Open IE with OLLIE

Max Lotstein

Information Extraction

Winter 2013

Outline

Inspiration

Architecture

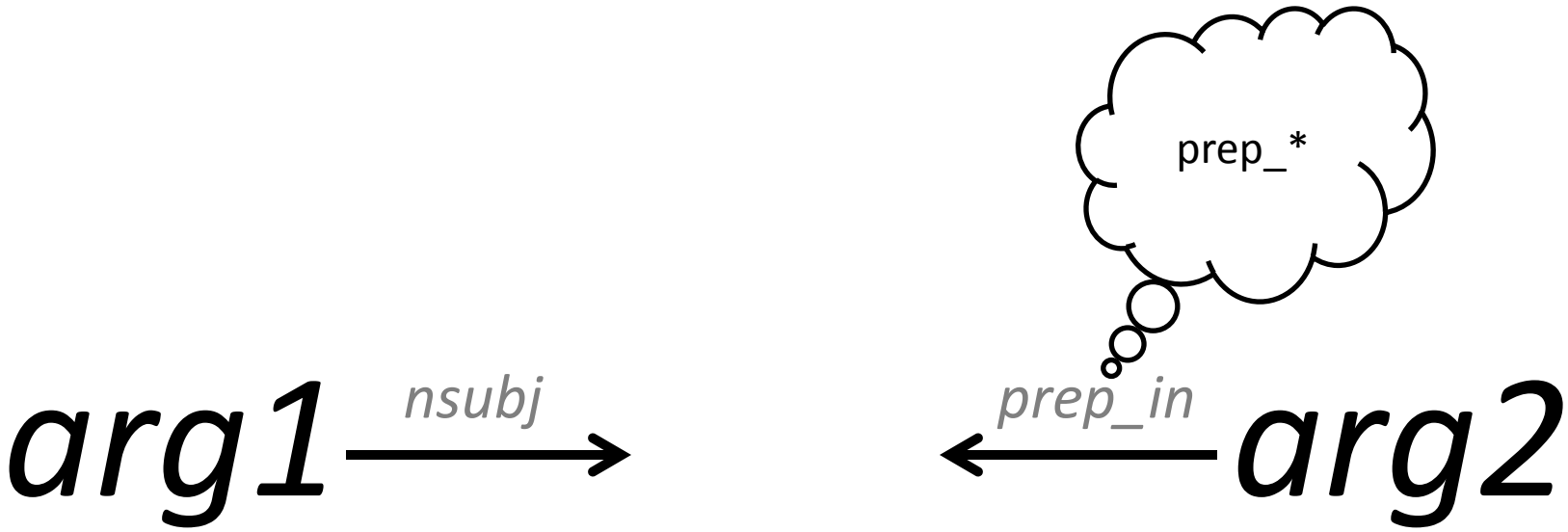
Performance

Conclusion

Outline

ReVerb

WOE, TextRunner*

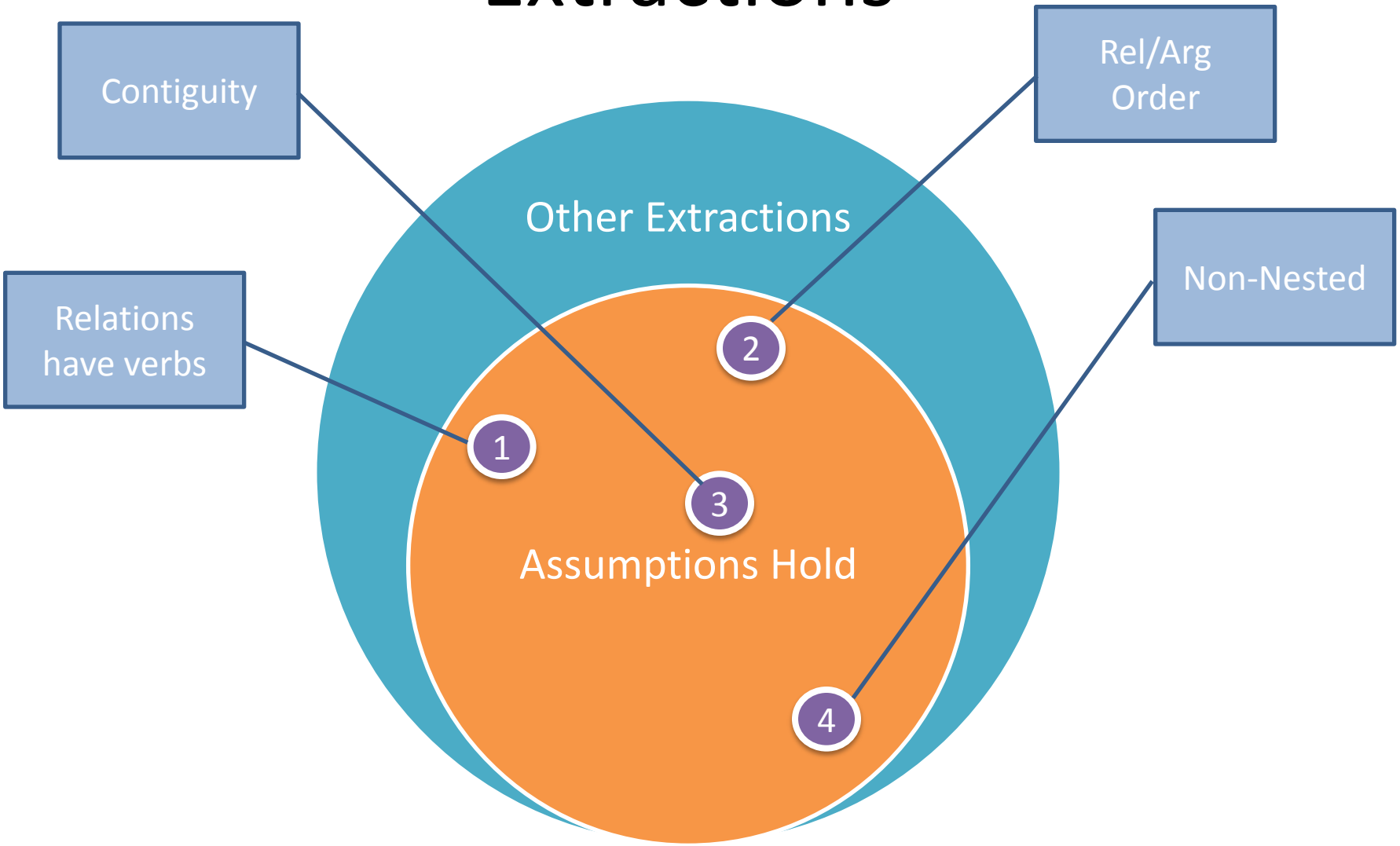


*NOTE: WOE^{pos} uses this information to train a classifier, as does TextRunner. Only WOE^{parse} uses it during extraction.

