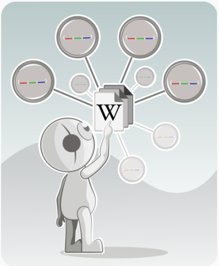
Compact Open Information Extraction on Large Corpora

Kiril Gashteovski

University of Mannheim
Data and Web Science Group



Motivation

Making the knowledge from unstructured text more accessible

Open Information Extraction

- Extract relations and their arguments from unstructured text in unsupervised manner

"AT&T, which is based in Dallas, is a telecommunication company."

("AT&T"; "is based in"; "Dallas")

("AT&T"; "is"; "telecommunication company")

- Common problem: relations/arguments can be overly specific

("~~The great~~ R. Feynman"; "worked ~~jointly~~ with"; "F. Dyson")

("Richard Feynman"; "worked with"; "Freeman Dyson")

MinIE - OIE system for minimizing and annotating facts

MinIE: Open Information Extraction System

https://github.com/uma-pi1/minie

MinIE: Minimize by Annotating and Structuring Information

- **Polarity:** is the triple positive (+) or negative (–)?
  - *("H. Clinton"; "is not president of"; "U.S.")* ⇒
    ("H. Clinton"; "is president of"; "U.S.") **(–)**
- **Modality:** is the triple a certainty (CT) or a possibility (PS)?
  - *("Bill Cosby"; "may go to"; "jail")* ⇒
    ("Bill Cosby"; "go to"; "jail") **(PS)**
- **Attribution:** consider the source of information
  - *("**D. T.**"; "said that"; "B. O. may have been born in Kenya")* ⇒
    ("B. Obama"; "have been born in"; "Kenya") **(+, PS)**
    Attribution: (Donald Trump, +, CT)
- **Quantities:** phrases expressing an amount of something
  - e.g. 9 cats, all cats, almost about 100 cats
    ⇒ *QUANT cats*

MinIE on Wikipedia: Example Triples

- Combinations of **factuality + frequency**
    - *("Barack Obama";"be";"president")* **(+, CT)**: 8,930
    - *("Barack Obama";"be";"president")* **(–, CT)**: 1
    - *("Barack Obama";"be";"president")* **(–, PS)**: 1
- Consider the source of information: **factuality + attribution**
    - *("Barack Obama"; "be born in"; "U.S.")* **(–, CT)**: 1
        - **attribution:** Orly Taitz **(+, CT)** ← conspiracy theorist
    - *("Barack Obama"; "be born in"; "U.S.")* **(+, CT)**: 1
        - **attribution:** Joshua A. Wisch **(+, CT)** ← asst. attorney gen.
- How reliable is the **attribution**?
    - E.g: what is attributed to **Donald Trump**?
    - *("Donald Trump"; "be"; "pro-choice")* **(+,CT)**: 1
    - *("Donald Trump"; "be"; "pro-life")* **(+,CT)**: 1
    - *("Barack Obama"; "be born in"; "Kenya")* **(+, PS)**: 1
    - *("Barack Obama"; "be born in"; "United States")* **(+, CT)**: 1

MinIE: Minimize by Dropping Overly Specific Words

("The great R. Feynman"; "worked jointly with"; "F. Dyson") ⇒
("Richard Feynman"; "worked with"; "Freeman Dyson")

- MinIE-S: drop words considered to be safe
  - determiners, adverbs modifying verbs, ...
  - e.g. *"the President"* ⇒ *"President"*
- MinIE-D: drop words modifying noun phrases
  - adjectives, adverbs, etc.
  - keep constituents that are found frequently in the corpus
  - additional fuel: enrich dictionary with domain knowledge
  - *"very long cold war"* ⇒ *"cold war"*
- MinIE-A: aggressive strategy of dropping of words
  - drop constituents (prepositional attachments, quantities, ...)
  - *(J. Cleese; starred as Lancelot in national tour of; Spamalot) ⇒ (John Cleese; starred in tour of; Spamalot)*

MinIE: Comparative Analysis

	MinIE-S	MinIE-D	MinIE-A
precision^{wiki}	0.73	0.72	0.60
precision^{nyt}	0.79	0.77	0.68
triple length ($\mu \pm \sigma$)	7.2 $\pm$ 4.2	6.9 $\pm$ 4.0	4.7 $\pm$ 1.9

Applications of MinIE: SalIE

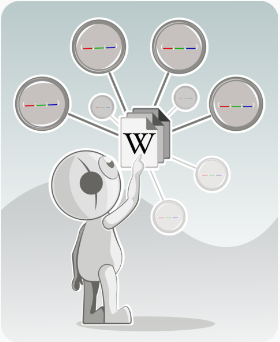
Fact saliance: *"generate machine-readable representation of the most prominent info. in text document as a set of facts"* (Ponza et al., 2018)

