MANCHESTER 1824

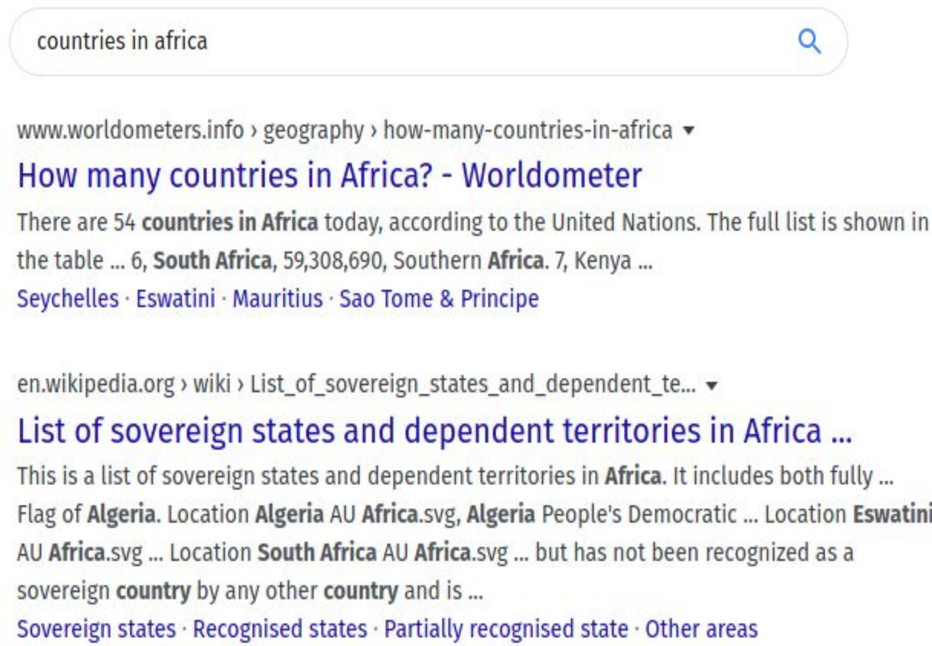
The University of Manchester

Application of the classical NLP pipeline: Open Information Extraction

Viktor Schlegel

Why OpenIE?

- Most information is expressed in textual form
- How do we access it (at scale)?

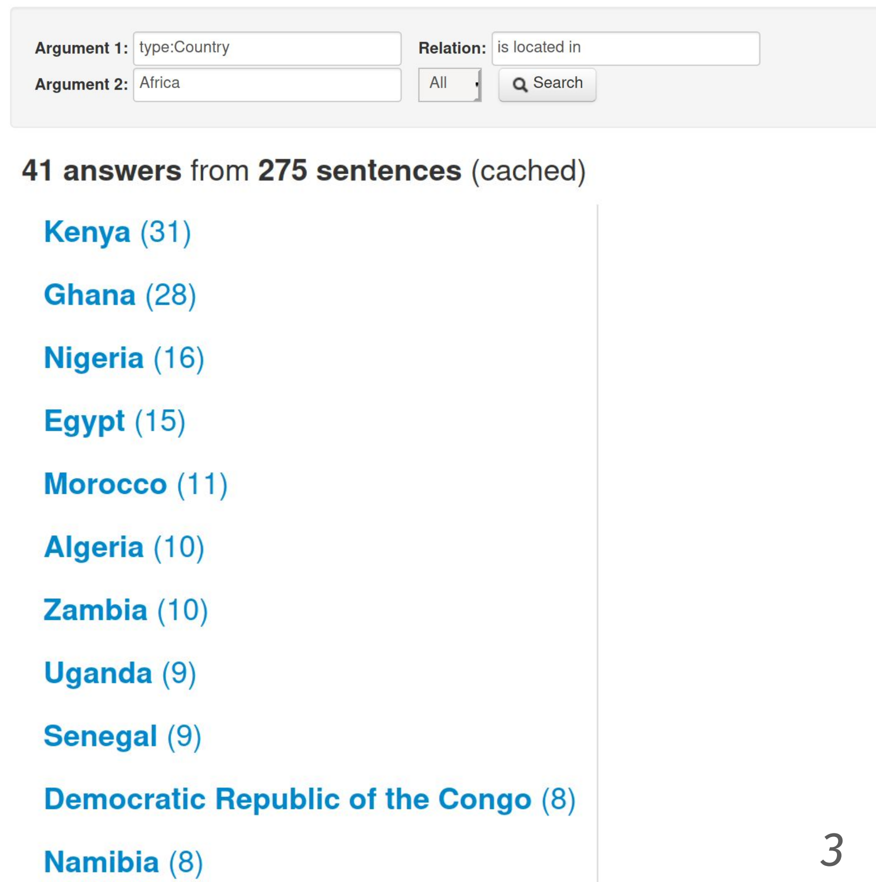


Oren Etzioni. 2011. Search needs a shake-up. Nature

MANCHESTER 1824
The University of Manchester

Why OpenIE?

