

Open Information Extraction

Mausam

Associate Professor

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)

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!

Paradigm Shift: from retrieval to reading

Who won Bigg Boss 12?

Dipika Kakar

What sport teams are based in Arizona?

Phoenix Suns, Arizona Cardinals,...



Information Food Chain



Paradigm Shift: from retrieval to reading

Quick view of today's news



Science Report

Finding: beer that doesn't give a hangover

Researcher: Ben Desbrow

Country: Australia

Organization: Griffith Health Institute

Google

World Wide Web



on Food Chain

Paradigm Shift: from retrieval to reading

Compare Roku vs Fire



most apps but not iTunes
remote
good UI
works perfectly
needs laptop during travel

most apps but not Vudu, iTunes
voice-controlled remote
good UI
blames router
connects easily during travel



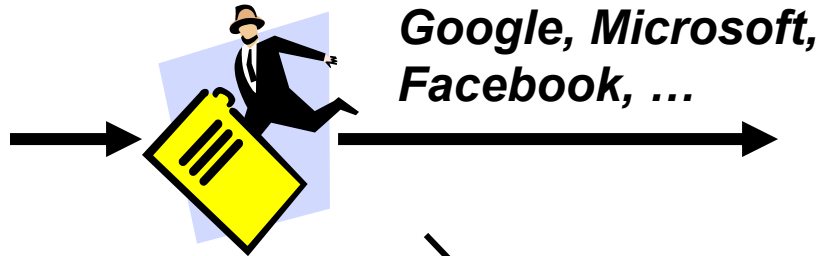
World Wide Web



Food Chain

Paradigm Shift: from retrieval to reading

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

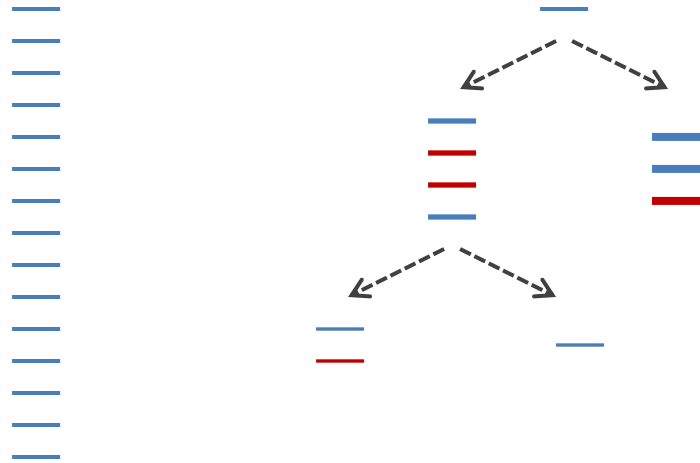


Information Food Chain



Information Systems Pipeline

Data → Information → Knowledge → Wisdom



Text → Facts → Knowledge Base → Applications



(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)

...

Lessons from DB/KR Research

- Declarative KR is expensive & difficult
- Formal semantics is at odds with
 - Broad scope
 - Distributed authorship
- KBs are brittle: “can only be used for tasks whose knowledge needs have been anticipated in advance” (Halevy IJCAI '03)

Motivation

- General purpose
 - hundreds of thousands of relations
 - thousands of domains
- Scalable: computationally efficient
 - huge body of text on Web and elsewhere
- Scalable: minimal manual effort
 - large-scale human input impractical
- Knowledge needs not anticipated in advance
 - rapidly retargetable



Open IE Guiding Principles

- Domain independence
 - Training for each domain/fact type not feasible
- Scalability
 - Ability to process large number of documents fast
- Coherence
 - Readability important for human interactions



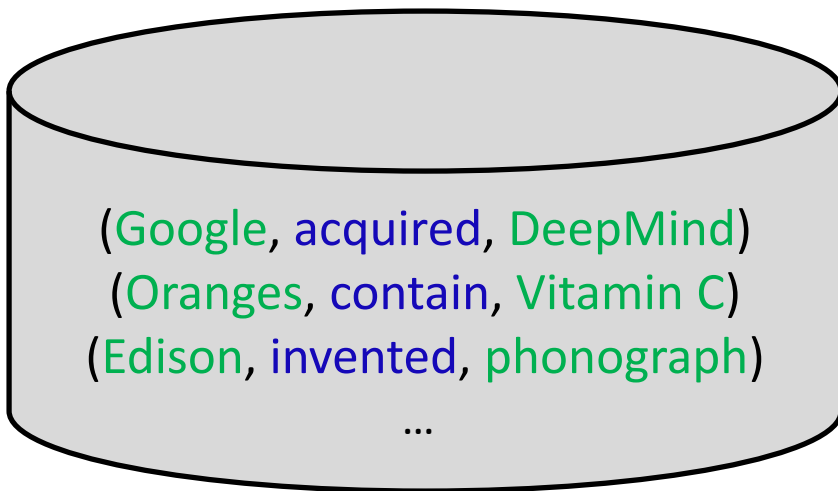
Open Information Extraction

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)



Argument 1: Relation: kills Argument 2:

antibiotics (381)
Chlorine (113)
Ozone (61)
Heat (60)
Honey (55)
Benzoyl peroxide (45)

The heat kills the bacteria .
Heat kills the bacteria .
The heat kills bacteria .
Only heat 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 .

Open vs. Closed IE


	Closed IE	Open IE
Input:	Corpus + Hand-labeled Data	Corpus + Existing resources
Relations:	Specified in Advance	Discovered Automatically
Complexity:	$O(D * R)$ R relations	$O(D)$ D documents
Output:	R relations	all relations
Consistency:	semantic rels	textual rel phrases

Demo

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

Open Information Extraction

- 2007: Texrunner (~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 (under development)
 - neural model for Open IE



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 **Relations** from **Verbs**.

1. Find longest phrase matching a simple syntactic constraint:

$$V \mid VP \mid VW^*P$$

V = verb particle? adv?

W = (noun | adj | adv | pron | det)

P = (prep | particle | inf. marker)

Sample of ReVerb Relations

invented

**inhibits tumor
growth in**

**has a maximum
speed of**

gained fame as

**was the first
person to**

acquired by

voted in favor of

**died from
complications of**

**granted political
asylum to**

**identified the cause
of**

has a PhD in

won an Oscar for

mastered the art of

**is the patron
saint 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 \approx { Anne mothers
George England
Hubbins quality footwear
....

