ClausIE: Clause-Based Open Information Extraction
Luciano Del Corro, Rainer Gemulla

Hannah Seitz

Open Information Extraction

December 3, 2015
Overview

- What?
  - Extract triplets (subject, relation, argument(s))
  - Domain independent, unsupervised, scaleable
Overview

What?
- Extract triplets (subject, relation, argument(s))
- Domain independent, unsupervised, scaleable

How?
- Use dependency parsing
- Use syntactic information (clauses)
- Separate detection of clauses from proposition generation
- Use small domain independent lexica
Overview

- **What?**
  - Extract triplets (subject, relation, argument(s))
  - Domain independent, unsupervised, scalable

- **How?**
  - Use dependency parsing
  - Use syntactic information (clauses)
  - Separate detection of clauses from proposition generation
  - Use small domain independent lexica

- **Don’t:**
  - Use training data
  - Do global post-processing
Different Approaches to OIE

- Semantic parsing:
  - Shallow (TextRunner, Reverb)
  - Deep (WOE, OLLIE)
Different Approaches to OIE

- Semantic parsing:
  - Shallow (TextRunner, Reverb)
  - Deep (WOE, OLLIE)

- Patterns:
  - Handcrafted (KrakeN)
  - Learned (WOE, OLLIE)
Different Approaches to OIE

- Semantic parsing:
  - Shallow (TextRunner, Reverb)
  - Deep (WOE, OLLIE)
- Patterns:
  - Handcrafted (KrakeN)
  - Learned (WOE, OLLIE)

Main Differences of ClausIE:
- Use parsing and lexica
- Extract useful pieces of information (clauses)
- Reason about the set of clauses
Clause:
Part of a sentence, that expresses some coherent piece of information
Clause:

Part of a sentence, that expresses some coherent piece of information

- Consists of:
  - Subject (S)
  - Verb (V)
  - Optionally:
    - Indirect object (O)
    - Direct object (O)
    - Complement (C)
    - One or more adverbials (A)

7 possible clause types
Clauses

Clause:

Part of a sentence, that expresses some coherent piece of information

- Consists of:
  - Subject (S)
  - Verb (V)

- Optionally:
  - Indirect object (O)
  - Direct object (O)
  - Complement (C)
  - One or more adverbials (A)
Clauses

Clause:
Part of a sentence, that expresses some coherent piece of information

- Consists of:
  - Subject (S)
  - Verb (V)
- Optionally:
  - Indirect object (O)
  - Direct object (O)
  - Complement (C)
  - One ore more adverbials (A)

- 7 possible clause types
Verb Types

- Intransitive, monotransitive, ditransitive, copular, extended-copular
- Use of lexica:
  - Copular (be, prove, sound, ...)
  - Extended-copular (love, be, prove, ...)
  - Not extended-copular (die, walk)
  - Complex transitive: (set, lay, bring, ...)
- Determines type of clause and vice versa
- Combined approach
## Clause Types

