NestIE
NESTED PROPOSITIONS IN OPEN INFORMATION EXTRACTION

Nikita Bhutani, H V Jagadish, Dragomir Radev

EMNLP, Nov 2 2016

NestIE

NESTED PROPOSITIONS IN OPEN INFORMATION EXTRACTION

Nikita Bhutani, H V Jagadish, Dragomir Radev

EMNLP, Nov 2 2016

EXTRACTING KNOWLEDGE FROM TEXT

Closed KB

Ontology

co-founder death_date birth_date

"Steve Jobs, who co-founded Apple.."

(Steve Jobs, **co-founder**, Apple)

* [Yates et al., ACL 2007]

EXTRACTING KNOWLEDGE FROM TEXT

Closed KB

Ontology

co-founder death_date birth_date

"Steve Jobs, who co-founded Apple.."

↓

(Steve Jobs, **co-founder**, Apple)

- expensive, not-scalable
- pre-defined relations

* [Yates et al., ACL 2007]

EXTRACTING KNOWLEDGE FROM TEXT

Closed KB

Ontology

co-founder death_date birth_date

"Steve Jobs, who co-founded Apple.."

↓

(Steve Jobs, **co-founder**, Apple)

- expensive, not-scalable
- pre-defined relations

Open KB*

~~Ontology~~

~~relation schema~~

"8.8 million have lost their jobs.."

↓

(8.8 million people, **lost**, their jobs)

* [Yates et al., ACL 2007]

EXTRACTING KNOWLEDGE FROM TEXT

Closed KB

Ontology

co-founder death_date birth_date

"Steve Jobs, who co-founded Apple.."

↓

(Steve Jobs, **co-founder**, Apple)

- expensive, not-scalable
- pre-defined relations

Open KB*

~~Ontology~~

~~relation schema~~

"8.8 million have lost their jobs.."

↓

(8.8 million people, **lost**, their jobs)

- broad coverage
- light-weight structure

* [Yates et al., ACL 2007]

PROPOSITIONS FOR BINARY RELATIONS

Binary

"8.8 million people have lost their jobs since the start of the recession."

Proposition: (8.8 million people, **lost**, their jobs)

PROPOSITIONS FOR BINARY RELATIONS

Binary

"8.8 million people have lost their jobs since the start of the recession."

Proposition: (8.8 million people, **lost**, their jobs)

Subject Relation Object

PROPOSITIONS FOR BINARY RELATIONS

Binary++: Contextual Information

"The Bureau of Labor Statistics believes that 8.8 million people have lost their jobs since the start of the recession."

Proposition[1]: ((8.8 million people, **lost**, their jobs)
attributedTo believe, Bureau of Labor Statistics)

[1] [Schmitz et al.]

PROPOSITIONS FOR BINARY RELATIONS

Binary++: Contextual Information

"The Bureau of Labor Statistics believes that 8.8 million people have lost their jobs since the start of the recession."

Proposition[1]: ((8.8 million people, **lost**, their jobs)
attributedTo believe, Bureau of Labor Statistics)

- few argument types: conditional, attribution, temporal..

1 [Schmitz et al.]

PROPOSITIONS FOR BINARY RELATIONS

Binary++: n-ary relation

"8.8 million people lost their jobs <u>in the Great Depression.</u>"

Proposition[2]: (8.8 million people, **lost**, their jobs, <u>in the Great Depression</u>)

[2] [Del Corro et al.]

PROPOSITIONS FOR BINARY RELATIONS

Binary++: n-ary relation

"8.8 million people lost their jobs <u>in the Great Depression.</u>"

Proposition[2]: (8.8 million people, **lost**, their jobs, <u>in the Great Depression</u>)

- few grammatical constructs to identify constituents: limited coverage

2 [Del Corro et al.]

UNINFORMATIVE & INCOMPLETE PROPOSITIONS

Long arguments are not informative

"Sheryl Sandberg is the <u>COO of Facebook and author of Lean In.</u>"

Proposition[3]: (Sheryl Sandberg, **be**, COO of Facebook and author of *Lean In*)

[3] [Fader et al.]

UNINFORMATIVE & INCOMPLETE PROPOSITIONS

Long arguments are not informative

"Sheryl Sandberg is the COO of Facebook and author of Lean In."

Proposition[3]: (Sheryl Sandberg, **be**, COO of Facebook and author of *Lean In*)

- an accurate fact may itself contain another accurate fact

[3] [Fader et al.]

UNINFORMATIVE & INCOMPLETE PROPOSITIONS

Long arguments are uninformative

"Sheryl Sandberg is the COO of Facebook and author of Lean In."

Proposition[3]: (Sheryl Sandberg, **be**, COO of Facebook and author of *Lean In*)

- an accurate fact may itself contain another accurate fact

Proposition: (Sheryl Sandberg, **be**, COO of Facebook)

Proposition: (Sheryl Sandberg, **be**, author of Lean In)

[3] [Fader et al.]

UNINFORMATIVE & INCOMPLETE PROPOSITIONS

Missing context makes propositions incomplete

"8.8 million people have lost their jobs <u>since the start of the recession.</u>"

Proposition[1]: (8.8 million people, **lost**, their jobs)

[1] [Schmitz et al.]

UNINFORMATIVE & INCOMPLETE PROPOSITIONS

Missing context makes propositions incomplete

"8.8 million people have lost their jobs since the start of the recession."

Proposition[1]: (8.8 million people, **lost**, their jobs)

- limited expressivity

[1] [Schmitz et al.]

NESTED PROPOSITIONS

Complex Assertions	n-ary relations nested relations subordinate clauses	Triples	limited expressivity non-minimality lost context

Complex User Information Needs

Challenges

NESTED PROPOSITIONS

Complex Assertions	n-ary relations nested relations subordinate clauses	Triples	limited expressivity non-minimality lost context