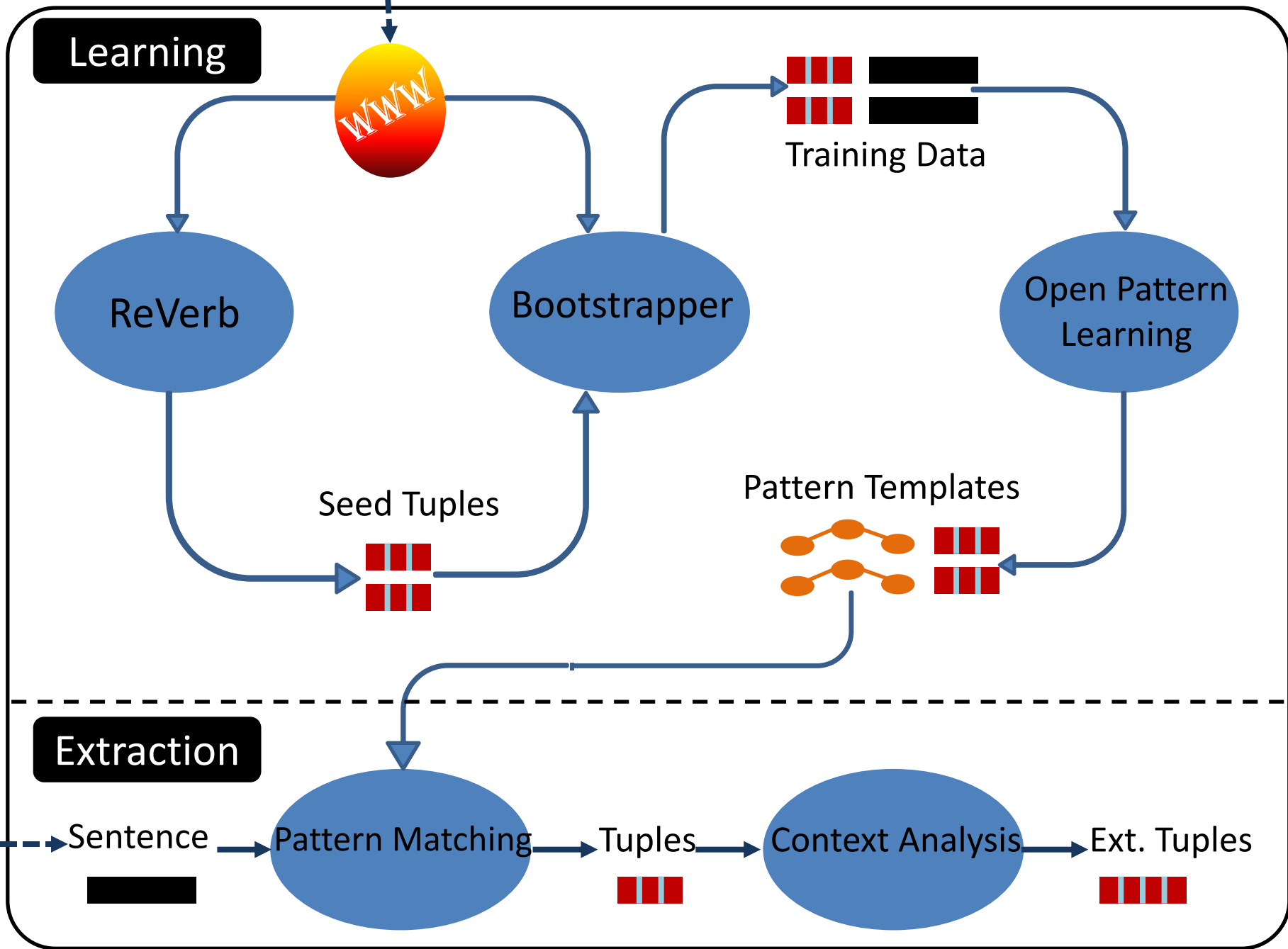
Number of Relations

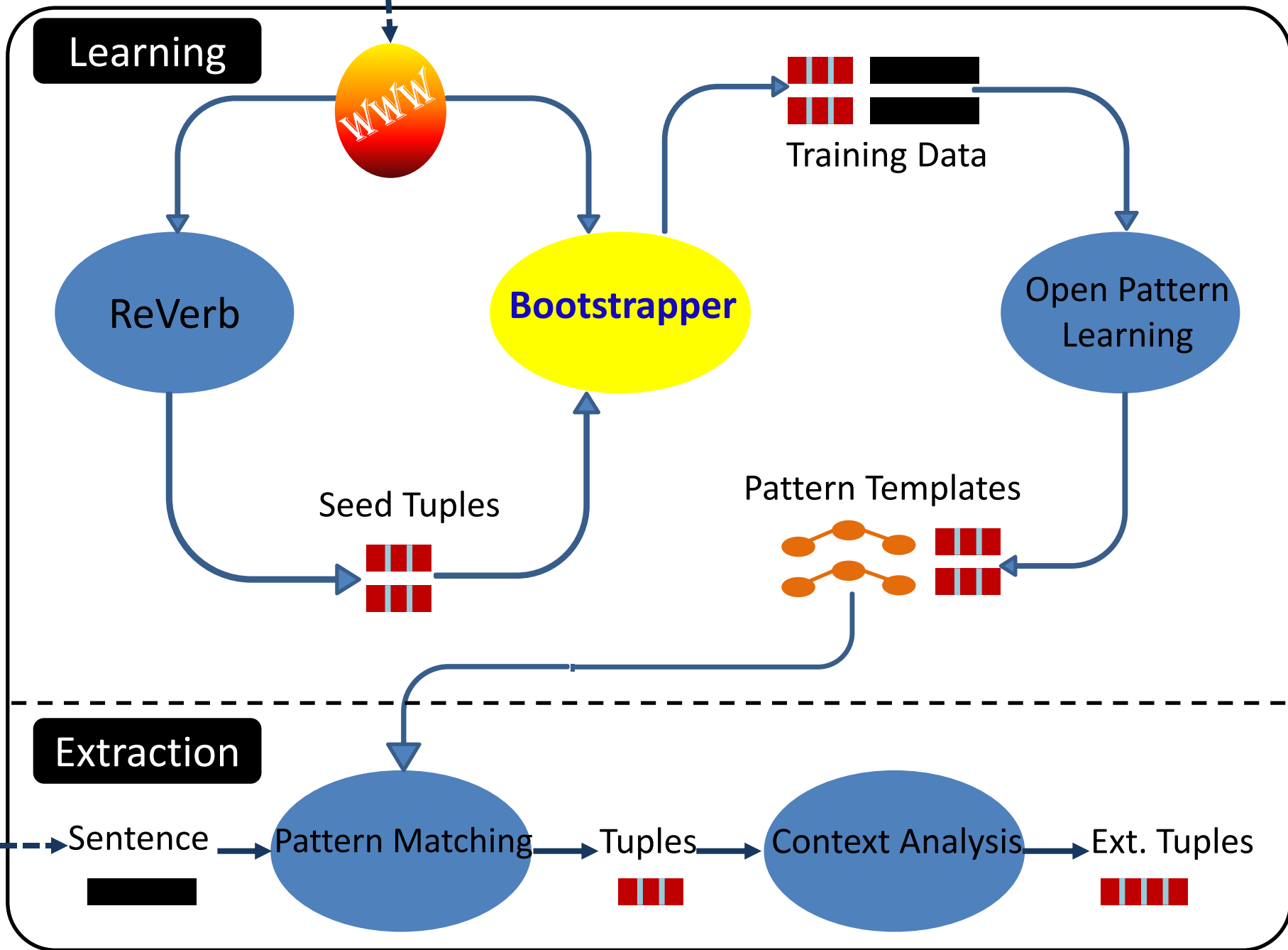
DARPA MR Domains	<50
NYU, Yago	<100
NELL	~500
DBpedia 3.2	940
PropBank	3,600
VerbNet	5,000
WikiPedia InfoBoxes, $f > 10$	~5,000
TextRunner (phrases)	100,000+
ReVerb (phrases)	1,500,000+

ReVerb: Error Analysis

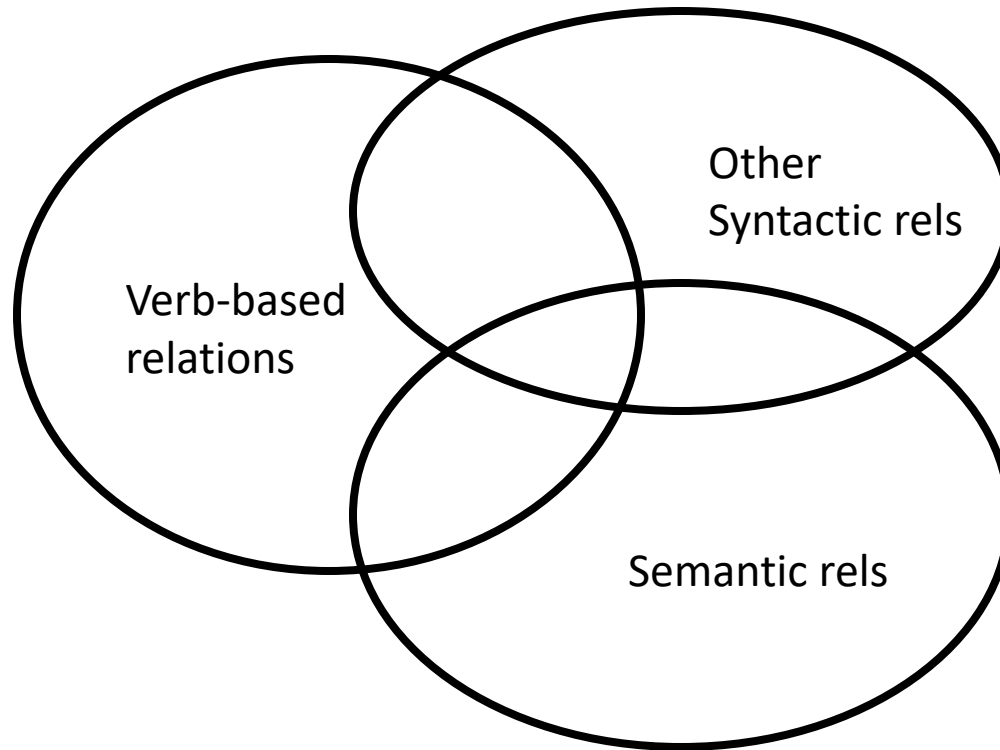
- Ginni Rometty, the CEO of IBM, talks about artificial intelligence.
- After winning the Superbowl, the Giants are now the top dogs of the NFL.
- Ahmadinejad was *elected* as the new President of Iran.

