# Supervised Learning based

- Why no end2end-learning based approaches?
  - Requires annotated data, no big enough dataset
  - Dataset "hard" to create
  - Solution: Transform existing SRL dataset to OIE dataset (Who did what to whom, when, where, why)
  - Transform task of OIE to sequence labelling

(more in week 4)

Stanovsky, G et al   2018.   Supervised Open Information Extraction. NAACL

The University of Manchester

Define as sequence labeling task

- For an input sentence
  - For each verb
    - Expand predicate (P) (rule based)
    - For each word:
   - label as Argument (ARG_i), or non-participating (O)
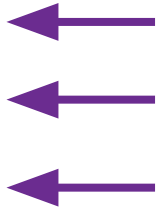
(a) The *president* *claimed* that he won *the* *majority* *vote*.

Create a supervision dataset

- OpenIE task bears resemblance to Semantic role Labelling
  (Who did what to whom/where/when)
- Automatically convert an existing high quality SRL dataset to an OpenIE dataset

Create a supervision dataset

Mercury filling, particularly prevalent in the USA, was banned in the EU, partly because it causes antibiotic resistance. *Text*

Predicate	QA-SRL	QAMR *Question*	Open IE
made	-	What is the **filling made of**? mercury	-
prevalent	-	What was **particularly prevalent in the USA**? mercury filling	(mercury filling; **particularly prevalent**; in the USA) *Transformed Extraction*
banned	What was **banned**? mercury filling Where was something **banned**? the EU Why was something **banned**? partly because it causes antibiotic resistance	What was **banned in the EU partly because it causes antibiotic resistance**? mercury filling	(mercury filling; **was banned**; in the EU; partly because it causes antibiotic resistance)
causes	What **caused** something? mercury filling What did something **cause**? antibiotic resistance	What did **mercury filling cause**? antibiotic resistance	(mercury filling; **caused**; antibiotic resistance)

4
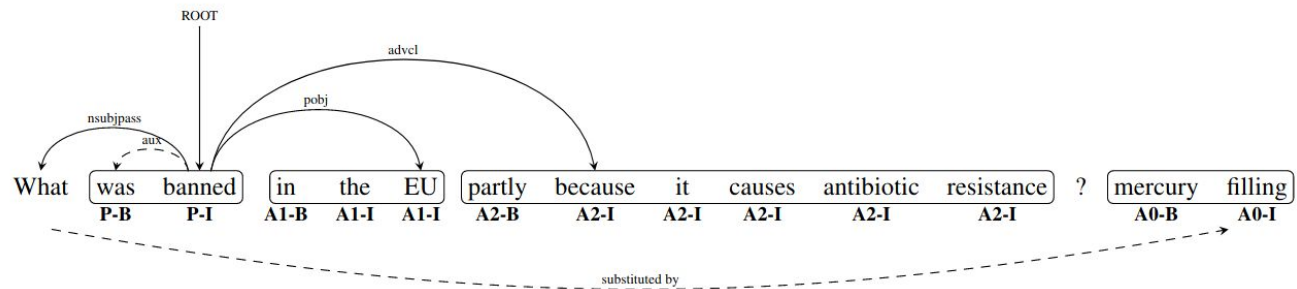
(mercury filling; **was banned**; in the EU; partly because it causes antibiotic resistance)

Create a supervision dataset

- Discard questions that:
  - Introduce words not in text
  - Have more than one wh-word
  - Do not ask for what/who/when/where
  - Ask for the predicate ← *"What did X do?"*
- Run dependency parse on question
  - annotate predicate's arguments as arguments
  - Substitute the wh-word with the answer
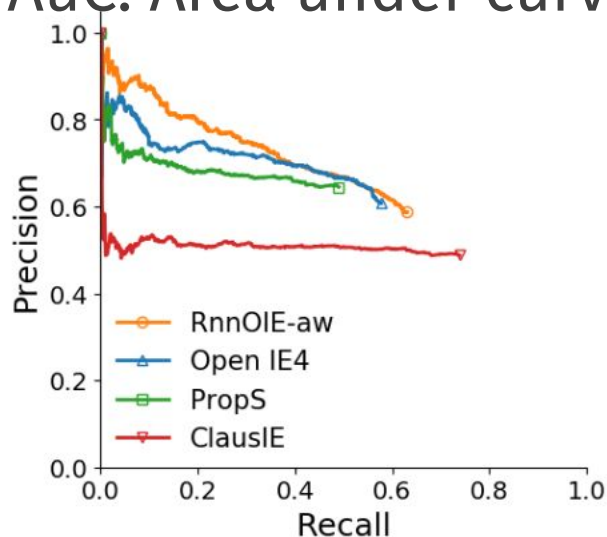
Train neural sequence labelling model

- Resulting dataset is big enough to train a neural sequence labelling model
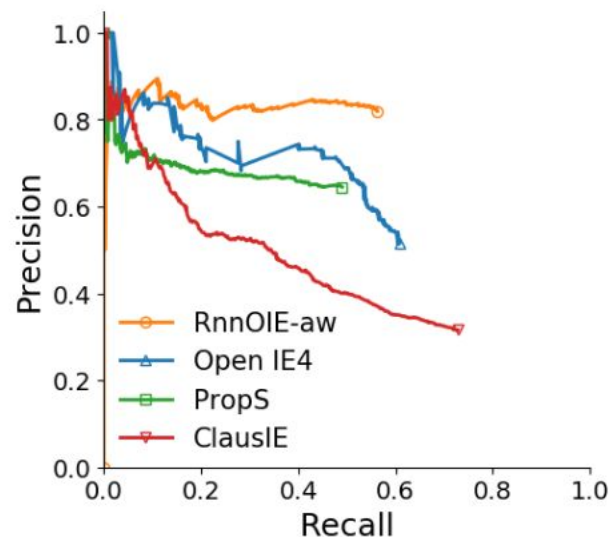- What is that?