Complex User Information Needs

Challenges

Nested Representation	(X, reported, (Y, be, Z)) ((X, lost, Y), in, Z)……

Solution

NESTED PROPOSITIONS

Complex Assertions	n-ary relations nested relations subordinate clauses

Triples	limited expressivity non-minimality lost context

Complex User Information Needs

Challenges

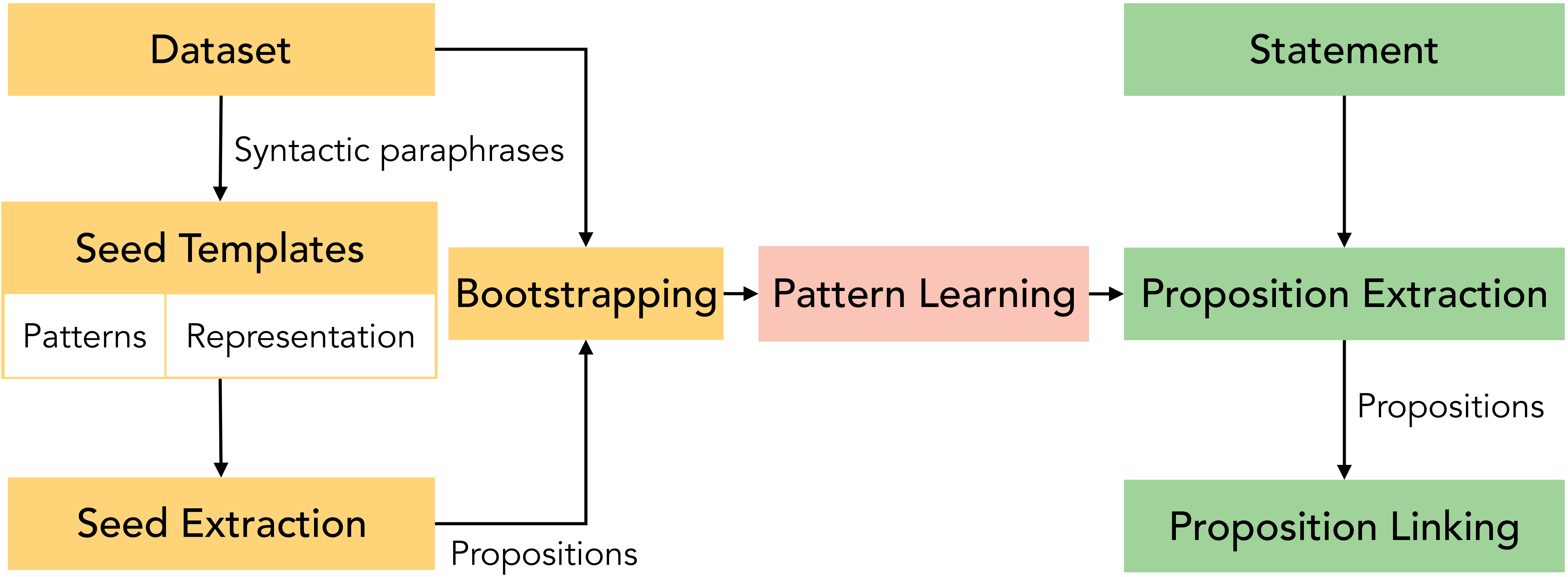
Nested Representation	(X, reported, (Y, be, Z)) ((X, lost, Y), in, Z)……

Solution

KB with light-weight nested structure: NestIE

OUTLINE

- System Architecture
  - Seed Set Construction
  - Pattern Learning
  - Proposition Extraction
  - Proposition Linking
- Experiments
- Analysis



1. Seed Fact Extraction and Bootstrapping
2. Pattern Learning
3. Proposition Extraction and Linking

SEED EXTRACTION AND BOOTSTRAPPING

RTE (Recognizing Textual Entailment) Dataset*

Hypothesis
simple, short sentences hand-written templates

* [de Marneffe et al.]

SEED EXTRACTION AND BOOTSTRAPPING

RTE (Recognizing Textual Entailment) Dataset*

Hypothesis	Template
simple, short sentences	dependency sub-tree
hand-written templates	nested representation

* [de Marneffe et al.]

SEED EXTRACTION AND BOOTSTRAPPING

RTE (Recognizing Textual Entailment) Dataset*

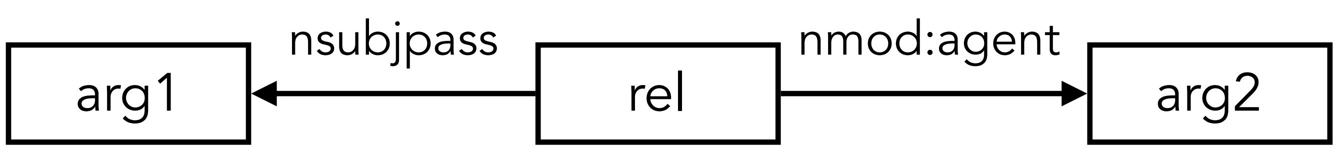
Hypothesis	Template	Statement
simple, short sentences hand-written templates	dependency sub-tree nested representation	long, complex sentences learn syntactic variants

* [de Marneffe et al.]

SEED EXTRACTION AND BOOTSTRAPPING

RTE (Recognizing Textual Entailment) Dataset*

Hypothesis	Template	Statement
simple, short sentences	dependency sub-tree	long, complex sentences
hand-written templates	nested representation	learn syntactic variants



(arg1, [rel, by], arg2)

A body has been found by police.

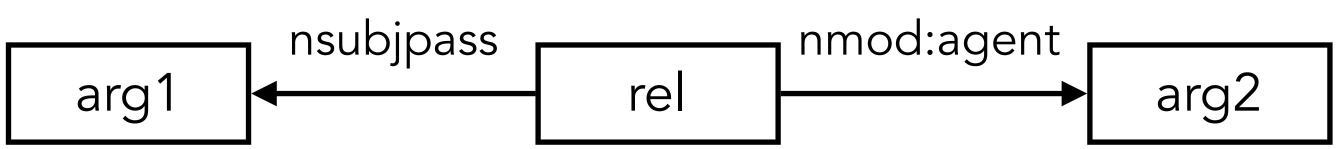
(body, [found, by], police)

* [de Marneffe et al.]

SEED EXTRACTION AND BOOTSTRAPPING

RTE (Recognizing Textual Entailment) Dataset*

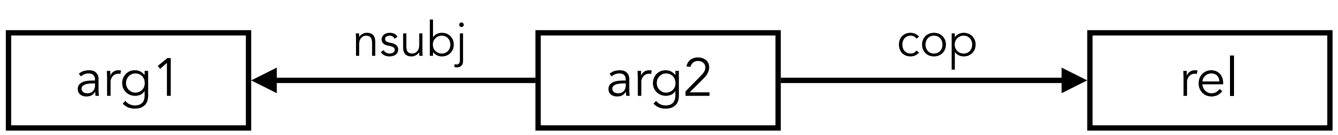
Hypothesis	Template	Statement
simple, short sentences	dependency sub-tree	long, complex sentences
hand-written templates	nested representation	learn syntactic variants



A body has been found by police.

(arg1, [rel, by], arg2)

(body, [found, by], police)



Fallujah is an Iraqi city

(arg1, be, arg2)

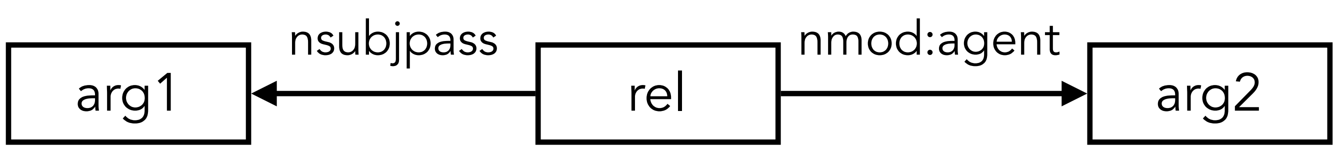
(Fallujah, be, city)

* [de Marneffe et al.]

SEED EXTRACTION AND BOOTSTRAPPING

RTE (Recognizing Textual Entailment) Dataset*

Hypothesis	Template	Statement
simple, short sentences hand-written templates	dependency sub-tree nested representation	long, complex sentences learn syntactic variants



A body has been found by police.