### Table 1: Patterns and clause types (based on [15]).

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Clause type</th>
<th>Example</th>
<th>Derived clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic patterns</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_1$: $SV_1$</td>
<td>$SV$</td>
<td>AE died.</td>
<td>(AE, died)</td>
</tr>
<tr>
<td>$S_3$: $SV_cC$</td>
<td>$SVC$</td>
<td>AE is smart.</td>
<td>(AE, is, smart)</td>
</tr>
<tr>
<td>$S_4$: $SV_mO$</td>
<td>$SVO$</td>
<td>AE has won the Nobel Prize.</td>
<td>(AE, has won, the Nobel Prize)</td>
</tr>
<tr>
<td>$S_5$: $SV_{dt}O$</td>
<td>$SVOO$</td>
<td>RSAS gave AE the Nobel Prize.</td>
<td>(RSAS, gave, AE, the Nobel Prize)</td>
</tr>
<tr>
<td>$S_6$: $SV_{ct}OA$</td>
<td>$SVOA$</td>
<td>The doorman showed AE to his office.</td>
<td>(The doorman, showed, AE, to his office)</td>
</tr>
<tr>
<td>$S_7$: $SV_{ct}OC$</td>
<td>$SVOC$</td>
<td>AE declared the meeting open.</td>
<td>(AE, declared, the meeting, open)</td>
</tr>
<tr>
<td><strong>Some extended patterns</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_8$: $SV_{1AA}$</td>
<td>$SV$</td>
<td>AE died in Princeton in 1955.</td>
<td>(AE, died)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AE, died, in Princeton)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AE, died, in 1955)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AE, died, in Princeton, in 1955)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AE, remained, in Princeton, until his death)</td>
</tr>
<tr>
<td>$S_{10}$: $SV_cCA$</td>
<td>$SVC$</td>
<td>AE is a scientist of the 20th century.</td>
<td>(AE, is, a scientist)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AE, is, a scientist, of the 20th century)</td>
</tr>
<tr>
<td>$S_{11}$: $SV_mO A$</td>
<td>$SVO$</td>
<td>AE has won the Nobel Prize in 1921.</td>
<td>(AE, has won, the Nobel Prize)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AE, has won, the Nobel Prize, in 1921)</td>
</tr>
<tr>
<td>$S_{12}$: $ASV_mO$</td>
<td>$SVO$</td>
<td>In 1921, AE has won the Nobel Prize.</td>
<td>(AE, has won, the Nobel Prize)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AE, has won, the Nobel Prize, in 1921)</td>
</tr>
</tbody>
</table>

S: Subject, V: Verb, C: Complement, O: Direct object, O$_i$: Indirect object, A: Adverbial, V$_i$: Intransitive verb, V$_c$: Copular verb, V$_e$: Extended-copular verb, V$_{mt}$: Monotransitive verb, V$_{dt}$: Ditransitive verb, V$_{ct}$: Complex-transitive verb
Coherence

- Clause type conveys minimal unit of coherent information
Coherence

- Clause type conveys minimal unit of coherent information
- Adverbials:
  - Coherent without adverbial: optional
  - Coherent only with adverbial: essential
Coherence

- Clause type conveys minimal unit of coherent information
- Adverbials:
  - Coherent without adverbial: optional
  - Coherent only with adverbial: essential
- Other Clauses:
  - Always essential, always part of the clause type
Coherence

- Clause type conveys minimal unit of coherent information
- Adverbials:
  - Coherent without adverbial: optional
  - Coherent only with adverbial: essential
- Other Clauses:
  - Always essential, always part of the clause type

Coherent:

AE remained in Princeton. → SVA
AE died (in Princeton). → SV
Coherence

- Clause type conveys minimal unit of coherent information
- Adverbials:
  - Coherent without adverbial: optional
  - Coherent only with adverbial: essential
- Other Clauses:
  - Always essential, always part of the clause type

Coherent:

AE remained in Princeton.  $\rightarrow$ SVA
AE died (in Princeton).  $\rightarrow$ SV

Not coherent:

AE remained.
Figure 1: An example sentence with dependency parse, chunks, and POS tags (chunks by Apache OpenNLP)
From Dependencies to Clauses

- Identify head words
From Dependencies to Clauses

- Identify head words
- Map:
  - Subject - S
  - Main verb - V
  - dobj, iobj - O
  - xcomp, comp - C
  - advmod, advcl, prep_in - A
  - Dependant of the verb that occur to the right - candidate adverbial
From Dependencies to Clauses

- Identify head words
- Map:
  - Subject - S
  - Main verb - V
  - dobj, iobj - O
  - xcomp, comp - C
  - advmod, advcl, prep_in - A
  - Dependant of the verb that occur to the right - candidate adverbial

- Synthetic clauses for:
  - Relative pronouns
  - Appositions
  - Possessives
  - Participal modifiers
Identifying Clause Type

- **Decision tree**
  - Knowledge of properties of verbs (from lexica)
  - Knowledge of structure of the input clause (from dependency parse)
- **SV/SVA**: If undetermined, choose SVA
- **SVO/SVOA**: Lexicon-based heuristics, if undetermined, choose SVO

