# A Survey on Open Information Extraction

COLING 2018

Christina Niklaus[1], Matthias Cetto[1], André Freitas[2]
and Siegfried Handschuh[1]

[1]Natural Language Processing and Semantic Computing
University of Passau

[2]School of Computer Science
University of Manchester

August 24, 2018

# Information Extraction

- Task of IE: distill **semantic relations** from NL text

*"Barack Obama was born in 1961."*

$\langle$*Barack Obama*; *was born in*; *1961*$\rangle$

- **Traditional IE**:
  - hand-labeled data
  - pre-defined set of target relations (supervised approach)
  - small, homogeneous corpora
- scalability to large, heterogeneous corpora?

- Task of IE: distill **semantic relations** from NL text

"*Barack Obama was born in 1961.*"

⟨*Barack Obama*; *was born in*; *1961*⟩

- **Traditional IE**:
  - hand-labeled data
  - pre-defined set of target relations (supervised approach)
  - small, homogeneous corpora
- scalability to large, heterogeneous corpora?

UNIVERSITÄT PASSAU

- Task of IE: distill **semantic relations** from NL text

"*Barack Obama was born in 1961.*"

⇓

⟨*Barack Obama*; *was born in*; *1961*⟩

- **Traditional IE**:
  - hand-labeled data
  - pre-defined set of target relations (supervised approach)
  - small, homogeneous corpora
- scalability to large, heterogeneous corpora?

# Information Extraction

- Task of IE: distill **semantic relations** from NL text

  "*Barack Obama was born in 1961.*"

  ⟨*Barack Obama*; *was born in*; *1961*⟩

- **Traditional IE**:
  - hand-labeled data
  - pre-defined set of target relations (supervised approach)
  - small, homogeneous corpora

- scalability to large, heterogeneous corpora?

# Information Extraction

- Task of IE: distill **semantic relations** from NL text

“*Barack Obama was born in 1961.*”

⟨*Barack Obama*; *was born in*; *1961*⟩

- **Traditional IE**:
  - hand-labeled data
  - pre-defined set of target relations (supervised approach)
  - small, homogeneous corpora
- scalability to large, heterogeneous corpora?

# Open Information Extraction

- Introduction of a new extraction paradigm: **Open IE** (Banko et al., 2007)
- **Challenges** of Open IE systems:
  1. **automation**: automatic discovery of relations (unsupervised approach)
  2. **corpus heterogeneity**: domain-independent usage
  3. **efficiency**: readily scale to large amounts of text

# Open Information Extraction

- Introduction of a new extraction paradigm: **Open IE** (Banko et al., 2007)
- **Challenges** of Open IE systems:
  1. **automation**: automatic discovery of relations (unsupervised approach)
  2. **corpus heterogeneity**: domain-independent usage
  3. **efficiency**: readily scale to large amounts of text

# Early Approaches in Open IE

1. **hand-crafted extraction patterns**:
   A human manually defines a set of extraction rules.
2. **self-supervised learning**:
   The system automatically finds and labels its own training examples.

# Early Approaches in Open IE

1. **hand-crafted extraction patterns**:
   A human manually defines a set of extraction rules.
2. **self-supervised learning**:
   The system automatically finds and labels its own training examples.

# Early Approaches in Open IE

1. **hand-crafted extraction patterns**:
   A human manually defines a set of extraction rules.
2. **self-supervised learning**:
   The system automatically finds and labels its own training examples.

# TextRunner (Banko et al., 2007)

1. **self-supervised learner**: heuristically identifies and labels a set of extractions as examples to train a model of relations using unlexicalized features
2. **extractor**: makes a single pass over the corpus to extract tuples for *all* possible relations
3. **redundancy-based assessor**: assigns a probability to each tuple based on the number of sentences from which each extraction was found

# TextRunner (Banko et al., 2007)

1. **self-supervised learner**: heuristically identifies and labels a set of extractions as examples to train a model of relations using unlexicalized features
2. **extractor**: makes a single pass over the corpus to extract tuples for *all* possible relations
3. **redundancy-based assessor**: assigns a probability to each tuple based on the number of sentences from which each extraction was found

