ifis

Institut für Informationssysteme
Technische Universität Braunschweig

Open Information Extraction in Digital Libraries: Current Challenges and Open Research Questions

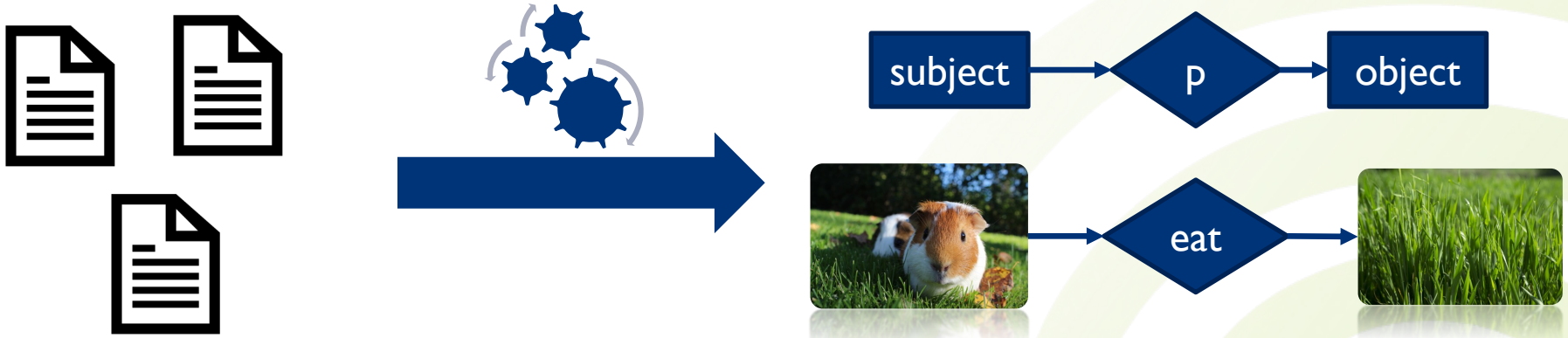
DISCO@JCDL2021

Hermann Kroll, Judy Al-Chaar and Wolf-Tilo Balke

Institut für Informationssysteme
Technische Universität Braunschweig

Information Extraction

- Transform **unstructured** into **structured information**
  - E.g., extract statements from natural language texts

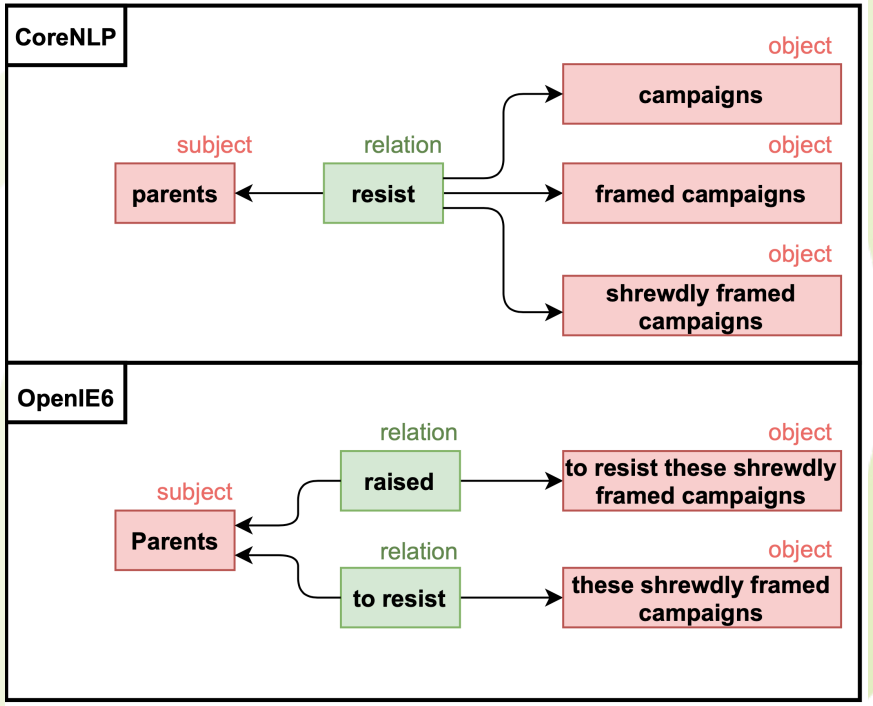


- Information extraction can be done by:
  - **Supervised** methods (relation extraction)
  - **Unsupervised** methods (open information extraction)