Figure 2: Flow chart for verb-type and clause-type detection
Proposition Generation

- Can be customized to the application
Proposition Generation

- Can be customized to the application
- Triples or n-ary propositions
Proposition Generation

- Can be customized to the application
- Triples or n-ary propositions
- Multiple triples: 
  - (S:AE, V:died)
  - (S:AE, V:died, A: [in] Princeton)
  - (S:AE, V:died, A: [in] 1955)
Proposition Generation

- Can be customized to the application
- Triples or n-ary propositions
- Multiple triples: 
  - $(S:AE, V:died)$
  - $(S:AE, V:died, A: [in] Princeton)$
  - $(S:AE, V:died, A: [in] 1955)$
- Select at most 1 optional adverbial (avoid over-specification)
Coordinated Conjunctions

- and, or
- Replace CC by its conjoints:
  - (S:Bell, V:makes and distributes, O: [electronic and computer and building] products):
    - (S:Bell, V:makes, O: [electronic] products)
    - (S:Bell, V:makes, O: [computer] products)
    - (S:Bell, V:makes, O: [building] products)
    - (S:Bell, V:distributes, O: [electronic] products)
    - (S:Bell, V:distributes, O: [computer] products)
    - (S:Bell, V:distributes, O: [building] products)
Coordinated Conjunctions

- and, or
- Replace CC by its conjoints:
- (S:Bell, V:makes and distributes, O: [electronic and computer and building] products):
  - (S:Bell, V:makes, O: [electronic] products)
  - (S:Bell, V:makes, O: [computer] products)
  - (S:Bell, V:makes, O: [building] products)
  - (S:Bell, V:distributes, O: [electronic] products)
  - (S:Bell, V:distributes, O: [computer] products)
  - (S:Bell, V:distributes, O: [building] products)

Problem: Combinatory CCs:
Anna and Bob married each other \neq Anna married each other; Bob married each other
Coordinated Conjunctions

- and, or
- Replace CC by its conjoints:
  - \((S:\text{Bell}, V:\text{makes and distributes}, O: \text{[electronic and computer and building] products})\):
    - \((S:\text{Bell, V:makes, O: [electronic] products})\)
    - \((S:\text{Bell, V:makes, O: [computer] products})\)
    - \((S:\text{Bell, V:makes, O: [building] products})\)
    - \((S:\text{Bell, V:distributes, O: [electronic] products})\)
    - \((S:\text{Bell, V:distributes, O: [computer] products})\)
    - \((S:\text{Bell, V:distributes, O: [building] products})\)

Problem: Combinatory CCs:
Anna and Bob married each other \(\neq\) Anna married each other; Bob married each other

- Solution: Optional processing of CCs
Proposition Generation

- Generate one proposition for each selected subset of constituents
Proposition Generation

- Generate one proposition for each selected subset of constituents
- Obtain textual representations by concatenating:
  - Bell makes, electronic products
  - Bell makes computer products
  - Bell makes building products
  - Bell distributes electronic products
  - Bell distributes computer products
  - Bell distributes building products
Datasets

- Reverb (500 sentences)
  - noisy
- Wikipedia (200 sentences)
  + short
  + simple
  - may contain mistakes
- New York Times (200 sentences)
  + clean
  - long
  - complex
- Manually labeled by 2 annotators
<table>
<thead>
<tr>
<th></th>
<th>ClausIE</th>
<th>ClausIE w/o CCs</th>
<th>ClausIE (non-redundant)</th>
<th>ClausIE w/o CCs (non-redundant)</th>
<th>OLLIE</th>
<th>Reverb</th>
<th>WOE</th>
<th>TextRunner</th>
<th>TextRunner (Reverb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverb dataset</td>
<td>1706/2975</td>
<td>1466/2344</td>
<td>1221/2161</td>
<td>1050/1707</td>
<td>547/1242</td>
<td>547/1242</td>
<td>388/727</td>
<td>447/1028</td>
<td>343/837</td>
</tr>
<tr>
<td>Wikipedia dataset</td>
<td>598/1001</td>
<td>536/792</td>
<td>424/727</td>
<td>381/569</td>
<td>234/565</td>
<td>234/565</td>
<td>165/249</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NYT dataset</td>
<td>696/1303</td>
<td>594/926</td>
<td>508/926</td>
<td>444/685</td>
<td>211/497</td>
<td>211/497</td>
<td>149/271</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(a) Reverb dataset
Table 2: Number of correct extractions and total number of extractions

