

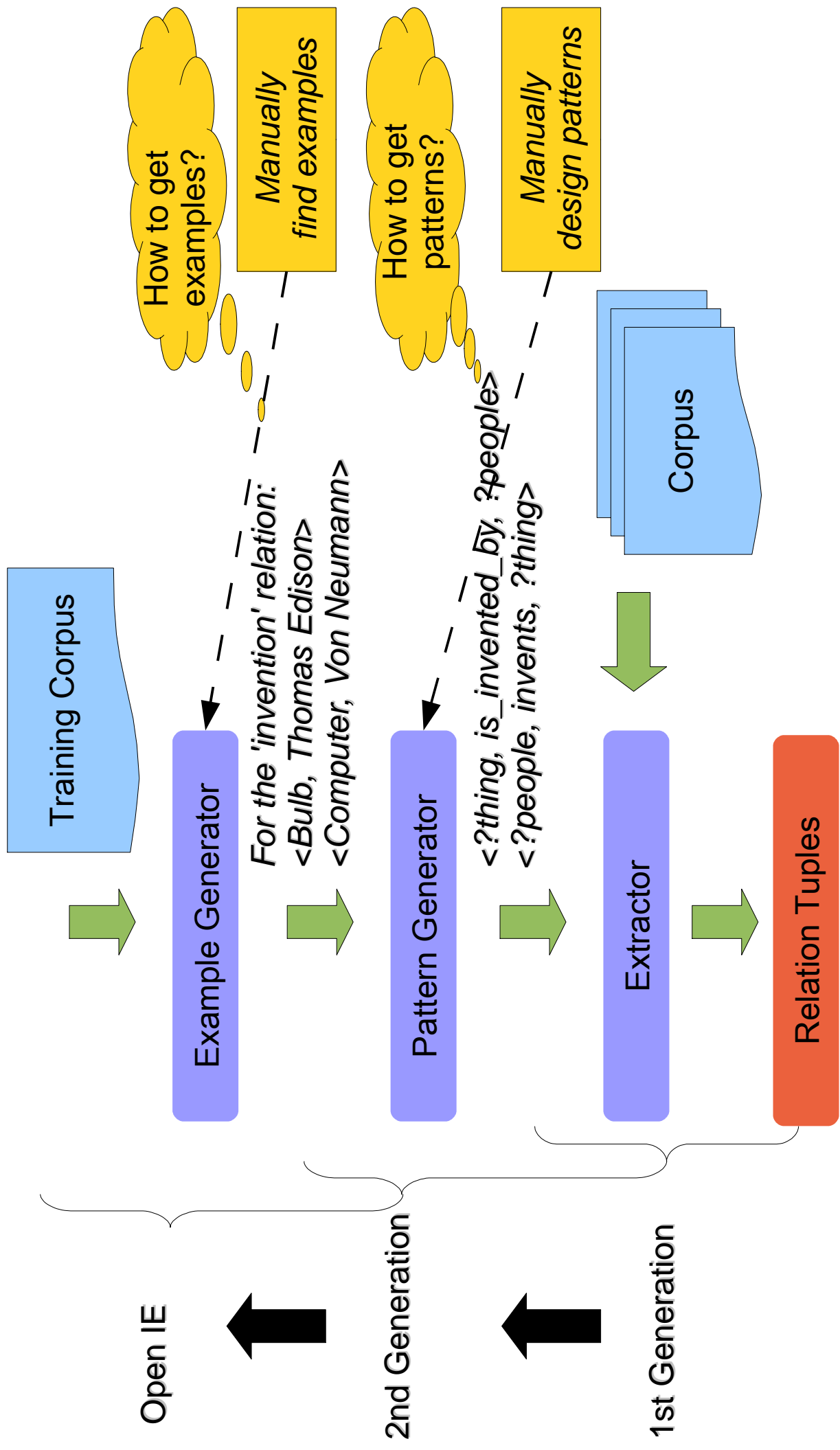
Open Information Extraction

Ning YAN
Xiaonan LI

Outline

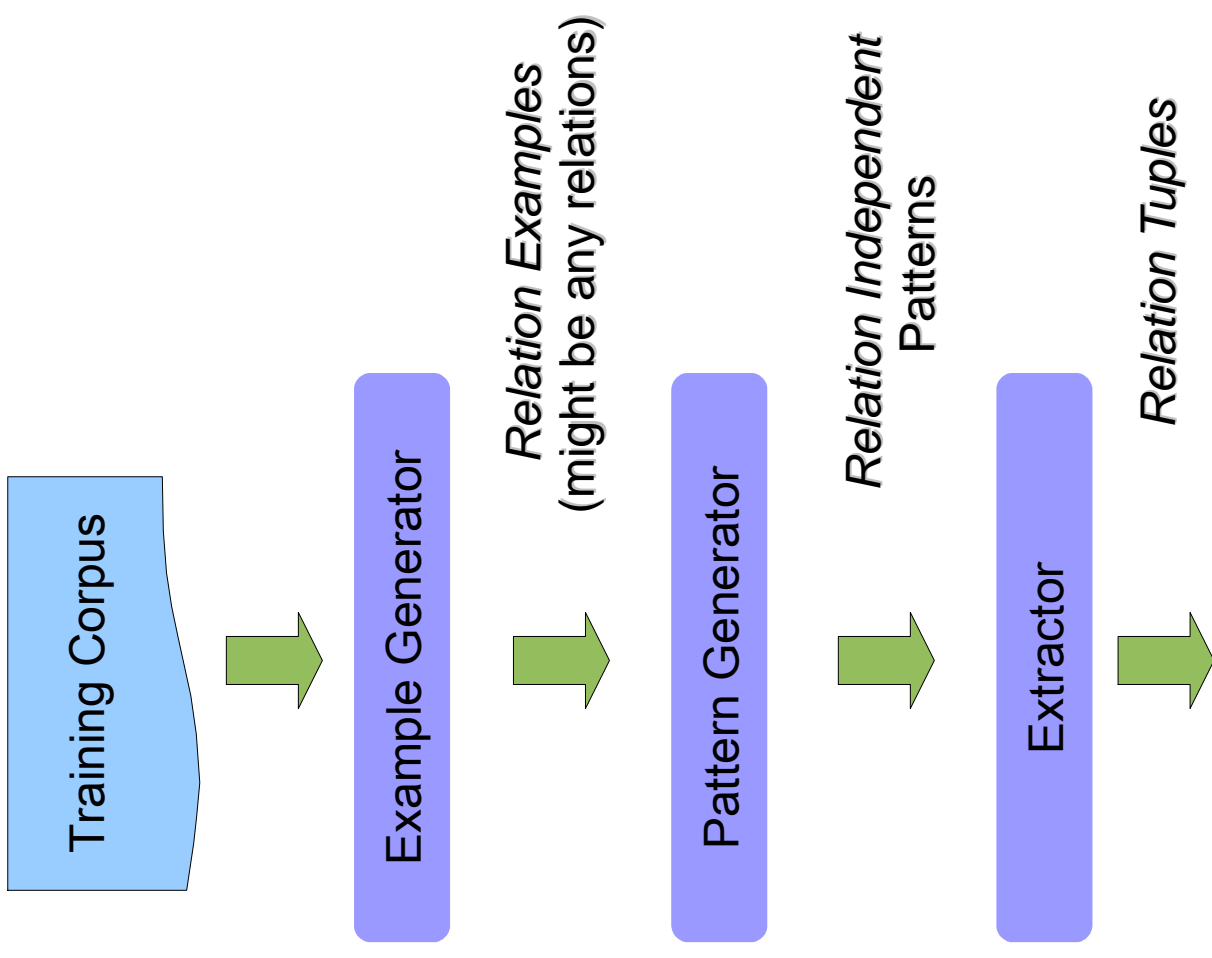
- Traditional IE → Open IE (OIE)
- Is OIE Applicable?
- TextRunner – The First OIE System
- O-CRF – A Better OIE System
- Summary