Open Information Extraction

- **Open information extraction** extracts statements from texts **without knowing** entities and relations a-priori
  - We understood OpenIE as unsupervised extraction method
  - OpenIE systems may build upon:
    - Rule-based architectures
    - Neural architectures

- No pre-known relations, works out-of-the-box
  - sounds great, but…



Research Questions

- OpenIE systems are usually **precision-oriented**
  - Claimed in the literature
  - Reflected in our personal experiences
- OpenIE extractions are **not canonicalized**
  - Several noun phrases might refer to the same concept, e.g., NY, New York, New York City
  - Several verb phrases might describe the same relation, e.g., has birthplace, is born, etc.
- OpenIE might extract **complex arguments**
  - A complex argument involves multiple concepts, e.g., (Einstein, won, the Nobel Prize in 1921)

Qualitative Evaluation

- We analyzed two OpenIE systems in two domains:
  - Stanford CoreNLP and OpenIE 6
  - 10 news articles from the New York Times
  - 17 biomedical articles from PubMed

- We asked the following questions:
  - How **precise** will OpenIE systems keep the original information?
    - Fully retained, partially retained, or not retained
  - How does OpenIE react to **different kinds of sentences**?
    - Simple, compound, complex, nested and negated sentences
  - How **complex** will OpenIE **arguments** be?
    - Count how many arguments are complex and how many are simple

Observations (I/II)

- ”India has about 10 million coronavirus cases now, and schools have been offering online instruction since March.”
  - Both: (India; has; about 10 million coronavirus cases now)
  - OpenIE 6: (schools; have been offering; online instruction since March.)

- “Relentless advertising campaigns are telling Indian parents that coding is critical because making children code will develop their cognitive skills.”
  - OpenIE 6: (Relentless advertising campaigns; are telling; Indian parents that coding is critical because making children code will develop their cognitive skills)

https://www.nytimes.com/2021/01/02/opinion/teaching-coding-schools-india.html

Observations (II/II)

- "As a result, many marine species are impeccably adapted to detect and communicate with sound."
  - Both: (many marine species; are impeccably adapted; to communicate with sound)
  - OpenIE 6: (many marine species; are impeccably adapted; to detect with sound)

https://www.nytimes.com/2021/02/04/science/ocean-marine-noise-pollution.html

- "Recent studies show that man was not always the hunter."
  - CoreNLP: (Recent studies; show; man)
  - OpenIE 6: (Recent studies; show; that man was not always the hunter)
  - OpenIE 6: (man; was not; always the hunter)

https://www.nytimes.com/2021/01/01/opinion/women-hunter-leader.html

Evaluation Results (I/II)

- How **precise** will OpenIE systems keep the original information?
  - Fully retained, partially retained, or not retained
- How does OpenIE react to **different kinds** of **sentences**?
  - Simple, compound, complex, nested and negated sentences

Corpus	Sent. Category	#Sent.	CoreNLP			OpenIE6		
			Full	Partial	Not	Full	Partial	Not
NY Times	Simple	20	62%	19%	19%	**100%**	0%	0%
	Compound	20	24%	41%	35%	**81%**	19%	0%
	Complex	20	15%	53%	32%	**78%**	18%	4%
	Nested	20	4%	54%	42%	**80%**	18%	2%
	Negation	20	5%	5%	90%	**73%**	10%	17%
PubMed	Simple	20	52%	38%	10%	**100%**	0%	0%
	Compound	20	15%	44%	41%	**76%**	14%	10%
	Complex	20	38%	48%	14%	**56%**	13%	31%
	Nested	20	22%	63%	15%	**89%**	11%	0%
	Negation	20	5%	33%	62%	**81%**	15%	4%