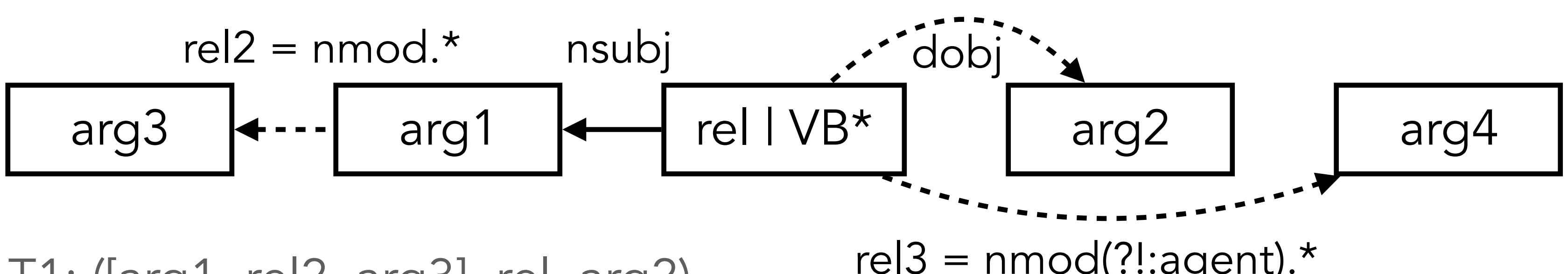
(arg1, [rel, by], arg2)

(body, [found, by], police)

Fallujah is an Iraqi city

(arg1, be, arg2)

(Fallujah, be, city)



10,000 people in Africa died of Ebola

T1: ([arg1, rel2, arg3], rel, arg2)
T2: (T1, rel3, arg4)

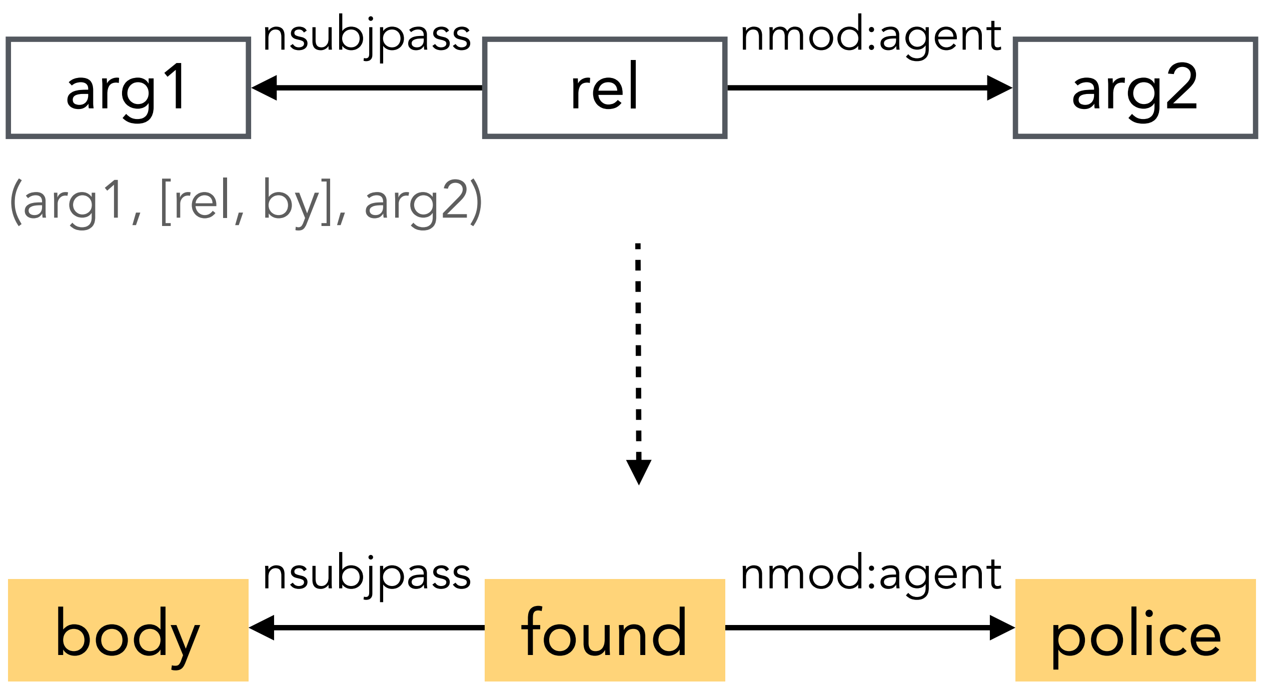
T1: (people in Africa, died, ∅)
T2: (T1, of, Ebola)

… 13 seed templates

* [de Marneffe et al.]

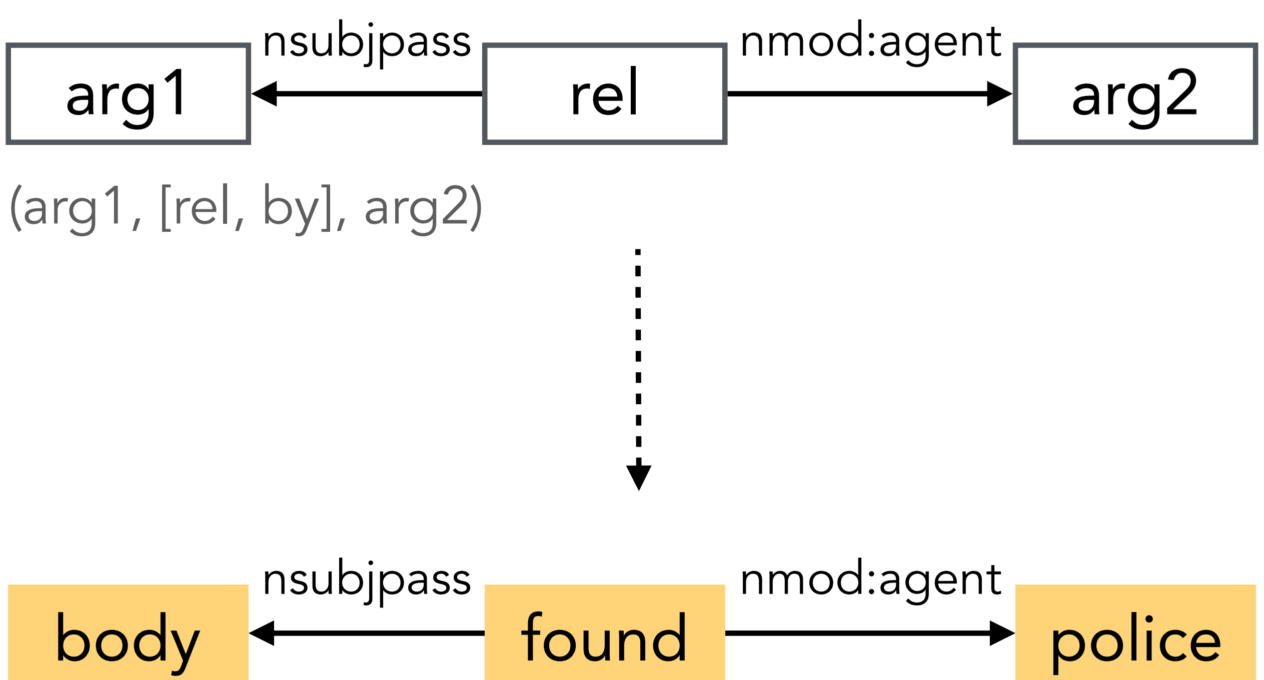
BOOTSTRAPPING

"A body was found by U.S. military police."