**OLLIE: Open Language Learning
for Information Extraction**



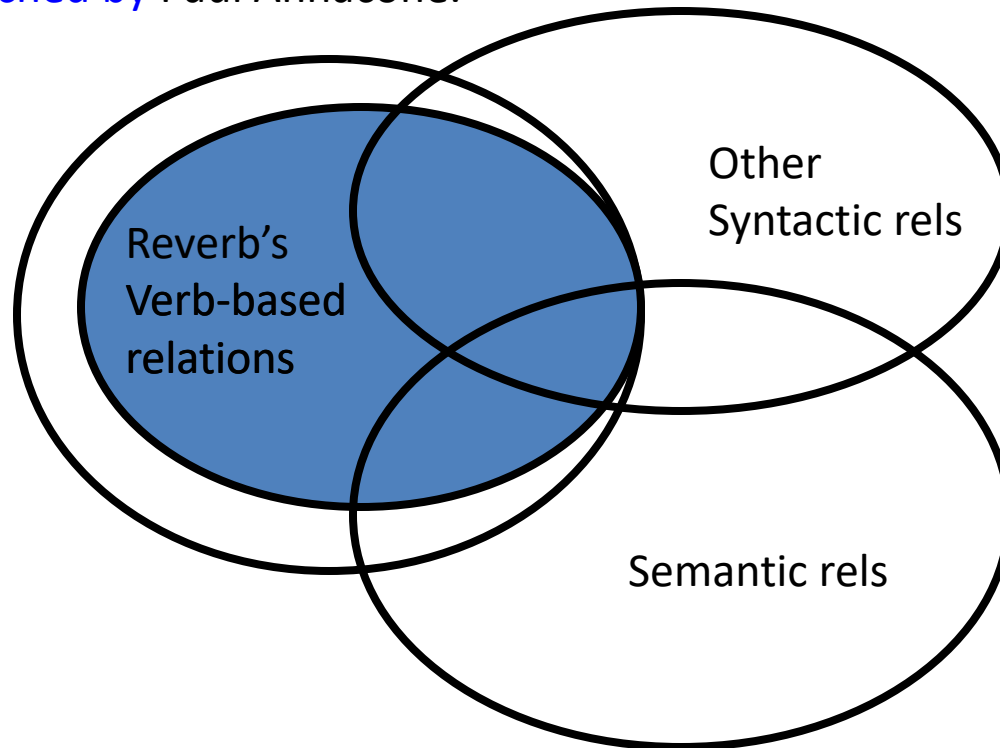


Bootstrapping Approach



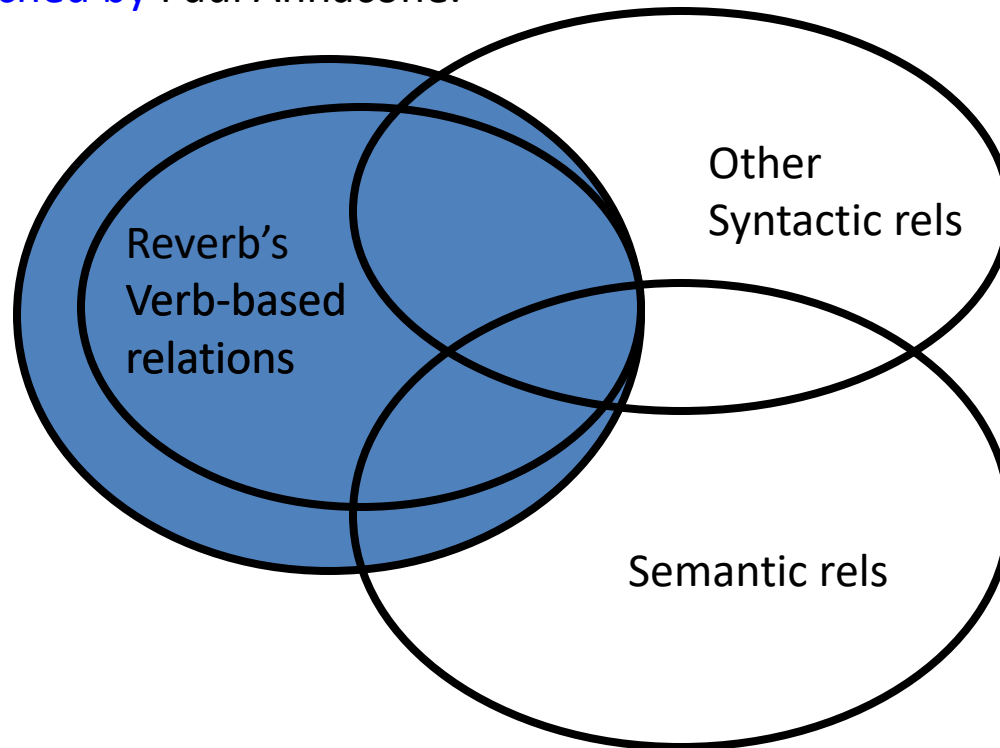
Bootstrapping Approach

Federer *is coached by* Paul Annacone.



Bootstrapping Approach

Federer *is coached by* Paul Annacone.

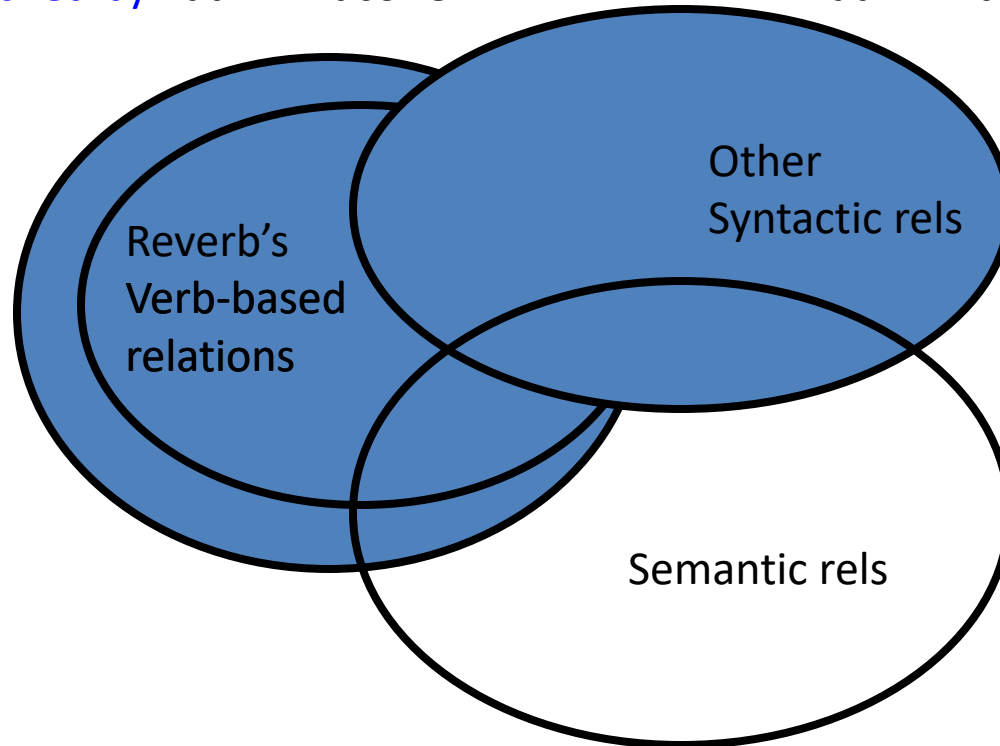


Now *coached by* Paul Annacone, Federer has ...

Bootstrapping Approach

Federer *is coached by* Paul Annacone.

Paul Annacone, *the coach of* Federer,

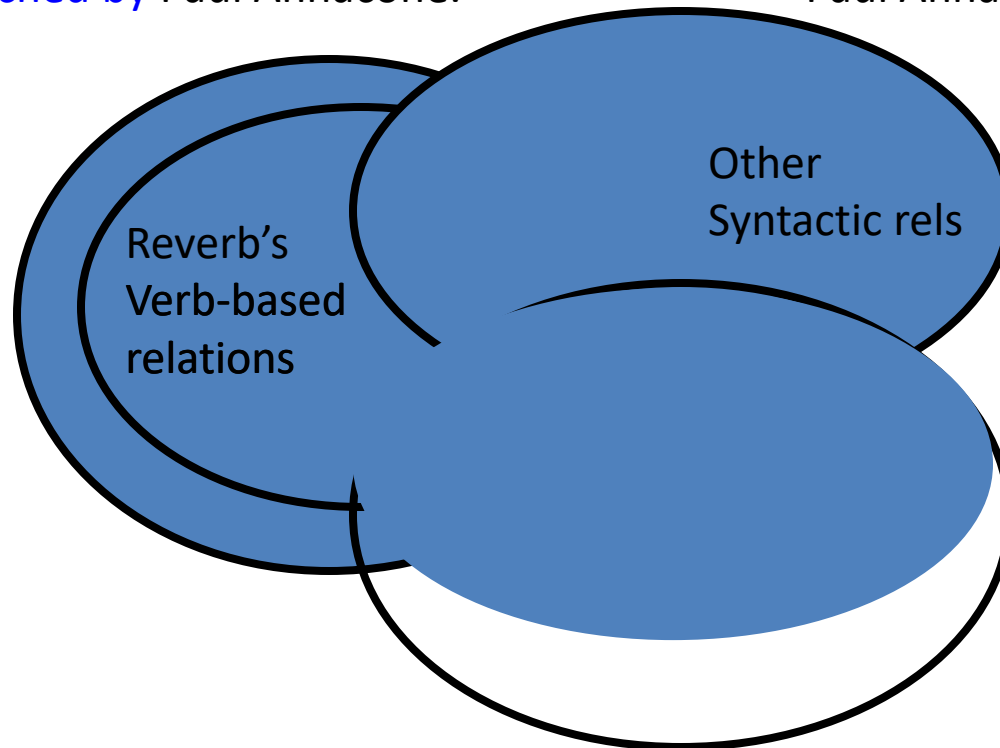


Now *coached by* Paul Annacone, Federer has ...

