ClausIE: Clause-Based Open Information Extraction

Luciano Del Corro Rainer Gemulla
Max-Planck-Institut für Informatik
May 2013
GOAL: Extract information from natural text
**GOAL:** Extract information from natural text

**Sentence**
Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.
GOAL: Extract information from natural text

Sentence
Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Extractions/Propositions
(Bell, ’is’, a telecommunication company)
(Bell, is based in, Los Angeles)
(Bell, makes, electronic products)
(Bell, distributes, electronic products)

...
GOAL: Extract information from natural text

Sentence
Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Extractions/Propositions
(Bell, 'is', a telecommunication company)
(Bell, is based in, Los Angeles)
(Bell, makes, electronic products)
(Bell, distributes, electronic products)

Most OIE extractors
- Propositions expressed as triples \((arg_1, relation, arg_2)\)
- Verb based relation
- Arguments restricted to noun phrases
Challenges/Requirements

- Domain independent
- Unbounded set of relations
- No filtering of information
- Structured output
- Scalable
Open Information Extraction: challenges and applications

Challenges/Requirements

- Domain independent
- Unbounded set of relations
- No filtering of information
- Structured output
- Scalable

Applications

- Structured search
- Automatic ontology construction
- Question answering
- Semantic role labeling, discourse parsing, ...?
Outline

1. Information and Representation
2. Open Information Extractors and Language Technology
3. ClausIE
   - Clauses in the English Language
   - From clauses to propositions
4. Results
5. Conclusions and Future Directions
Information and Representation: a two-step approach

Information

- 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

Representation

- What is the form of the relation?
  - Messi plays in Barcelona → plays or plays in
- Triples or n-ary propositions?
  - (Messi, plays football in, Barcelona)
  - (Messi, plays, football, in Barcelona)
- What should be the scope of the arguments?
  - Gandhi was vegetarian
Information and Representation: a two-step approach

Information

- 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

Representation

- What is the form of the relation?
  - Messi plays in Barcelona $\rightarrow$ plays or plays in
- Triples or n-ary propositions?
  - (Messi, plays football in, Barcelona) or (Messi, plays, football, in Barcelona)
- What should be the scope of the arguments?
  - Gandhi was vegetarian
Information and Representation: a two-step approach

Information
- 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

Representation
- What is the form of the relation?
  - Messi plays in Barcelona $\rightarrow$ plays or plays in
- Triples or n-ary propositions?
  - (Messi, plays football in, Barcelona) or (Messi, plays, football, in Barcelona)
- What should be the scope of the arguments?
  - Gandhi was vegetarian

We aim to separate these two phases
Outline

1. Information and Representation

2. Open Information Extractors and Language Technology

3. ClausIE
   - Clauses in the English Language
   - From clauses to propositions

4. Results

5. Conclusions and Future Directions
Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.
Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.
Outline

1 Information and Representation

2 Open Information Extractors and Language Technology

3 ClausIE
   - Clauses in the English Language
   - From clauses to propositions

4 Results

5 Conclusions and Future Directions
Clause Essentials

- A clause is like a simple sentence
  - Paul eats a chocolate bar
Clause Essentials

- A clause is like a simple sentence
  - Paul eats a chocolate bar

- A sentence can be composed by more than one clause
  - Anna drinks coffee and Bob plays football
Clause Essentials

- A clause is like a simple sentence
  - Paul eats a chocolate bar

- A sentence can be composed by more than one clause
  - Anna drinks coffee and Bob plays football

- Each clause encodes one or more propositions
Clause Essentials

- A clause is like a simple sentence
  - Paul eats a chocolate bar

- A sentence can be composed by more than one clause
  - Anna drinks coffee and Bob plays football

- Each clause encodes one or more propositions

- Clauses can have optional adverbials
  - He will take the exam in May
Clause Essentials

- A clause is like a simple sentence
  - Paul eats a chocolate bar

- A sentence can be composed by more than one clause
  - Anna drinks coffee and Bob plays football

- Each clause encodes one or more propositions

- Clauses can have optional adverbials
  - He will take the exam in May

- A minimal clause is a clause without its optional adverbials
  - He will take the exam
The seven clauses

1. $SV_i \rightarrow \text{Albert Einstein died.}$

**S:** Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, **O_i:** Indirect Object, **O:** Direct Object
The seven clauses

1. **SV_i** → Albert Einstein died.

2. **SV_e A** → Albert Einstein remained in Princeton.
The seven clauses

1. \(SV_i\) \rightarrow Albert Einstein died.

2. \(SV_e\ A\) \rightarrow Albert Einstein remained in Princeton.