Traditional IE → Open IE



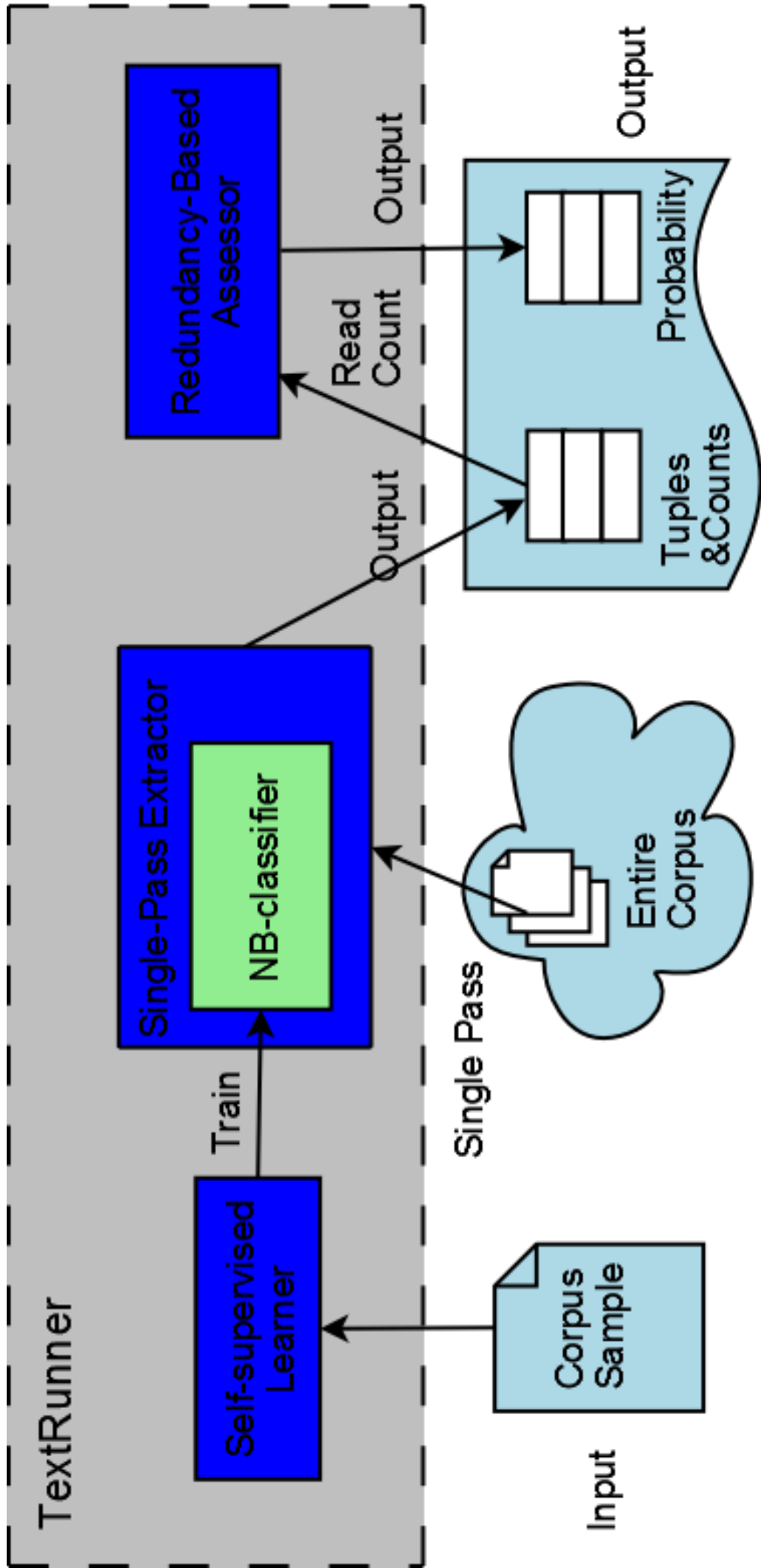
Is OIE Applicable?