Bootstrapping Approach

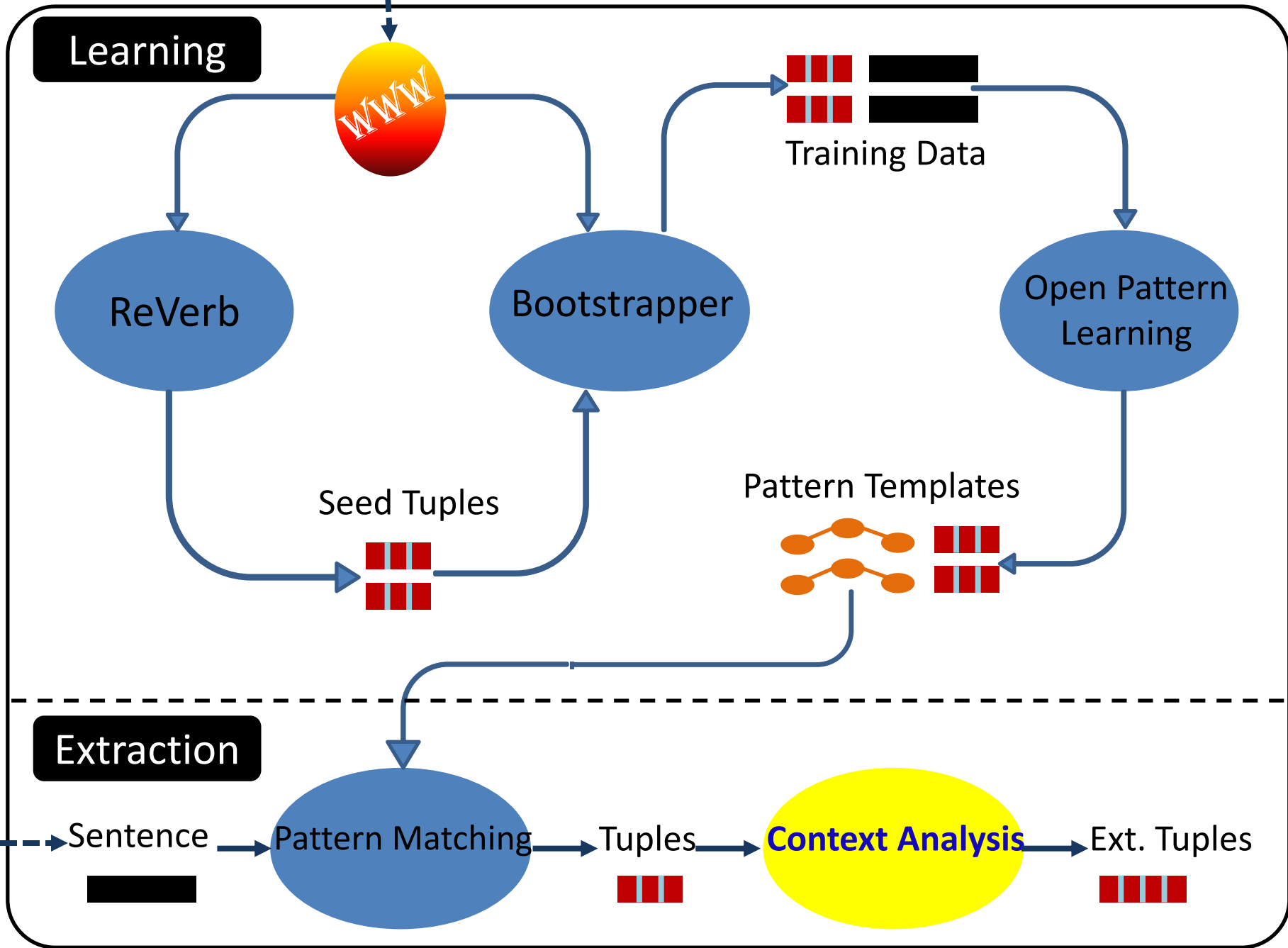
Federer *is coached by* Paul Annacone.

Paul Annacone, *the coach of* Federer,



Now *coached by* Paul Annacone, Federer has ...

Federer *hired* Annacone as his new *coach*.



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]

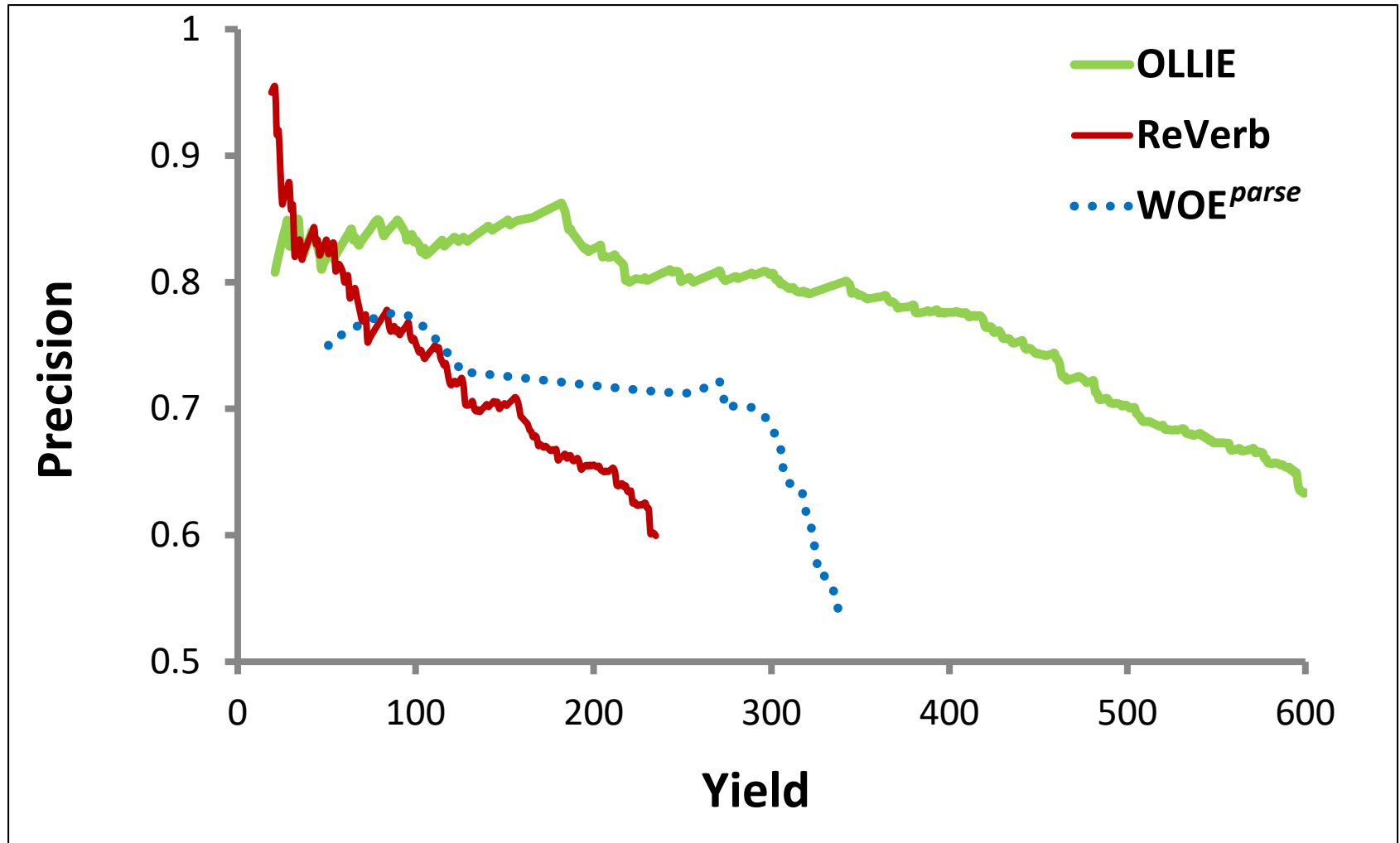
“If she wins California, Hillary will be the nominated presidential candidate.”



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


Evaluation

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



Open Information Extraction

- 2007: Texrunner (~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 (under development)
 - neural model for Open IE



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	<i>NIH director Francis Collins</i>	<i>(Francis Collins; [is] director [of]; NIH)</i>
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