Many Systems, One Product: Tuples

$$(arg1\ ,\ relation\ ,\ arg2)$$

Extractions



#1: Relations w/o Verbs

"*Microsoft* <u>co-founder</u> *Bill Gates* …"

(*Bill Gates*, *be co-founder of*, *Microsoft*)

#2: Relation/Argument Order

After <u>winning</u> *the election*, *Obama* celebrated.

(*Obama*, *win*, *the election*)

#3: Non-Contiguous Elements

There <u>are</u> plenty of *taxis* <u>available at</u> *Bali airport*.

(*taxis*, *be available at*, *Bali airport*)

#4: Nested Relations

Early astronomers believed that *the earth* <u>is the center of</u> *the universe*.

((*the Earth*, *be the center of*, *the universe)* AttributedTo *believe, Early astronomers)*

If he makes this shot, *Tiger Woods* <u>will win</u> *the championship*.

((*Tiger Woods*, *will win*, *the championship*) ClausalModifier *if*, *he makes this shot*)

OLLIE uses deep syntactic analysis to extract these new relations, and uses a new form of representation when appropriate.

Information Density and Emphasis

- Many ways to encode information textually
- $\frac{relations}{sentence} > 1$

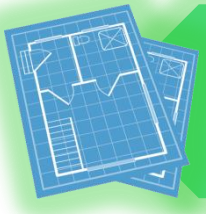
Bill Gates <u>is the co-founder of</u> *Microsoft*. *Bill Gates* <u>is</u> *a billionaire*. *Bill Gates* <u>owns</u> a dog named Bucks.

vs.

Microsoft <u>co-founder</u>, *Bill Gates*, *who* <u>is</u> a *billionaire*, <u>owns</u> *a dog named Bucks*.