Lecture in Week 4!

Evaluation: **More in week 5**

How to compare different systems?

- Precision: how many of our extractions are true?
- Recall: how many of all true extractions did we get?
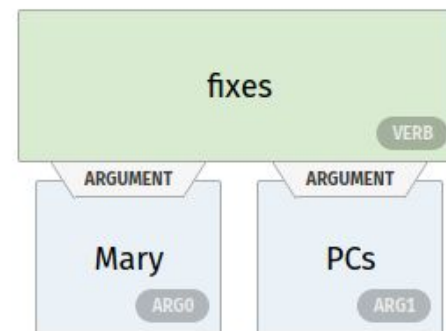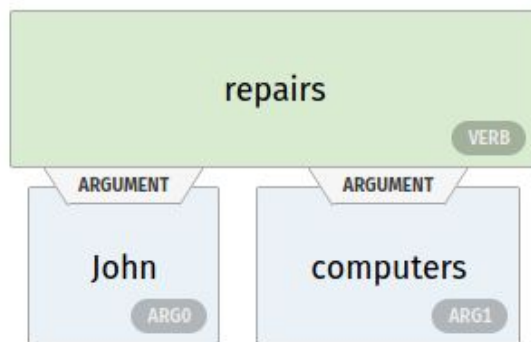- AuC: Area under curve

(a) OIE2016

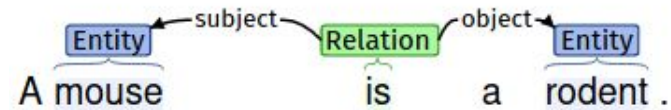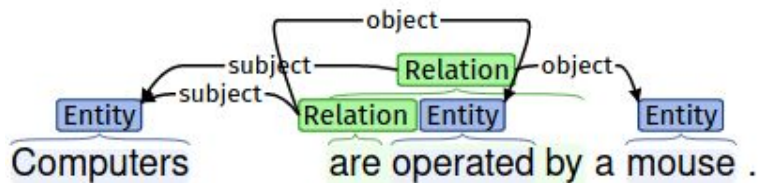(b) WEB

7

Pitfalls: Synonyms

- Who should I go to with a broken computer?
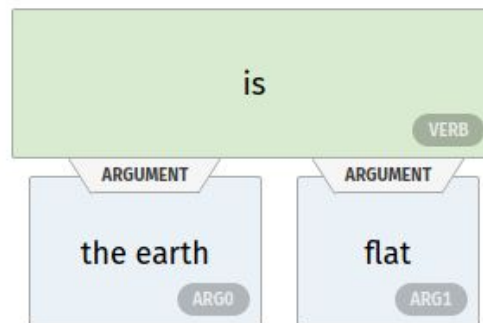
Pitfalls: Ambiguous Entities

- Are computers operated by rodents?

Pitfalls: Trustworthiness

- Is the earth flat?

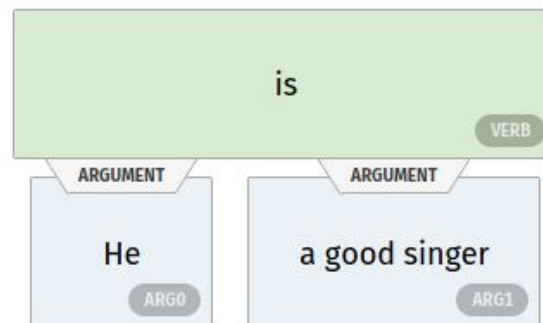Some people believe that the earth is flat .

(The earth, be, flat)
(AttributedTo, some people, believe)

Corpus extraction statistics!

10

Pitfalls: Coreference

- Who is a good singer?

Coreference Resolution!

What I learned in this lecture is...

- ... What is Open Information Extraction

  ...Domain-independent discovery of relations extracted from text...

- ... What is it good for

  structured, machine processable knowledge representation; semantic search

- ... How was it approached

  Pattern & templates; linguistic analysis & natural logic; deep learning

- ... NLP "Bread and butter" always useful

  ...even for the "deepest" of learning

What I learned in this lecture is…

- TextRunner
  - First OpenIE system
  - "Quick and dirty"
- ReVerb
  - Syntactic and lexical constraints
  - More informative, coherent and general extractions
- OLLIE
  - Based on dependency patterns
  - Noun-mediated relations
  - Relations don't have to be in between arguments

What I learned in this lecture is…

- Stanford OpenIE
  - Clause based
  - Natural language logic
  - Pattern based extraction

- Deep learning for OpenIE
  - Based on sequence labelling
  - N-ary relations
  - GPU optimised

Tools

- [Stanford Open Information Extraction](#)

- [AllenNLP Open Information Extraction](#)

- [OLLIE, ReVerb and TextRunner](#)

- [OpenIE 5.1](#)

Downstream Application:
KB Construction

Mostly [this](#) paper

If you think there is time for it

Things i didn't include

- Inter-proposition relationships (i.e. "some people believe the earth is flat will" ----> (earth; be; flat)
- There are systems that capture them
- Maybe loop in when talking about pitfalls, i.e. trustworthiness

Motivation

- Why openIE? (something about machine-readable/searchable representation from unstructured text)
- Downstream tasks: KG construction, data integration/augmentation
- Maybe show example from some medical/bio ontology + natural language text description
- Also [this](#)

Pattern-based

- TextRunner + ReVeRB
- OLLIE