COMPOUND

– GM Vice Chairman Bob Lutz

RELATIONAL NOUNS

Experiments

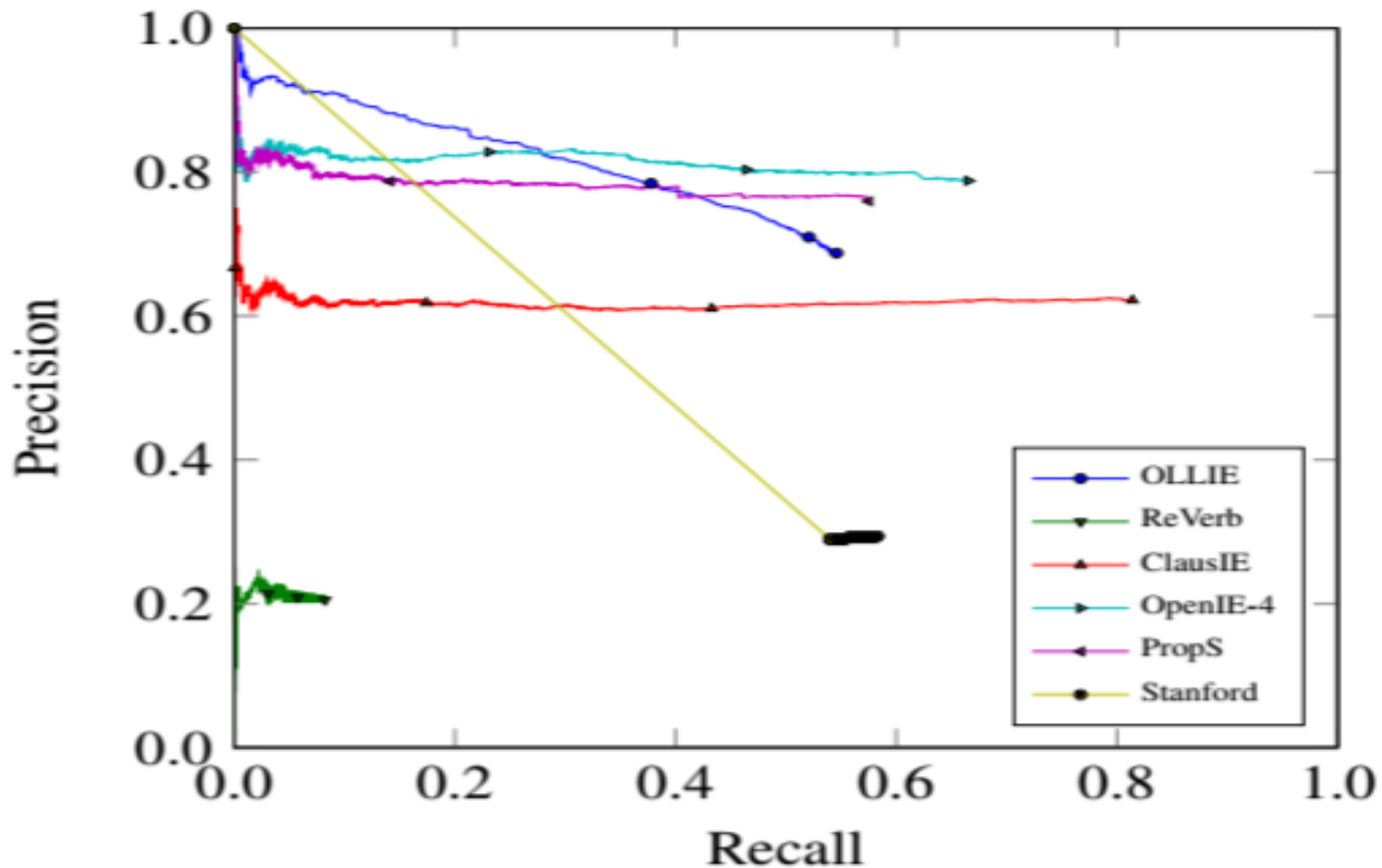
[Pal & Mausam AKBC'16]

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
+ compound relational nouns	0.69	209

RelNoun 2.0 →

Third Party Evaluation

[Stanovsky & Dagan ACL 2016]



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)

Open Information Extraction from Conjunctive Sentences

Swarnadeep Saha

IBM Research – India

and

Mausam

Indian Institute of Technology, Delhi

Nested Lists in Open IE

[Saha, Mausam COLING'18]

“President Trump met the leaders of India and China.”

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

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

“Barack Obama visited India, Japan and South Korea.”

Open IE 4: (Barack Obama, visited, India, Japan and South Korea)

Open IE 5: (Barack Obama, visited, India)
(Barack Obama, visited, Japan)
(Barack Obama, visited, South Korea)

Contributions

- CALM (Coordination Analyzer using Language Model)
 - Disambiguates conjunct boundaries
 - by correcting typical errors from dependency parses.
 - Single Coordinating Conjunction: use of language model
 - Multiple Coordinating Conjunction
 - Use of Hierarchical Coordination Tree (HCTree)
- CALMIE
 - New Open IE system
 - Uses output generated by CALM
 - Outperforms state-of-the-art Open IE systems on conjunctive sentences

Language Model for Disambiguation

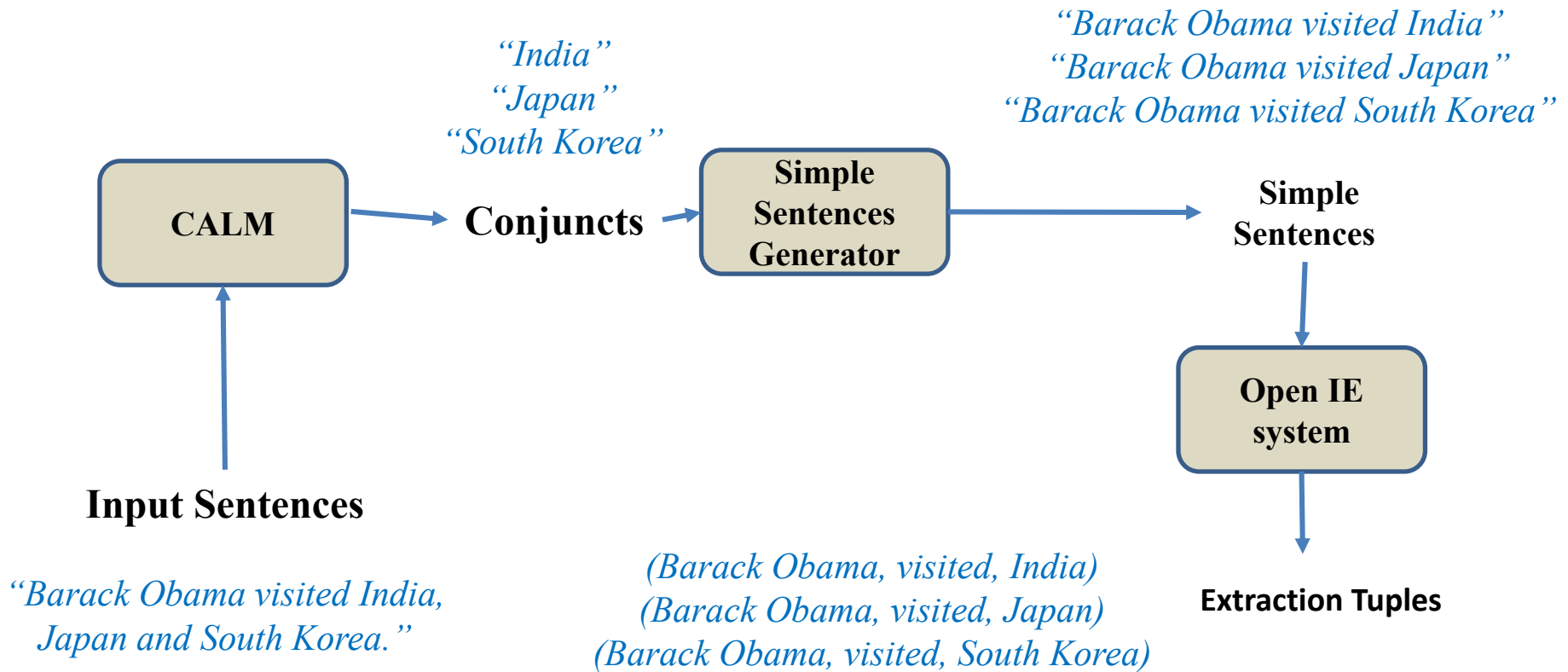
“President Trump met (the leaders of India) and (China).”

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

“President Trump met the leaders of (India) and (China).”

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

Flow Diagram



CALM: Only One Conjunction in Sentence

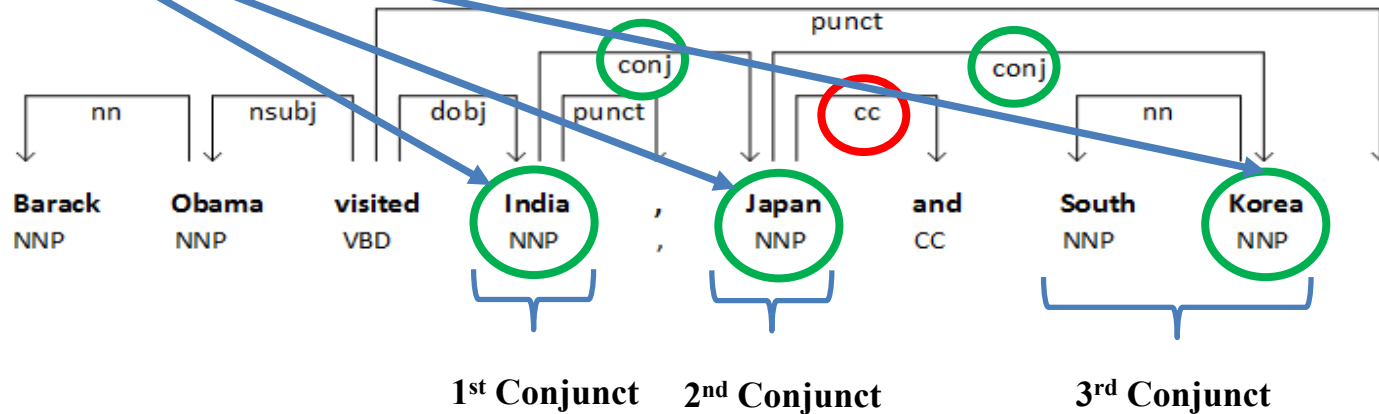
Rule Based Baseline

1. Generate the dependency parse of the sentence.
2. Identify all conjunctions “and”, “or”, etc.
 - a. Check for “cc” edges in the dependency parse
3. For each conjunction, identify corresponding conjunct headwords.
 - a. Node connected by a “cc” edge with conjunctive word is the first headword.
 - b. Subsequent headwords are connected by “conj” edges.
4. Form each conjunct by expanding the subtree(excluding “cc” and “conj” edges) under the conjunct headword.
5. Split sentence about each conjunction → all possible simple sentences.