Outline

Inspiration

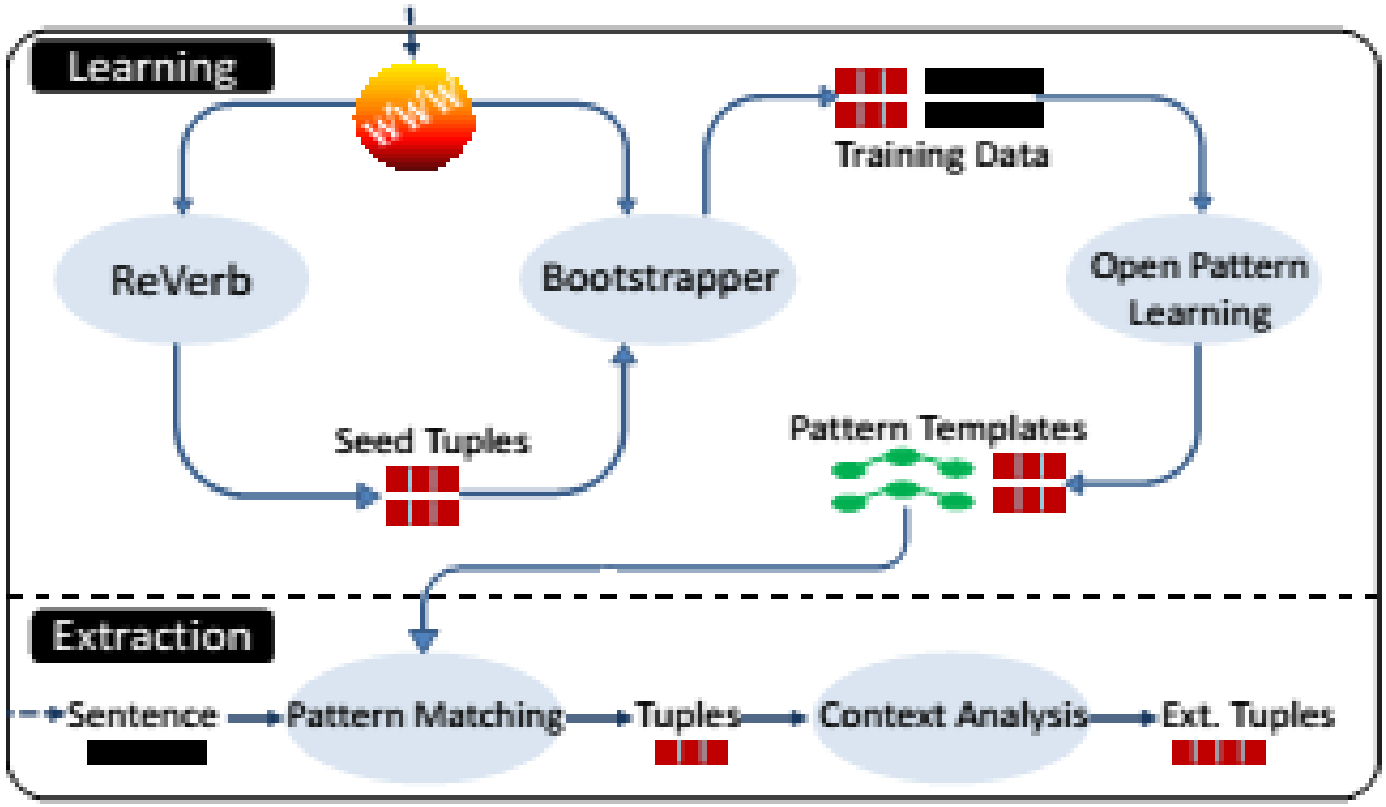


Architecture

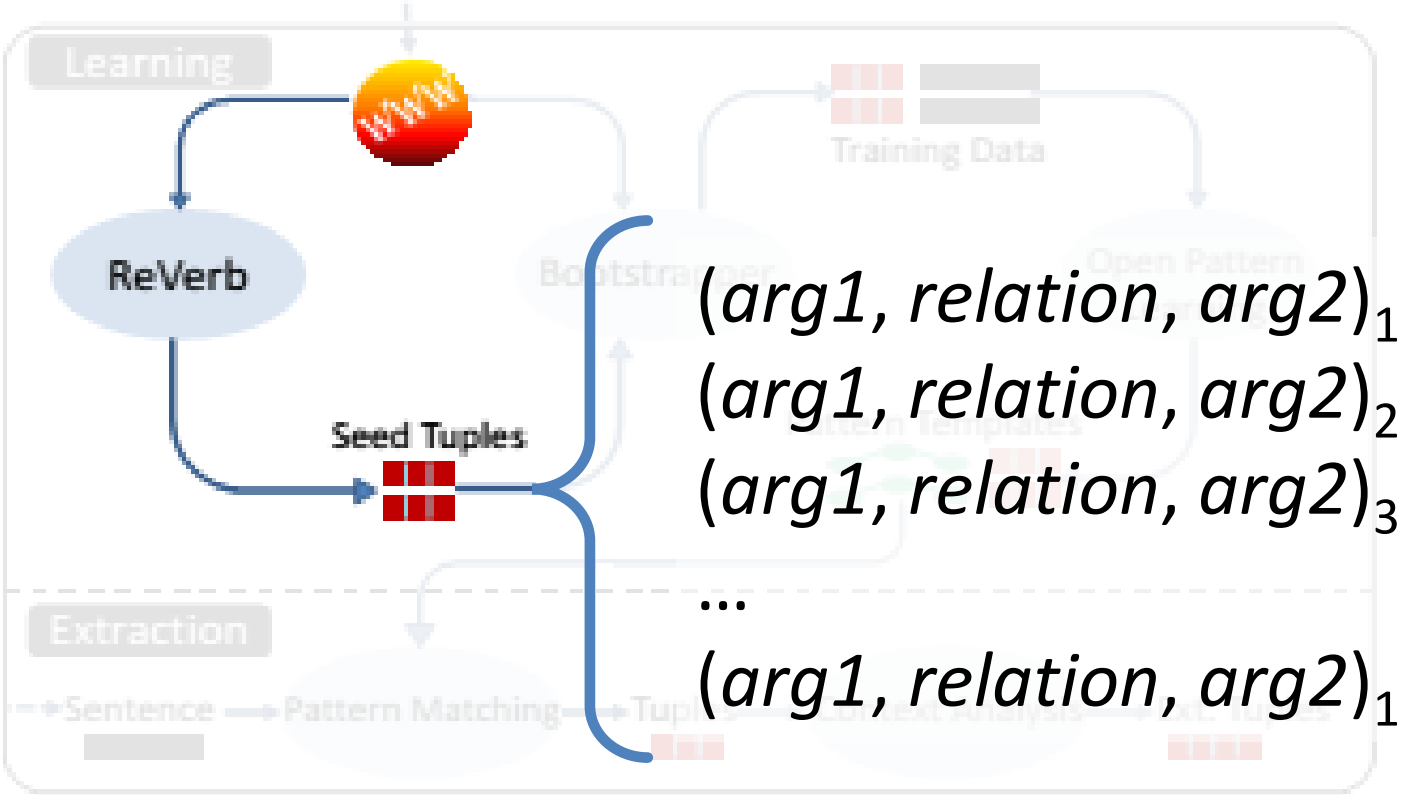
Performance

Conclusion

Architecture

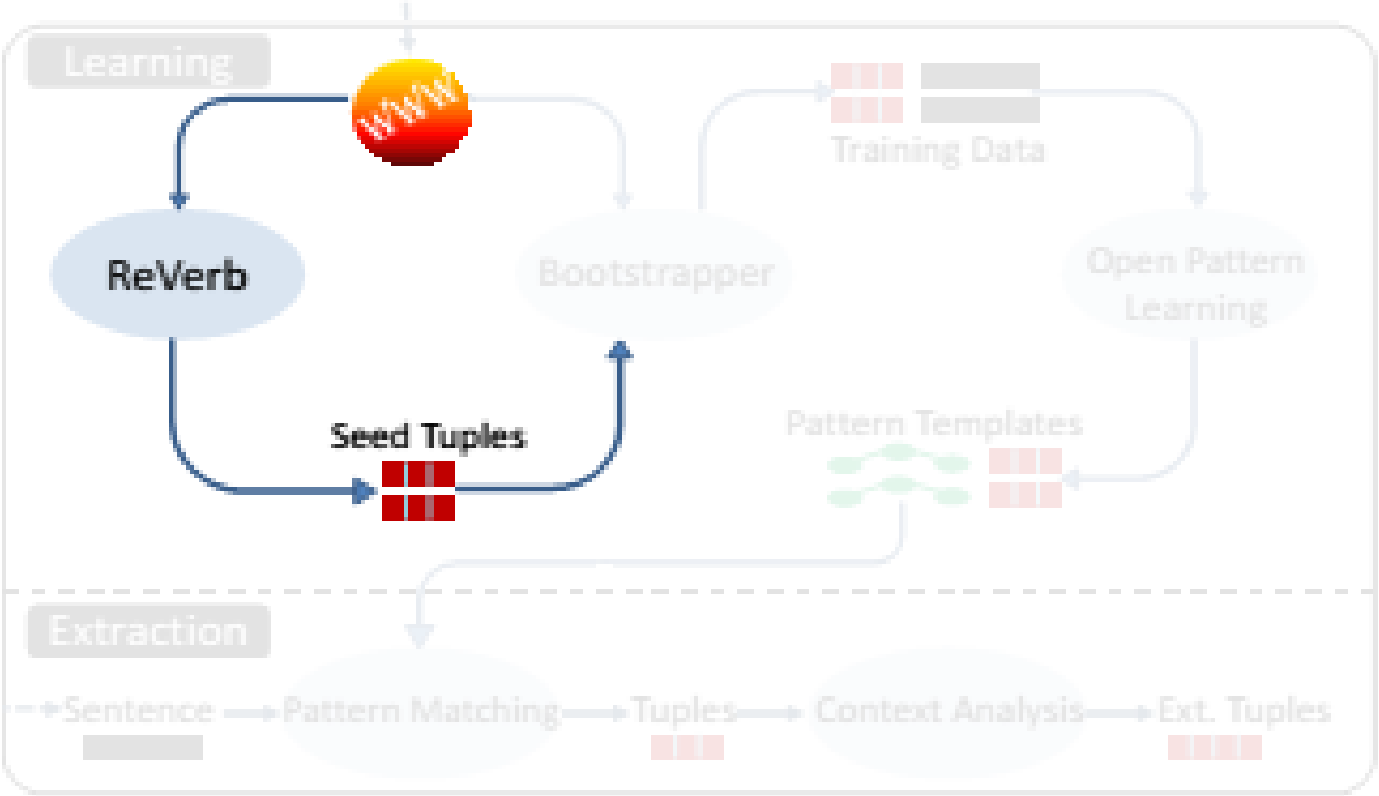


Seed Tuples



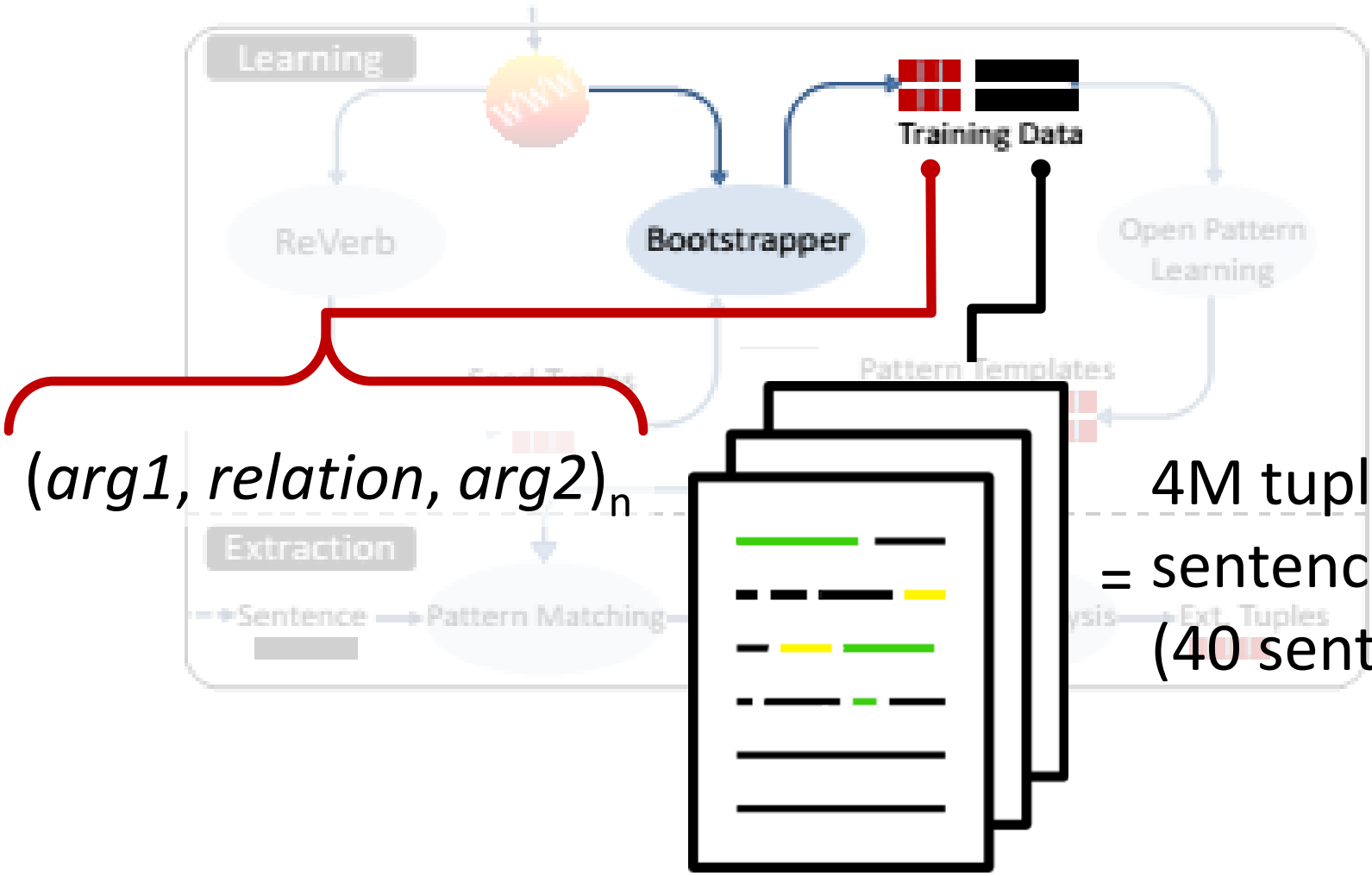