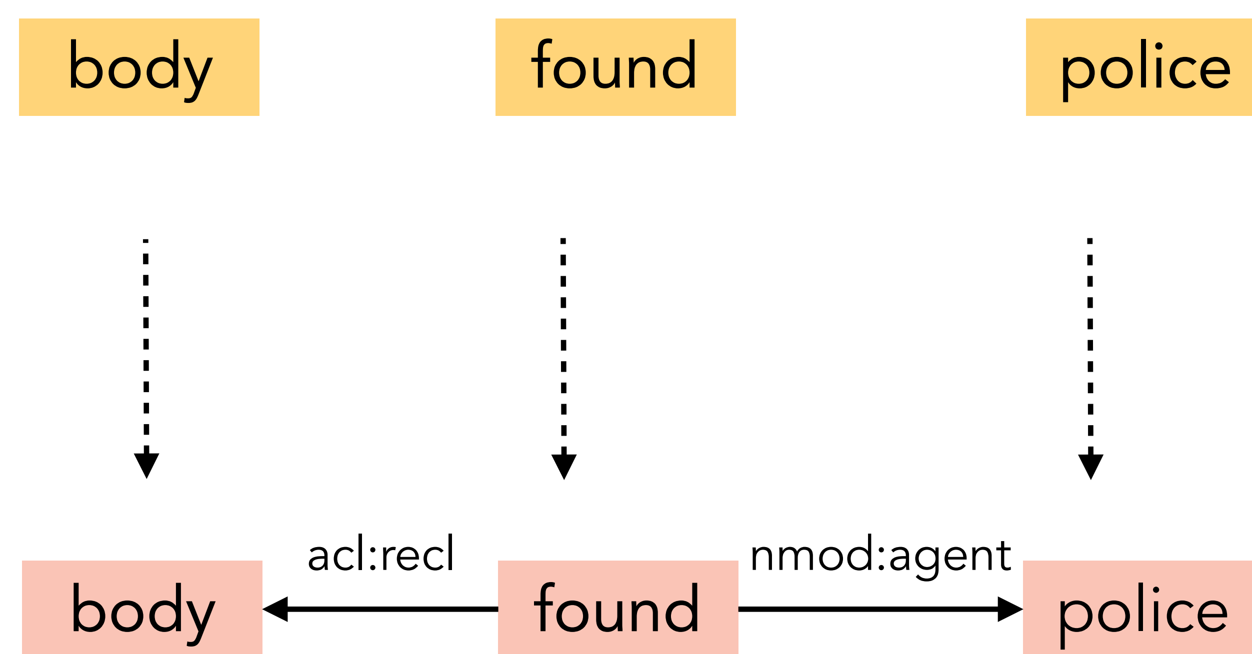


BOOTSTRAPPING

"A body was found by U.S. military police."

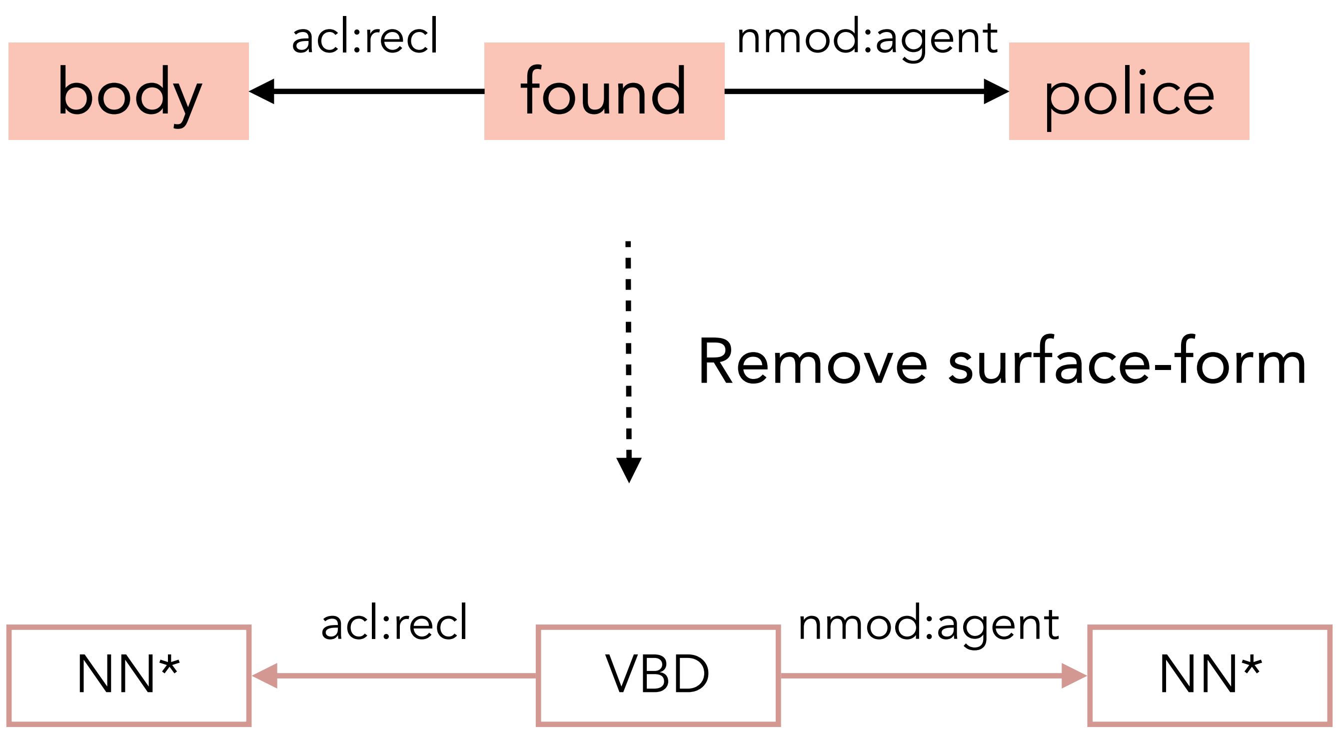


"A senior official in Iraq said the body, which was found by U.S. military police, was thrown from a vehicle."