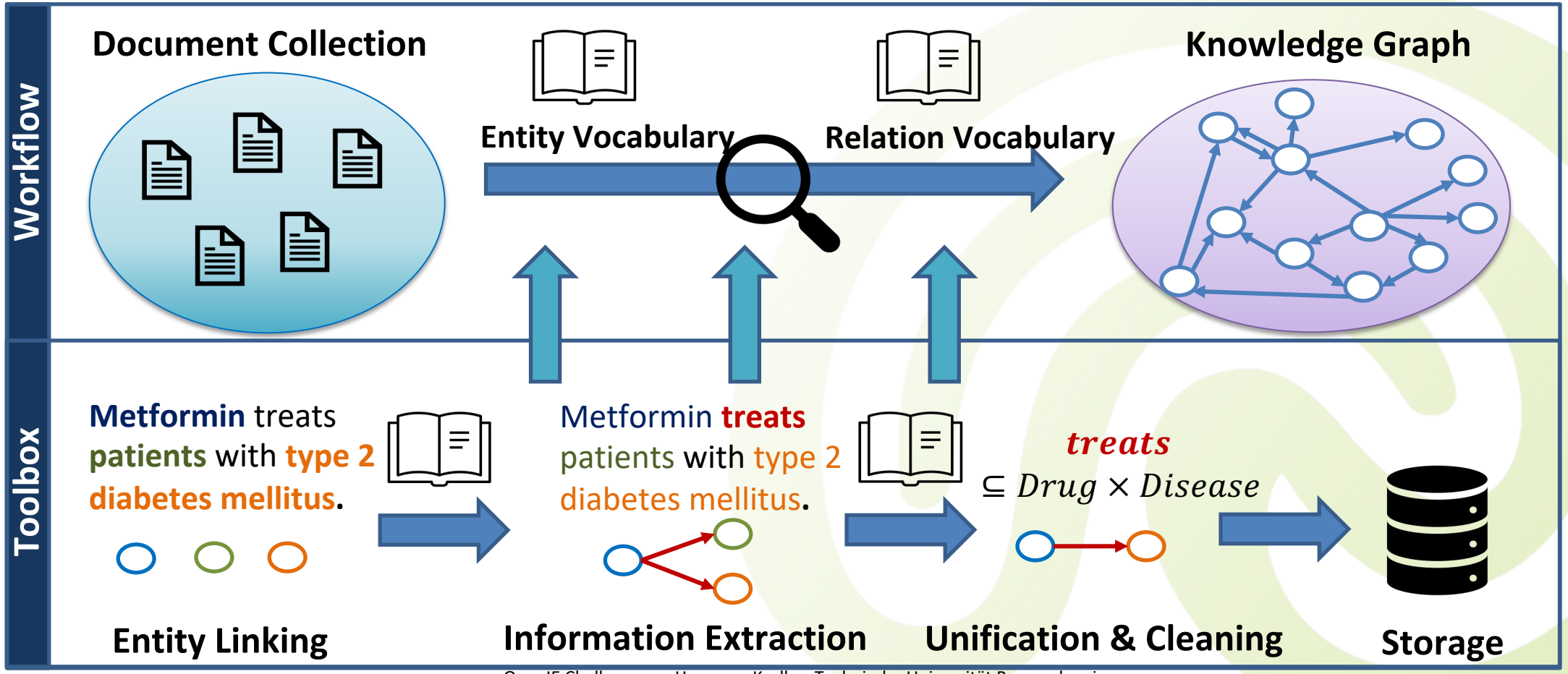
Evaluation Results (II/II)

- How **complex** will OpenIE **arguments** be?
  - Count how many arguments are complex and how many are simple
  - A complex argument is an argument that involves multiple concepts, e.g., (Nobel Prize in 1921) contains a prize and a date information

Corpus	Argument Type	CoreNLP		OpenIE6	
		Single	Complex	Single	Complex
NY Times	Subject	98%	2%	89%	11%
	Object	80%	20%	32%	68%
PubMed	Subject	99%	1%	76%	24%
	Object	75%	25%	47%	53%

JCDL2021 Toolbox

- A Toolbox for the Nearly-Unsupervised Construction of DL Knowledge Graphs:
  - https://github.com/HermannKroll/KGExtractionToolbox
  - Shared as **Open Source**, written in **Python** and published with an **MIT license**



Conclusion

- We believe that OpenIE helps to bring more structure into otherwise unstructured collections
  - Supervised methods would require cost-intensive training data

- RQ1: **Trade-Off: Accuracy vs. Runtime?**
  - Achieving the best precision requires to use the latest neural extraction architectures
  - Additional filtering & cleaning will take even more time
- RQ2: How can **complex arguments** be **handled**?
  - E.g., (Einstein, won, the Nobel Prize in 1921)
- RQ3: How can **OpenIE outputs** be **canonicalized**?
  - Linking synonymous noun and verb phrases to precise concepts / relations

Thank You!

FACHINFORMATIONSDIENST
PHARMAZIE
TU Braunschweig

Technische Universität Braunschweig

If you have any questions, contact me via:

kroll@ifis.cs.tu-bs.de

@HermannKroll