<table>
<thead>
<tr>
<th></th>
<th>ClausIE</th>
<th>ClausIE w/o CCs</th>
<th>ClausIE (non-redundant)</th>
<th>ClausIE w/o CCs (non-redundant)</th>
<th>OLLIE</th>
<th>Reverb</th>
<th>WOE</th>
<th>TextRunner (Reverb)</th>
<th>TextRunner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverb dataset</td>
<td>1706/2975</td>
<td>1466/2344</td>
<td>1221/2161</td>
<td>1050/1707</td>
<td>547/1242</td>
<td>388/727</td>
<td>447/1028</td>
<td>343/837</td>
<td>286/798</td>
</tr>
<tr>
<td>Wikipedia dataset</td>
<td>598/1001</td>
<td>536/792</td>
<td>424/727</td>
<td>381/569</td>
<td>234/565</td>
<td>165/249</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NYT dataset</td>
<td>696/1303</td>
<td>594/926</td>
<td>508/926</td>
<td>444/685</td>
<td>211/497</td>
<td>149/271</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(b) Wikipedia dataset
### Table 2: Number of correct extractions and total number of extractions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ClausIE</th>
<th>ClausIE w/o CCs</th>
<th>ClausIE (non-redundant)</th>
<th>ClausIE w/o CCs (non-redundant)</th>
<th>OLLIE</th>
<th>Reverb</th>
<th>WOE</th>
<th>TextRunner (Reverb)</th>
<th>TextRunner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverb dataset</td>
<td>1706/2975</td>
<td>1496/2344</td>
<td>1221/2161</td>
<td>1050/1707</td>
<td>547/1242</td>
<td>388/722</td>
<td>447/1028</td>
<td>343/837</td>
<td>286/798</td>
</tr>
<tr>
<td>Wikipedia dataset</td>
<td>598/1001</td>
<td>536/792</td>
<td>424/727</td>
<td>381/569</td>
<td>234/565</td>
<td>165/249</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NYT dataset</td>
<td>696/1303</td>
<td>594/926</td>
<td>508/926</td>
<td>444/685</td>
<td>211/497</td>
<td>149/271</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(c) New York Times dataset
Results

- 2.5 – 3.5 (1.8 – 2.4) times more correct extraction than OLLIE
Results

- 2.5 – 3.5 (1.8 – 2.4) times more correct extraction than OLLIE
- 27 – 29% redundancy
Results

- 2.5 – 3.5 (1.8 – 2.4) times more correct extraction than OLLIE
- 27 – 29% redundancy
- Advantages of ClausIE:
  - Considers all adverbials
  - Extracts non-verb-mediated propositions,
  - Detects non-consecutive constituents
  - Processes coordinated conjunctions
Improvements

- Incorrect dependency parses
  - Better parses will give better results
Improvements

- Incorrect dependency parses
  - Better parses will give better results
- Processing of coordinated clauses
  - Returns false extractions for combinatory conjunctions
    (-5 – 10.7% precision)
  - Returns overall more correct extractions
    (+11 – 27%) correct extractions
Improvements

- Incorrect dependency parses
  - Better parses will give better results
- Processing of coordinated clauses
  - Returns false extractions for combinatory conjunctions
    (-5 – 10.7% precision)
  - Returns overall more correct extractions
    (+11 – 27%) correct extractions
- SVOA - SVO
  - Hard to distinguish
  - Implementation tends to miss essential adverbials
  - Solution: Improved dictionaries
Any questions?

Otherwise, let’s have a look at it!