3. \(SV_c\ C\) \rightarrow Albert Einstein is smart.

**S:** Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, **O_i:** Indirect Object, **O:** Direct Object
The seven clauses

1. \( SV_i \) → Albert Einstein died.
2. \( SV_e \ A \) → Albert Einstein remained in Princeton.
3. \( SV_c \ C \) → Albert Einstein is smart.
4. \( SV_{mt} \ O \) → Albert Einstein has won the Nobel Prize.
5. \( SV_{dt} \ O_i \ O_d \) → RSAS gave Albert Einstein the Nobel Prize.
6. \( SV_{ct} \ O \ A \) → The doorman showed Albert Einstein to his office.
7. \( SV_{ct} \ O \ C \) → Albert Einstein declared the meeting open.

**S**: Subject, **V**: Verb, **A**: Adverbial, **C**: Complement, **O_i**: Indirect Object, **O**: Direct Object
The seven clauses

1. $SV_i$ → Albert Einstein died.
2. $SV_e A$ → Albert Einstein remained in Princeton.
3. $SV_c C$ → Albert Einstein is smart.
4. $SV_{mt} O$ → Albert Einstein has won the Nobel Prize.
5. $SV_{dt} O_i O_d$ → RSAS gave Albert Einstein the Nobel Prize.
6. $SV_{ct} O A$ → The doorman showed Albert Einstein to his office.
7. $SV_{ct} O C$ → Albert Einstein declared the meeting open.

By identifying each minimal clause in a sentence we can identify the essential information.
# The seven clauses: optional adverbials

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Clause Type</th>
<th>Example</th>
<th>Derived clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV_AA</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>
</tbody>
</table>

**S**: Subject, **V**: Verb, **A**: Adverbial, **C**: Complement, **O_i**: Indirect Object, **O**: Direct Object
# The seven clauses: optional adverbials

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Clause Type</th>
<th>Example</th>
<th>Derived clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV₁IA</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>SVEAA</td>
<td>SVA</td>
<td>AE remained in Princeton until his death.</td>
<td>(AE, remained, in Princeton)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AE, remained, in Princeton, until his death)</td>
</tr>
</tbody>
</table>

**S**: Subject, **V**: Verb, **A**: Adverbial, **C**: Complement, **O₁**: Indirect Object, **O**: Direct Object
## The seven clauses: optional adverbials

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Clause Type</th>
<th>Example</th>
<th>Derived clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>(SV_e AA)</td>
<td>SVA</td>
<td>AE remained in Princeton until his death.</td>
<td>(AE, remained, in Princeton) (AE, remained, in Princeton, until his death)</td>
</tr>
<tr>
<td>(SV_c CA)</td>
<td>SVC</td>
<td>AE is a scientist of the 20th century.</td>
<td>(AE, is, a scientist) (AE, is, a scientist, of the 20th century)</td>
</tr>
<tr>
<td>(SV_{mt} OA)</td>
<td>SVO</td>
<td>AE has won the Nobel Prize in 1921.</td>
<td>(AE, has won, the Nobel Prize) (AE, has won, the Nobel Prize, in 1921)</td>
</tr>
<tr>
<td>(ASV_{mt} O)</td>
<td>SVO</td>
<td>In 1921, AE has won the Nobel Prize.</td>
<td>(AE, has won, the Nobel Prize) (AE, has won, the Nobel Prize, in 1921)</td>
</tr>
</tbody>
</table>

**S:** Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, **O_i:** Indirect Object, **O:** Direct Object
Gandhi was vegetarian.

Gandhi was vegetarian.

DP
From clauses to clause types (I)

Gandhi was vegetarian.

DP → Clause
Gandhi was vegetarian.

NNP VBD JJ.
From clauses to clause types (I)

Gandhi was vegetarian.
NNP VBD JJ.

Q1
DP → Clause →
Object?

Q2
Complement?

No
From clauses to clause types (I)

Gandhi was vegetarian.
NNP VBD JJ.

1. DP $\rightarrow$ Clause $\rightarrow$ Object?
2. Q$_1$ $\rightarrow$ Complement?
3. No $\rightarrow$ Q$_2$
4. Yes $\rightarrow$ Copular (SVC)

Gandhi was vegetarian.
From clauses to clause types (I)

Gandhi was vegetarian.

(NNP VBD JJ.

(S: Gandhi, V: was, C: vegetarian)

DP → Clause → Object?

Complement?

Yes

Q2

Copular (SVC)

No

Q1
Albert Einstein died in Princeton.
B-NP I-NP B-VP B-PP B-NP.
NNP NNP VBD IN NNP.