$(arg1, relation, arg2)_1$
$(arg1, relation, arg2)_2$
$(arg1, relation, arg2)_3$
…
$(arg1, relation, arg2)_{110k}$

Seed Tuple, Example



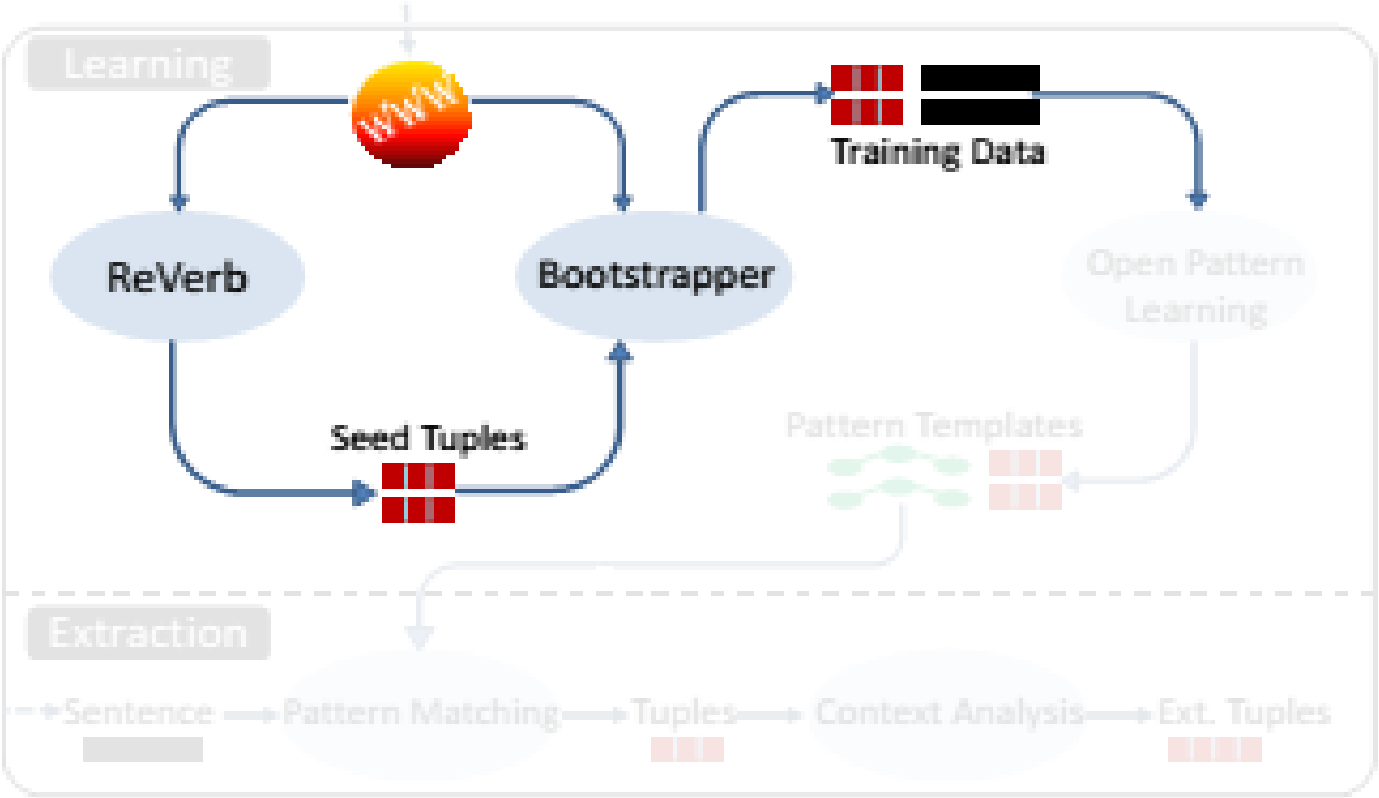
(*Obama*, *win*, *the election*)

Training Data



4M tuple-sentence pairs (40 sent./tup.)