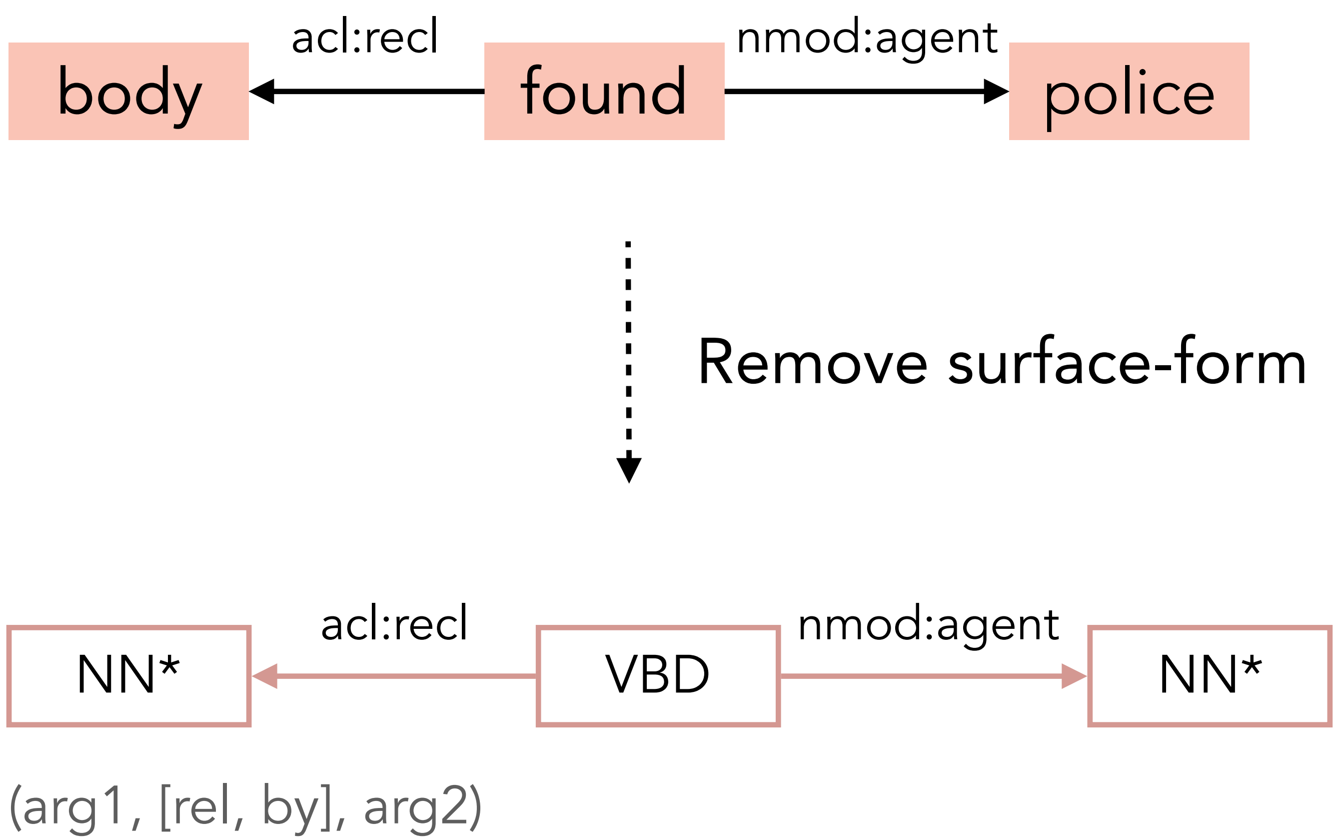
PATTERN LEARNING

PATTERN LEARNING



(arg1, [rel, by], arg2)

PATTERN LEARNING

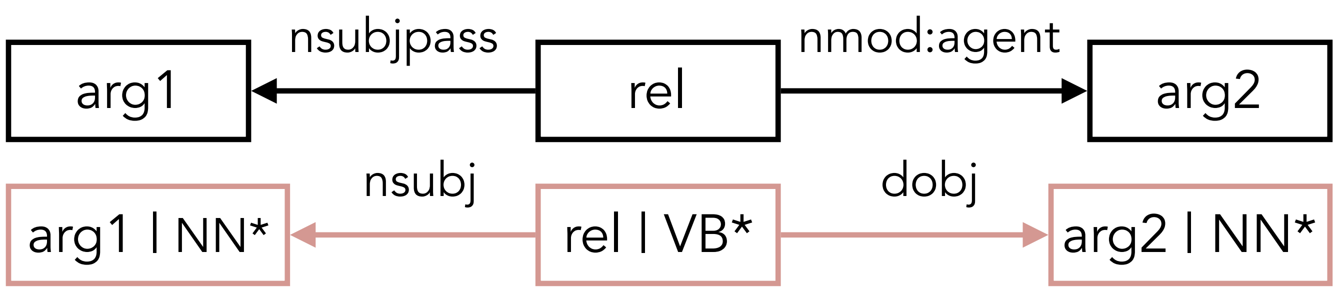


Extend existing bootstrapping approaches:

- Match all nodes in the template and not just two arguments (and relation)
- Learn nested extraction patterns

LEARNED PATTERNS

LEARNED PATTERNS

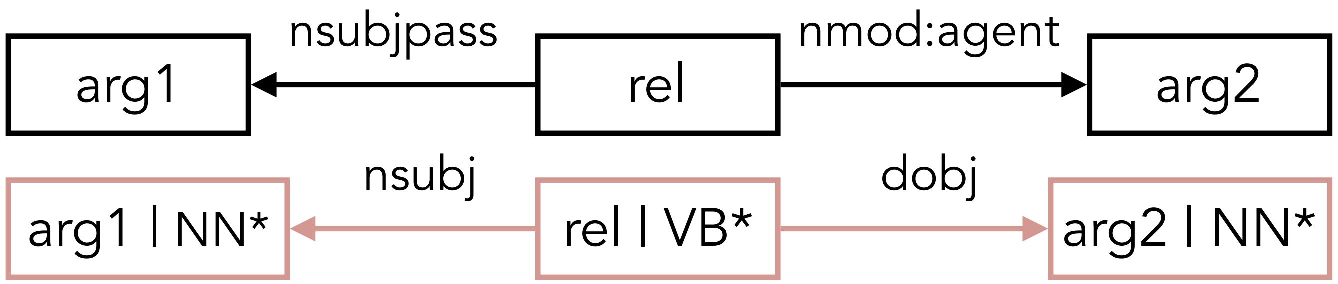


(arg1, [rel, by], arg2)

A body has been found by police.

Police found a body.

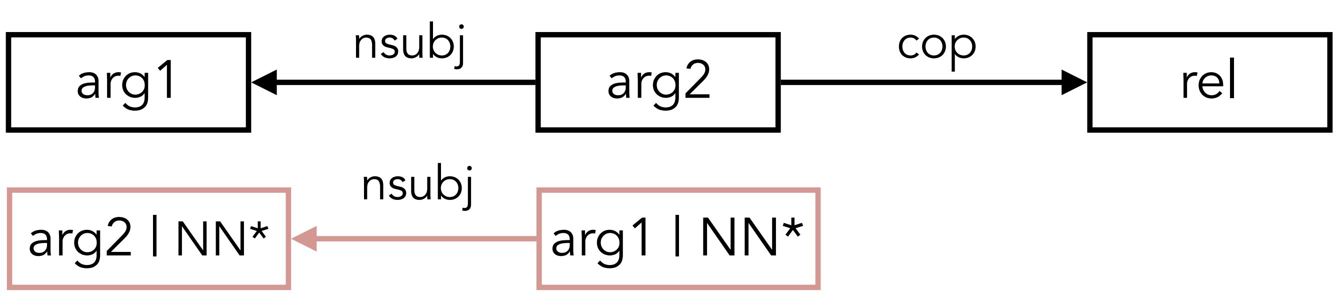
LEARNED PATTERNS



A body has been found by police.

Police found a body.