DP
Albert Einstein died in Princeton.
B-NP I-NP B-VP B-PP B-NP.
NNP NNP VBD IN NNP.

DP → Clause
From clauses to clause types (II)

Albert Einstein died in Princeton.
B-NP I-NP B-VP B-PP B-NP.
NNP NNP VBD IN NNP.

DP → Clause → Object?
From clauses to clause types (II)

Albert Einstein died in Princeton.
B-NP I-NP B-VP B-PP B-NP.
NNP NNP VBD IN NNP.
Albert Einstein died in Princeton.
B-NP I-NP B-VP B-PP B-NP.
NNP NNP VBD IN NNP.

From clauses to clause types (II)

DP → Clause → Object?

\[ Q_1 \]

Complement?

\[ \text{No} \]

Candidate adverbial?

\[ \text{No} \]

\[ \text{No} \]
From clauses to clause types (II)

Albert Einstein died in Princeton.
B-NP I-NP B-VP B-PP B-NP.
NNP NNP VBD IN NNP.
From clauses to clause types (II)

Albert Einstein died in Princeton.
B-NP I-NP B-VP B-PP B-NP.
NNP NNP VBD IN NNP.

Intransitive (SV)

Q2

Complement?
No

Candidate adverbial?
Yes

Known non-ext. copular?
Yes

Q1

Object?

DP → Clause
From clauses to clause types (II)

Albert Einstein died in Princeton.
B-NP I-NP B-VP B-PP B-NP.
NNP NNP VBD IN NNP.

(S: AE, V: died,)
(S: AE, V: died, A: in Princeton)
From clauses to clause types (II)
We first identify the information and then generate the proposition.
We first identify the information and then generate the proposition.
Del Corro, Gemulla (MPI)
Example

Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.
Example

Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Reverb  \rightarrow  (a telecommunication company, is based in, Los Angeles)

Ollie  \rightarrow  (Bell, distributes, electronic, computer and building products)
Example

Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Reverb → (a telecommunication company, is based in, Los Angeles)

Ollie → (Bell, distributes, electronic, computer and building products)

ClausIE → (S: Bell, V: ’is’, C: a telecommunication company)
(S: Bell, V: is based, A: in Los Angeles)
(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)
Example

Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Reverb → (a telecommunication company, is based in, Los Angeles)

Ollie → (Bell, distributes, electronic, computer and building products)

ClausIE → (S: Bell, V: 'is', C: a telecommunication company)
(S: Bell, V: is based, A: in Los Angeles)
(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)
Example

Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Reverb → (a telecommunication company, is based in, Los Angeles)

Ollie → (Bell, distributes, electronic, computer and building products)

ClausIE → (S: Bell, V: 'is', C: a telecommunication company)
(S: Bell, V: is based, A: in Los Angeles)
(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)
Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Reverb → (a telecommunication company, is based in, Los Angeles)

Ollie → (Bell, distributes, electronic, computer and building products)

ClausIE → (S: Bell, V: 'is', C: a telecommunication company)
(S: Bell, V: is based, A: in Los Angeles)
(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)
Example

Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Reverb → (a telecommunication company, is based in, Los Angeles)

Ollie → (Bell, distributes, electronic, computer and building products)

ClausIE → (S: Bell, V: 'is', C: a telecommunication company)
(S: Bell, V: is based, A: in Los Angeles)
(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)
Identifying information

- ClausIE separates the identification of the information from its representation
Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause
Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause
- No training data
Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause
- No training data
- Initial support non-verb mediated relations

⋆ Messi and Iniesta play in Barcelona
   → (Messi, plays, in Barcelona), (Iniesta, plays, in Barcelona)

⋆ I saw the man whose house you like
   → (I, saw, the man), (You, like, the man's house)
Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause
- No training data
- Initial support non-verb mediated relations
- Processing of conjunctions (in verbs and subject/arguments)
  - Messi and Iniesta play in Barcelona $\rightarrow$ \((Messi, \text{plays, in Barcelona}), \ (Iniesta, \text{plays, in Barcelona})\)
Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause
- No training data
- Initial support non-verb mediated relations
- Processing of conjunctions (in verbs and subject/arguments)
  - Messi and Iniesta play in Barcelona → (Messi, plays, in Barcelona), (Iniesta, plays, in Barcelona)
- Resolution of relative clauses
  - I saw the man whose house you like → (I, saw, the man), (You, like, the man’s house) ...
Proposition Generation: a flexible process

Arbitrary form of relations

- *(Messi, plays football in, Barcelona) or (Messi, plays, football in Barcelona)*
Proposition Generation: a flexible process