Bootstrap, Example

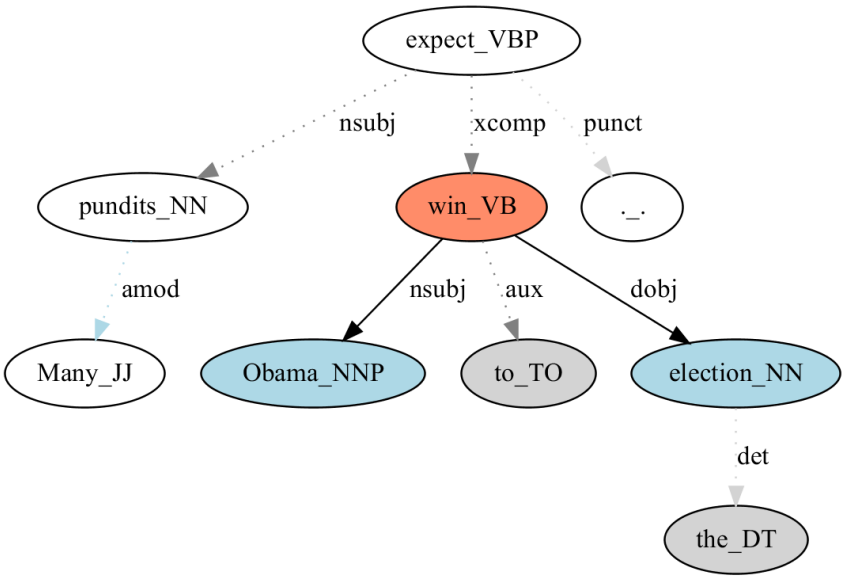


Many pundits expect *Obama* <u>to win</u> the *election*.

(*Obama*, *win*, *the election*)

Creating an Open Pattern

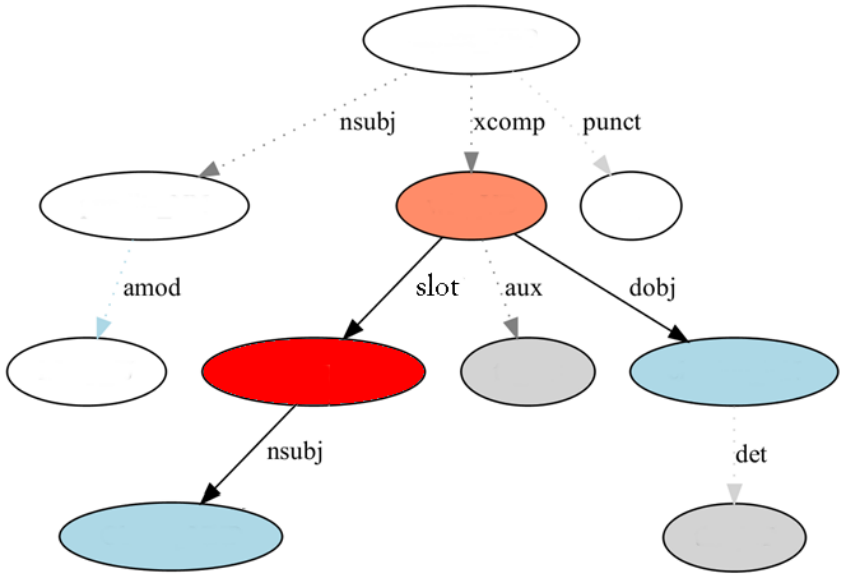
1. Extract path
2. Annotate relation node with word and POS
3. Normalize copula



{Obama} ↑nsubj↑ {<u>win</u>:postag=VB}↓dobj↓{the election}

Slot Node

A node on the dependency path that *isn't* part of the extraction



Can We De-Lexicalize?

If *all* of the following:

- NO slot node on path
- Relation node between arguments
- $\text{Preposition}_{\text{pattern}} = \text{Preposition}_{\text{tuple}}$
- Path has no *nn* or *amod* edges

Then: *syntactic pattern*

Else: *lexical/semantic pattern*

Purely Syntactic Patterns

Aggressively generalize:

- Relations, remove lexical constraints
- Prepositions, convert to {prep_*}

Consider sentences:

1. "*Michael* <u>appeared on</u> *Oprah*…"
2. "… when *Alexander the Great* <u>advanced to</u> *Babylon*."

Both have the pattern:

{arg1} ↑nsubj↑ {rel:postag=VBD} ↓{prep_*}↓ {arg2}

Lexical/Semantic Patterns, Example

"*Microsoft* co-founder *Bill Gates*…"

(*Bill Gates*, *is co-founder of*, *Microsoft)*

"*Chicago* Symphony *Orchestra"**

(*Orchestra*, *is symphony of*, *Chicago*)*

Can we still generalize to unseen words?

Lexical/Semantic Patterns

People = WordNet's People class

Location = WordNet's Location class

L = list of lexical items

$l_{people} = L \cap People$

$l_{location} = L \cap Location$

If l_{people} (or $l_{location}$) > 3/4*L:

Then: Use l_{people}, drop L (Use $l_{location}$)

