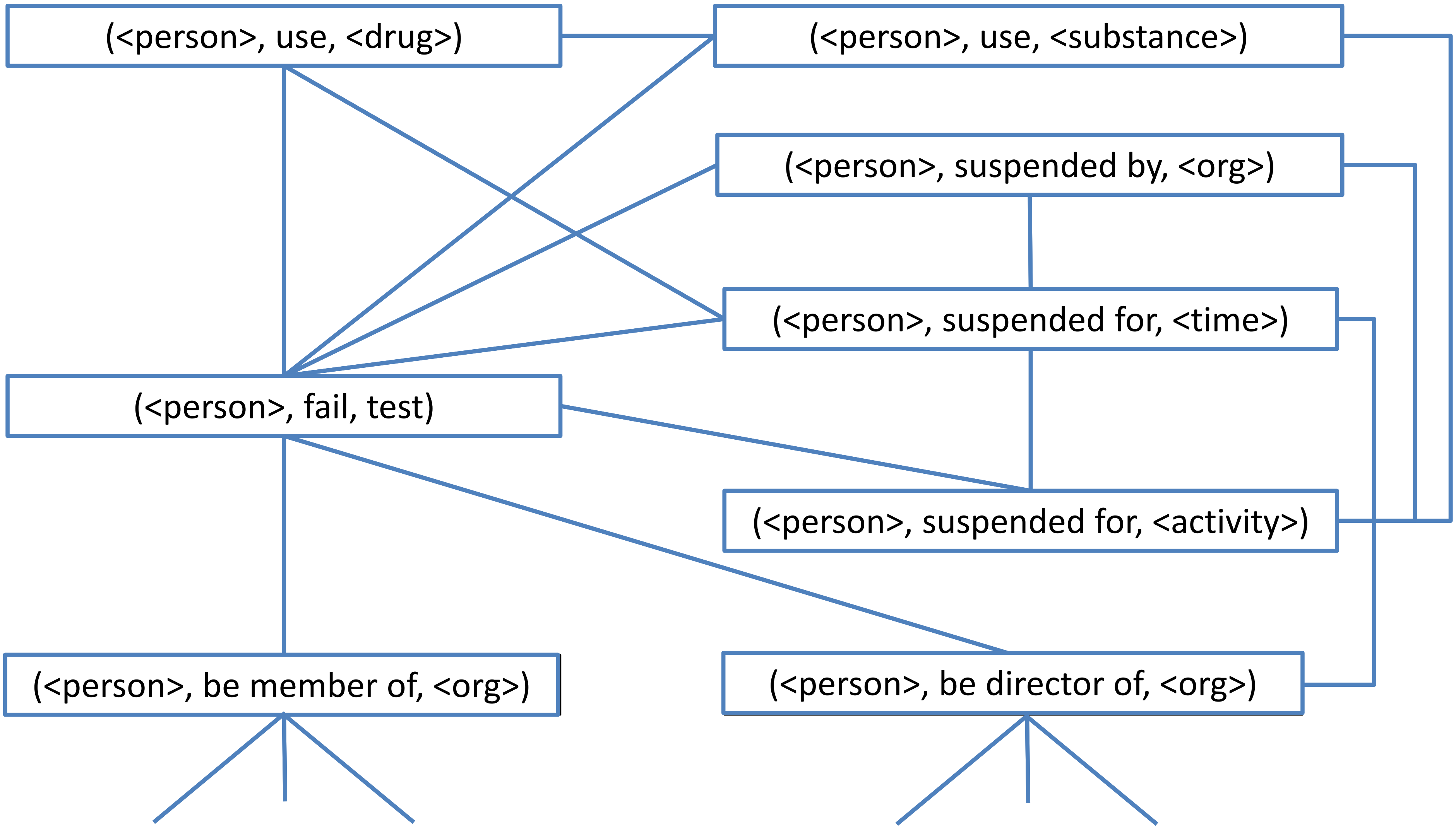
(Y, develop, drug)
(Y, cause, disease)
(Y, used to treat, condition)
...

Found 65 rel-grams.

First Tuple (f)	Second Tuple (s)	$P(R_{i+10}=s \mid R_i=f)$	$\#(R_i=f,\ldots,R_{i+10}=s)$	$\#(R_i=f,\ldots,R_{i+10}=*)$
(X, treat, disease)	(Y, develop, drug)	0.017	4.0	221.0
(X, treat, disease)	(Y, cause, disease)	0.017	4.0	221.0
(X, treat, disease)	(Y, use to treat, condition)	0.013	3.0	221.0
(X, treat, disease)	(Y, trigger response from, muscle)	0.013	3.0	221.0
(X, treat, disease)	(Y, treat, patient)	0.013	3.0	221.0
(X, treat, disease)	(Y, show that, protease inhibitor)	0.013	3.0	221.0
(X, treat, disease)	(Y, reach by, e-mail)	0.013	3.0	221.0
(X, treat, disease)	(Y, know, it)	0.013	3.0	221.0