- Top-3 salient facts automatically extracted from a sample of two NYT documents
- **Human Summary:** *Body of Toni Grossi Abrams, widow and Staten Island socialite, is found in warehouse on outskirts of Panama City, Panama, where she had moved to begin career in real estate; Debra Ann Ridgley, one of her tenants, is charged with stabbing Abrams to death in her apartment on April 9.*
- **SalIE:**
    1. (Abrams, had been stabbed to death in, apartment)
    2. (Remains, were discovered beside warehouse at, edge of cinder-topped soccer field on outskirts of Panama City)
    3. (Apartment, tending wounds at time of, murder)

MinScIE: OIE w/h Semantic Information about Citations

- Current OIE systems perform significantly worse when applied to the science domain (Groth et al. 2018)
- MinScIE: OIE system based on MinIE
  - provides fixes for most of the issues identified by Groth et al. (2018)
  - designed to handle citations
  - semantically enrich triples when applied to scientific content
- Provides triples enriched with semantic information about citations
  - citation polarity
  - purpose
- Code on GitHub: `https://github.com/gkiril/MinSCIE`

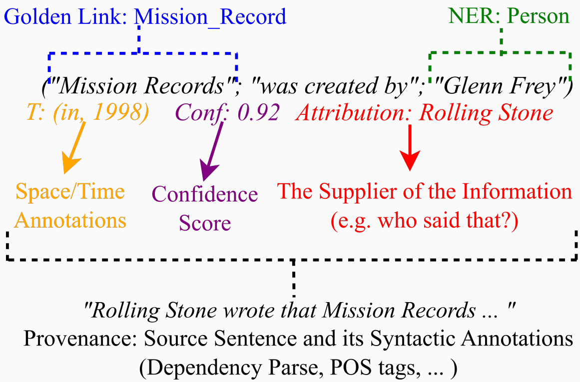
OPIEC: An Open Information Extraction Corpus



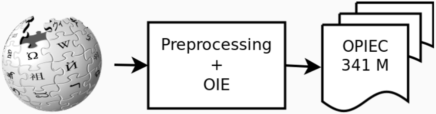
https://www.uni-mannheim.de/dws/research/resources/opiec/

OPIEC: An Open Information Extraction Corpus

- The **largest OIE corpus** to date (341M triples)
- **Rich with meta-data:** many syntactic/semantic annotations
- Ran MinIE-SpaTe on the **entire English Wikipedia**
  - the original **golden links** from a Wikipedia article are kept

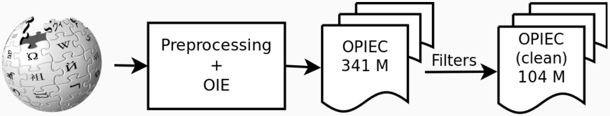


OPIEC: Underspecific Triples



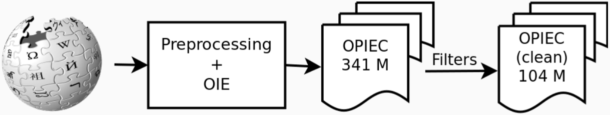
- Large portion of the triples are underspecific ($\sim$ 25%)
  ("he"; "founded"; "Microsoft")
  ("this"; "leads to"; "controversy")
- Entity mentions are broken up ($\sim$ 1%)
  ("Zip"; "Goes"; "a Million")

OPIEC-Clean



- Triple filters
  - entity mentions are broken up → e.g. *("Zip"; "Goes"; "a Million")*
  - triple has an empty object → e.g. *("Albert Einstein"; "died")*
  - are underspecific → e.g. *("He"; "co-founded"; "Microsoft")*
- Argument constraints
  - fully linked
  - recognized named entity (person, location, organization, ...)
  - matches a Wikipedia page title (e.g. *Super Bowl, biology, ...*)

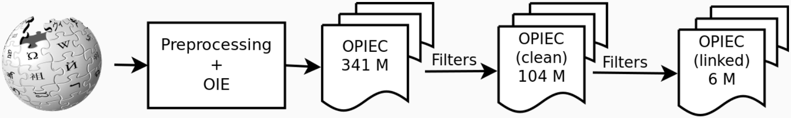
OPIEC-Clean



PERSON-PERSON		LOCATION-LOCATION		PERSON-LOCATION	
"have"	(130,019)	*"be in"*	(2,126,562)	*"be bear in"*	(203,091)
"marry"	(49,405)	*"have"*	(40,298)	*"die in"*	(37,952)
"be son of"	(40,265)	*"be village in adminis-trative district of"*	(9,130)	*"return to"*	(36,702)
"be daughter of"	(37,089)	*"be north of"*	(3,816)	*"move to"*	(36,072)
"be bear to"	(29,043)	*"be suburb of"*	(3,291)	*"be in"*	(25,847)
"be know as"	(25,607)	*"be west of"*	(3,238)	*"live in"*	(22,399)
"defeat"	(22,151)	*"be part of"*	(3,188)	*"grow up in"*	(17,571)