- Arbitrary form of relations
  - *(Messi, plays football in, Barcelona) or (Messi, plays, football in Barcelona)*

- Propositions can be customized (e.g. triple, n-ary, etc)
  - *(Messi, plays, football in Barcelona) or (Messi, plays, football, in Barcelona)*
**Proposition Generation: a flexible process**

- **Arbitrary form of relations**
  - *(Messi, plays football in, Barcelona)* or *(Messi, plays, football in Barcelona)*

- **Propositions can be customized (e.g. triple, n-ary, etc)**
  - *(Messi, plays, football in Barcelona)* or *(Messi, plays, football, in Barcelona)*

- **Arbitrary argument types (e.g. noun phrases, adjectives, etc)**
  - *(Gandhi, was, vegetarian)* or *(Gandhi, was, a vegetarian)* or *(Gandhi from Porbandar, was, a vegetarian)*
Proposition Generation: a flexible process

- Arbitrary form of relations
  - \((Messi, \text{plays football in}, \text{Barcelona})\) or \((Messi, \text{plays, football in Barcelona})\)

- Propositions can be customized (e.g. triple, n-ary, etc)
  - \((Messi, \text{plays, football in Barcelona})\) or \((Messi, \text{plays, football, in Barcelona})\)

- Arbitrary argument types (e.g. noun phrases, adjectives, etc)
  - \((Gandhi, \text{was, vegetarian})\) or \((Gandhi, \text{was, a vegetarian})\) or \((Gandhi \text{from Porbandar, was, a vegetarian})\)

- Optional arguments can be used to generate new propositions
  - \((Paul, \text{takes, a shower, in the morning})\) or \((Paul, \text{takes, a shower})\)
Outline

1. Information and Representation
2. Open Information Extractors and Language Technology
3. ClausIE
   - Clauses in the English Language
   - From clauses to propositions
4. Results
5. Conclusions and Future Directions
Evaluation

- 3 datasets
  - Reverb: Web, very noisy (500 sentences)
  - New York Times: Complex, written by experts (200 sentences)
  - Wikipedia: Simple, written by non-experts (200 sentences)

- 2 labelers, pessimistic approach.

- Agreement 57%-68%.

- High precision, high recall.
Results I: Reverb Sentences

![Graph showing precision vs. number of extractions for various systems including ClausIE, ClausIE (non-red.), ClausIE w/o CCs, ClausIE w/o CCs (non-red.), Reverb, OLLIE, TextRunner, and TextRunner (Reverb)].

<table>
<thead>
<tr>
<th>System</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClausIE</td>
<td>Blue</td>
</tr>
<tr>
<td>ClausIE (non-red.)</td>
<td>Blue dashes</td>
</tr>
<tr>
<td>ClausIE w/o CCs</td>
<td>Cyan</td>
</tr>
<tr>
<td>ClausIE w/o CCs (non-red.)</td>
<td>Cyan dashes</td>
</tr>
<tr>
<td>Reverb</td>
<td>Green</td>
</tr>
<tr>
<td>OLLIE</td>
<td>Orange</td>
</tr>
<tr>
<td>TextRunner</td>
<td>Red</td>
</tr>
<tr>
<td>TextRunner (Reverb)</td>
<td>Red dashes</td>
</tr>
<tr>
<td>WOE</td>
<td>Gray</td>
</tr>
</tbody>
</table>
Results II: Wikipedia and New York Times

Wikipedia (200 sentences)  
New York Times (200 sentences)
Outline

1. Information and Representation
2. Open Information Extractors and Language Technology
3. ClausIE
   - Clauses in the English Language
   - From clauses to propositions
4. Results
5. Conclusions and Future Directions
Conclusions and Future Directions

Conclusions

- ClausIE is a principled approach for OIE
- Separates identification and representation
- No training needed
- DP based

Future Directions

- Build dictionaries
- Incorporate context analysis
- Post processing of arguments
- Input to other tasks: discourse processing, SRL, targeted IE, ontology learning, QA, ...
Conclusions and Future Directions

Conclusions

- ClausIE is a principled approach for OIE
- Separates identification and representation
- No training needed
- DP based

Future Directions

- Build dictionaries
- Incorporate context analysis
- Post processing of arguments
- Input to other tasks: discourse processing, SRL, targeted IE, ontology learning, QA, ...
Conclusions and Future Directions

Conclusions

- ClausIE is a principled approach for OIE
- Separates identification and representation
- No training needed
- DP based

Future Directions

- Build dictionaries
- Incorporate context analysis
- Post processing of arguments
- Input to other tasks: discourse processing, SRL, targeted IE, ontology learning, QA, ...

Thank You!