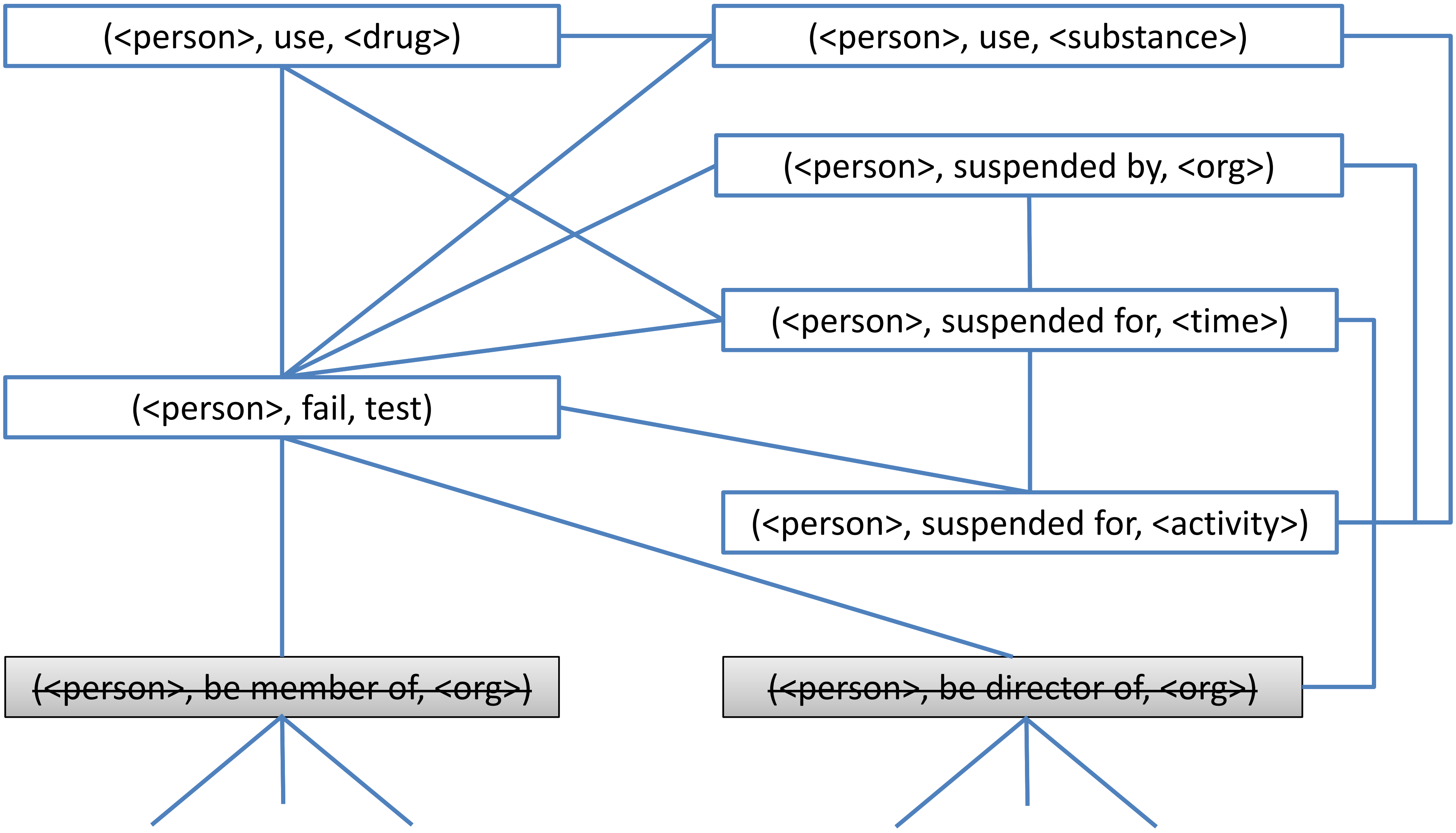
NASSCOM

Personalized PageRank over RelGram Graph



NASSCOM

Personalized PageRank over RelGram Graph



Extract Actors → Event Schemas

[Balasubramanian, Soderland, Mausam, Etzioni – EMNLP’13]

Actor	**Rel**	**Actor**
A1:<person>	failed	A2:test
A1:<person>	was suspended for	A3:<time period>
A1:<person>	used	A4:<substance, drug>
A1:<person>	was suspended for	A5:<game, activity>
A1:<person>	was in	A6:<location>
A1:<person>	was suspended by	A7:<organization, person>
Actor Instances:		
A1: {Murray, Morgan, Governor Bush, Martin, Nelson}		
A2: {test}		
A3: {season, year, week, month, night}		
A4: {cocaine, drug, gasoline, vodka, sedative}		
A5: {violation, game, abuse, misfeasance, riding}		
A6: {desert, Simsbury, Albany, Damascus, Akron}		
A7: {Fitch, NBA, Bud Selig, NFL, Gov Jeb Bush}		