- Input
 - Corpus
- Output
 - Tuples of any relations
- Challenges
 - Relations unknown
 - Entities unknown



TextRunner - The First OIE System

- An overview of 3 component

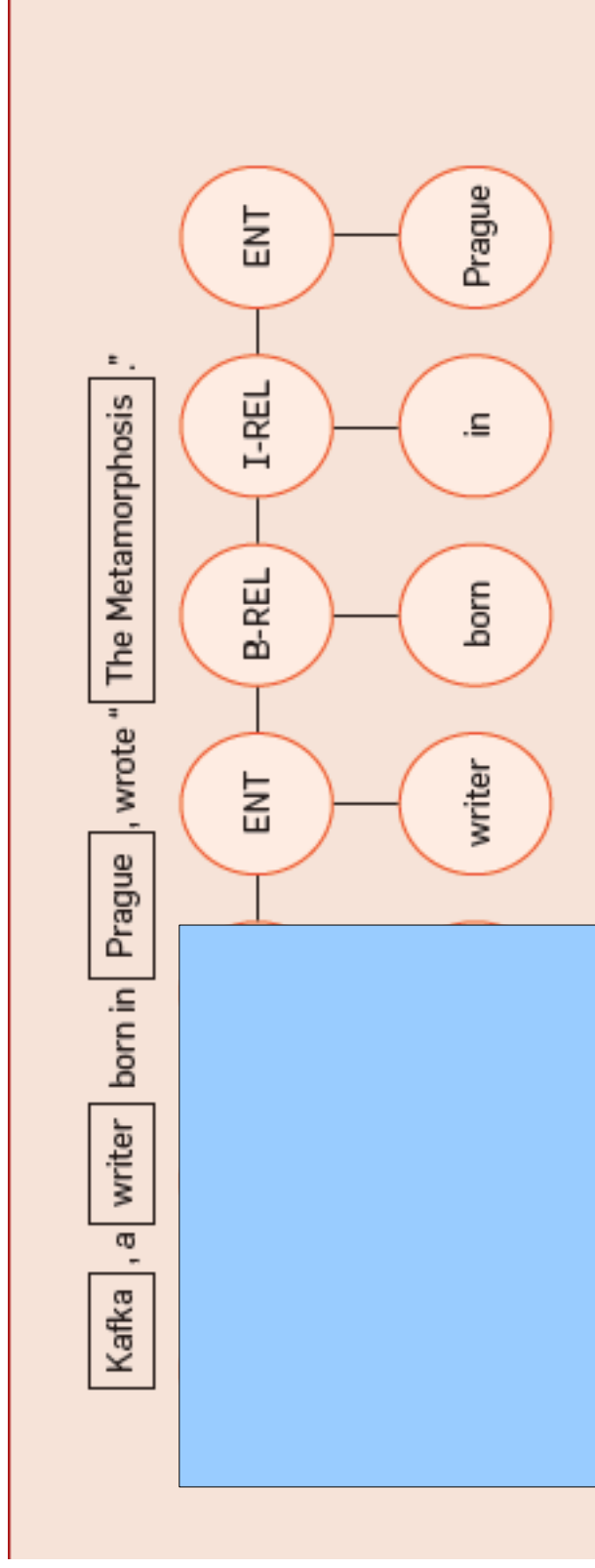


TextRunner - The First OIE System

- Self-Supervised Learner
 - Purpose: To train a **Naive Bayes classifier**
 - NB-Classifer(features-of-sentence) →
relation-pattern
 - How it works:
 - Tag a given sentence with POS tags
 - Using **linguistic parser** to **parse** relations
[Klein and Manning, 2003]
 - Successfully parsed <sentence, relation-pattern>
pairs are labeled as positive example.
 - Compute features-of-sentence
 - Train the classifier based on
 - <features-of-sentence, relation-pattern>

TextRunner - The First OIE System

- Self-Supervised Learner - Example



- POS tag (noun, verb, prep, noun)
- Parsed relation-pattern
 - (ENT, B-REL, I-REL, ENT)

TextRunner - The First OIE System

- Self-Supervised Learner - Example
 - Several features are selected
 - POS tag sequence
 - Number of tokens between the two ENT
 - Stop words between ... , etc.
 - as the training input of Naive Bayes classifier
- Example features-of-sentence:
 - <POS-sequence, #tokens, #stop>
- Label <(noun, verb, prep, noun), 2, 1>

