NestIE
NESTED PROPOSITIONS IN
OPEN INFORMATION EXTRACTION

Nikita Bhutani, H V Jagadish, Dragomir Radev
NestIE

NESTED PROPOSITIONS IN OPEN INFORMATION EXTRACTION

Nikita Bhutani, H V Jagadish, Dragomir Radev
Closed KB

Ontology
co-founder death_date birth_date

“Steve Jobs, who **co-founded** Apple..”

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

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

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

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

Proposition: (8.8 million people, lost, their jobs)
Proposition 1: ((8.8 million people, lost, their jobs) attributedTo believe, Bureau of Labor Statistics)
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..
PROPOSITIONS FOR BINARY RELATIONS

Binary++: n-ary relation

“8.8 million people lost their jobs in the Great Depression.”

Proposition$^2$: (8.8 million people, lost, their jobs, in the Great Depression)
Binary++: n-ary relation

“8.8 million people lost their jobs in the Great Depression.”

Proposition²: (8.8 million people, lost, their jobs, in the Great Depression) 

- few grammatical constructs to identify constituents: limited coverage

² [Del Corro et al.]
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)
Long arguments are not informative

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

Proposition: (Sheryl Sandberg, be, COO of Facebook and author of Lean In)

- an accurate fact may itself contain another accurate fact
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.]
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)

$^1$ [Schmitz et al.]
Missing context makes propositions incomplete

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

Proposition¹: (8.8 million people, lost, their jobs)

- limited expressivity

¹ [Schmitz et al.]
# NESTED PROPOSITIONS

<table>
<thead>
<tr>
<th>Complex Assertions</th>
<th>n-ary relations</th>
<th>nested relations</th>
<th>subordinate clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triples</td>
<td>limited expressivity</td>
<td>non-minimality</td>
<td>lost context</td>
</tr>
</tbody>
</table>

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

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

Challenges

Solution
NESTED PROPOSITIONS

Complex Assertions
- n-ary relations
- nested relations
- subordinate clauses

Triples
- limited expressivity
- non-minimality
- lost context

Complex User Information Needs

Nested Representation

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*

<table>
<thead>
<tr>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple, short sentences</td>
</tr>
<tr>
<td>hand-written templates</td>
</tr>
</tbody>
</table>

* [de Marneffe et al.]
SEED EXTRACTION AND BOOTSTRAPPING

RTE (Recognizing Textual Entailment) Dataset*

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Template</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple, short sentences</td>
<td>dependency sub-tree</td>
</tr>
<tr>
<td>hand-written templates</td>
<td>nested representation</td>
</tr>
</tbody>
</table>

* [de Marneffe et al.]
## SEED EXTRACTION AND BOOTSTRAPPING

### RTE (Recognizing Textual Entailment) Dataset*

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Template</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple, short sentences</td>
<td>dependency sub-tree</td>
<td>long, complex sentences</td>
</tr>
<tr>
<td>hand-written templates</td>
<td>nested representation</td>
<td>learn syntactic variants</td>
</tr>
</tbody>
</table>

* [de Marneffe et al.]
# RTE (Recognizing Textual Entailment) Dataset*

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Template</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple, short sentences</td>
<td>dependency sub-tree</td>
<td>long, complex sentences</td>
</tr>
<tr>
<td>hand-written templates</td>
<td>nested representation</td>
<td>learn syntactic variants</td>
</tr>
</tbody>
</table>

```
A body has been found by police.

(arg1, [rel, by], arg2)
```

* [de Marneffe et al.]

---

**SEED EXTRACTION AND BOOTSTRAPPING**

```
(arg1, [rel, by], arg2)
```

```
A body has been found by police.

(body, [found, by], police)
```
### SEED EXTRACTION AND BOOTSTRAPPING

**RTE (Recognizing Textual Entailment) Dataset**

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Template</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple, short sentences</td>
<td>dependency sub-tree</td>
<td>long, complex sentences</td>
</tr>
<tr>
<td>hand-written templates</td>
<td>nested representation</td>
<td>learn syntactic variants</td>
</tr>
</tbody>
</table>

- **Template**
  - A body has been found by police.
  - (body, [found, by], police)
  
- **Template**
  - Fallujah is an Iraqi city
  - (Fallujah, be, city)

* [de Marneffe et al.](#)
RTE (Recognizing Textual Entailment) Dataset*

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Template</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple, short sentences</td>
<td>dependency sub-tree</td>
<td>long, complex sentences</td>
</tr>
<tr>
<td>hand-written templates</td>
<td>nested representation</td>
<td>learn syntactic variants</td>
</tr>
</tbody>
</table>

A body has been found by police.
(body, [found, by], police)

Fallujah is an Iraqi city
(Fallujah, be, city)

10,000 people in Africa died of Ebola
T1: (people in Africa, died, ∅)
T2: (T1, of, Ebola)

* [de Marneffe et al.]

... 13 seed templates
"A body was found by U.S. military police."

```
(arg1, [rel, by], arg2)
```

```
body nsubjpass found nmod:agent police
```
"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

body \(\text{acl:recl}\) found \(\text{nmod:agent}\) police
PATTERN LEARNING

Remove surface-form

(body, found, police)

(arg1, [rel, by], arg2)
Extend existing bootstrapping approaches:

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

PATTERN LEARNING

- ACL: recl
- NMOD: agent

(body, found, police)

Remove surface-form

(NN*, VBD, NN*)

(arg1, [rel, by], arg2)
A body has been found by police.

Police found a body.
A body has been found by police.

Police found a body.

Fallujah is an Iraqi city

... in Iraqi city, Fallujah.
A body has been found by police.

Police found a body.

Fallujah is an Iraqi city

... in Iraqi city, Fallujah.

10,000 people in Africa died of Ebola

Several people are reported to have died

... 183 learned patterns
EXTRACTING PROPOSITIONS

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

(body, [found, by], 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
"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, Ø)

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

Template: too long, too complex, difficult to define
"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)

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)*
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

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

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

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

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

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

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

- 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

- Incorrect dependency parsing affects pattern matching and linking: 33-35%
- Null arguments: disabling null arguments improves precision by 4%-6%, 27%
- Aggressive generalization patterns to unseen relations, 24%
- Unidentified dependency types: typical in long and complex sentences, 21%
- Linking errors: proposition linking rules are too generic, 20%
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
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

Code to be released soon    |    Nikita Bhutani    |    nbhutani@umich.edu
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