# TextRunner (Banko et al., 2007)

1. **self-supervised learner**: heuristically identifies and labels a set of extractions as examples to train a model of relations using unlexicalized features
2. **extractor**: makes a single pass over the corpus to extract tuples for *all* possible relations
3. **redundancy-based assessor**: assigns a probability to each tuple based on the number of sentences from which each extraction was found

# TextRunner (Banko et al., 2007)

1. **self-supervised learner**: heuristically identifies and labels a set of extractions as examples to train a model of relations using unlexicalized features
2. **extractor**: makes a single pass over the corpus to extract tuples for *all* possible relations
3. **redundancy-based assessor**: assigns a probability to each tuple based on the number of sentences from which each extraction was found

# TextRunner (Banko et al., 2007)

1. **self-supervised learner**: heuristically identifies and labels a set of extractions as examples to train a model of relations using unlexicalized features
2. **extractor**: makes a single pass over the corpus to extract tuples for *all* possible relations
3. **redundancy-based assessor**: assigns a probability to each tuple based on the number of sentences from which each extraction was found

**Example**

*The novelist Franz Kafka is the author of a short story entitled "The Metamorphosis".*

# TextRunner (Banko et al., 2007)

1. **self-supervised learner**: heuristically identifies and labels a set of extractions as examples to train a model of relations using unlexicalized features
2. **extractor**: makes a single pass over the corpus to extract tuples for *all* possible relations
3. **redundancy-based assessor**: assigns a probability to each tuple based on the number of sentences from which each extraction was found

**Example**

*The novelist Franz Kafka is the author of a short story entitled "The Metamorphosis".*

---

⟨*The novelist Franz Kafka; is; the author of a short story*⟩

# Problems with TextRunner

- **incoherent extractions**: relational phrase has no meaningful interpretation

| Sentence | Incoherent Relation |
|---|---|
| The guide *contains* dead links and *omits* sites. | contains omits |
| The Mark 14 *was central* to the *torpedo* scandal of the fleet. | was central torpedo |
| They *recalled* that Nungesser *began* his career as a precinct leader. | recalled began |

- **uninformative extractions**: omit critical information

# Problems with TextRunner

- **incoherent extractions**: relational phrase has no meaningful interpretation

| Sentence | Incoherent Relation |
|---|---|
| The guide *contains* dead links and *omits* sites. | contains omits |
| The Mark 14 *was central* to the *torpedo* scandal of the fleet. | was central torpedo |
| They *recalled* that Nungesser *began* his career as a precinct leader. | recalled began |

- **uninformative extractions**: omit critical information

| | |
|---|---|
| is | is an album by, is the author of, is a city in |
| has | has a population of, has a Ph.D. in, has a cameo in |
| made | made a deal with, made a promise to |
| took | took place in, took control over, took advantage of |
| gave | gave birth to, gave a talk at, gave new meaning to |
| got | got tickets to, got a deal on, got funding from |

# Problems with TextRunner

- **incoherent extractions**: relational phrase has no meaningful interpretation
- **uninformative extractions**: omit critical information

# Problems with TextRunner

- **incoherent extractions**: relational phrase has no meaningful interpretation
- **uninformative extractions**: omit critical information


ReVerb (Fader et al., 2011)

# ReVerb (Fader et al., 2011)

- find longest phrase matching a simple **syntactic constraint**

$$V \mid VP \mid VW^*P$$
$$V = \text{verb particle? adv?}$$
$$W = (\text{noun} \mid \text{adj} \mid \text{adv} \mid \text{pron} \mid \text{det})$$
$$P = (\text{prep} \mid \text{particle} \mid \text{inf. marker})$$

- to avoid overspecified relational phrases, a **lexical constraint** is introduced: $|\text{args(rel)}| > k$