Else: Keep L

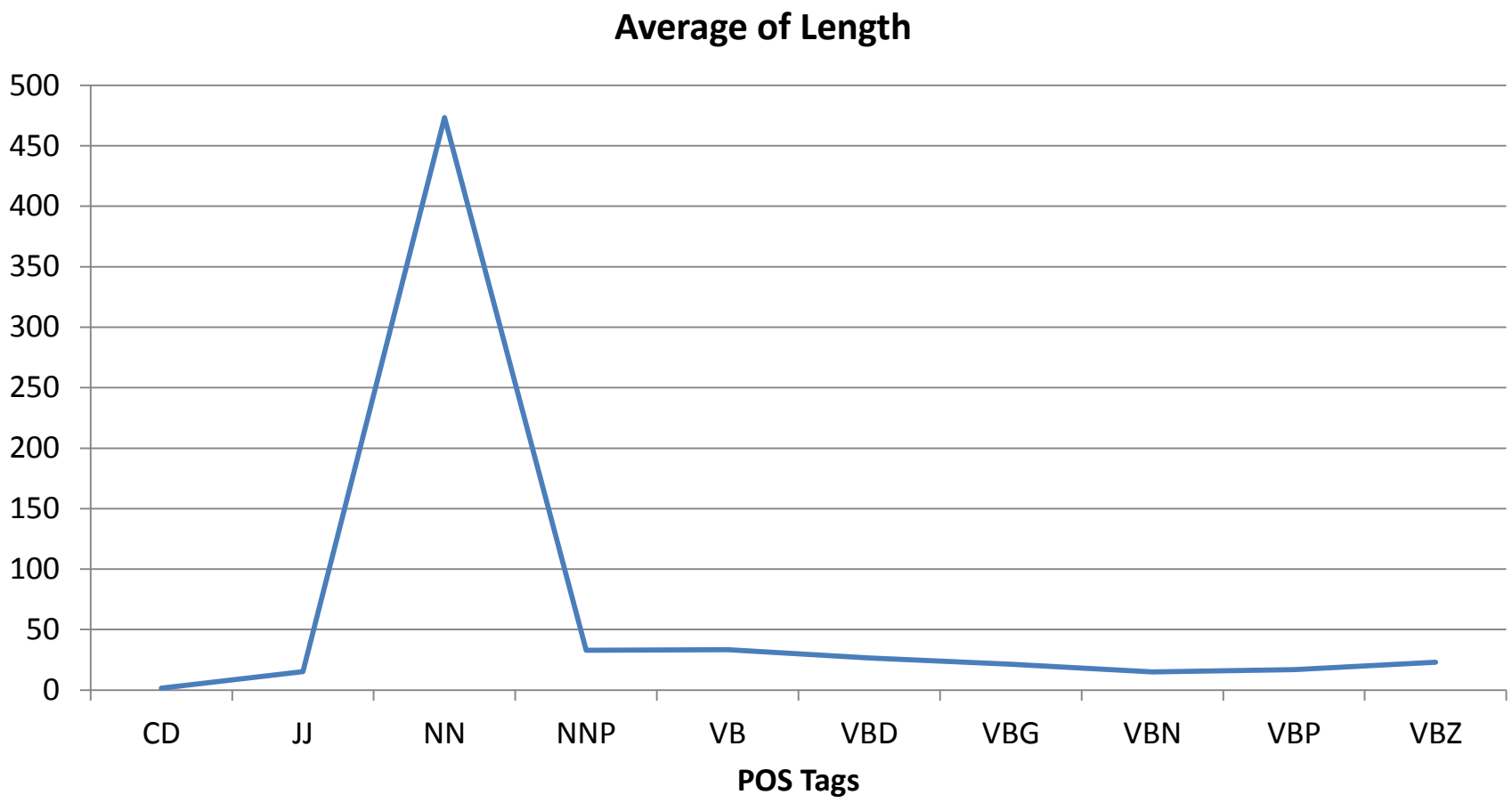
Some Open Pattern Templates

Extraction Template	**Open Pattern**
1. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓{prep_*}↓ {arg2}
2. (arg1; {rel}; arg2)	{arg1} ↑nsubj↑ {rel:postag=VBD} ↓dobj↓ {arg2}
3. (arg1; be {rel} by; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓agent↓ {arg2}
4. (arg1; be {rel} of; arg2)	{rel:postag=NN;**type=Person**} ↑nn↑ {arg1} ↓nn↓ {arg2}
5. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {slot:postag=VBN;**lex ∈announce|name|choose...**} ↓dobj↓ {rel:postag=NN} ↓{prep_*}↓ {arg2}

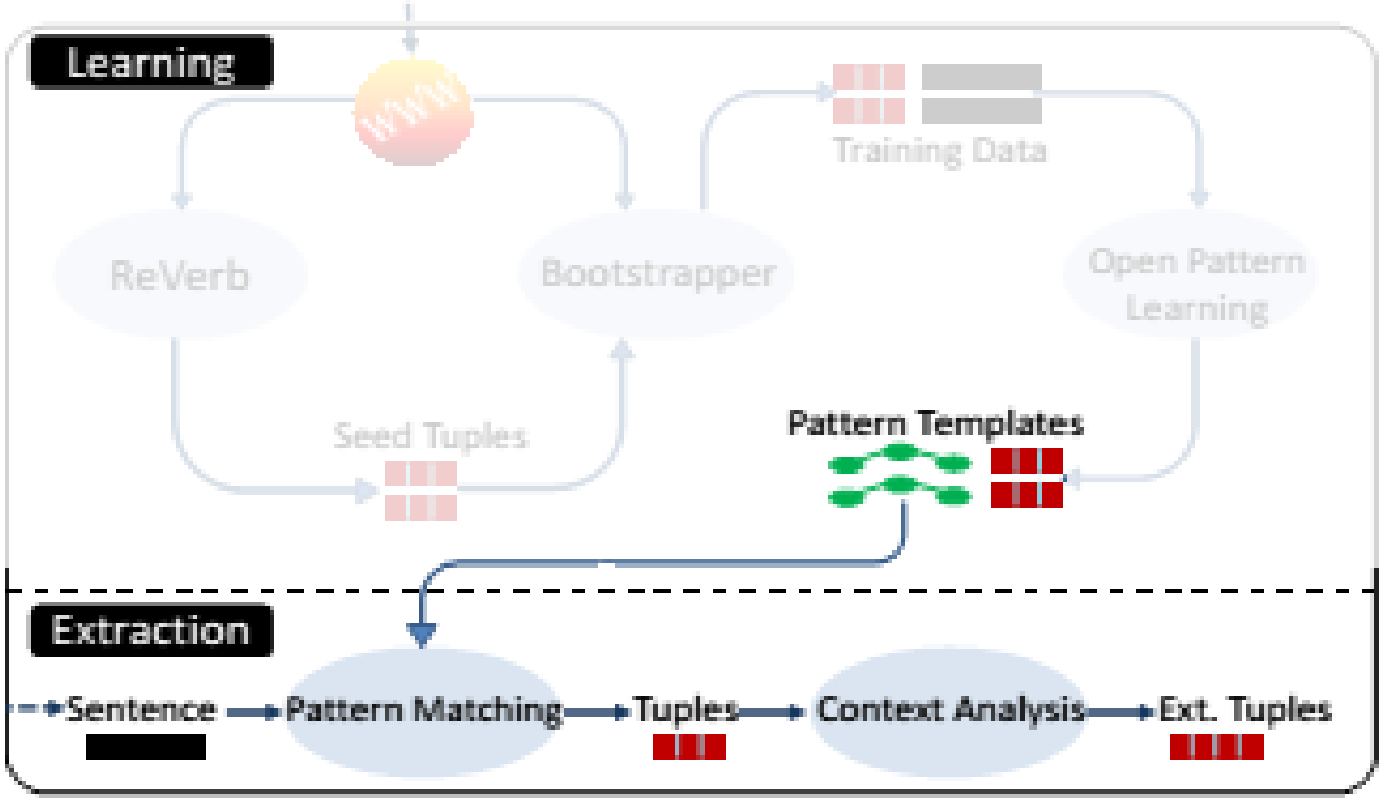
Open Pattern Template Statistics

PatternType	Average Rank	Frequency
Lexical/Semantic	344	515
Purely Syntactic	186	114
Grand Total		**629**

Lexical Constraint Statistics



Extraction


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

Pattern Matching

1. Apply Pattern Template
2. Expand on relevant edges
   e.g. “election” → “the election” (det)
3. Use word order from sentence to make tuple

Context Analysis

- Attribution
  - Marked by *ccomp* edges
    - E.g. “He says that you like to swim” (says, like)
  - Communication/cognition verbs, e.g. ‘believe’
- Clausal Modifier: when dependent clause modifies main extraction
  - Marked by *advcl*
    - “The accident occurred as night fell” (occurred, fell)
  - *If, when, although, because …*