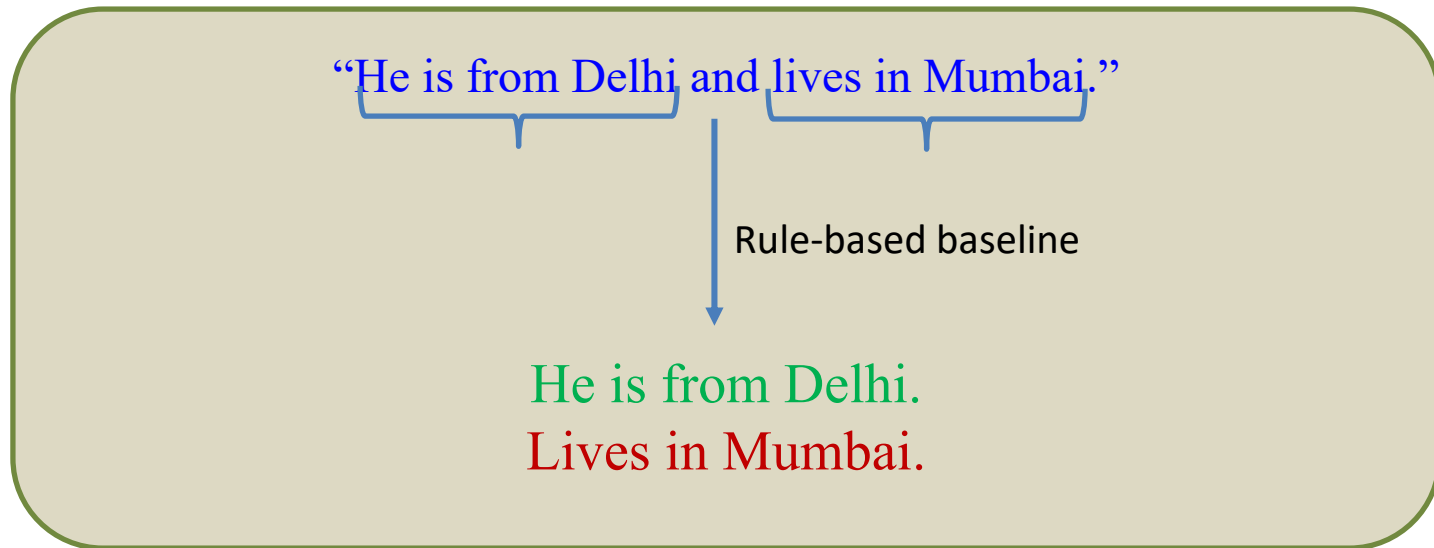
Rule-based Baseline

Conjunct
Heads

Dependency Parser



Rule-based Baseline - Errors



> 80 % of incorrect conjunct boundaries are longer than necessary

“He is from Delhi and lives in Mumbai.” -->



Correct conjunct

- He is from Delhi. ✓
- He lives in Mumbai. ✓

Rule Based Baseline - Errors

Correct conjunct

“He rejoices at the fact that she started off with smalltown views, and began thinking globally.”

Wrong conjunct

- *“He rejoices at the fact that she started off with smalltown views.”* ✓
- *“He rejoices at the fact began thinking globally.”* ✗

Rule Based Baseline - Observations

- Each conjunct is a contiguous span of words.
- The conjuncts are separated by commas or a conjunctive word.
- **Boundaries** - Start of first conjunct and end of last conjunct.
- A better algorithm should fix the incorrect boundaries.
- Baseline generates grammatically incorrect simple sentences.
- **Idea** - Choose boundaries such that the resultant simple sentences are grammatically correct.

Language Model and Its Use

- Language Model computes the probability of a sentence or a sequence of words.

$$\begin{aligned} P(W) &= P(w_1, w_2, w_3, w_4, w_5 \dots w_n) \\ &= P(w_1) * P(w_2 | w_1) * P(w_3 | w_1 w_2) * \dots * P(w_n | w_1, w_2, \dots, w_{n-1}) \end{aligned}$$

- $P(\text{Correct Sentence}) > P(\text{Incorrect Sentence})$.
- Shift the boundaries given by the Rule Based Algorithm.
- Choose that boundary that gives the highest average Language Model score for the simple sentences.

Language Model-based Algorithm

“He is from Delhi and lives in Mumbai.”

S1: Lives in Mumbai.

S2: He lives in Mumbai.

S3: He is lives in Mumbai.

S4: He is from lives in Mumbai.

$P(S2) > P(S1)$

$P(S2) > P(S3)$

$P(S2) > P(S4)$

- Use Language Model to compute probabilities.
 - correction for length of simple sentences
- Pick the configuration with the highest value.

Similarly to fix the end of last conjunct, we do a left shift of the last conjunct.

Problem 1

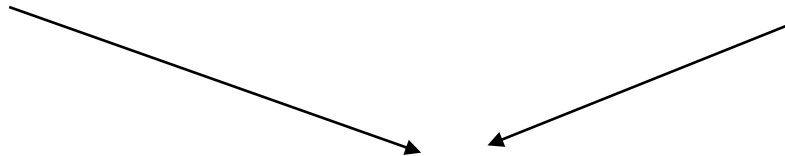
- Unequal number of n-grams in the sentences.

S_1 : "Well"

S_2 : "She sings really well"

$$P(S_1) = P(\text{Well})$$

$$P(S_2) = P(\text{She}) * P(\text{sings} | \text{She}) * P(\text{really} | \text{She sings}) * P(\text{well} | \text{She sings really well})$$



How to compare??

- **Approach 1** - Take $|S|$ -th root of the probability value if there are $|S|$ words in the sentence.

Problem 2

“To the best of my knowledge he is from Delhi and lives in Mumbai.”



- S1: To the best of my knowledge lives in Mumbai.
- S2: To the best of my knowledge he lives in Mumbai.
- S3: To the best of my knowledge he is lives in Mumbai.
- S4: To the best of my knowledge he is from lives in Mumbai.

Not obvious which
has highest root-prob.

Approach 1 doesn't work well - For considerably longer sentence, higher probability values of certain n-grams increases the overall score of the sentence.

- Consider only those n-grams at the point of intersection.
- For incorrect sentences, their probability values will be less.
- Remove common n-grams among the sentences.

Solution

“To the best of my knowledge he is from Delhi and lives in Mumbai.”

S1: To the best of my knowledge lives in Mumbai.

$$p(\text{lives}|\text{my knowledge}) * p(\text{in}|\text{knowledge lives})$$

S2: To the best of my knowledge he lives in Mumbai.

$$p(\text{lives}|\text{knowledge he}) * p(\text{in}|\text{he lives})$$

S3: To the best of my knowledge he is lives in Mumbai.

$$p(\text{lives}|\text{he is}) * p(\text{in}|\text{is lives})$$

S4: To the best of my knowledge he is from lives in Mumbai.

$$p(\text{lives}|\text{is from}) * p(\text{in}|\text{from lives})$$

Use of Linguistic Constraints

- Each simple sentence must have a subject.
- Named Entities should not be split.
- If two verbs are adjacent, they must be light verb.
- Verb categories VBD, VBZ and VBP must precede pre-defined POS tags.

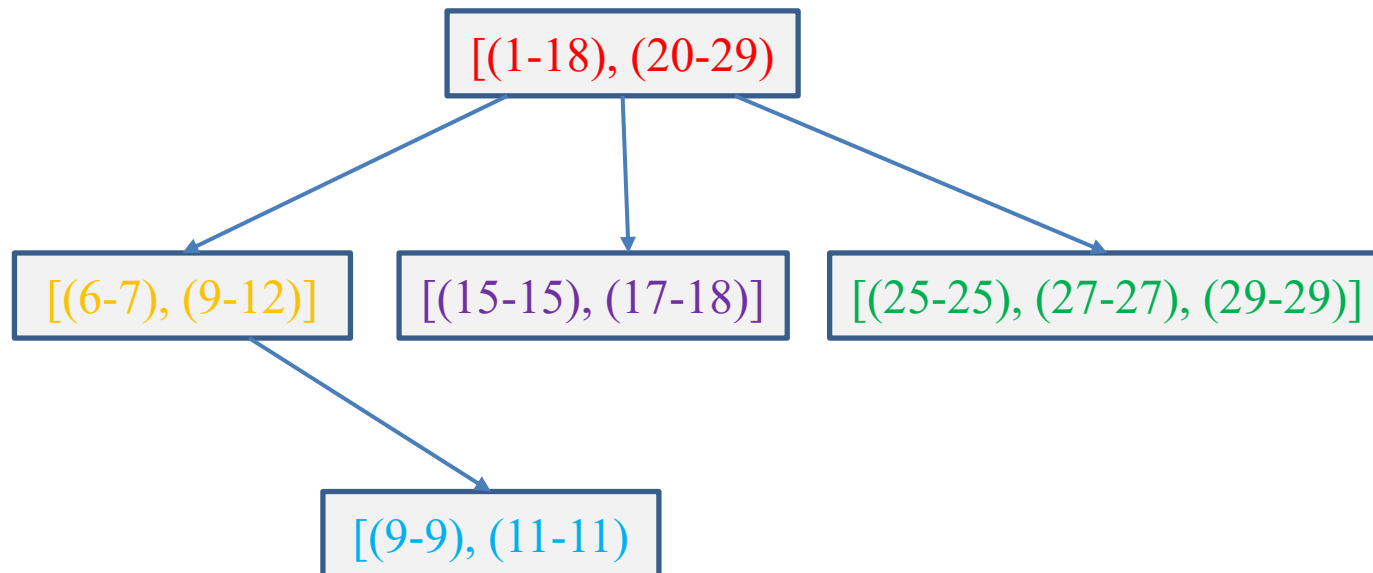
CALM: Multiple Conjunctions in Sentence

Multiple Coordinating Conjunctions

- **Coordination Structure:** Conjuncts associated with each conjunction.
- Two coordination structures have to be either disjoint or nested.
 - **Disjoint** – No word in common.
 - **Nested** – One coordination structure is contained entirely within the span of one conjunct of the other coordination structure.
- Partial intersections are **ungrammatical:** hence **not** possible
- Joint disambiguation of all coordination structures.
- **Hierarchical Coordination Tree**

Hierarchical Coordination Tree (HCTree)

“[(Jeff Bezos, an American [(electrical engineer) and [(technology) and (retail)] entrepreneur]), founded [(Amazon.com) and (Blue Origin)] and (his diversified business interests include [(books), (aerospace) and (newspapers)])].”