Table 1: Most frequent open relations between persons and locations

OPIEC-Linked



Triples containing only linked arguments

Example

("Michael Jordan"; "grew up in"; "Wilmington")

Analysis: OPIEC and KBs

- DBpedia and YAGO: KBs constructed from same resource as OPIEC (Wikipedia)
- KB hit: exploiting the distant supervision assumption
  - for each OIE triple: *("subject", "relation", "object")*
  - search for any KB triple (`db:subject, db:rel, db:object`) or (`db:object, db:rel, db:subject`)
  - ∼30% of OPIEC-Linked has KB hits in either DBpedia or YAGO
- Relation alignment: *relation* is aligned with `db:rel`
- Example:
  ("Kachin Independence Army"; "has headquarters near"; "Laiza") ⇔ (dbr:Kachin_Independence_Army; dbp:headquarters; dbr:Laiza)

Analysis: OPIEC and KBs

- No 1:1 correspondence between open relations and KB relations
- Open relations can be more specific than KB relations
- Although "semantically correlated", open relations may be *semantically different* than their KB rel. counterparts

location		associatedMusicalArtist		spouse	
"be in"	(35,754)	*"be"*	(5,521)	*"be wife of"*	(1,580)
"have"	(2,500)	*"have"*	(3,248)	*"be"*	(980)
"be"	(1,473)	*"be guitarist of"*	(619)	*"marry"*	(551)
"be at"	(793)	*"be drummer of"*	(433)	*"be widow of"*	(392)
"be of"	(525)	*"be feature"*	(377)	*"be marry to"*	(246)
"be historic home located at"	(520)	*"be frontman of"*	(367)	*"have"*	(244)

Table 2: Most frequent open relations aligned to DBpedia relations

OPIEC-Linked: Alignment with KBs

- We selected top-100 most freq open rels (38% of OPIEC-Clean)
  - the fraction of KB hits is low (averaging at 14.7%)
  - on average, there are about 40 KB relations per open relation

Open relation	Frequency in OPIEC-Link	# KB hits	# distinct KB rel.s	Top-3 aligned DBpedia rel. and hit frequency	
"be"	1,475,332	162,438 (11.0%)	403	type	72,542
				occupation	12,254
				isPartOf	6,776
"have"	216,332	118,625 (54.8%)	320	author	11,678
				director	9,429
				writer	8,751
"be in"	1,150,667	734,330 (63.8%)	221	country	269,189
				isPartOf	200,851
				state	65,242
"include"	14,746	1,364 (9.2%)	127	type	376
				associatedBand	75
				associatedMusicalArtist	75
"be bear in"	7,138	1,478 (20.7%)	31	birthPlace	1,172
				isPartOf	68
				deathPlace	60

Take-aways

- Motivation: make natural language text data **more structured** and useful
- MinIE: OIE System for compact extractions
  - Improve content by adding **semantic annotations** on extracted triples (polarity, modality, attribution, quantities)
  - Minimize relations and arguments by **dropping overly specific words**
  - Different **levels of minimization**: complete, safe, dictionary and aggressive
- OPIEC: the **largest OIE corpus** to date
  - Aims to spur research in **AKBC, open Q&A**, ...
  - **Rich with meta-data:** many syntactic/semantic annotations
  - Multiple **sub-corpora** from noisy to clean
  - Analyzed and compared with **Wikipedia-based KBs**

Thank you for your attention!

Time: temporal information about facts

Fixed time points

- Explicit date
  - ("J.F.K."; "was assassinated by L.H.O. on"; "22.11.1963")⇒ ("John F. Kennedy"; "was assassinated by"; "Lee H. Oswald") **Time:** 22.11.1963
- Textual temporal expression
  - ("John F. Kennedy"; "was assassinated by"; "Lee H. Oswald") **Time:** "yesterday"; "many years ago", etc.
- Discretized temporal references (past, present, future)
  - *"In times past, Donald Trump was a Democrat."* ⇒
  - ("Donald Trump"; "was Democrat in"; "times past") ⇒
  - ("Donald Trump"; "was"; "Democrat") **Time ref:** past

Time: temporal information about the arguments

Arguments can contain temporal information of their own

"Isabella II opened the 17th-century Parque del Retiro in 1868."

↓

("Isabella II"; "opened 17th-century Parque del Retiro in"; "1868")

↓

("Isabella II"; "opened"; "17th-century Parque del Retiro")

Time: 1868

↓

("Isabella II"; "opened"; "Parque del Retiro")

Time: 1868

↓

17th-century