Demonstration Time

Outline

Inspiration

Architecture

Performance

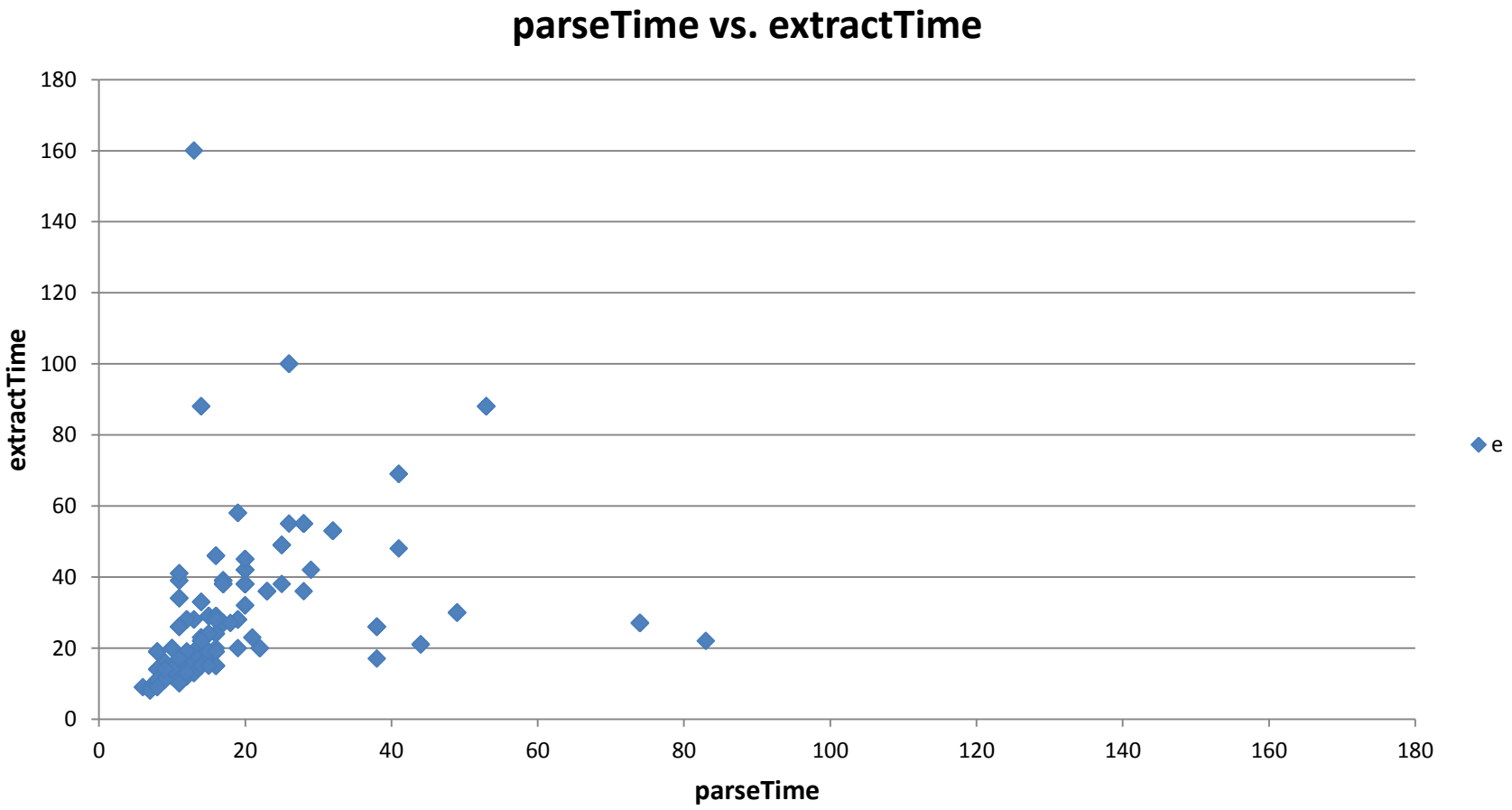
Conclusion

Speed: Conflicting Reports

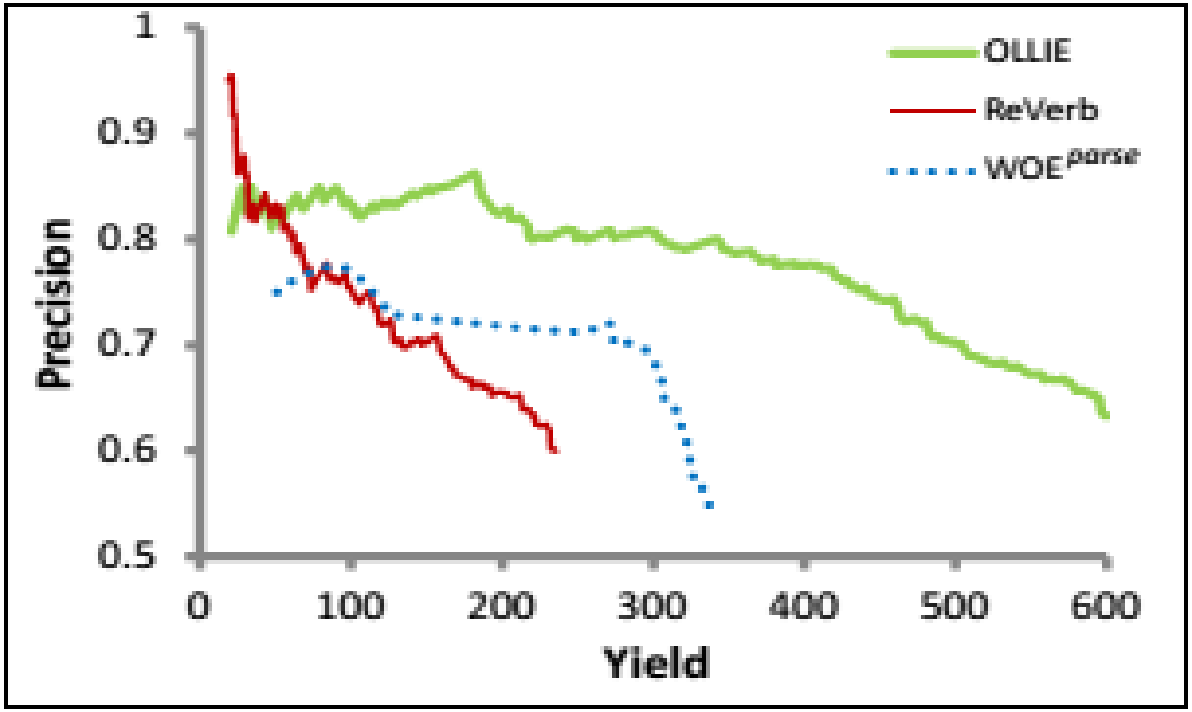
System	Sent/sec	ReportedIn
OLLIE	89	OLLIE
ReVerb	104	ReVerb
TextRunner	662	TextRunner
TextRunner	79	ReVerb
TextRunner	2727	WOE
WOE_{parse}	3	ReVerb
WOE_{parse}	88	WOE
WOE_{pos}	79	ReVerb
WOE_{pos}	2727	WOE