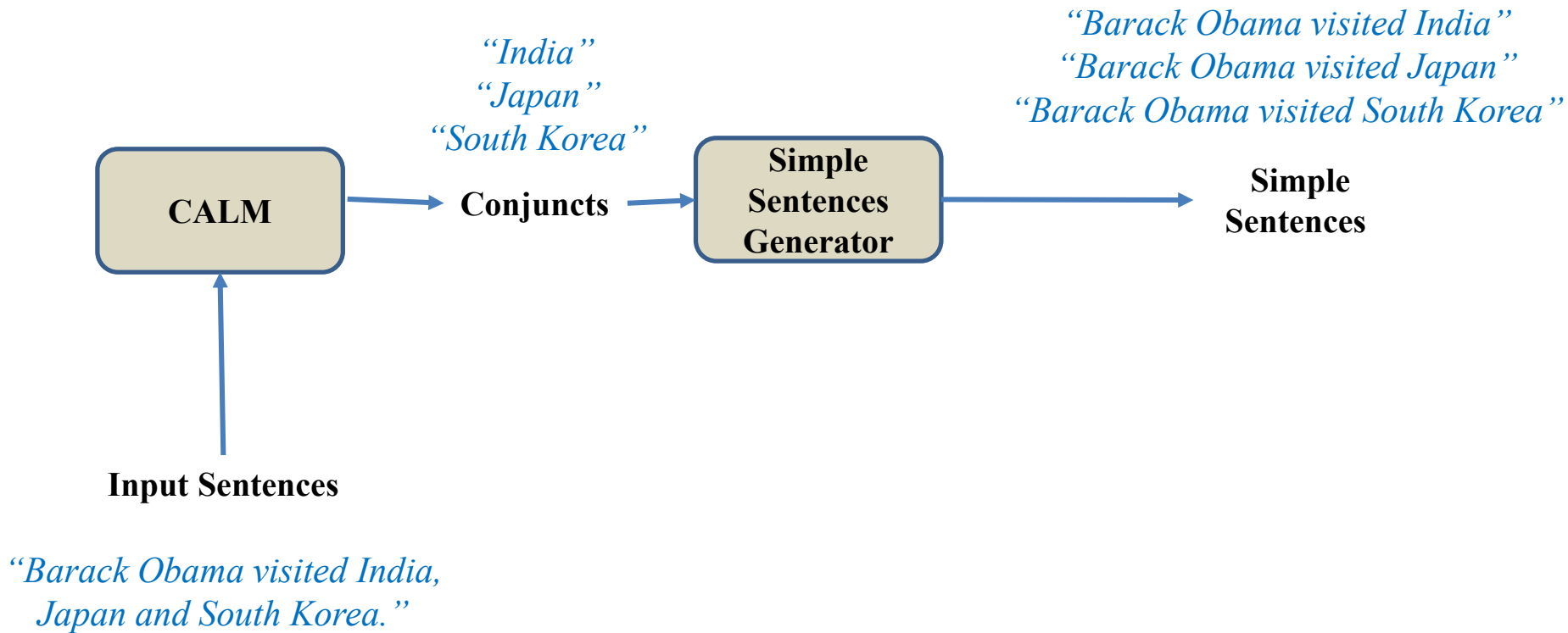
Multiple Conjunction Constraint

- Create an initial HCTree from the parse.
- In a bottom-up pass, fix the coordination structures.
 - Smaller conjuncts are easier to fix.
- Search space is reduced by keeping the structure of HCTree unchanged.
- Shortening of conjuncts ensure that the consistency of HCTree is not violated.

CALMIE:

Open IE over Conjunctive Sentences

Flow Diagram



Simple Sentence Generator

- Process the HCTree in a top-down order.
- At each level, generate all possible sentences from sentences in the previous level by concatenating parts of sentences that are not in any conjunct.
- No duplication of sentences.

Breaking into Simple Sentences

Simple sentences can be generated by processing the conjunct structures in a level order manner.

She [(wandered back into the living-room , with its [(rugged stone walls) and [(polished wood) and (leather)]])] , and (looked out again at the [(darkened skies) and (pouring rain)]].

1st Level

- *She wandered back into the living-room , with its [(rugged stone walls) and [(polished wood) and (leather)]].*
- *She looked out again at the [(darkened skies) and (pouring rain)].*

Breaking into Simple Sentences

- *She wandered back into the living-room , with its [(rugged stone walls) and ((polished wood) and (leather))].*
- *She looked out again at the [(darkened skies) and (pouring rain)].*

2nd Level

- *She wandered back into the living-room , with its rugged stone walls.*
- *She wandered back into the living-room , with its [(polished wood) and (leather)].*
- *She looked out again at the darkened skies.*
- *She looked out again at the pouring rain.*

Breaking into Simple Sentences

- *She wandered back into the living-room , with its rugged stone walls.*
- *She wandered back into the living-room , with its [(polished wood) and (leather)].*
- *She looked out again at the darkened skies.*
- *She looked out again at the pouring rain.*

3rd Level

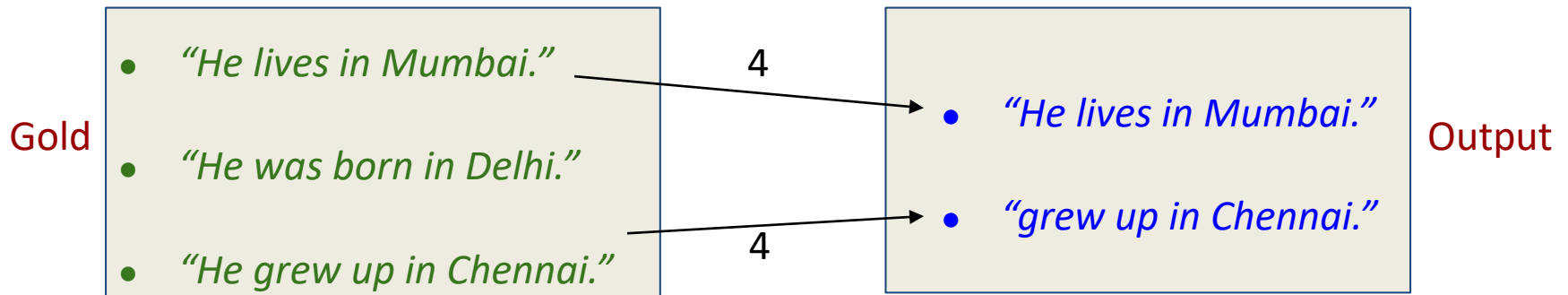
- *She wandered back into the living-room , with its rugged stone walls.*
- *She wandered back into the living-room , with its polished wood.*
- *She wandered back into the living-room , with its leather.*
- *She looked out again at the darkened skies.*
- *She looked out again at the pouring rain.*

Un-splittable Conjunctive Sentences

- Non-distributive conjunctions – “**or**”, “**nor**”.
 - “*Adam’s nationality is French or German.*”
- Paired conjunctions – “**either-or**”, “**neither-nor**”.
 - “*You will neither giggle nor smile.*”
- Non-distributive triggers like “**between**”, “**among**”, “**sum**”, etc.
 - “*The world cup final was played between Germany and Argentina.*”
 - “*The average of 3 and 5 is 4.*”
 - “*We ‘ve been humping away for a whole two and a half pages .*”