(arg1, [rel, by], arg2)

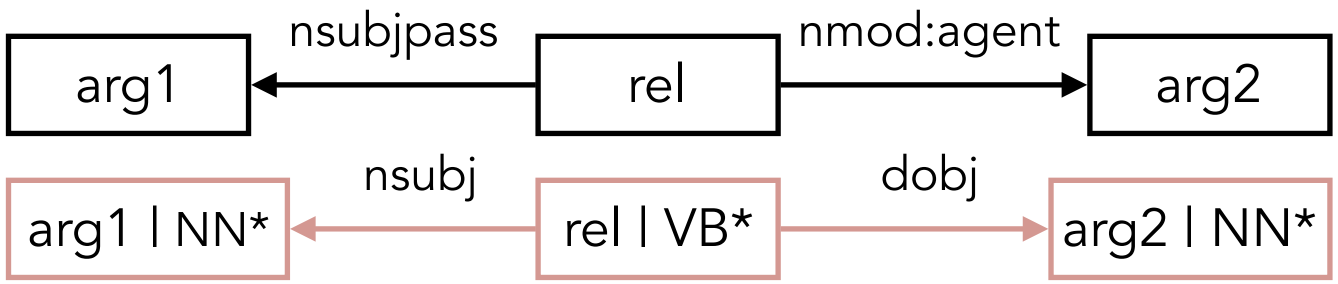


Fallujah is an Iraqi city

… in Iraqi city, Fallujah.

(arg1, be, arg2)

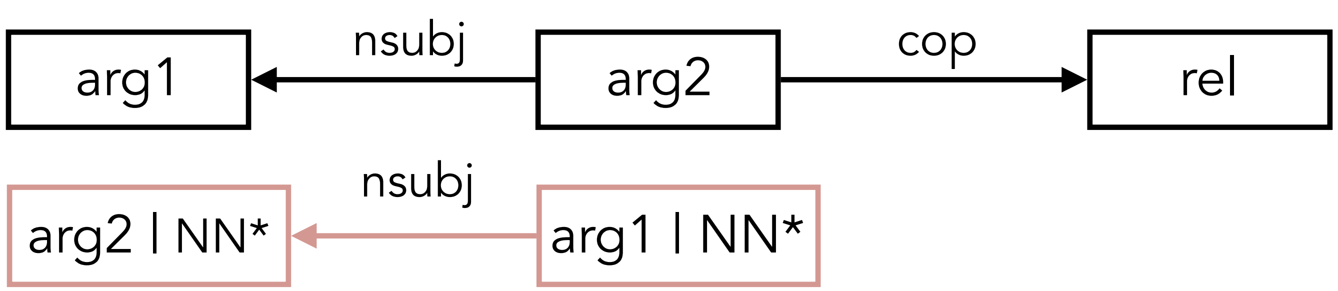
LEARNED PATTERNS



A body has been found by police.

Police found a body.

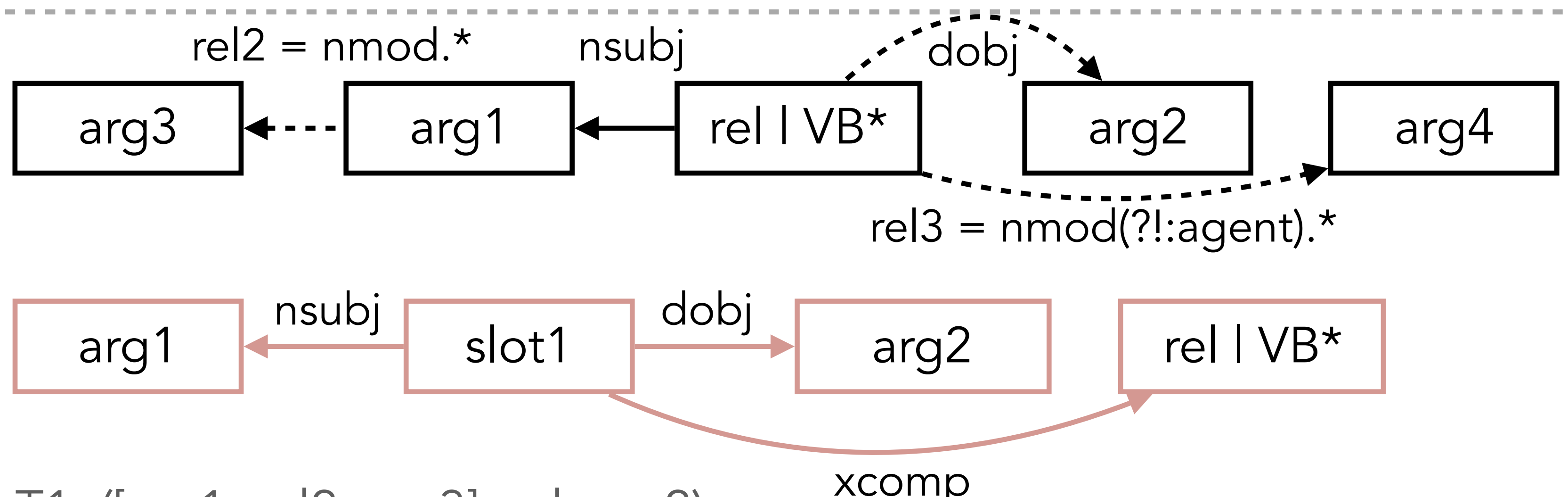
(arg1, [rel, by], arg2)



Fallujah is an Iraqi city

… in Iraqi city, Fallujah.

(arg1, be, arg2)



10,000 people in Africa died of Ebola

Several people are reported to have died

T1: ([arg1, rel2, arg3], rel, arg2)
T2: (T1, rel3, arg4)

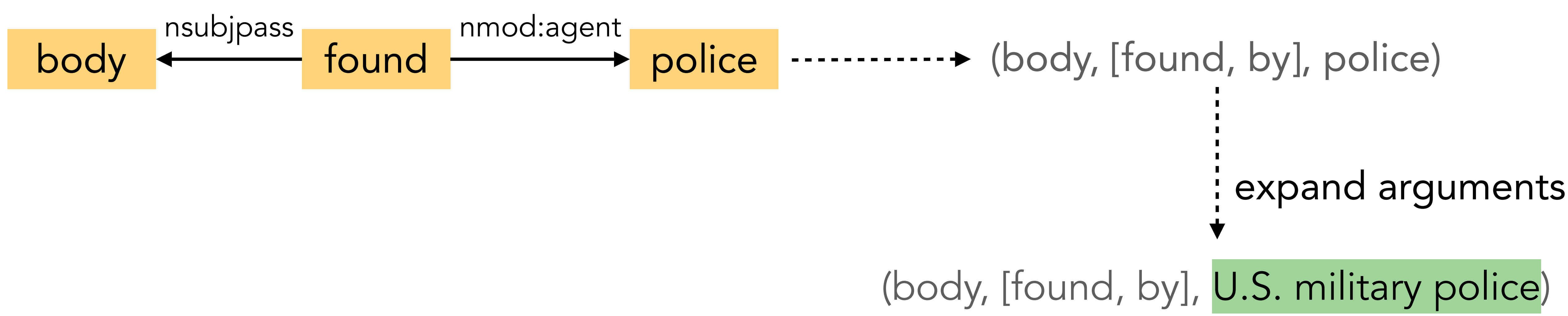
… 183 learned patterns

EXTRACTING PROPOSITIONS

"A body was found by U.S. military police."

EXTRACTING PROPOSITIONS

"A body was found by U.S. military police."