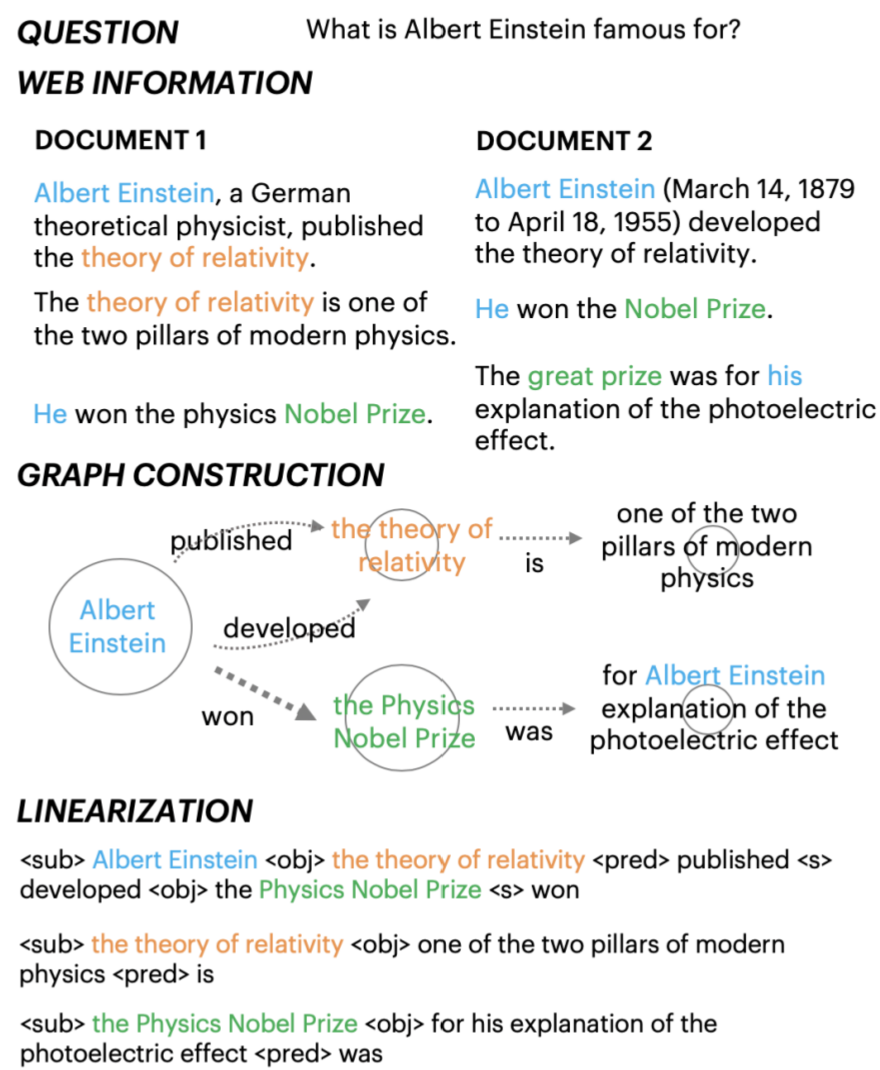
NASSCOM

Multi-document Summarization

[Fan, Gardent, Braud, Bordes, EMNLP'19]

- Use OpenIE to create dynamic Knowledge Graphs from multiple documents

- Use graph summarization



Complex Question Answering

[Khot, Sabharwal, Clark, ACL'17]

- Science Questions are often complicated and require background knowledge
- OpenIE converts background knowledge into tuples to help answer the question

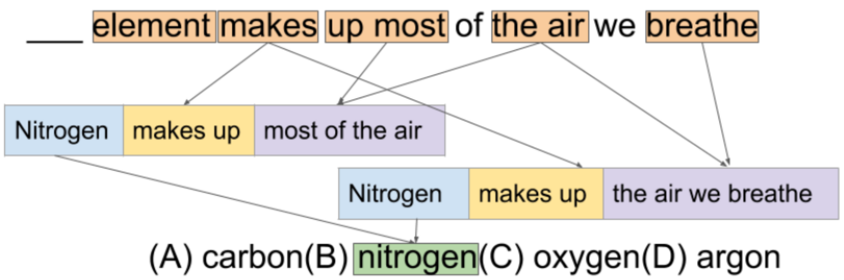


Figure 1: An example support graph linking a question (top), two tuples from the KB (colored) and an answer option (nitrogen).

Conclusions

- Populating a KB: starting to achieve some maturity
  - still many phenomena waiting to be modeled
- KBs adds tremendous value to end-user apps
  - summarization, data exploration, q/a
  - Complex QA, dialog
- KBs valuable for downstream NLP tasks
  - event schema induction
  - sentence similarity
  - text comprehension
  - vector embeddings
- Exciting research challenges in inference, QA, dialog space

Thanks

THANK YOU