TextRunner - The First OIE System

- Single-Pass Extractor
 - Tag given sentence with a POS tagger
 - Identify noun phrases with a **noun phrases chunker** [Ratnaparkhi, 1998])
 - Noun phrases are treated as entities
 - Compute features-of-sentence
 - Determine relation with **Naive Bayes classifier**
 - Relations are found examining the text between the noun phrases, and eliminating non-essential phrases (prep. phrases like 'from XXXX')

TextRunner - The First OIE System

- Single-Pass Extractor - Example
 - What are features-of-sentence?
 - “Jerry Yang graduates from Stanford”
 - POS tag: (noun, noun, verb, prep, noun)
 - Chunker: ENT=”Jerry Yang”, ENT=”Stanford”
 - features-of-sentence
 - =<(noun, noun, verb, prep, noun), 5, 1>
 - NB-Classifier(features-of-sentence)
 - =(ENT, B-REL, I-REL, ENT) -> '+'
 - Relation identified

TextRunner - The First OIE System

- Redundancy-Based Assessor
 - For each noun phrase, the chunker above will provide a **probability** associated with each words
 - Identical Tuples are merged and counts from distinct sentences where they are found
 - Assessor will assign a **probability** to each of the identical tuples using probabilistic model PMI-based (Pointwise mutual information)
[KnowItAll], [Downey, 2005]

TextRunner - The First OIE System

- Redundancy-Based Assessor - Example
 - Suppose: “Writer born in Prague XXX”
- Chunker probability: $P(\text{XXX})=0.1$ ->discard XXX
- “Writer born in Prague”
- “Writer was born in Prague” -> “born”
- “Writer to be born in Prague”
- Merge the 3 sentences to a normalized form, and count is 3, assign proper probability for estimation of correctness

TextRunner - The First OIE System

- The Demo



TextRunner Search

Retrieved 2849 results for kills in the predicate and bacteria in argument 2.

Grouping results by predicate. Group by: [argument_1](#) | [argument_2](#)

kills - 31 results

the new antibiotics (69), Benzoyl peroxide (47), Chlorine (36), [119 more...](#) **kills** bacteria

<https://curingc.cs.washington.edu:7125 - TextRunner Search Result>

Heat (27) **kills** the beneficial bacteria

Amoxicillin (26) **kills** bacteria

ozone (24) **kills** any bacteria

Penicillin (23) **kills** the pneumococcal bacteria

Oxygen (16) **kills** anaerobic bacteria

Honey (12) **kills** the bacteria

Cooking (12) **kills** these bacteria

The process (11) **kills** other bacteria

Irradiation (11) **kills** most harmful bacteria

Garlic (9) **kills** the bad bacteria

Alcohol (9) **kills** the bacteria

Search again:

Argument 1

Predicate

kills

Argument 2

bacteria

Search

examples of "bacteria":

e. coli (13)


salmonella (12)

Jump to:

[kills \(31\)](#)

[will kill \(7\)](#)

O-CRF – A Better OIE System

- The system
 - Based on TextRunner
 - Substitute CRF classifiers for NB classifier in TextRunner  CRF: Conditional Random Field
 - Achieve better precision and recall

Category	O-CRF			O-NB		
	P	R	F1	P	R	F1
Verb	93.9	65.1	76.9	100	38.6	55.7
Noun+Prep	89.1	36.0	51.3	100	9.7	55.7
Verb+Prep	95.2	50.0	65.6	95.2	25.3	40.0
Infinitive	95.7	46.8	62.9	100	25.5	40.6
Other	0	0	0	0	0	0
All	88.3	45.2	59.8	86.6	23.2	36.6

$$precision = \frac{\text{correctly extracted tuples}}{\text{all extracted tuples}}$$

$$recall = \frac{\text{correctly extracted tuples}}{\text{all relations expressed}}$$

Summary

- Open Information Extraction – open ended method scales to entire Web
 - Unlimited types of relations
 - Extraction in single pass
- Future work
 - Integrate with inference will enable it to reason based on information extracted
 - Unify Open IE with information from WordNet, Cyc, OpenMind, Freebase ...

**What kind of relations cannot be
extract by OIE?**