A Less Precise Consensus

TextRunner/ReVerb > WOE_{pos} > WOE_{parse}/OLLIE


parseTime vs. extractTime
extractTime
parseTime
0
20
40
60
80
100
120
140
160
180
e

Precision vs. Yield



- 300 Sentences, selected randomly from News, Wikipedia and a Biology textbook
- Hand corrected extractions by multiple humans

Comparing Comparisons

$$\frac{Precision}{Yield} \quad vs. \quad \frac{Precision}{Recall}$$

- Requires natural order
  - Confidence Values

- Requires full set
  - Allows false negative detection

Noun-Mediated Relations

Relation	OLLIE	REVERB	incr.
is capital of	8,566	146	59x
is president of	21,306	1,970	11x
is professor at	8,334	400	21x
is scientist of	730	5	146x

"Obama, the president of the US"
"Obama, the US president" >> "Obama is the president of the US"
"US President Obama"

OLLIE vs. SRL

	LUND	OLLIE	union
Verb relations	0.58 (0.69)	0.49 (0.55)	0.71 (0.83)
Noun relations	0.07 (0.33)	0.13 (0.13)	0.20 (0.33)
All relations	0.54 (0.67)	0.47 (0.52)	0.67 (0.80)

- SRL performs well
  - Bad at grammatical complexity
- OLLIE deals with co-reference better
- Noun-mediated relations are harder, rarer
- Union is higher than both: everybody wins!

Sources of Error

Source of Error	%
Parser Error	32
Aggressive generalization	18
Incorrect application of lexical pattern	12
Missed Context	13
Limitations of Binary Representation	12

Illuminating Errors

- The rotation of the planet causes it to take the shape of an oblate spheroid; that is, *it* <u>is flattened at</u> *the poles and bulges* at the equator.[1]
  - (it, is flattened at, the poles and bulges)*
- *Saturn* <u>is the only planet of</u> *the Solar System* that is less dense than water--about 30% less.[2]
  - (Saturn, is the only planet of, the Solar System)*
- I shot the man with the gun.
  - (I, shot the man with, the gun)*
  - (I, shot, the man)*

http://en.wikipedia.org/wiki/Saturn[1]
http://simple.wikipedia.org/wiki/Saturn[2]

Two Observations About Language

1. In English, words can act in groups
   - I <u>like ice cream</u>. Do you (like ice cream)?
   - I <u>like ice cream</u> and <u>hate bananas</u>.
   - I said I would hit Fred and <u>hit Fred</u> I did.

Phrase Driven Grammar

2. Words also depend on other words by
   - Verbs have *agents, objects*, etc.
     - I (subj) throw (verb) the (det) ball (obj)

Dependency Grammar

Neither approach is perfect.

Outline

Inspiration

Architecture

Performance

Conclusion

Conclusions, Methodology

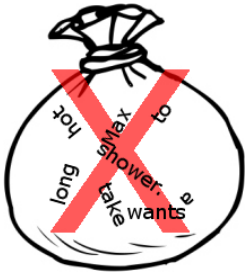
- How big must a sample be in order to be representative?
  - Bootstrapping hypothesis, only 100
  - 50 sentences in SRL comparison
- ‘Gold standard’ annotation (support recall)
  - Potentially more reliable inter-system comparison
    - “Hey look, our system is better! What are the odds!”
  - Better false negative detection
  - Ahem … grad students are cheap.

Conclusion, Theoretical

- Generalization Techniques
  - Syntactic: a bit too aggressive
  - Lexical/Semantic: a bit too tame
    - Many other options. See Angeli, Gabor, and Manning 2013
- OLLIE lives and dies by its parser
  - Responsible for sig. % of errors
  - Accounts of sig. % of time
- Relations still assumed binary
  - Many are n-ary, have optional arguments
    - See KrakeN, ClausIE
- Contextual Relations are limited, flawed
  - What really are relations, anyway?

Our Work Isn't Done

Words are not bags of characters:
Opposites, synonyms, entailment, classes …



Sentences are not bags of words:
Syntactic structure, semantic frames ….

Are documents bags of sentences?
Coreference disambiguation

References

Schmitz, Michael, et al. "Open language learning for information extraction." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012.

OLLIE Readme, Accessed on 12-3-2013. URL *https://github.com/knowitall/ollie*

Etzioni, Oren, et al. "Open information extraction: The second generation." *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume One*. AAAI Press, 2011.

Fader, Anthony, Stephen Soderland, and Oren Etzioni. "Identifying relations for open information extraction." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.

Banko, Michele, et al. "Open Information Extraction from the Web." *IJCAI*. Vol. 7. 2007.

Wu, Fei, and Daniel S. Weld. "Open information extraction using Wikipedia." *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010.

De Marneffe, Marie-Catherine, and Christopher D. Manning. "Stanford typed dependencies manual." *URL http://nlp. stanford. edu/software/dependencies manual. pdf* (2008).

"Dependency Grammar." Advanced Natural Language Processing, University of Edinburgh. URL *http://www.inf.ed.ac.uk/teaching/courses/anlp/slides/anlp15.pdf*

Angeli, Gabor, and Christopher D. Manning. "Philosophers are mortal: Inferring the truth of unseen facts." *CoNLL-2013* (2013): 133.

Akbik, Alan, and Alexander Löser. "Kraken: N-ary facts in open information extraction." *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics, 2012.

Del Corro, Luciano, and Rainer Gemulla. "ClausIE: clause-based open information extraction." *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013.

Wikipedia, *That*, http://en.wikipedia.org/wiki/That (as of Dec. 4, 2013).

GitHub | This repository | Search or type a command | Explore | Features | Enterprise | Blog | Sign up | Sig

PUBLIC knowitall / ollie | Star 28 | Fo

branch: master | ollie / Commits

Nov 18, 2013

Merge pull request #22 from mlotstein/master ···
rbart authored 15 days ago
5ca8c4ec15
Browse code

Nov 17, 2013

Update README.md
schmmd authored 16 days ago
e04b3b49a7
Browse code

Nov 15, 2013

Fix package name in command line call in README file ···
mlotstein authored 18 days ago
4f5cb51d6f
Browse code

Merge pull request #21 from mlotstein/master ···
schmmd authored 18 days ago
6d76094088
Browse code

Fixed import statements in Java Example File ···
mlotstein authored 18 days ago
0faf3b8386
Browse code

Confidence Function

- Top Positive Features:
  - nn edges in pattern 0.91
  - rel contains verb 0.48
  - openparse confidence 0.43
- Top Negative Features:
  - if right before arg1 -1.22
  - vacuous extraction -0.64
  - semantic constraints in pattern -0.43