- Extend arguments on: nmod, amod, compound, nummod, det, neg
- Extend relations on: advmod, neg, aux, auxpass, cop, nmod

LINKING PROPOSITIONS

"A senior official in Iraq said the body, which was found by U.S. military police, was thrown from a vehicle."

P1: (the body, found by, U.S. military police)

P2: (A senior official in Iraq, said, ∅)

P3: ((the body, was thrown, ∅), from, a vehicle)

Missing Link

LINKING PROPOSITIONS

"A senior official in Iraq said the body, which was found by U.S. military police, was thrown from a vehicle."

P1: (the body, found by, U.S. military police)

P2: (A senior official in Iraq, said, ∅)

P3: ((the body, was thrown, ∅), from, a vehicle)

Missing Link

Template	too long too complex difficult to define

LINKING PROPOSITIONS

"A senior official in Iraq said the body, which was found by U.S. military police, was thrown from a vehicle."

P1: (the body, found by, U.S. military police)

P2: (A senior official in Iraq, said, ∅)

P3: ((the body, was thrown, ∅), from, a vehicle)

Missing Link

Template	too long too complex difficult to define

Use syntactic cues to identify missing links

more details in paper

EXPERIMENTAL SETUP

Dataset(s):

- 200 random sentences from Wikipedia*
- 200 random sentences from New York Times (NYT)*

* Datasets released with ClausIE

EXPERIMENTAL SETUP

Dataset(s):

- 200 random sentences from Wikipedia*
- 200 random sentences from New York Times (NYT)*

Baseline Systems:

- Reverb
- ClausIE
- Ollie

* Datasets released with ClausIE

EXPERIMENTAL SETUP

Dataset(s):

- 200 random sentences from Wikipedia*
- 200 random sentences from New York Times (NYT)*

Baseline Systems:

- Reverb
- ClausIE
- Ollie

Two annotators (CS graduate students) manually label the propositions for minimality, correctness, completeness: **pessimistic approach**

Inter-annotator agreement: 0.59 (kappa score)

* Datasets released with ClausIE

EVALUATION CRITERIA

INFORMATIVENESS

Set of propositions is ranked on a scale of 0-5, based on whether the set captured the meaning of the statement.

CORRECTNESS

A proposition is correct if it was asserted in the text and if it correctly captured the contextual information.

MINIMALITY

A proposition is minimal if the arguments or relation are not excessively long.

RESULTS

Dataset	Metric	Reverb	Ollie	ClausIE	NestIE
NYT	Informativeness	1.437/5	2.09/5	2.32/5	**2.762/5**
	Correct	187/275 (0.680)	359/529 (0.678)	527/882 (0.597)	469/914 (0.513)
	Minimal (of correct)	161/187 (0.861)	238/359 (0.663)	199/527 (0.377)	**355/469 (0.757)**
Wikipedia	Informativeness	1.63/5	2.267/5	2.432/5	**2.602/5**
	Correct	194/258 (0.752)	336/582 (0.577)	453/769 (0.589)	415/827 (0.501)
	Minimal (of correct)	171/194 (0.881)	256/336 (0.761)	214/453 (0.472)	**362/415 (0.872)**

RESULTS

Dataset	Metric	Reverb	Ollie	ClausIE	NestIE
NYT	Informativeness	1.437/5	2.09/5	2.32/5	$\boxed{\text{2.762/5}}$
	Correct	187/275 (0.680)	359/529 (0.678)	527/882 (0.597)	469/914 (0.513)
	Minimal (of correct)	161/187 (0.861)	238/359 (0.663)	199/527 (0.377)	**355/469 (0.757)**
Wikipedia	Informativeness	1.63/5	2.267/5	2.432/5	$\boxed{\text{2.602/5}}$
	Correct	194/258 (0.752)	336/582 (0.577)	453/769 (0.589)	415/827 (0.501)
	Minimal (of correct)	171/194 (0.881)	256/336 (0.761)	214/453 (0.472)	**362/415 (0.872)**

- NestIE has 1.1-1.9 times higher informativeness score than other systems

RESULTS

Dataset	Metric	Reverb	Ollie	ClausIE	NestIE
NYT	Informativeness	1.437/5	2.09/5	2.32/5	**2.762/5**
	Correct	187/275 (0.680)	359/529 (0.678)	527/882 (0.597)	469/914 (0.513)
	Minimal (of correct)	161/187 (0.861)	238/359 (0.663)	199/527 (0.377)	**355/469 (0.757)**
Wikipedia	Informativeness	1.63/5	2.267/5	2.432/5	**2.602/5**
	Correct	194/258 (0.752)	336/582 (0.577)	453/769 (0.589)	415/827 (0.501)
	Minimal (of correct)	171/194 (0.881)	256/336 (0.761)	214/453 (0.472)	**362/415 (0.872)**

- NestIE has 1.1-1.9 times higher informativeness score than other systems
- NestIE has more correct propositions than Ollie and Reverb

RESULTS

Dataset	Metric	Reverb	Ollie	ClausIE	NestIE
NYT	Informativeness	1.437/5	2.09/5	2.32/5	**2.762/5**
	Correct	187/275 (0.680)	359/529 (0.678)	527/882 (0.597)	469/914 (0.513)
	Minimal (of correct)	161/187 (0.861)	238/359 (0.663)	199/527 (0.377)	$\boxed{355/469\ (0.757)}$
Wikipedia	Informativeness	1.63/5	2.267/5	2.432/5	**2.602/5**
	Correct	194/258 (0.752)	336/582 (0.577)	453/769 (0.589)	415/827 (0.501)
	Minimal (of correct)	171/194 (0.881)	256/336 (0.761)	214/453 (0.472)	$\boxed{362/415\ (0.872)}$

- NestIE has 1.1-1.9 times higher informativeness score than other systems
- NestIE has more correct propositions than Ollie and Reverb
- NestIE has comparable precision, higher minimality and informativeness than ClauseIE

DISCUSSION

Do nested propositions improve minimality of any extractor?

DISCUSSION

Do nested propositions improve minimality of any extractor?

Ollie propositions

DISCUSSION

Do nested propositions improve minimality of any extractor?

Ollie propositions

Filter: n-ary relations, long arguments

Candidates for nested propositions

DISCUSSION

Do nested propositions improve minimality of any extractor?