- Ultimate goal: structured, machine processable representation of knowledge



Open Information Extraction

"Domain-independent discovery of relations extracted from text and readily scale to the diversity and size of the Web corpus."

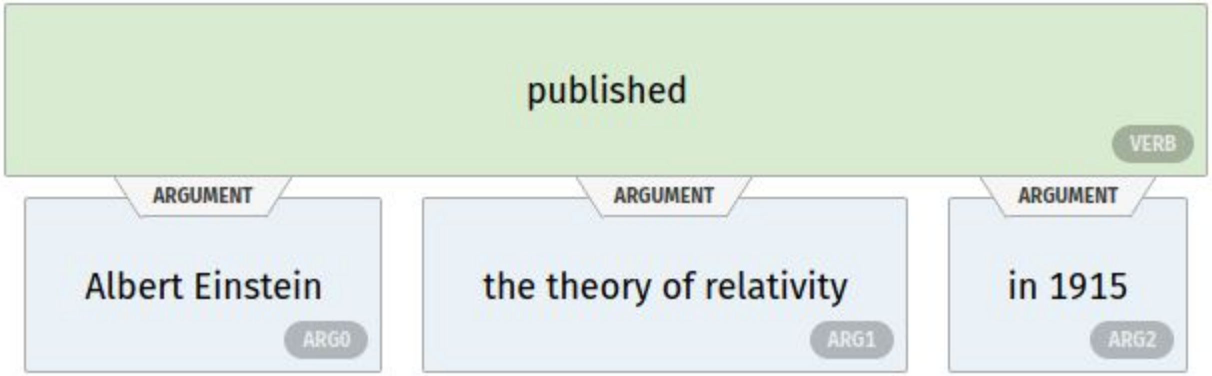
- Input: a corpus of documents

- Output: a set of extracted relations

Banko, M. et al. 2007. Open Information Extraction from the Web. IJCAI

Example

- "Albert Einstein, a German theoretical physicist, published the theory of relativity in 1915."



The University of Manchester

Like RE but different

- Goal is to find any relation in text data
  - vs relation extraction, where we know what relations we're looking for beforehand

- Applicable to a lot of (heterogeneous) documents
  - Cannot resort to specific domain knowledge
  - Cannot wait ages for a single extraction

The University of Manchester

Resulting requirements

- Automated
- Domain-agnostic
- Scalable and efficient

"Open domain"

How to solve it?

- Statistical and linguistic analysis and patterns (TextRunner, ReVerb, OLLIE)

- Clause based (Stanford OpenIE)

- Deep Learning based (Supervised Open Information Extraction)

The University of Manchester

TextRunner: Idea

- Based on rules, generate triples from dependency parses
- But: dependency parsing computationally expensive (at least back then)
- So: train Naive Bayes classifier on triples
- Apply classifier on the web

Banko, M. et al. 2007. Open Information Extraction from the Web. IJCAI

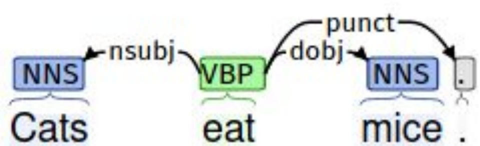
- There exists a dependency chain between e_i and e_j that is no longer than a certain length.
- The path from e_i to e_j along the syntax tree does not cross a sentence-like boundary (e.g. relative clauses).
- Neither e_i nor e_j consist solely of a pronoun.

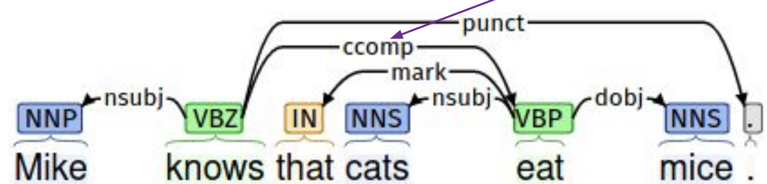
etc...