# ReVerb (Fader et al., 2011)

- find longest phrase matching a simple **syntactic constraint**

$$
\boxed{\begin{array}{l}
\qquad\qquad V \mid VP \mid VW^*P \\
V = \text{verb particle? adv?} \\
W = (\text{noun} \mid \text{adj} \mid \text{adv} \mid \text{pron} \mid \text{det}) \\
P = (\text{prep} \mid \text{particle} \mid \text{inf. marker})
\end{array}}
$$

- to avoid overspecified relational phrases, a **lexical constraint** is introduced: $|\text{args(rel)}| > \text{k}$

# ReVerb (Fader et al., 2011)

- find longest phrase matching a simple **syntactic constraint**

$$\boxed{\begin{array}{l} \qquad\qquad V \mid VP \mid VW^{*}P \\ V = \text{verb particle? adv?} \\ W = (\text{noun} \mid \text{adj} \mid \text{adv} \mid \text{pron} \mid \text{det}) \\ P = (\text{prep} \mid \text{particle} \mid \text{inf. marker}) \end{array}}$$

- to avoid overspecified relational phrases, a **lexical constraint** is introduced: $|\text{args(rel)}| > \text{k}$

# ReVerb (Fader et al., 2011)

- find longest phrase matching a simple **syntactic constraint**

$$\boxed{\begin{array}{c} V \mid VP \mid VW^*P \\ \begin{array}{l} V = \text{verb particle? adv?} \\ W = (\text{noun} \mid \text{adj} \mid \text{adv} \mid \text{pron} \mid \text{det}) \\ P = (\text{prep} \mid \text{particle} \mid \text{inf. marker}) \end{array} \end{array}}$$

- to avoid overspecified relational phrases, a **lexical constraint** is introduced: $|\text{args(rel)}| > \text{k}$

**Example**

*The novelist Franz Kafka is the author of a short story entitled "The Metamorphosis".*

# ReVerb (Fader et al., 2011)

- find longest phrase matching a simple **syntactic constraint**

$$V \mid VP \mid VW^*P$$

$V$ = verb particle? adv?
$W$ = (noun | adj | adv | pron | det)
$P$ = (prep | particle | inf. marker)

- to avoid overspecified relational phrases, a **lexical constraint** is introduced: $|\text{args(rel)}| > \text{k}$

**Example**

*The novelist Franz Kafka is the author of a short story entitled "The Metamorphosis".*

---

⟨*The novelist Franz Kafka; is the author of; a short story*⟩

# Problem with ReVerb

limited to relations that are **mediated by verbs**

# Problem with ReVerb

limited to relations that are **mediated by verbs**

**Example**

*The novelist Franz Kafka* $\boxed{\textit{is the author of}}$ *a short story entitled "The Metamorphosis".*

# Problem with ReVerb

limited to relations that are **mediated by verbs**

**Example**

*The $\boxed{\textit{novelist}}$ Franz Kafka $\boxed{\textit{is the author of}}$ a short story $\boxed{\textit{entitled}}$ "The Metamorphosis".*

# Problem with ReVerb

limited to relations that are **mediated by verbs**

**Example**

*The* $\boxed{\textit{novelist}}$ *Franz Kafka* $\boxed{\textit{is the author of}}$ *a short story* $\boxed{\textit{entitled}}$ *"The Metamorphosis".*

# Problem with ReVerb

limited to relations that are **mediated by verbs**

**Example**

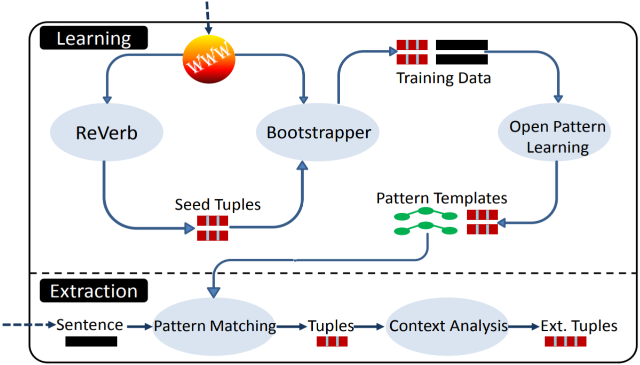
*The novelist Franz Kafka is the author of a short story entitled "The Metamorphosis".*


identification of relationships mediated by **nouns** and **adjectives**:

OLLIE (Mausam et al., 2012)

# OLLIE (Mausam et al., 2012)

- applies a set of high precision **seed tuples** from ReVerb
- to **bootstrap** a large training set
- over which it learns a set of extraction pattern templates using dependency parses



# OLLIE (Mausam et al., 2012)

- sample 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} |

- applied to individual sentences at extraction time

# OLLIE (Mausam et al., 2012)

**Example**

*The* $\boxed{\textit{novelist}}$ *Franz Kafka* $\boxed{\textit{is the author of}}$ *a short story* $\boxed{\textit{entitled}}$ *"The Metamorphosis".*

# OLLIE (Mausam et al., 2012)

**Example**

*The* $\boxed{\textit{novelist}}$ *Franz Kafka* $\boxed{\textit{is the author of}}$ *a short story* $\boxed{\textit{entitled}}$ *"The Metamorphosis".*

---

- ⟨*Franz Kafka; is the author of; a short story*⟩
- ⟨*Franz Kafka; is; a novelist*⟩
- ⟨*a short story; be entitled; "The Metamorphosis"*⟩

# Problem with OLLIE (and previous approaches)

limited to the extraction of **binary relations**

# Problem with OLLIE (and previous approaches)

limited to the extraction of **binary relations**

**Example**

*Franz Kafka was born in Prague in 1883.*

# Problem with OLLIE (and previous approaches)

limited to the extraction of **binary relations**

**Example**

*Franz Kafka was born into a Jewish family in Prague in 1883.*

---

ReVerb: ⟨*Franz Kafka; was born into; a Jewish family*⟩

# Problem with OLLIE (and previous approaches)

limited to the extraction of **binary relations**

**Example**

*Franz Kafka was born into a Jewish family* $\boxed{\textit{in Prague}}$ $\boxed{\textit{in 1883}}$.

---

ReVerb: ⟨*Franz Kafka; was born into; a Jewish family*⟩

# Problem with OLLIE (and previous approaches)

UNIVERSITÄT PASSAU

limited to the extraction of **binary relations**

**Example**

*Franz Kafka was born into a Jewish family* $\boxed{\textit{in Prague}}$ $\boxed{\textit{in 1883}}$.

---

REVERB: ⟨*Franz Kafka; was born into; a Jewish family*⟩


capture **complete facts** from sentences by gathering the full set of arguments for each relational phrase (**n-ary relations**):

- KRAKEN (Akbik and Löser, 2012)
- EXEMPLAR (Mesquita et al., 2013)

# KrakeN and Exemplar

- hand-crafted extraction rules over dependency parses

# KrakeN and Exemplar

- hand-crafted extraction rules over dependency parses

**Example**

*Franz Kafka was born* $\boxed{\textit{into a Jewish family}}$ $\boxed{\textit{in Prague}}$ $\boxed{\textit{in 1883}}$.

---

*⟨Franz Kafka; was born; (into) a Jewish family; (in) Prague; (in) 1883⟩*

Previous approaches often produce **erroneous extractions on syntactically complex sentences**

Previous approaches often produce **erroneous extractions on syntactically complex sentences**


generation of an **intermediate representation** using a sentence restructuring stage:

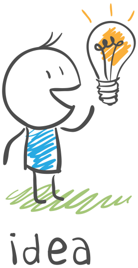
- ClausIE (Del Corro and Gemulla, 2013)
- Schmidek and Barbosa (2014)
- Stanford Open IE (Angeli et al., 2015)

# Problem with Paraphrase-based Approaches

**lack the expressiveness** needed for a proper interpretation of complex assertions taking into account the context under which a proposition is *complete* and *correct*

# Problem with Paraphrase-based Approaches

**lack the expressiveness** needed for a proper interpretation of complex assertions taking into account the context under which a proposition is *complete* and *correct*




systems that capture **inter-proposition relationships**

## OLLIE and its Successors Open IE 4 and 5

- OLLIE: extra field to distinguish between **information asserted** in a sentence and information that is only **hypothetical or conditionally true**
- Open IE 4 and 5: mark up **temporal and local context**