Ollie propositions

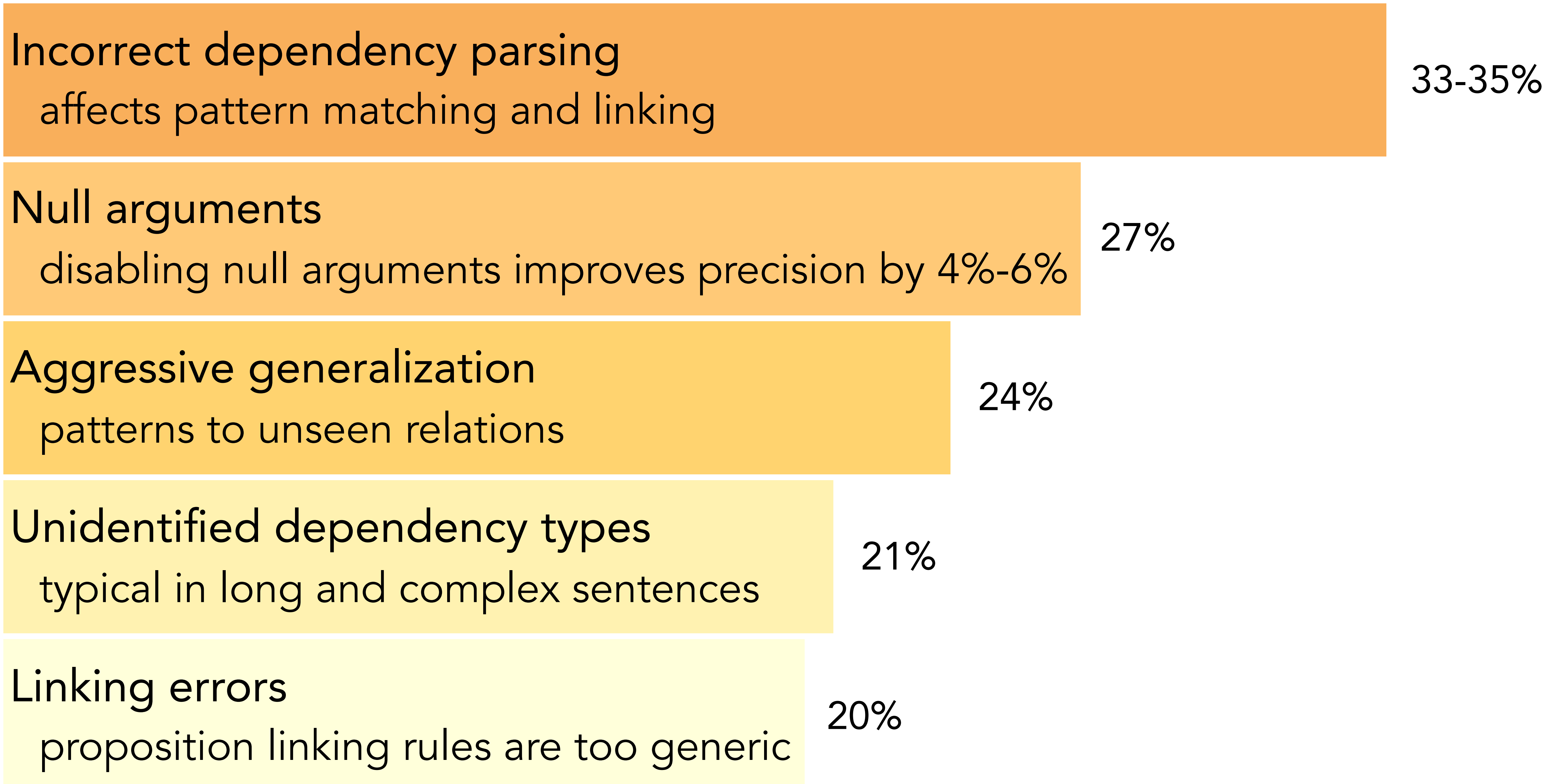
↓ Filter: n-ary relations, long arguments

Candidates for nested propositions

↓ Post-processing

Minimality: **76.1%** to **93.75%**

ERROR ANALYSIS



CONTRIBUTIONS AND FUTURE WORK

- Proposed a novel **nested representation** to express complex assertions
- Nested representation helps achieve higher **minimality** and **informativeness**
- Extended existing bootstrapping techniques to **learn dependency-based extraction patterns** for nested representation

CONTRIBUTIONS AND FUTURE WORK

- Proposed a novel **nested representation** to express complex assertions
- Nested representation helps achieve higher **minimality** and **informativeness**
- Extended existing bootstrapping techniques to **learn dependency-based extraction patterns** for nested representation

Future directions:

- Nested propositions for other tasks: **open question answering, SRL, ontology learning**
- Bootstrapping with bigger and nosier datasets
- **Sentence simplification** to under longer sentences correctly

UNIVERSITY OF MICHIGAN

CONTRIBUTIONS AND FUTURE WORK

- Proposed a novel **nested representation** to express complex assertions
- Nested representation helps achieve higher **minimality** and **informativeness**
- Extended existing bootstrapping techniques to **learn dependency-based extraction patterns** for nested representation

Future directions:

- Nested propositions for other tasks: **open question answering, SRL, ontology learning**
- Bootstrapping with bigger and nosier datasets
- **Sentence simplification** to understand longer sentences correctly

REFERENCES

- Open Question Answering over Curated and Extracted Knowledge Bases
  Fader et al., 2014, KDD
- Paraphrase-Driven Learning for Open Question Answering
  Fader et al., 2013, ACL
- ClausIE: Clause-based Open Information Extraction
  Corro et al., 2013, WWW
- Open Language Learning for Information Extraction
  Mausam et al., 2012, EMNLP
- Natural Language Questions for the Web of Data
  Mohamed, 2012, EMNLP-CoNLL
- Identifying Relations for Open Information Extraction
  Fader et al., 2011, EMNLP
- Open Information Extraction using Wikipedia
  Wu et al., 2010, ACL
- Open Information Extraction from the Web
  Banko et al., 2007, IJCAI

BACKUP SLIDES

- What information is expressed?

- How much to retain?

- How to identify it? e.g. non-verb mediated propositions, Messi, a golden ball winner, plays in Barcelona

RELATED WORK - Ollie

- Unlike previous extractors, can **capture relations not mediated by verbs**

"There are plenty of taxis available at Bali airport."

- Extend propositions to **include contextual information**

AttributedTo: who hopes, believes, said or doubts the information

ClausalModifier: extract information that is conditionally true

- Use **Reverb** extractions to bootstrap a training corpus that includes dependency path, relation words and sentence

- Learn open patterns to extract binary relations from unseen text

RELATED WORK - Ollie

- Unlike previous extractors, can **capture relations not mediated by verbs**

"There are plenty of taxis available at Bali airport."

- Extend propositions to **include contextual information**

AttributedTo: who hopes, believes, said or doubts the information

ClausalModifier: extract information that is conditionally true

- Use **Reverb** extractions to bootstrap a training corpus that includes dependency path, relation words and sentence

- Learn open patterns to extract binary relations from unseen text

- **NestIE** doesn't focus exclusively on binary relations. Uses seed templates that are more expressive.

RELATED WORK - ClausIE

- Dependency-based extractor
- Exploits knowledge of English grammar to detect clause constituents and type of each clause in a sentence
- Derive triples (possibly n-ary) from constituents
- Requires no training data, labeled or unlabeled
- Captures a subset of grammatical constructs to identify constituents
- Minimality is not the primary goal of the system

RELATED WORK - ClausIE

- Dependency-based extractor
- Exploits knowledge of English grammar to detect clause constituents and type of each clause in a sentence
- Derive triples (possibly n-ary) from constituents
- Requires no training data, labeled or unlabeled
- Captures a subset of grammatical constructs to identify constituents
- Minimality is not the primary goal of the system

- **NestIE** uses known grammatical constructs for generating seed set with minimal arguments. Bootstraps to learn more constructs that map to similar representation

Questions to answer:

- Why is the problem worth solving?

- Core difference between your method and all those that came before

- what does your method accomplishes

- why accomplish more?

- what is the evidence that it works better?

- one message