TextRunner: Dataset

NP

- Identify noun phrases (e_i, e_j)
- Use syntactic rules to generate triples $(e_i, r_{i,j}, e_j)$ from dependency parses





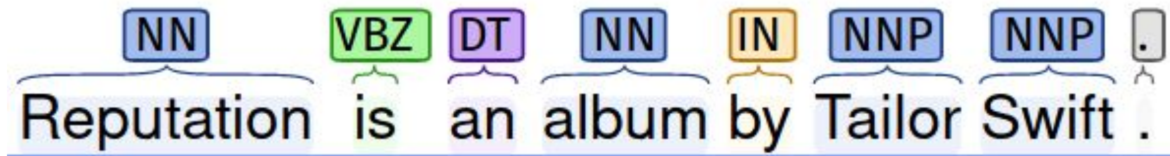
(cats, eat, mice) ✓

(cats, eat, mice) ✓
(mike, knows eat, cats) ✗
(mike, knows eat, mice) ✗

MANCHESTER
1824
The University of Manchester

TextRunner: Classifier

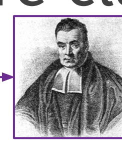
- Train Naive Bayes classifier to distinguish between positive and negative tuples *(cats - eat - mice)* ✓ *(mike - knows, eat - cats)* ✗
- Classifier is not using dependency features as input -> no dependency parser needed at application time

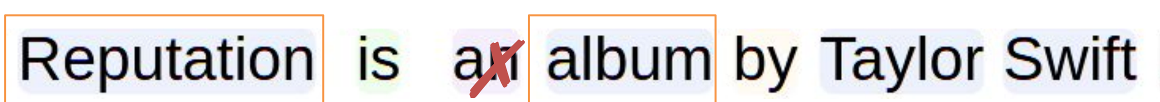


TextRunner: Application

- For an input sentence:
  - Identify Noun Phrases (Based on POS tags)
  - For words between each pair of noun phrases:
    - Remove ‘over-specific’ words
      (e.g. prepositional phrases, adverbs)
    - Classify remaining words with NB classifier
  - Return triples that are classified positively

(Reputation, is, album)





MANCHESTER
1824
The University of Manchester

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

ReVerb: Empirical analysis

- Textrunner's extractions uninformative and incoherent
- Extractions too specific

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

```
The Obama administration is offering only
modest greenhouse gas reduction targets at
the conference.
```

(Obama administration, offering only modest greenhouse gas reductions targets at, conference)

Fader, A. et al 2011. Identifying Relations for Open Information Extraction. EMNLP

The University of Manchester

Binary Verbal Relation Phrases	
85%	Satisfy Constraints
8%	Non-Contiguous Phrase Structure
	Coordination: X is produced and maintained by Y
	Multiple Args: X was founded in 1995 by Y
	Phrasal Verbs: X turned Y off
4%	Relation Phrase Not Between Arguments
	Intro. Phrases: Discovered by Y, X ...
	Relative Clauses: ...the Y that X discovered
3%	Do Not Match POS Pattern
	Interrupting Modifiers: X has a lot of faith in Y
	Infinitives: X to attack Y

ReVerb: Remedy 1

- Introduce a syntactic constraint for relations (based on POS tags)

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

(*, got in a, *) ✓
(*, gave a talk at, *) ✓
(*, was central torpedo, *) ✗

The University of Manchester

ReVerb: Remedy 2

- To avoid overly specific relations: introduce a "lexical constraint"
- What that means: Only keep relations that are "common enough", i.e. observed with at least *k* different arguments in the data (such as *k*=20)

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

Weight	Feature
1.16	(x, r, y) covers all words in s
0.50	The last preposition in r is *for*
0.49	The last preposition in r is *on*
0.46	The last preposition in r is *of*
0.43	$len(s) \leq 10$ words
0.43	There is a WH-word to the left of r
0.42	r matches VW*P from Figure 1
0.39	The last preposition in r is *to*
0.25	The last preposition in r is *in*
0.23	10 words $< len(s) \leq$ 20 words
0.21	s begins with x
0.16	y is a proper noun
0.01	x is a proper noun
-0.30	There is an NP to the left of x in s
-0.43	20 words $< len(s)$
-0.61	r matches V from Figure 1
-0.65	There is a preposition to the left of x in s
-0.81	There is an NP to the right of y in s
-0.93	Coord. conjunction to the left of r in s

ReVerb: Application

- For an input sentence:
  - For each verb in sentence:
    - Find the longest candidate word sequence *r* satisfying both constraints
  - Merge adjacent candidates
  - For each relation candidate *r*:
    - Find nearest Noun Phrase left and right of *r*
    - Assign confidence score with a classifier

(Reputation, is an album by, Taylor Swift)