# OLLIE and its Successors Open IE 4 and 5

- OLLIE: extra field to distinguish between **information asserted** in a sentence and information that is only **hypothetical or conditionally true**

**Example**

*Romney will be elected President if he wins five key states.*

---

*(⟨Romney; will be elected; President⟩;*
CLAUSALMODIFIER *if; he wins five key states)*

- Open IE 4 and 5: mark up **temporal and local context**

# OLLIE and its Successors Open IE 4 and 5

- OLLIE: extra field to distinguish between **information asserted** in a sentence and information that is only **hypothetical or conditionally true**
- Open IE 4 and 5: mark up **temporal and local context**

# OLLIE and its Successors Open IE 4 and 5

- OLLIE: extra field to distinguish between **information asserted** in a sentence and information that is only **hypothetical or conditionally true**
- Open IE 4 and 5: mark up **temporal and local context**

**Example**

*Franz Kafka was born into a Jewish family* $\boxed{\textit{in Prague}}$ $\boxed{\textit{in 1883}}$.

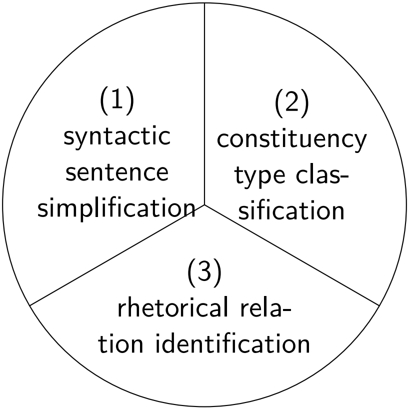
---

⟨*Franz Kafka; was born; into a Jewish family;* ***L:*** *in Prague;* ***T:*** *in 1883*⟩

# Graphene (Cetto et al., 2018)

**Lightweight semantic representation**:

- **two-layered hierarchy** of core relational tuples and accompanying contextual information that are
- **semantically linked** via rhetorical relations

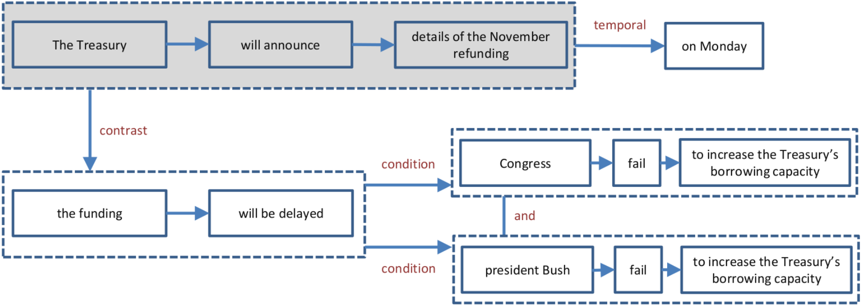


# Graphene (Cetto et al., 2018)

"*Although the Treasury will announce details of the November refunding on Monday, the funding will be delayed if Congress and President Bush fail to increase the Treasury's borrowing capacity.*"

# Graphene (Cetto et al., 2018)

"*Although the Treasury will announce details of the November refunding on Monday, the funding will be delayed if Congress and President Bush fail to increase the Treasury's borrowing capacity.*"



# Problems in the Evaluation of Open IE Systems

- no clear formal specification of what constitutes a **valid relational tuple**
- no established, large-scale annotated corpus serving as a **gold standard dataset**

# Problems in the Evaluation of Open IE Systems

- no clear formal specification of what constitutes a **valid relational tuple**
- no established, large-scale annotated corpus serving as a **gold standard dataset**

- **scalability** to large amounts of text?
- **portability** to various genres of text?
- objective and reproducible **cross-system comparison**?

# Open Research Questions

- large-scale gold standard **evaluation dataset** allowing for an objective and reproducible cross-system comparison
- applicability and transferability of the proposed Open IE approaches to **languages other than English**
- **canonicalization** of relational phrases and arguments