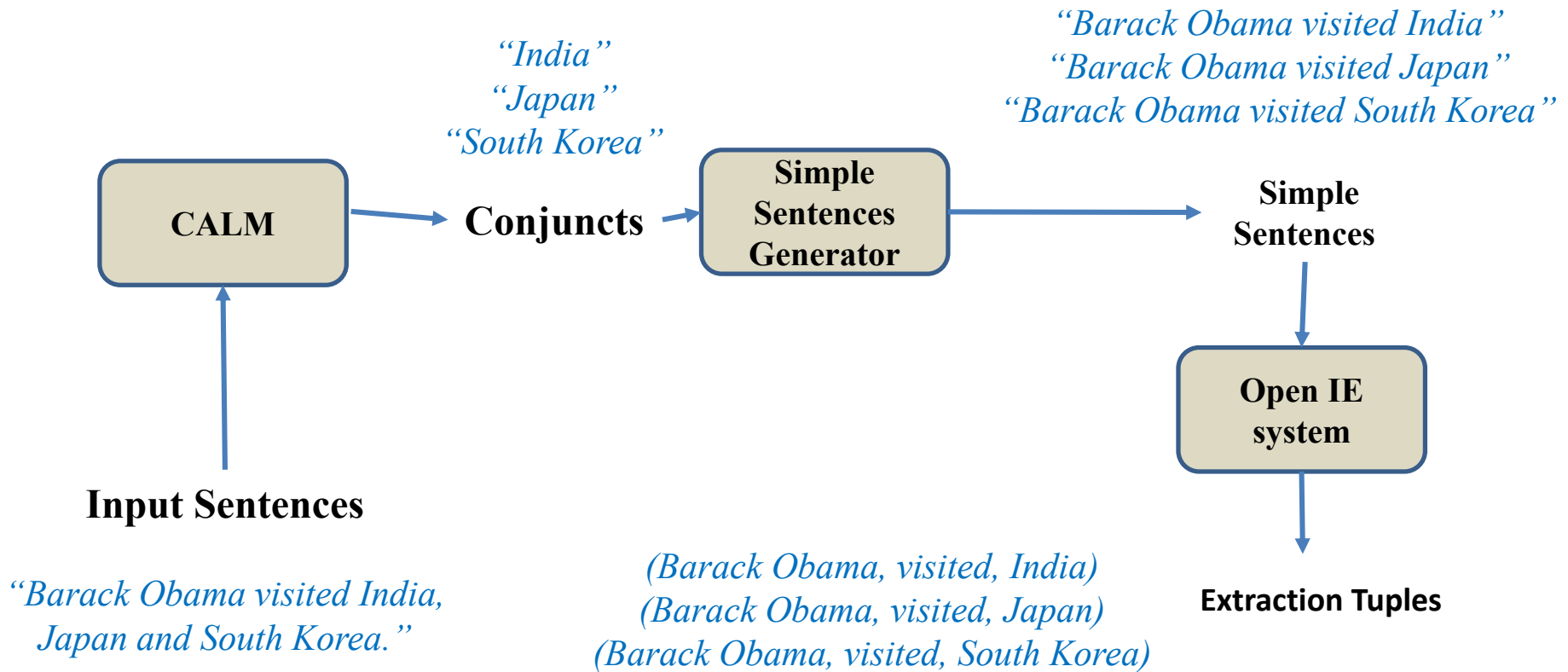
Precision and Recall

- Computed by doing a best match between the gold and output sentences and then calculating the number of common words.



- Precision = $1/2 * (4/4 + 4/4) = 1.0$
- Recall = $1/3 * (4/4 + 0/5 + 4/5) = 0.6$

Flow Diagram



CALM - Evaluation

- Previous work (Ficler and Goldberg, 2016) gives credit when the conjuncts for a sentence match exactly.
- This is not ideal!

“Obama visited India and Japan and South Korea.”

Multiple correct interpretations depending on which “and” is considered the top level conjunction.

- Compare resultant simple sentences, using traditional word overlap precision and recall.

CALM Results – BNC Test Set

	Parser Baseline (Clear Parser)			+ Language Model			+ Constraints		
	SC	MC	SC+MC	SC	MC	SC+MC	SC	MC	SC+MC
Precision	94.69	86.78	92.14	94.33	87.85	92.24	94.22	88.00	92.21
Recall	90.22	78.34	86.39	91.36	82.75	88.58	92.97	83.23	89.83
F-score	92.40	82.34	89.17	92.82	85.22	90.37	93.59	85.55	91.00

- British News Corpus test set (publicly available).
- 577 conjunctive sentences.
- 391 Single Conjunction sentences.
- 186 Multiple Conjunction sentences.

Over 3 pt improvement in multiple-conjunction case.

CALM Results – Penn Treebank

	(Ficler and Goldberg, 2016)	CALM
Precision	72.81	75.12
Recall	72.61	70.64
F1	72.7	72.81

- Comparison with SOA system on Penn Treebank dataset.
- Comparison on *only last two* conjuncts
- Evaluate using their metric – exact matches of conjunct boundaries.

CALM – Error Analysis

- Inaccuracy of parsers (absence of ‘cc’ edge).
- Missing contexts.

“Two years ago, we were carrying huge inventories and that was the big culprit.”

↓ Missing prefix context

“Two years ago, we were carrying huge inventories.”

“That was the big culprit.”

CALMIE Results: ClueWeb and News+Wiki

	ClueWeb12				News+Wikipedia			
	[C]	[Cm[C]]	[O4]	[Cm[O]]	[C]	[Cm[C]]	[O4]	[Cm[O]]
Precision	62.50		70.04		67.17		79.12	
Yield	267		199		204		172	

- 100 conjunctive sentences from ClueWeb12.
- 100 conjunctive sentences from an Open IE benchmarking dataset (Stanovsky and Dagan, 2016).
- 2 manual annotators.

- [C] = ClausIE, [Cm[C]] = CALM + ClausIE.
- [O4] = Open IE 4, [Cm[O]] = CALM + Open IE 4.

CALMIE Results – Penn Treebank

	Two Conjuncts		More than Two Conjuncts	
	[FG]	[Cm[O]]	[FG]	[Cm[O]]
Precision	72.71	72.35	74.50	74.78
Yield	323	330	346	445

- 100 sentences with two conjuncts, 95 with > two conjuncts.
- [FG] = Ficler + Open IE 4.
- [Cm[O]] = CALM + Open IE 4.
- Ficler's system always outputs only two conjuncts.
- CALMIE outputs all conjuncts.

CALMIE - Error Analysis

- Difficulty in figuring out cases when not to split.
- “Japan’s domestic sales of cars, trucks and buses in October rose by 18%.”
- “The Perch and Dolphin fields moved their headquarters.”
- “Germany and Argentina beat Brazil and Netherlands in the semis respectively.”
- Fixing these can further improve CALMIE.

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]

Critique

- Why I like this paper?
 - Emphasizes the importance of linguistics today
 - Paper writing provides intuitions every step of the way
 - First (?) paper to carefully study multi-conjunction case
- Why I dislike this paper?
 - Too dependent on the parser
 - Too dependent on the language model
 - Cannot benefit from training data directly

Pros

- [Vaibhav] exemplifies methodology to approach a research problem!
- [Keshav, Soumya, Siddhant..] Applicable to any Open IE system
- [Soumya] linguistic constraints easily transferable to many languages!
- [Shubham] no training data needed

Critique -- Quality

- [Keshav] is it high enough quality?
- [Atishya] negation handling
 - “Ajay plays football but not cricket”
- [Vaibhav] study when not to split
 - “It is raining cats and dogs”
 - [Vipul] handle common phrases by lookup?
- [Deepanshu] how to handle “respectively”
- [Lovish] “Donald Trump” still got split

Extensions -- techniques

- [Keshav] bootstrap training data
 - (Sentence, split sentence) pairs
- [Keshav] use neural similarity with large corpus to determine when to split
 - [Soumya] use language model?
 - [Soumya] needs semantics.
 - [Rajas] LM may not have enough of it
 - [Saransh] LM may not preserve meaning only grammaticality
 - [Deepanshu] language model across sentence lengths!
- [Vipul] sequence labeler to predict whether to split or not
 - Training data?
- [Vaibhav] increase conjunct lengths
 - Do we need this?

Critique – use of constraints

- [Sankalan] Do we need constraints?
 - Shouldn't language model automatically handle them?
- [Vaibhav, Pratyush, Shubham] use neural language model
 - Maybe then we don't need constraints.

Critique -- Pipeline

- [Soumya] Errors could multiply
 - “The boy who loves rock and roll bet a hundred dollars and won.”
- [Keshav] make it end to end?

Extensions – rewrites

- [Sankalan] “The man, still dazed” → (the man, was still dazed)
- [Sankalan] “The match played between Germany and Argentina” → “The match played by Germany”
 - [Atishya] How to do this in general?
 - [Keshav] use an NLI system
 - [Soumya] “Adam is possibly German” is boring
 - [Rajas, Pratyush] no its ok!

Extensions -- techniques

- [Siddhant] how to use ML ideas + CALM
 - [Shubham] use Ficler's annotations
- Good research idea!

Critique

- [Sankalan] comma-separated clauses handled?
- [Rajas] why reject “similar syntactic structure” hypothesis

Critique -- evaluation

- [Soumya] small dataset
- [Siddhant] ablation study
- [Shubham] more insights against [Ficler & Goldberg]

Other Interesting Comments

- [Sankalan] ~ converting Boolean expressions to SOP canonical form
- [Atishya] subordinating conjunctions?
 - word that connects an independent clause to a dependent clause
 - although, because, if, even if, unless, while, before..
 - “I am happy because you love me”
 - “I am happy if you love me”
- [Vaibhav] adversative conjunction?
 - but, still, yet, whereas, while, nevertheless
- [Siddhant, Deepanshu] apply it to other NLU tasks

Conclusion

- SOA Open IE systems lose substantial recall due to ineffective conjunction processing.
- Introduced **CALM**, a coordination analyzer that corrects conjunct boundaries from dependency parses.
 - Significant improvement in conjunction analysis
- Developed **CALMIE**, which uses CALM generated simple sentences to improve SOA Open IE systems.
 - Huge boost in Open IE recall

Conclusion

- Integrated CALMIE into Open IE 4.2 to
 - release Open IE 5.
- Code available at <https://github.com/dair-iitd/OpenIE-standalone>.
- Demo available at <http://www.cse.iitd.ac.in/nlpdemo/web/oieweb/OpenIE5/>.
- Not much followup work. Worth investigating as a project